Hans-Christoph Hobohm (Hrsg.)

Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten

Information Science between Virtual Infrastructure and Material Lifeworlds

Unter Mitarbeit von Judith Pfeffing

Proceedings des 13. Internationalen Symposiums
für Informationswissenschaft (ISI 2013)

Potsdam, 19. bis 22. März 2013

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Identifikation von Kundenrezensionen im WWW als Basis eines Crawlers für das Opinion Mining

Sebastian Kastner, Thomas Mandl, Julia Maria Struß

Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Marienburger Platz 22, D-31141 Hildesheim
{kastners|mandl|julia.struss}@uni-hildesheim.de

Zusammenfassung

Diese Arbeit befasst sich mit der Klassifikation von Webseiten in solche, die Kundenrezensionen zu Produkten oder Dienstleistungen enthalten und solche, die keine enthalten. Die derart gewonnenen Rezensionen können anschließend als Input für Opinion-Mining-Systeme dienen, die sich mit der Extraktion und Klassifikation von Meinungen z. B. in der genannten Textsorte beschäftigen. Erste Evaluierungsergebnisse deuten mit einer Accuracy von 91% auf einen vielversprechenden Ansatz hin.

Abstract

This paper describes an approach for classifying web pages into those containing customer reviews and those not containing customer reviews. The reviews found in this way can then be used as input for opinion-mining systems, which deal with the extraction and classification of opinions in written text. A first evaluation with 91% accuracy indicates a promising approach.

In: H.-C. Hobohm (Hrsg.). Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten. Tagungsband des 13. Internationalen Symposiums für Informationswissenschaft (ISI 2013), Potsdam, 19.–22. März 2013. Glückstadt: Verlag Werner Hülsbusch, 384–389.

1 Einleitung

Viele Kunden nutzen heutzutage das Internet, um sich über Produkte und Dienstleistungen zu informieren. Insbesondere die Meinungen anderer Kunden, die über verschiedene Portale und Händlerseiten verfügbar sind, werden dazu herangezogen (vgl. Yahoo! Deutschland 2010). Aber auch für herstellende oder vertreibende Unternehmen dienen solche Kundenrezensionen als Informationsquelle (vgl. Pang/Lee 2008: 13 f.).

Das Opinion Mining beschäftigt sich mit der Herausforderung der automatischen Zusammenfassung und Aufbereitung von Dokumenten hinsichtlich der darin geäußerten Meinungen. Als Textsorte werden in vielen Arbeiten Kundenrezensionen verwendet, daher wird in dieser Arbeit ein Ansatz entwickelt, der später in einem Crawler für das Sammeln dieser Art von Texten genutzt werden kann. Als Motivation für Opinion Mining Systeme wird u. a. die für viele Produkte große Anzahl verfügbarer Rezensionen genannt, die eine manuelle Erfassung sehr zeit- und somit auch kostenintensiv werden lässt (vgl. u. a. Qiu et al. 2011; Hu/Liu 2004). Viele Arbeiten in diesem Bereich vernachlässigen jedoch die Beschaffung der zu analysierenden Dokumente weitgehend und setzen diese als gegeben voraus. In diesem Aufsatz wird nun ein Ansatz vorgestellt, der diese Lücke schließen soll.

Im Folgenden werden zunächst einige verwandte Arbeiten vorgestellt, bevor der Aufbau des Korpus und die durchgeführten Experimente beschrieben werden. Abschließend werden erste Evaluierungsergebnisse präsentiert und ein Ausblick auf das weitere Vorgehen gegeben.

2 Verwandte Arbeiten

Nur wenige Arbeiten haben sich bisher mit der Beschaffung der Dokumente für ein Opinion-Mining-System beschäftigt. Morinaga et al. stellen Suchanfragen mit Produktnamen an Suchmaschinen und extrahieren aus den gelieferten Ergebnissen die meinungstragenden Sätze (vgl. Morinaga et al. 2002: 302). Na und Thet (2009) bedienen sich ebenfalls der Ergebnislisten von Suchmaschinen. Als Suchanfrage verwenden sie gleichermaßen den Produktnamen, fügen diesem aber jeweils den Begriff „review“ hinzu. Für

die Erkennung von Seiten mit Produktrezensionen auf Basis der Ergebnislisten entwickeln sie einen Klassifikationsalgorithmus. Sie erzielen unter Einbeziehung von semi-automatisch erstellten Heuristiken eine Klassifikationsgenauigkeit von bis zu 89% (ebd.: 6 ff.). Ein Ansatz, der den Quelltext von Webseiten direkt nutzt, ist nicht bekannt.

Ein weiterer verwandter Bereich ist die Subjektivitätsanalyse. Hier werden Dokumente oder Textstellen als subjektiv bzw. objektiv klassifiziert (vgl. u. a. Wiebe et al. 2004). Die hier vorgestellte Arbeit unterscheidet sich von jenen im Bereich der Subjektivitätsanalyse insofern, dass die Klassifikation nicht zwischen subjektiven und objektiven Dokumenten erfolgt, sondern zwischen Texten, die eine Rezensionen enthalten oder nicht. Nicht jeder subjektive Text stellt auch eine Rezension dar.

3 Korpus-Erstellung

Zunächst wird mithilfe eines Webcrawlers (crawler4j[1]) ein Korpus aus Webdokumenten zusammengestellt. Der Crawler wurde durch Programmierung an die Aufgabe angepasst und verwendet als Ausgangspunkt Reviewportale wie alatest.com, die neben eigenen Rezensionen auch Links zu Rezensionen anderer Seiten enthalten. Diese werden anschließend manuell in solche mit und ohne Rezensionen eingeteilt. Ein Dokument zählt dabei als eines mit Rezensionen, wenn mindestens eine komplette Rezensionen enthalten ist. Als Rezension gilt eine textuelle Äußerung einer Meinung zu Produkten oder Dienstleistungen (vgl. Kastner 2011: 20). Das so erstellte Korpus enthält 464 Webseiten, davon 68 mit Rezensionen. Um die Ungleichverteilung der Dokumente hinsichtlich der Anzahl der enthaltenen Rezensionen auszugleichen, wird das Korpus in einem zweiten Schritt erweitert. Es werden Suchanfragen mit den Namen von 50 Produkten an eine Suchmaschine gestellt. Dabei wurde auf Zusätze wie „review“ verzichtet (vgl. Na/Thet 2009), damit von der Suchmaschine nicht bevorzugt Seiten mit dem entsprechenden Begriff an prominenten Stellen wie Titel oder URL zurückgegeben werden, um eine Verzerrung der Lernmenge zu vermeiden.

1 http://code.google.com/p/crawler4j/ <20.02.2013>

Nach mehreren Durchgängen wird eine Zahl von 250 Dokumenten mit Rezensionen erreicht, die zusammen mit 250 zufällig ausgewählten Dokumenten aus der Dokumentensammlung im ersten Schritt zu einem Korpus zusammengefasst werden (vgl. Kastner 2011: 21 f.).

4 Merkmalsauswahl und Ergebnisse

Im ersten Schritt erfolgt die Auswahl geeignete Merkmale für die Klassifikationsaufgabe der Dokumente in solche mit oder ohne Rezensionen. Für die Klassifikationsaufgabe wird auf Support Vector Machines (SVM) zurückgegriffen, da diese sich als gut geeignet für Textkategorisierung und die Klassifikation von Webseiten erwiesen haben (vgl. Joachims 1998: 139 ff.). Für die Merkmalsauswahl wird ein Teil des Korpus, bestehend aus 400 zufällig ausgewählten Dokumenten, genutzt. Die Ergebnisse der besten einzelnen Merkmale sowie Merkmalskombination sind in nachfolgender Tabelle dargestellt. Da das Korpus teilweise mithilfe von Suchmaschinen erstellt wurde, welche als Anchor-Text der Links zumeist den Seitentitel verwenden, ist das Ergebnis hier kritisch zu sehen, weshalb dieses Merkmal für die nachfolgenden Experimente nicht verwendet wird.

Tab. 1: Ergebnisse der Merkmalsauswahl unter Verwendung von 10-Fold-Crossvalidation (links Accuracy für einzelne Merkmale, rechts für Merkmalskombinationen) (vgl. Kastner 2011: 31, 34)

Merkmale	Acc.
Anchor Text	89,95
URL	85,42
Content	84,07
Part-of-Speech (POS)	80,75
Meta Description	72,52
Unique Words	72,25
Titel	71,27
Text Complexity	70,75

Merkmale	Acc.
URL+POS+Title+Unique Words	86,75
URL+POS	86,50
URL+POS+Title+Link Density	86,00
URL+POS+Title	86,00
URL+POS+Title+Text Complexity	85,75
URL	85,42
URL+POS+Meta Description	85,25
URL+Titel	84,30

Die zuvor ermittelte beste Merkmalskombination wird abschließend auf den verbliebenen 100 Dokumenten des Korpus mit einer Document Frequency von 25 und einem Cost Parameter von 500, die zuvor ebenfalls experi-

mentell ermittelt wurden, erneut getestet. Das erzielte Ergebnis mit einer Accuracy von 91% liegt dabei knapp 2%-Punkte über dem von Na und Thet (2009) erreichten Wert, verwendet im Gegensatz zu diesen jedoch nicht die Ergebnislisten von Suchmaschinen, sondern nutzt die Webseiten direkt (vgl. Kastner 2011: 35). Zukünftig soll der beschriebene Ansatz auf einer breiteren Dokumentenbasis evaluiert werden, um die Aussagekraft der Evaluierungsergebnisse zu erhöhen. Weiter ist es auch interessant, die Verwendung der Anchor-Texte als Merkmal zu untersuchen, da diese bei der Merkmalsauswahl bei isolierter Betrachtung die besten Ergebnisse liefern, jedoch aufgrund der Methode der Korpus-Erstellung in der weiteren Evaluierung nicht mit berücksichtigt wurden.

Literaturverzeichnis

Ganjisaffar, Y. (o.J.). crawler4j – Open Source Web Crawler for Java. http://code.google.com/p/crawler4j/ <08.11.2012>.

Hu, M.; Liu, B. (2004). Mining and summarizing customer reviews. In: Proc. 10th ACM SIGKDD Intl Conf. on Knowledge discovery and data mining. New York: ACM, 168–177.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proc. 10th European Conference on Machine Learning. Berlin u.a.: Springer, 137–142.

Kastner, S. (2011). Klassifikation von Webdokumenten in Dokumente mit und ohne Kundenrezensionen als Basis eines Crawlers für das Opinion Mining. B.A. Thesis. Universität Hildesheim.

Morinaga, S.; Yamanishi, K.; Tateishi, K.; Fukushima, T. (2002). Mining Product Reputations on the Web. In: Proc. 8th ACM SIGKDD Intl Conf. on Knowledge Discovery and Data Mining. ACM, 341–349.

Na, J.; Thet, T. (2009). Effectiveness of web search results for genre and sentiment classification. In: Journal of Information Science 35 (6), 709–726.

Pang, B.; Lee, L. (2008). Opinion Mining and Sentiment Analysis. In: Foundations and Trends in Information Retrieval 2 (1-2), 1–135.

Qiu, G.; Liu, B.; Bu, J.; Chen, C. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. In: Computational Linguistics 37 (1), 9–27.

Wiebe, J.; Wilson, T.; Bruce, R.; Bell, M.; Martin, M. (2004). Learning Subjective Language. In: Computational Linguistics 30 (3), 277–308.

Yahoo! Deutschland (2010). Neue Yahoo!-Branchenstudie zum Einzelhandel: Von Bekleidung bis Spiele, von Unterhaltungselektronik bis zum Heimwerkerbedarf – Ohne Web fällt kaum eine Kaufentscheidung. München. http://yahoo.enpress.de/pressemeldungen.aspx?p=2483 <12.11.2012>.