Die Suche nach medizinischen Experten
Eine Studie zur Angabe von Autoren in medizinischen Datenbanken

Stepanowsky, Petra [a], Kulczycki, Peter [a], Schierl Michael [b], Lirk, Gerald [a]

[a] FH OÖ Forschungs- & Entwicklungs GmbH, Softwarepark 11, A-4232 Hagenberg, AUSTRIA
[b] Franz-Josef-Platz 16, A-4810 Gmunden, AUSTRIA

KURZFASSUNG/ABSTRACT:
Die Suche nach medizinischen ExpertInnen ist sowohl für Forschungsinstitutionen als auch für die Industrie bedeutsam. Die Zuordnung der AutorInnen biomedizinischer Publikationen zu einzelnen Personen ist jedoch, auf Grund der Heterogenität der Angaben, äußerst diffizil. Mit dem Computerprogramm ALBERT wurde eine Analyse der AutorInnen in der medizinischen Literaturdatenbank PubMed vorgenommen und die AutorInnen zu Personen geclustert. Damit konnte die Anzahl der potentiellen AutorInnen um einen Faktor 4 reduziert werden. Weiters wurden Vorschläge erarbeitet um die für eine automatisierte Literatursuche entscheidenden Parameter *recall* und *precision* zu verbessern.

1 EINLEITUNG

Das medizinische Wissen nimmt exponentiell zu. Quantifizierbar ist dies durch die Anzahl der monatlich in der medizinischen Literaturdatenbank PubMed eingetragen Artikel. Innerhalb von 10 Jahren stieg dieser Wert von 67.000 auf 105.000 Publikationen. Etwa 5% davon sind die Ergebnisse klinischer Studien (CT). Eine Weiterbildung medizinisch tätigen Personals ist auf Grund dieser Menge an Primärliteratur aus zeitlichen Gründen nicht möglich. Daher werden in speziellen Einrichtungen (z.B. Cochrane [1]) entsprechende Zusammenfassungen von ExpertInnen erstellt. Das Auffinden von SpezialistInnen für diese speziellen medizinischen Fragestellungen ist die erste Aufgabe, der man dabei gegenüber steht. Auch die Pharmaindustrie ist ständig auf der Suche nach Forschungstrends und nach entsprechenden Fachleuten, die ihnen bei der Suche nach neuen Medikamenten helfen könne. Daher beschäftigt dieser Industriezweig ganze Abteilungen, welche u.a. diese Datenbanken durchforsten.

2 ZIEL

PubMed [2] bietet die Möglichkeit die Suchergebnisse nach bestimmten Kriterien wie Autor, Adresse, Publikationsdatum oder Studientyp, zu filtern. Obwohl die Datenbank-Einträge prinzipiell strukturiert sind, gibt es doch gewaltige Differenzen in der Angabe der AutorInnen und deren Adresse. Ziel dieses Projekt ist es die AutorInnen bestimmten Personen zuzuordnen und auf diese Weise „FachspezialistInnen“ zu identifizieren Mit Hilfe der von unserer Projektgruppe entwickelten Software ALBERT sollte es in Folge auch möglich sein die Expertise der AutorInnen zu bestimmen und damit die ExpertInnen zu entdecken.

3 METHODE

ALBERT ist in der Lage sowohl strukturierte als auch unstrukturierte Informationen aus PubMed-Artikel zu extrahieren und in einer eigenen Datenbank zu speichern [3]. Diese Informationen werden analysiert und mit verschiedenen Parametern, z.B. den Oxford Level of Evidence, die Güte klinischer Studien bestimmt. Diese Datenbasis wurde in dem vorliegenden Projekt verwendet um die AutorInnen, welche in unterschiedlicher Weise hinterlegt sind, zu bestimmten Personen zusammenfassen. Mit Hilfe des Statistik-Pakets R wurden die PubMed-AutorInnen analysiert.

4 ERGEBNISSE

Für die beschriebenen Fragestellungen wurde mit einem Datensatz von 1.158.254 klinischen Studien der Jahre 1975-2014 gearbeitet. Diese wurden von ca. 900.000 verschiedenen ErstautorInnen verfasst. Während alle Publikationen durchschnittlich 5,0 (Median = 4) AutorInnen haben, ist die durchschnittliche AutorInnenzahl bei den CTs mit 6,2 (Median 6) etwas höher.
Das Clustering mehrerer AutorInnennamen zu einzelnen Personen bzw. deren Identifizierung, wie es für eine exakte Festlegung der Expertise notwendig erscheint, ist nicht einfach: ca. 250.000 AutorInnen haben den gleichen Nachnamen. Da die Speicherung des vollständigen Vornamens in PubMed erst seit 2002 möglich ist, dies jedoch häufig nicht gemacht wird, benötigt es zusätzliche Parameter zur Identifikation. Hier bietet sich die angegebene Adresse der AutorInnen an, wobei zusätzlich zu berücksichtigen ist, dass eine Person die Adresse während im Laufe der Karriere ändern wird. Die Zuordnung zu Adressen ist ebenfalls nicht einfach, da diese uneinheitlich angegeben werden kann:

- Sprachspezifische Schreibweise der Adresse vs. Übersetzung ins Englische
- Unterschiedliche Reihenfolge der Adressteile
- Verschiedene Namen für „Abteilung“ (Department, Dept, Division, Facility, Faculty, etc.)

Mit Hilfe der Adresse ist es gelungen die Spezifikation der AutorInnen um einen Faktor 3,5 zu verbessern.

Weiters konnte gezeigt werden, dass im Laufe der letzten Jahre größere Teams die Ergebnisse klinischer Studien publizieren. Der Trend zu längeren Autorenlisten, den bereits Wuchty et al [4] und Papatheodorou et al [5] beschrieben haben, setzte sich somit fort.

5 DISKUSSION

Die AutorInnenschaft klinischer Publikationen einzelnen Personen zuzuordnen und damit medizinische ExpertInnen zu lokalisieren, ist schwierig, aber nicht unmöglich. Mit Hilfe von Länder- und Universitätslisten kann eine spezifische Zuordnung der Adressteile erfolgen. Unterschiedliche Schreibweisen von z.B. Institutsnamen, aber auch der AutorInnennamen (mit und ohne 2. Vornamen) können mit Alignment-Algorithmen gelöst werden [6]. Weiters können zusätzliche, teilweise händisch annotierte Datenbanken wie Research Gate oder Microsoft Science zur Identifikation verwendet werden. Ein Zugriff auf die Adressteile in den Volltextversionen der Publikationen wird die Spezifikation weiter erleichtern. Trotz bereits überzeugenden Clusterings der AutorInnen werden immer weitere Optimierungsmöglichkeiten des Algorithmus bleiben. Spezialfälle wie Namensänderungen werden wohl nie zur Gänze ermittelt werden können. Zudem müssen Sensitivität (recall) und positiver Vorhersagewert (precision) noch händisch bestimmt werden.

LITERATURVERWEISE

[1] Cochrane (2014): http://www.thecochranelibrary.com/ (25.01.2015).
[2] PubMed (2014): http://www.ncbi.nlm.nih.gov/pubmed/ (25.01.2015).
[3] Stepanowsky P., Brunner S., Lirk G. (2014): Automatische Identifikation der besten Evidenz für klinische Fragestellungen - Automatische Identifikation der besten Evidenz für klinische Fragestellungen. Tagungsband des 8. Forschungsforum der österreichischen Fachhochschulen, Kufstein.
[4] Wuchty S., Jones, B.F., Uzzi, B. (2007): The Increasing Dominance of Teams in Production of Knowledge. Science, 316, 1036-1039.
[5] Papatheodorou, S.I., Trikalinos, T.A., Ioannidis, J.P.A. (2008): Inflated numbers of authors over time have not been just due to increasing research complexity. J.Clin.Epid., 61, 546-551.
[6] Huber I. (2014): Anwendung der Bioinformatik in Online-Übersetzungstools – Masterarbeit, FH OÖ.