Multimedia- und Web 2.0- Daten zur frühzeitigen Erkennung von Krankheitsausbrüchen

Denecke, Kerstin; Eckmanns, Tim; Dolog, Peter; Stewart, Avaré

Publication date:
2010

Document Version
Early version, also known as pre-print

gmds
55. Jahrestagung
MANNHEIM

Abstract ID: 18
Korrespondenz: Kerstin DENECKE, Universität Hannover
Thema: A99. Freie Themen

Originalvortrag

Multimedia- und Web 2.0- Daten zur frühzeitigen Erkennung von Krankheitsausbrüchen

Kerstin Denecke[0], Tim Eckmanns[1], Peter Dolog[2], Avaré Stewart[0]

[0] Forschungszentrum L3S (Hannover)
[1] Robert Koch Institut (Berlin)
[2] Aalborg University (Dänemark)

Onlinemedien, Weblogs, wissenschaftliche und nicht-wissenschaftliche Diskussionsforen sowie elektronische Kommunikation können als Ergänzung zu traditionellen Methoden der Berichterstattung gesehen werden. Sie stellen zunehmend eine wertvolle Informationsquelle in verschiedenen Bereichen dar. Auch Gesundheitsorganisationen sind sich bewusst, dass traditionelle Informationswege und quellen für eine frühzeitige Erkennung und Reaktion auf Krankheitsausbrüche nicht ausreichen, sondern zusätzliche Informationsquellen herangezogen werden müssen [1].
Diese Arbeit beschäftigt sich mit der Nutzung von Multimediadaten und neuen Medien des Web 2.0 (z.B. Blogs, Forums, Twitter) zur Erkennung von Hinweisen auf potentielle Krankheitsausbrüche. Diese Datenquellen stellen besondere Anforderungen an die zur Verarbeitung zu verwendenden Technologien. Daten bzw. Dokumente sind in großen Mengen vorhanden, daher gilt es darin die relevanten zu identifizieren. Sprachlich reicht das Spektrum von offiziellen Berichten in Fachsprache über Blog- und Foreneinträge in Alltagssprache bis hin zu Satz- und Wortfragmenten mit Abkürzungen wie z.B. in Twitternachrichten. Die besonderen Herausforderungen dieser Daten liegen daher in der (1) Sammlung relevanter Dokumente, (2) Filterung irrelevanter Informationen, (3) Extraktion von Information zu Krankheiten und Krankheitsausbrüchen, (4) der sensitiven Interpretation der extrahierten Daten und deren (5) nutzerspezifische Selektion und Präsentation.
Zur Extraktion von Information zu Krankheiten und Krankheitsausbrüchen aus Social Media Daten ist es wichtig, Hypothesen von Fakten, historische Ereignisse von aktuellen etc. zu unterscheiden. In verschiedenen Datenquellen können ähnliche oder verwandte Ereignisse beschrieben werden. Extrahierte Informationen müssen daher ggf. zusammengeführt werden sowie Redundanzen vermieden werden.
Um relevante Dokumente zu identifizieren, können manuell spezifizierte Schlüsselworte genutzt werden [2]. Um jedoch zusätzlich in der Lage zu sein, auch Hinweise auf potentielle Ausbrüche auf Krankheiten zu entdecken, deren Namen noch nicht bekannt ist (z.B. die Bezeichnung Schweinegrippe war vor April 2009 nicht gebräuchlich), sind weitere Methoden notwendig. Unser Ansatz kombiniert daher Methoden des maschinellen Lernens, um Informationen zu Krankheiten und Krankheitsausbrüchen in natürlichsprachigen Dokumenten zu identifizieren. Insbesondere entwickeln wir einen Ansatz, der Event Extraction Technologien aus dem Bereich Data Mining mit Verfahren aus dem Text Mining (u.a. Cross-Classification, Bootstrapping [3,4]) kombiniert. Diese Herangehensweise erlaubt den Umgang mit den großen Daten, mit denen wir es hier zu tun haben und reduziert den manuellen Aufwand, z.B. zum Erstellen von Extraktionsregeln.
Es ist zu erwarten, dass die Informationen, die mit diesen Verfahren extrahiert werden sehr zahlreich

sind. Es muss daher vermieden werden, Nutzer mit großen Mengen an potentiell relevanten Informationen zu überfluten. Mittels sensitiver Tuning- und Personalisierungsmechanismen versuchen wir die Anzahl an falschen Alarmen zu reduzieren und das dem Nutzer präsentierte Analyseergebnis auf die Nutzerbedürfnisse zuzuschneiden.

[1] Paquet C, Coulombier D, Kaier R, Ciotti M: Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill. 2006;11(12)
[2] Yangarber, R., et al.: Combining information about epidemic threats from multiple sources. In: RANLP-2007, Borovets, Bulgaria (2007)
[3] Chen, Z., Ji, H.: Can one language bootstrap the other: a case study on event extraction. In: SemiSupLearn '09: Proceedings of the NAACL HLT 2009Workshop on Semi-Supervised Learning for Natural Language Processing, Morristown, NJ, USA, Association for Computational Linguistics (2009) 66-74
[4] Zhen, Y., Li, C.: Cross-domain knowledge transfer using semi-supervised classification. In: AI '08: Proceedings of the 21st Australian Joint Conference on Artificial Intelligence, Berlin, Heidelberg, Springer-Verlag (2008) 362-371

Keywords: Epidemic Intelligence, Natural Language Processing