Richard Zowalla
Dr. sc. hum.

Erfassung, Auswertung und Analyse des deutschsprachigen Gesundheitsweb

Fach/Einrichtung: Medizinische Biometrie und Informatik
Doktorvater: Prof. Dr. Thomas Wetter

Hintergrund: Das Internet ist zu einer immer wichtigeren Quelle für Gesundheitsinformationen geworden. Angesichts der wachsenden Zahl von Webseiten ist es für Menschen jedoch nahezu unmöglich, die sich entwickelnden und ständig ändernden Inhalte im Gesundheitsbereich vollständig zu erfassen. Um die Natur aller webbasierten Gesundheitsinformationen in einer bestimmten Sprache besser zu verstehen, ist es wichtig, (1) Informationsknotenpunkte im Gesundheitsbereich, (2) Anbieter von Inhalten mit hoher Bedeutsamkeit, (3) die Lesbarkeit von Gesundheitsinformationen und (4) wichtige Themen und Trends im gesundheitsbezogenen Web zu identifizieren. In diesem Zusammenhang kann ein Web-Crawling-Ansatz die Daten für eine computergestützte und statistische Analyse zur Beantwortung der Fragen (1) bis (4) liefern.

Zielsetzung: Das primäre Ziel dieser Arbeit ist es, einen methodischen Ansatz zu entwickeln, mit dem das deutschsprachige Gesundheitsweb (GHW), das alle gesundheitsbezogenen Webinhalte der drei mehrheitlich deutschsprachigen Länder Deutschland, Österreich und Schweiz umfasst, d.h. die Top-Level Domains `.de`, `.at` und `.ch`, erfasst und analysiert werden kann. Dabei sollen Anbieter von Gesundheitsinformationen mit hoher Bedeutsamkeit identifiziert werden. Bedeutsam sind dabei solche gesundheitsbezogenen Webseiten, die mit besonders hoher Wahrscheinlichkeit von vielen Webanwendern aufgesucht werden, so dass die Seiteninhalte das Denken und Verhalten vieler Anwender potentiell beeinflussen. Zu diesem Zweck soll ein fokussierter Web-Crawler entwickelt werden, welcher das Verhalten von menschlichen Benutzern, in Webseiten vorgefundene Links zu einem bestimmten Thema zu verfolgen, systematisch nachgebildet. Als sekundäres Ziel dieser Arbeit sollen die durch diesen Web-Crawler gewonnenen Strukturinformationen für einen Vergleich zwischen den Ländern der D-A-CH-Region genutzt werden. Darauf aufbauend soll die Lesbarkeit der erfassten Gesundheitsinformationen mittels syntaktischer und vokabularbasierter Lesbarkeitsmaße untersucht werden. Die wichtigsten Themenbereiche sollen durch Topic Modelling identifiziert werden.

Methoden: Ein Support-Vector-Machine-Klassifikator wurde auf einem großen Datensatz trainiert, der mittels der Inhalte verschiedener Informationsanbieter zusammengestellt wurde, um zwischen gesundheitsbezogenen und nicht gesundheitsbezogenen Webseiten zu unterscheiden. Der Klassifikator wurde anhand von Accuracy, Recall und Precision auf einem 80/20-Trainings-/Test-Split (TD1) und anhand eines mittels Crowd-Sourcing validierten Datensatzes (TD2) bewertet. Um den Web-Crawler zu implementieren, wurde das Open-Source-Framework StormCrawler erweitert. Der eigentliche Web-Crawl wurde über 370 Tage durchgeführt. Der Web-Crawler wurde anhand der Harvest Rate evaluiert und dessen Recall mit einem Seed-Target-Ansatz geschätzt. Zudem wurde die Abdeckung des Web-Crawls gegenüber der Suchmaschine Google ermittelt. Wichtige Websites wurden durch die Anwendung von PageRank auf dem Web-Graphen des GHW identifiziert. Latent Dirichlet Allocation wurde angewendet, um Themen innerhalb der bestplatzierten Websites zu entdecken. Als nächstes wurde eine computergestützte Lesbarkeits- und

Vokabularanalyse für jede gesundheitsbezogene Website durchgeführt. Für die Lesbarkeitsanalyse wurde die an die deutsche Sprache angepasste Flesch Reading Ease (FRE) Metrik und die vierte Wiener Sachtextformel (WSTF) verwendet. Um die Laienfreundlichkeit des Vokabulars zu beurteilen, wurde dessen fachsprachliches Niveau mittels eines speziell trainierten Support-Vector-Machine-Klassifikators berechnet.

Ergebnisse: Insgesamt wurden während des Untersuchungszeitraums von 370 Tagen $n = 14.193.743$ gesundheitsbezogene Webseiten gesammelt. Der Textklassifikator erreichte auf $TD1$ eine Genauigkeit von $0,937$ ($TD2 = 0,966$), eine Precision von $0,934$ ($TD2 = 0,954$) und einen Recall von $0,944$ ($TD2 = 0,989$). Die durchschnittliche Harvest-Rate betrug $21,77\%$; der Recall wurde gemäß Seed-Target-Ansatz auf $0,821$ ($4.105/5.000$) geschätzt. Zudem konnten $5.944/6.829(87,04\%)$ der über eine *Custom Google Search Engine* aufgefunden relevanten Websites im Web-Crawler wiedergefunden werden. Der resultierende Host-aggregierte Web-Graph umfasst 231.733 Knoten, die über 429.530 Kanten verbunden sind (Netzwerkdurchmesser=25; durchschnittliche Pfadlänge=$6,804$; durchschnittlicher Grad=$1,854$; Modularität=$0,723$). Von den 3.000 bestplatzierten Websites (1.000 pro Top-Level Domain) gehören $18,50\%$ ($555/3.000$) zu Websites von staatlichen oder öffentlichen (Gesundheits-)Einrichtungen, $18,03\%$ ($541/3.000$) zu Non-Profit-Organisationen, $54,03\%$ ($1.621/3.000$) zu privaten Organisationen, $4,07\%$ ($122/3.000$) zu Nachrichtenagenturen, $3,87\%$ ($116/3.000$) zu pharmazeutischen Unternehmen, $0,90\%$ ($27/3.000$) zu privaten Bloggern und $0,60\%$ ($18/3.000$) zu anderen. Mittels Latent Dirichlet Allocation konnten 50 Themen identifiziert und in elf Themenbereichen gruppiert werden: „Forschung & Wissenschaft“, „Krankheit & Verletzung“, „Der Staat“, „Strukturen des Gesundheitswesens“, „Ernährung“, „Medizinische Fachgebiete“, „Wirtschaft“, „Lebensmittelproduktion“, „Gesundheitskommunikation“, „Familie“ und „Sonstiges“. Die häufigsten Themenbereiche waren „Forschung & Wissenschaft“ und „Krankheit & Verletzung“ mit $21,04\%$ bzw. $17,92\%$ aller Themen über alle Domains und Informationsanbieterkategorien hinweg. Die Lesbarkeitsanalyse zeigt, dass der Großteil der analysierten Websites strukturell schwer oder sehr schwer lesbar ist: $93,35\%$ ($2.539/2.720$) erreichten einen WSTF ≥ 12, $98,93\%$ ($2.691/2.720$) einen FRE ≤ 49. Darüber hinaus zeigt die Vokabularanalyse, dass $48,55\%$ ($1.320/2.719$) der Websites Vokabular verwenden, das für ein Laienpublikum gut geeignet ist. Die Satzkomplexitätsmaße FRE und WSTF sind zudem stark korreliert, sodass sie austauschbar verwendet werden können. Außerdem korreliert ein hoher Schwierigkeitsgrad des Vokabulars mäßig mit der Satzkomplexität. Insgesamt konnte festgestellt werden, dass sich die Ergebnisse für die untersuchten Top-Level-Domains nur unwesentlich unterscheiden.

Schlussfolgerungen: Die Ergebnisse zeigen, dass der vorgestellte fokussierte Web-Crawler eine geeignete Methode ist, um einen großen Teil des GHW zu erfassen. Es konnten die wichtigsten Informationsknotenpunkte und Themen ermittelt werden. Die Ergebnisse zeigen zudem, dass die Lesbarkeit innerhalb des GHW niedrig ist. Infolgedessen können Patienten auf Barrieren stoßen, auch wenn das verwendete Vokabular aus medizinischer Sicht angemessen erscheint. Zukünftige Forschungsprojekte könnten die Analysen ausweiten, um vertrauenswürdige Gesundheitsinformationsanbieter vollautomatisch zu identifizieren.