Datenanalyse

Albert Weichselbraun*, Philipp Kuntschik und Sandro Hörler

Optimierung von Unternehmensbewertungen durch automatisierte Wissensidentifikation, -extraktion und -integration

https://doi.org/10.1515/iwp-2020-2119

Zusammenfassung: Unternehmensbewertungen in der Biotech-Branche, Pharmazie und Medizintechnik stellen eine anspruchsvolle Aufgabe dar, insbesondere bei Berücksichtigung der einzigartigen Risiken, denen Biotech-Startups beim Eintritt in neue Märkte ausgesetzt sind. Unternehmen, die auf globale Bewertungsdienstleistungen spezialisiert sind, kombinieren daher Bewertungsmodelle und Erfahrungen aus der Vergangenheit mit heterogenen Metriken und Indikatoren, die Einblicke in die Leistung eines Unternehmens geben. Dieser Beitrag veranschaulicht, wie automatisierte Wissensidentifikation, -extraktion und -integration genutzt werden können, um (i) zusätzliche Indikatoren zu ermitteln, die Einblicke in den Erfolg eines Unternehmens in der Produktentwicklung geben und um (ii) arbeitsintensive Datensammelprozesse zur Unternehmensbewertung zu unterstützen.

Schlagworte: Company Valuation, Deep Web, Web Intelligence, Biotech companies, Knowledge acquisition, Knowledge extraction, Slot filling

Abstract: Performing company valuations within the domain of biotechnology, pharmacy and medical technology is a challenging task, especially when considering the unique set of risks biotech start-ups face when entering new markets. Companies specialized in global valuation services, therefore, combine valuation models and past experience with heterogeneous metrics and indicators that provide insights into a company’s performance. This paper illustrates how automated knowledge discovery, extraction and data fusion can be used to (i) obtain additional indicators that provide insights into the success of a company’s product development efforts, and (ii) support labor-intensive data curation processes.

Keywords: Company Valuation, Deep Web, Web Intelligence, Biotech companies, Knowledge acquisition, Knowledge extraction, Slot filling

Optimisation des évaluations d’entreprises grâce à l’identification, l’extraction et l’intégration automatisées des connaissances

Résumé: Les évaluations d’entreprises dans les secteurs des biotechnologies, des produits pharmaceutiques et des technologies médicales posent un défi, en particulier si l’on considère les risques uniques auxquels les startups biotechnologiques sont exposées lorsqu’elles entrent sur de nouveaux marchés. Les entreprises qui se spécialisent dans les services d’évaluation mondiale combinent donc des modèles d’évaluation et l’expérience passée avec des métriques et des indicateurs hétérogènes qui donnent un aperçu de la performance d’une entreprise. Cet article illustre comment l’identification, l’extraction et l’intégration automatisées des connaissances peuvent être utilisées pour (i) identifier des indicateurs supplémentaires qui donnent un aperçu du succès d’une entreprise dans le développement de produits et (ii) pour soutenir les processus de collecte de données à forte intensité de main-d’œuvre pour l’évaluation de l’entreprise.

Les mots-clés: Évaluation d’entreprise, Deep Web, Web Intelligence, Entreprises de biotechnologie, Acquisition de connaissances, Extraction de connaissances, Remplissage de fentes
Einleitung

Unternehmensbewertungen liefern unabhängige Hinweise zum wirtschaftlichen Wert von Startups, ihren Produkten und Technologien. Die Durchführung solcher Bewertungen erfordert umfassende Kenntnisse über die Produkt-Pipelines, die Finanzierung, Lizenzverträge, das Management und die Erfolgsmetriken des Unternehmens. Das Sammeln dieser Datenpunkte braucht erhebliche Ressourcen und Investitionen in Datenerfassungs- und Verarbeitungsprozessen.

Die vorgestellte Forschungsleistung wurde zusammen mit einem auf Unternehmensbewertungen spezialisierten Schweizer Unternehmen durchgeführt und konzentriert sich auf die Steigerung der Effizienz und Effektivität dieser Datenerfassungs- und Verarbeitungsprozesse durch die Entwicklung von Komponenten des maschinellen Lernens, die

1. strukturierte Daten über die Produktpipeline eines Unternehmens (d.h. Produkte und Medikamente, die an klinischen Studien teilnehmen) automatisch identifizieren und aus behördlichen Plattformen wie zum Beispiel die vom U.S. Gesundheitsministerium betriebene ClinicalTrials.gov Seite, dem EU-Register für klinische Studien oder der WHO-Plattform für klinische Studien extrahieren.

2. Änderungen in der Unternehmensführung und der zugehörigen Kontaktdaten erfassen, die über die Webpräsenz eines Unternehmens kommuniziert werden.

Das automatisiert gesammelte externalierte Wissen wird dann in die Datenbanken des Industriepartners integriert.

Forschungsstand

Deep Web Data Acquisition behandelt Strategien zur Er schließung jener Webdaten, die sich hinter proprietären Suchschnittstellen verbergen. Es wird angenommen, dass der Umfang der Daten im Deep Web dem des durch Suchmaschinen zugänglichen Surface Webs übersteigt (Khelghi et al., 2013; Noor et al., 2011). Eine umfassende Studie von (Chang et al., 2004) aus dem Jahr 2004 schätzt, dass das Deep Web über 450.000 Web-Datenbanken umfasst, von denen mehr als drei Viertel die strukturierten Inhalte über eine Abfrageschnittstelle zugänglich machen. He et al., 2013 betrachten die Generierung von Abfragen, das Filtern leerer Ergebnisse und die URL-Deduplierung als wichtige Unterprobleme bei der Gewinnung von Daten aus Deep Web-Quellen und stellen Techniken vor, mit denen diesen Herausforderungen begegnet werden kann.

Named Entity Linking beschreibt eine Methode der Wissensextraktion, die es ermöglicht Nennungen von Produkten-, Medikamenten- und Firmen in textuellen Inhalten zu identifizieren und diese auf eine Zielontologie abzubilden. Gangemi, 2013 bietet einen Überblick über Werkzeuge zur Wissensextraktion, zusammen mit spezifischen Anwendungen für die Erkennung und Verknüpfung von Entitäten. C. Wang et al., 2012 schlagen ein auf Graphen basierendes Modell (MentionRank) vor, das sich zunutze macht, dass homogene Gruppen von Entitäten häufig in ähnlichen Dokumenten vorkommen. Dieser gemeinsame Kontext ermöglicht es, Begriffe wie „Apple“ oder „HP“ besser zu disambiguieren, wenn diese in Dokumenten mit einem informationstechnologischen oder unternehmerischen Schwerpunkt vorkommen. Neuere Ansätze wie (Weichselbraun et al., 2019) verwenden maschinelles Lernen für verschiedene Teilaufgaben dieses Verlinkungsprozesses, einschließlich der Generierung von Namensvarianten und der eigentlichen Verlinkung.

Slot-Filling geht einen Schritt weiter, indem es die im Text vorhandenen Entitäten und deren Eigenschaften auf eine Datenbank abbildet. Dabei stützt es sich auf die zuvor extrahierten Entitäten und weist ihnen spezifische Rollen (z.B. chiefExecutiveOfficerOf, phoneNumberOf etc.) zu, sodass mehrerer Datenpunkte (z.B. Name, Adresse, Telefonnummer etc.) zu einem einzigen Datensatz kombiniert werden können. Da Slot Filling die Ergebnisse von mehreren Wissensextraktionsstrukturen vereint, können sich kleine Ungenauigkeiten vervielfachen, was in der Praxis zu einer hohen Fehleranfälligkeit führt. Dieses Problem spiegelt sich auch in den Ergebnissen von Wettbewerben, die sich mit dem automatischen Slot Filling beschäftigen. So erreichte z.B. bie dem TAC 2017 Cold Start Slot Filling Task selbst das beste System nur einen F1-Wert unter 20 Prozent (Lim et al., 2017). Beim weniger komplexen TAC KBP 2013 English Slot Filling Task erzielte das von Roth et al., 2014 entwickelte System den höchsten F1-Wert von 37,3 Prozent. Das Gewinnersystem konnte mittels eines Distant Supervision Ansatzes, der in Kombination mit einer abfrageorientierten Methode zur Relationsextraktion angewandt wurde, die Identifikation von Beziehungen als Klassifikationsaufgabe modellieren.

Durch eine Reduktion der Aufgabenkomplexität ist es möglich die Genauigkeit von Slot Filling erheblich zu steigern. (X. Wang et al., 2019) erreichen zum Beispiel einen F1-Wert von über 76,5 Prozent für eine temporale Slot Filling Aufgabe, die sich darauf beschränkt, die Präsidenten und Präsidentinnen eines Landes sowie deren Legislatur-
perioden aus Nachrichtenartikeln abzuleiten. Analog zu diesem Vorgehen setzt die in diesem Papier vorgestellte Forschung Slot Filling erfolgreich für abgegrenzte, gut definierte Extraktionsaufgaben ein, die für die Erstellung von Unternehmensbewertungen von Biotech-Unternehmen unerlässlich sind.

**Methode**

Wesentliche Teile des World Wide Web sind für Suchmaschinen schwer zu erfassen, da sie sich aus dynamischen Webseiten zusammensetzen, die auf themenspezifischen Datenbanken zurückgreifen und nur über bestimmte Portalseiten zugänglich sind. Diese Deep Web-Datenquellen sind jedoch oft für industrielle Anwendungen von hoher Bedeutung, da sie qualitativ hochwertige, redigierte und aktuelle Inhalte zu verschiedenen Themen wie wissenschaftlichen Publikationen, Bibliothekskatalogen und medizinischen Studien enthalten. Der im vorliegenden Beitrag skizzierte Prozess kombiniert Datenerfassungstechniken wie Deep Web Mirroring und gezieltes Web Crawling mit Methoden der Wissensextraktion und Datenfusion, um die aus diesen Quellen extrahierten Daten in bestehenden Domain-Ontologien und Datenbanken zu integrieren.

Abbildung 1 zeigt den entwickelten Prozess, der vorhandenes, kodiertes Domänwnissen des Industriepartners als Hintergrundwissen berücksichtigt. Die Deep Web Mirroring-Komponente (1) nutzt unter anderem dieses Wissen zur Optimierung der Reihenfolge, in welcher Daten in Deep Web-Ressourcen abgefragt werden. Sobald eine klinische Studie abgerufen wurde, sorgt das Recognyze Named Entity Linking Framework (Weichselbraun et al., 2015) in Verbindung mit Slot Filling (3) für die Extraktion von relevanten Inhalten wie zum Beispiel den an der vorliegenden klinischen Studie beteiligten Unternehmen und Forschungseinrichtungen, dem Titel der Studie, den Ansatz, Einschlusskriterien und den Studienfortschritt. Anschließend werden die Inhalte normalisiert und mit den bestehenden Datensätzen in die Ontologie des Industriepartners kombiniert und mit diesen integriert (4).

Eine zweite, parallele Datenerfassungspipeline verwendet gerichtete Web-Crawls (2), um Firmen-Webseiten auf Änderungen in der Managementstruktur und den zugehörigen Kontaktdaten zu überwachen. Auch hier nützt das System die bereits erwähnten Methoden der Wissensextraktion (3), um die Personal- und Adressdaten des Unternehmens zu identifizieren und integrieren diese in die Domänenontologie des Industriepartners (4).

Beide Slot Filling-Aufgaben profitieren stark von der Applikation des Hintergrundwissens, das aus der Domänenontologie des Industriepartners und anderen domänspezifischen Wissensquellen bezogen wird. Insbesondere wird durch diese Herangehensweise Folgendes ermöglicht:

1. Die Substitution von Entitätenidentifikation, die den Typ einer Entität (zum Beispiel Person, Produkt, etc.) identifiziert durch Entitätenverlinkung. Diese verknüpft erkannte Entitäten (d.h. Produkte, Organisationen, Namen von Schlüsselpersonen, etc.) mit den entsprechenden Einträgen der Wissensdatenbank.
und hat sich im definierten Kontext als wesentlich ge-
nauer und weniger fehleranfällig erwiesen;
2. Die Vereinheitlichung verschiedener Beschreibungen
klinischer Phasen innerhalb der Deep-Web-Ressour-
cen auf standardisierte Werte;
3. Die exakte Identifizierung von Telefonnummern, da
den Extraktionskomponenten Hintergründe über die
korrekte Zusammensetzung je nach Land und Region
zur Verfügung stehen.

Durch die Beschränkung der verwendeten Slot Filling
Komponente auf sehr spezifische Rollen wie clinicalPha-
seOf, performedBy, isPhoneNumberOf etc. konnte deren
Genauigkeit signifikant verbessert werden.

Ergebnis

Mit diesem Prototyp konnten über 480.000 klinische Stu-
dien erschlossen werden, von denen sich 61.865 erfolg-
reich mit Unternehmen aus der Ontologie des Industrie-
partners verknüpft ließen. Insgesamt 35.757 klinische
Studien (7,4 % aller abgerufenen Studien) wurden zum
Zeitpunkt dieser Analyse als abgeschlossen gekennzeich-
net.

Zusätzlich wurden durch fokussierte Web-Crawls über
1,3 Millionen Webseiten von 53.335 verschiedenen Unter-
nehmen erfasst und aus diesen über 233.000 Datensätze
von Schlüsselpersonen aus dem Biotech-Bereich extra-
hiert und mit bestehenden Datensätzen kombiniert. Diese
Werte veranschaulichen sehr gut, dass ein manuelles Ku-
rationsverfahren wirtschaftlich unrentabel wäre und de-
monstrieren das Potenzial der vorgestellten Methoden im
kommerziellen Umfeld.

Die im Rahmen des Projekts entwickelten Innovationen
liefern umfassende und zeitnahe Information über
klinische Studien und Unternehmen zu sehr niedrigen
Kosten und ermöglichen es dem Industriepartner, seine
Bewertungen von Biotech- und Pharma-Startups auf eine
noch solide Grundlage zu stellen. Der Datensatz zu kli-
nischen Studien wurde in die Domänenontologie des In-
dustriepartners integriert und dessen Kunden als Teil ei-
er Investoren- und Unternehmensentwicklungsdaten-
bank zur Verfügung gestellt.

Danksagung: Das Projekt DISCOVER (https://www.fhgr.
ch/discover) wurde durch Innosuisse finanziert. Diese Un-
terstützung ermöglichte die Integration von Methoden aus
der aktuellen Forschung der Fachhochschule Graubünden
in kommerzielle Prozesse eines innovativen Schweizer
Unternehmens.

Literatur

Chang, K. C.-C., He, B., Li, C., Patel, M., & Zhang, Z. (2004).
Structured Databases on the Web: Observations and Implicati-
ons. SIGMOD Rec., 33(3), 61–70. https://doi.org/10.1145/1031570.
1031584.

Gangemi, A. (2013). A Comparison of Knowledge Extraction Tools for
the Semantic Web. In P. Cimiano, O. Corcho, V. Presutti, L. Hol-
link, & S. Rudolph (Eds.), The Semantic Web: Semantics and Big
Data (pp. 351–366). Springer Berlin Heidelberg.

He, Y., Xin, D., Ganti, V., Rajaraman, S., & Shah, N. (2013). Crawling
Deep Web Entity Pages. Proceedings of the Sixth ACM Inter-
national Conference on Web Search and Data Mining, 355–364.
https://doi.org/10.1145/2433396.2433442.

Khelghati, M., Hiemstra, D., & Van Keulen, M. (2013). Deep Web
Entity Monitoring. Proceedings of the 22Nd International Confe-
ence on World Wide Web, 377–382. https://doi.org/10.1145/24
87788.2487946.

Lim, S., Kwon, S., Lee, S., & Choi, J. (2017). UNIST SAIL System for TAC
2017 Cold Start Slot Filling. TAC.

Noor, U., Rashid, Z., & Rauf, A. (2011). Article: A Survey of Automatic
Deep Web Classification Techniques. International Journal of
Computer Applications, 19(6), 43–50.

Roth, B., Barth, T., Wiegand, M., Singh, M., & Klakow, D. (2014).
Effective Slot Filling Based on Shallow Distant Supervision
Methods. ArXiv:1401.1158 [Cs]. http://arxiv.org/abs/1401.1158
[13.7.2020].

Wang, C., Chakrabarti, K., Cheng, T., & Chaudhuri, S. (2012).
Targeted disambiguation of ad-hoc, homogeneous sets of
named entities. Proceedings of the 21st International Confe-
ence on World Wide Web, 719–728. https://doi.org/10.1145/21
87836.2187934.

Wang, X., Zhang, H., Li, Q., Shi, Y., & Jiang, M. (2019). A Novel Unsu-
supervised Approach for Precise Temporal Slot Filling from Incom-
plete and Noisy Temporal Contexts. The World Wide Web Confe-
rence, 3328–3334. https://doi.org/10.1145/3308558.3313435.

Weichselbraun, A., Kuntschik, P., & Brasoveanu, A. M. P. (2019).
Name Variants for Improving Entity Discovery and Linking.
Second conference on Language, Data and Knowledge (LDK
2019), Leipzig, Germany.

Weichselbraun, A., Streiff, D., & Scharl, A. (2015). Consolidating
Heterogeneous Enterprise Data for Named Entity Linking and
Web Intelligence. International Journal on Artificial Intelligence
Tools, 24(2).

Prof. Dr. habil Albert Weichselbraun
Fachhochschule Graubünden
Schweizerisches Institut für
Informationswissenschaft
Pulvermühlestrasse 57
7000 Chur
Schweiz
albert.weichselbraun@fhgr.ch

Albert Weichselbraun ist Professor für Informationswissenschaft an
der Fachhochschule Graubünden. Er promovierte 2004 am Institut für
Informationswirtschaft der Wirtschaftsuniversität Wien und war zwischen 2005 und 2011 als wissenschaftlicher Assistent und Privatdozent am Departement für Informationsverarbeitung und Prozessmanagement der Wirtschaftsuniversität Wien tätig, wo er 2010 habilitierte. Seine Forschungsinteressen fokussieren sich auf Big Data, Text und Data Mining, Künstliche Intelligenz und die Anwendung von Semantischen Technologien im Kontext des Natural Language Processings.

Philipp Kuntschik, MSc
Fachhochschule Graubünden
Schweizerisches Institut für Informationswissenschaft
Pulvermühlestrasse 57
7000 Chur
Schweiz
phlipp.kuntschik@fhgr.ch

Philipp Kuntschik ist seit 2014 als Wissenschaftlicher Mitarbeiter im Forschungsfeld Data Analytics des Schweizerischen Instituts für Informationswissenschaft tätig. Seine Forschung gilt semantischen Technologien, Text und Data Mining sowie Künstlicher Intelligenz.

Sandro Hörler, BSc
Fachhochschule Graubünden
Schweizerisches Institut für Informationswissenschaft
Pulvermühlestrasse 57
7000 Chur
Schweiz
sandro.hoerler@fhgr.ch

Sandro Hörler arbeitet nach seinem Bachelorabschluss in Computerwissenschaften im Jahr 2016 als Software Engineer mit PLSQL und C#. Seit 2017 ist er Wissenschaftlicher Mitarbeiter am Schweizer Institut für Informationswissenschaft. Er hat lange Projekterfahrung und war involviert in das Requirement Engineering, die Entwicklung und Optimierung von Komponenten des maschinellen Lernens und der Web-Intelligence. Seine fachlichen Schwerpunkte sind das maschinelle Lernen, Datenanalyse und Web-Intelligence.