HUMBOLDT-UNIVERSITÄT ZU BERLIN

Mathematisch Naturwissenschaftliche Fakultät II
Institut für Informatik

Extraktion von durch PubMed verlinkten Volltexten mit Hilfe von Machine Learning

Exposé Studienarbeit

Dozent: Ulf Leser

Abgabetermin: Januar 2013

Martin Beckmann

18. Januar 2013

Inhaltsverzeichnis

1 Einleitung

Um in der Forschung zu neuen Erkenntnissen zu kommen, besteht der erste Schritt meist darin, sich einen Überblick über bestehende Publikationen zu dem gewählten Thema zu verschaffen. Im biomedizinischen Bereich wird dabei meist PubMed als Anlaufstelle genutzt [FPMP08]. Diese Datenbank verfügt über knapp 18 Millionen englische Artikel [1], zu denen jeweils die Abstracts vorhanden sind. Einen Mehrwert an Informationen gegenüber diesen Abstracts haben die Volltexte zu den Artikeln, wobei PubMed hierfür auf andere Seiten verlinkt mit sogenannten LinkOuts [ML01]. LinkOuts sind Verweise auf andere Seiten und Datenbanken, in denen weiterführende Informationen zu dem Thema des Artikels stehen. Volltexte sind aber nicht alle frei verfügbar, sondern können ein Abonnement benötigen oder Geld kosten. Die Seite PMC(PubMed Central) hat es sich zur Aufgabe gemacht, den Teil der frei zugänglichen Volltexte zentral verfügbar zu machen und hat momentan knapp 2,4 Millionen Artikel[2]. Sucht man aber bei PubMed nach allen englischen Artikeln mit frei zugänglichen Volltexten, so werden ungefähr 3,7 Millionen Einträge angezeigt.
Das liegt vermutlich daran, dass sich PubMed Central mit den Publishern der Journals absprechen muss, um die Artikel anzeigen zu dürfen, während von PubMed nur verwiesen wird auf die Seiten der Publisher. Außerdem muss beachtet werden, dass PubMed Central ausdrücklich vorschreibt, dass die Volltexte nicht automatisch von ihrer Seite heruntergeladen werden dürfen (das ist auch an der robots.txt [Pea98] erkennbar) und verweisen dafür auf einen FTP-Service, der nur ungefähr 400.000 Artikel zur Verfügung stellt. Demzufolge wäre es wünschenswert, die frei verfügbaren Volltexte zu nutzen, die von PubMed verlinkt werden, wobei auch hier die Forderungen der Anbieter berücksichtigt werden müssen. Das Problem besteht dabei aber darin, dass der Volltext auf der verlinkten Seite noch extrahiert werden muss. Dieser ist meist durch einen Link auf ein Pdf-Dokument erhältlich, kann aber auch in anderen Formen vorliegen, wie beispielsweise als Text direkt eingebettet im HTML-Dokument.

2 Idee

Die Aufgabe besteht somit darin, auf einer Seite, auf die durch einen LinkOut von PubMed verlinkt wird, den Link zur gewünschten Datei zu finden (sei es ein Pdf-Dokument oder eine Datei im XML-Format) oder festzustellen, dass kein solcher Link auf der Seite existiert [3]. Dazu ist ein Ranking sinnvoll, um die erfolgsversprechensten Links zuerst zu betrachten. So ist instinktiv klar, das eine verlinkte Datei, die auf „.pdf“ endet, der gewünschte Volltext im Pdf-Format sein könnte. Es ist auch ersichtlich, dass ein Anchor

[1] Stand 10.12.2012
[2] Stand 10.12.2012
[3] Was in dem Fall getan werden kann, wenn kein Link vorhanden ist, steht in Kapitel 7

mit dem Text „free full text“ mit erhöhter Wahrscheinlichkeit auf eine/die Datei mit dem Volltext verweist. Dass diese beide Eigenschaften nicht notwendig sind, sieht man auf folgender Internetseite [4]. Darin besitzt der Anchor keinen Text, sondern ein Bild, und der Link („http://www.dovepress.com/getfile.php?fileID=14210“) endet auch nicht auf „.pdf“, auch wenn es auf ein Pdf-Dokument verlinkt.
Um das Ranking automatisch durchführen zu können, soll anhand einer korrekt annotierten Trainingsmenge eine SVM gelernt werden. Dabei können zum Beispiel folgende Features hilfreich sein:

- Endet der Link auf „.pdf“?
- Ist in dem Anchor ein Bild statt eines Textes
- Welche Wörter sind in dem Anchortext und darum herum

3 Related Work

Nach eigener Auffassung wurde in der vergangenen Zeit noch keine Arbeit veröffentlicht, die sich ebenfalls mit dem Erkennen von Links zu Volltexten der in PubMed angebotenen biomedizinischen Artikel beschäftigt. Die Aufgabe lässt sich am ehesten mit der eines focussed Crawler vergleichen. Bei einem solchen geht es darum, für ein gegebenes Thema relevante Seiten aus dem Internet zu finden, indem nach und nach einzelne Internetseite heruntergeladen und darin befindliche Links weiter untersucht werden. Die Suche ist fokussiert, da nur solche Links weiter verfolgt werden sollen, die als relevant angenommen werden. Die in meiner Arbeit anfallende Aufgabe für eine gegebene Seite den relevanten Link (oder Links) zu finden, entspricht also einem focussed Crawler, der bei einer Tiefe von 1 abbricht.
Auch bei focussed Crawler wird oft der Text eines Anchors oder der darum befindliche Text genutzt, um die Relevanz des Links zu bestimmen. Einerseits kann sich der benutzte Text auf den Anchortext beschränken [McB94]. Oft wird auch ein Fenster einer bestimmten Größe um den Anchor herum gewählt [CDR+98]. Bezieht man die strukturellen Informationen, die durch die HTML-Tags gegeben sind, mit ein, so können die Wörter in den „umliegenden“ Textblöcken in Verbindung mit dem jeweiligen Abstand als Features benutzt werden [CPS02]. In [Pan03] wird anhand des DOM-Trees ein aggregation-path gebildet vom Anchor bis zum root-Element (dem HTML-Tag). Dabei stellte der Autor fest, dass beim Benutzen des Anchortext ein höherer durchschnittlicher Simularity Score berechnet wird, als wenn man höhere Knoten wählt, dafür besteht aber eine höhere Wahrscheinlichkeit für Zero-Simularity. Sie stellten fest, dass der Parent-Knoten des Anchors einen guten Kompromiss geben würde.

[4] http://www.dovepress.com/overcoming-resistance-and-barriers-to-the-use-of-local-estrogen-therap-peer-reviewed-article-IJWH

	Registrierung	Abonnement	weder noch
kostenlos	217	0	16519
nicht kostenlos	0	17226	0

Tabelle 1: Aufteilung der kostenlosen und nicht kostenlosen Artikel aus den Kategorien Registration, Abonnement oder Artikel ohne beides.

4 LinkOuts

Nutzt man die von PubMed zur Verfügung gestellten E-utilities [5], so lassen sich für einen gegebenen Artikel aus PubMed alle LinkOuts anzeigen. Diese sind in verschiedene Kategorien eingeteilt, von denen für diese Arbeit aber nur die „Full Text Sources“ wichtig sind. Jeder LinkOut verfügt weiterhin über Attribute, die angeben, ob es sich um eine freie Quelle handelt („free resource“), ob man sich dafür auf der Seite registrieren muss („registration required“) oder ob man sogar ein Abonnement haben muss beziehungsweise Geld dafür ausgeben („subscription/membership/fee required“). Bei der Untersuchung von ungefähr 33000 LinkOuts der gewünschten Kategorie für eine zufällig gewählte Teilmenge der 3,8 Millionen englischen Artikel mit frei verfügbarem Volltext, ergab sich eine folgende Verteilung, die in Tabelle 1 zu sehen ist. Darin kann man sehen, dass LinkOuts genau dann nicht das Attribut „free resource“ besitzen, wenn das Attribut „subscription/membership/fee required“ gesetzt ist. Resourcen, die nur eine Registration und kein Abonnement benötigen, werden als kostenlos eingestuft. Somit werden die freien Resourcen unterteilt in solche, die eine Registration erfordern und solche, die es nicht tun, wobei der erste Teil eher gering ausfällt.
Um eine automatische Extraktion des Volltextes zu gewährleisten, ist es von Nöten, die Quellen zu ignorieren, die eine Registrierung erfordern.
Demnach werden nur solche LinkOuts betrachtet, die zur Kategorie „Full Text Sources“ gehören, das Attribut „free resource“ und weder das Attribut „subscription/membership/-fee required“ noch das Attribut „registration required“ besitzen.
Auf einer Menge von 40000 zufällig gewählten, von PubMed als englisch und frei verfügbar angegebenen Artikeln wurde überprüft, ob auch alle diese Artikel einen für die automatische Extraktion nützlichen LinkOut besitzen. Zunächst wurden alle Artikel entfernt, die einen LinkOut zu PMC haben, da nur solche Artikel neu gefunden werden sollen, die nicht schon über PMC zugänglich sind. Dies waren 23395 Artikel, was dem Verhältnis von 2,4 Millionen Artikeln in PMC gegenüber 3,7 Millionen Artikeln in PubMed nahe kommt. Von den restlichen 16605 Artikeln hatten gerade einmal 249 Artikel keinen nützlichen LinkOut. Interessant ist dabei, dass, von den 249 Artikeln, 32 auch keine Volltextquellen hatten, wenn man die Quellen dazu nimmt, die eine Registration verlangen.

[5] Für eine Einführung, siehe http://www.ncbi.nlm.nih.gov/books/NBK25500/

5 Trainingsdaten

Um das in Kapitel 2 erwähnte Lernverfahren durchführen zu können, werden vorannotierte Trainingsdaten benötigt. Dafür werden zwei Wissenschaftler mit der Aufgabe beauftragt, für eine Menge von Internetseiten, die von PubMed über LinkOuts verlinkt sind, alle Anchor zu markieren, die auf Volltexte des Artikels verweisen (sei es ein Pdf, XML oder ein anderes Format wie beispielsweise ePub). Die Trainingsmenge wird dabei aus den frei verfügbaren Artikeln von PubMed gewählt, wobei mehrere Artikel von verschiedenen Anbietern ausgewählt werden. Dadurch wird die Trainingsmenge nicht durch einige stark vertretene Publisher dominiert und es kann später eine Cross Validation durchgeführt werden, bei dem von jedem Publisher Artikel sowohl in der Traingsmenge, als auch in der Validierungsmenge sind.
Nicht berücksichtigt werden dabei die verlinkten Seiten, die selbst schon ein Pdf sind und somit nicht weiter betrachtet werden müssen.
Neben der erwähnten Cross Validierung, bei der Seiten des gleichen Publishers sowohl in der Trainingsmenge, als auch in der Validierungsmenge enthalten sind, soll auch eine Cross-Learning Validierung durchgeführt werden, bei der ein gegebener Publisher entweder in der Trainingsmenge enthalten ist oder in der Validierungsmenge, aber nicht in beiden.
Dadurch sollen die beiden Fälle simuliert werden, in dem ein Volltext von einem bekannten Publisher oder aber von einem unbekannten Publisher extrahiert werden soll.

6 Umgang mit robots.txt

Wie bereits erwähnt, sollte beim automatischen Herunterladen der Inhalte von Webseiten auf die Forderungen der Anbieter geachtet werden, um sich so vor rechtlichen Maßnahmen zu schützen. In welcher Form diese Forderungen angegeben werden, kann unterschiedlich sein. Durchgesetzt hat sich aber größtenteils die Nutzung einer sogenannten robots.txt. Darin können Webseitenbetreiber angeben, welche Seiten automatisch von Webcrawlern heruntergeladen und indexiert werden dürfen.
Um zu überprüfen, wie viele Anbieter das automatische Herunterladen der frei zugänglichen Volltexte erlauben, soll für die als relevant annotierten Links überprüft werden, ob sie in der robots.txt vom Indizieren geblockt werden. Somit soll approximiert werden, wie viele der 3,8 Mio. von PubMed als frei zugänglich angegebenen Volltexte auch für das automatische Laden zugänglich sind.

7 Abgrenzung

Zwei mögliche Ansätze werden in dieser Studienarbeit nicht weiter betrachtet.
Einerseits könnten die Bilder, die in einem Anchor verwendet werden, mit einem OCR-Tool analysiert werden, um so zu erkennen, ob im Bild die Worte „PDF“ oder „XML“ verwendet werden [6]. Dies könnte die Erkennung der richtigen Links auf einer Seite verbessern.
Andererseits könnte der Fakt ausgenutzt werden, dass oft der eigentliche Artikel auf der von PubMed verlinkten Seite im HTML eingebettet ist. Somit besteht die Möglichkeit, mit Hilfe von Boilerplate Detection [KFN10], den Artikel zu isolieren und als Klartext zu speichern. Um die Klassifikation der relevanten Blöcke auf der Internetseite zu verbessern, könnte dabei auf das womöglich schon entdeckte Pdf-Dokument zurückgegriffen werden, so dass weder zu wenig, noch zu viel Text als zum Artikel gehörig erkannt wird. Diese zweite Herangehensweise ermöglicht es auch mit Seiten zu arbeiten, die gar keinen Link auf Pdf- oder XML-Dokumente besitzen[7].

Literatur

[CDR+98] Chakrabarti, S. ; Dom, B. ; Raghavan, P. ; Rajagopalan, S. ; Gibson, D. ; Kleinberg, J.: Automatic resource compilation by analyzing hyperlink structure and associated text. In: *Computer Networks and ISDN Systems* 30 (1998), Nr. 1, 65–74. `http://www.sciencedirect.com/science/article/pii/S0169755298000877`

[CPS02] Chakrabarti, S. ; Punera, K. ; Subramanyam, M.: Accelerated focused crawling through online relevance feedback. In: *Proceedings of the 11th international conference on World Wide Web*, 2002, 148–159

[FPMP08] Falagas, M. E. ; Pitsouni, E. I. ; Malietzis, G. A. ; Pappas, G.: Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. In: *The FASEB Journal* 22 (2008), Nr. 2, S. 338–342

[KFN10] Kohlschütter, C. ; Fankhauser, P. ; Nejdl, W.: Boilerplate detection using shallow text features. In: *Proceedings of the third ACM international conference on Web search and data mining*, 2010, 441–450

[McB94] McBryan, O. A.: GENVL and WWWW: Tools for taming the web. In: *Proceedings of the First International World Wide Web Conference* Bd. 341, 1994

[6] http://jkms.org/DOIx.php?id=10.3346/jkms.2003.18.1.69
[7] http://www.ajmc.com/pubMed.php?pii=3064

[ML01] McEntyre, J. ; Lipman, D.: PubMed: bridging the information gap. In: *Canadian Medical Association Journal* 164 (2001), Nr. 9, 1317–1319. `http://www.ecmaj.ca/content/164/9/1317.short`

[Pan03] Pant, G.: Deriving link-context from HTML tag tree. In: *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003, 49–55

[Pea98] Peacock, Ian: Showing Robots the Door. In: *Ariadne* (1998), Nr. 15. `http://www.ariadne.ac.uk/issue15/robots`