Syntaktische Identifizierung von Lobbyeinflüssen auf die EU-Datenschutz-Grundverordnung

Exposé zur Studienarbeit

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

bearbeitet von: Frank Bicking

Betreuer: Prof. Dr. Ulf Leser
Prof. Dr. Björn Scheuermann
Florian Tschorsch

Datum: 12. Dezember 2013

1 Motivation

Auf der Ebene der Europäischen Union entsteht derzeit eine neue Datenschutz-Grundverordnung mit dem Ziel, den Datenschutz innerhalb aller Mitgliedsstaaten zu harmonisieren. Sie soll eine im Jahr 1995 veröffentlichte Richtlinie 95/46/EG ablösen, die derzeit die Grundlage für die Datenschutzgesetze in den einzelnen Staaten bildet und in Deutschland 2001 durch ein Bundesgesetz (BGBl. I Nr. 23 S. 904) ausgestaltet wurde.

Sowohl im Vorfeld als auch begleitend zu einem solchen Gesetzgebungsverfahren fordert die Europäische Kommission staatliche und nichtstaatliche Einrichtungen und Organisationen, Unternehmen, aber auch interessierte Bürger anhand von vorgegebenen Fragenkatalogen zur Stellungnahme auf. Zum Thema Datenschutz fanden seit Mai 2009 in mehreren Wellen derartige Konsultationen statt [EC09b] [EC09a] [EC10a] [EC10b] [EC11], die zum Teil nichtöffentlich waren. Bis zum Januar 2011 gingen insgesamt 532 öffentlich einsehbare Dokumente ein. Anfang 2012 wurde durch die Kommission schließlich ein 119 Seiten starker Gesetzentwurf (COM(2012) 11 final. 2012/0011 (COD)) fertiggestellt und zur weiteren Ausgestaltung an das Europäische Parlament sowie den Rat der Europäischen Union übergeben. Federführend ist der parlamentarische Ausschuss für Binnenmarkt und Verbraucherschutz.

Im Februar 2013 machte der Wiener Jurastudent Max Schrems darauf aufmerksam, dass auf dem Entwurfstext basierende Änderungsvorschläge zahlreicher Interessenvertreter durch einzelne Parlamentsabgeordnete zum Teil eins zu eins übernommen worden sind [Sch13]. Der Journalist Richard Gutjahr schuf daraufhin gemeinsam mit Marco Maas vom Unternehmen OpenDataCity die Crowdsourcing-Plattform LobbyPlag[1]. Crowdsourcing hat sich als Begriff für das Auslagern aufwändiger Aufgaben an ehrenamtlich tätige Internetnutzer eingebürgert. Vorbild waren ähnliche Websites, auf denen die Dissertationen öffentlicher Personen auf mögliche Plagiate untersucht wurden. LobbyPlag möchte herausfinden, inwieweit das neue Datenschutzgesetz nicht aus der Feder von Parlamentariern stammt, sondern von Organisationen und Unternehmen mit unterstellten Eigeninteressen. Dazu werden im Gesetzestext Stellen gesucht, die sich mit den Vorschlägen aus Lobbypapieren decken. Bis dato nicht durchleuchtet wurde hingegen die Bedeutung der Konsultationen für den ursprünglichen Kommissionsentwurf.

[1] http://lobbyplag.eu/

2 Zielstellung

Ziel der Studienarbeit ist es, mit Hilfe von textanalytischen Methoden zu überprüfen, inwieweit die Konsultationen einen Einfluss auf den Entwurfstext der EU-Kommission hatten. Dazu ist eine Software mit graphischer Benutzeroberfläche zu entwickeln, die Dokumente einliest, vergleicht und Fundstellen geeignet visualisiert.

3 Verwandte Arbeiten

Aus rein technischer Sicht kann die Suche nach inhaltlichen Übernahmen als ein Plagiaterkennungsproblem angesehen werden, bei dem von außen an den Gesetzgeber herangetragene Stellungnahmen und Empfehlungen als potentielle Quellen aufgefasst werden, während Gesetzentwurf, eingebrachte Änderungen, oder das fertige Gesetz das Zieldokument bilden. Der Begriff des Plagiats ist dabei insofern schwierig, eine juristische Bewertung außen vor gelassen, als dass eine Übernahme von Texten von Seiten der Interessenvertreter natürlich gewollt war und nicht zwingend negativ zu bewerten ist.

An Parlamentarier versandte Lobbypapiere stellten dem Wortlaut des Kommissionsvorschlags meist gewünschte eigene Fassungen mit entsprechenden Begründungen gegenüber [Lob13b]. Im Unterschied zu den von LobbyPlag nachgewiesenen Kopien ganzer Textpassagen [Lob13a] kann bei den Stellungnahmen an die EU-Kommission, um die es in dieser Arbeit gehen soll, jedoch nicht davon ausgegangen werden, Übereinstimmungen in diesem Ausmaß zu finden. Der Öffentlichkeit lag vor 2012 kein entsprechender Entwurf vor, auf den sich Einsendungen hätten stützen können. Ebensowenig waren diese Kommentare im Rahmen der Konsultationen als konkrete Gesetzestexte formuliert, die einfach hätten übernommen werden können. Vielmehr ist zu erwarten, dass es beim Versuch der Beantwortung der Frage, welche Dokumente prägend für die Entstehung des Kommissionsentwurfs waren, auf einzelne Formulierungen oder Begriffe ankommen wird, die nur in bestimmten Dokumenten auftauchen. Ähnlichkeiten in der Wortwahl könnten dann als Hinweis darauf interpretiert werden, dass die inhaltlichen Forderungen dieser Dokumente ihren Weg in die EU-Verordnung gefunden haben. Bei dem vorliegenden Text kann daher nicht von einem möglichen Plagiat gesprochen werden. Dennoch können Methoden aus der Plagiatsuche eine Grundlage bilden.

Aufgaben der Plagiaterkennung lassen sich in zwei Katagorien aufteilen [Pot+10]. Beim intrinsischen Erkennen von Plagiaten liegt lediglich ein verdächtiges Dokument vor, anhand dessen Charakteristika bestimmt werden soll, ob es von mehr als einem Autor verfasst wurde. Hierzu kann der Text in Abschnitte unterteilt und diese einer

vergleichenden Analyse des Schreibstils unterzogen werden [ZES06]. Alternativ lässt sich intrinsische Plagiaterkennung als binäres Klassifikationsproblem auffassen, das Dokumente in Plagiate und Nichtplagiate zu unterteilen versucht. Dies kann beispielsweise mit Hilfe von Support Vector Machines und geeigneter Kernel geschehen [Bao+04].

Extrinsische Plagiatsuche dagegen beschäftigt sich mit der Identifizierung plagiierter Textstellen in einem Dokument anhand einer Sammlung von Quellen. Potthast et al. [Pot+10] beschreiben die im Rahmen eines Wettbewerbs häufig eingesetzte Vorgehensweise, für die zu vergleichenden Dokumente Mengen von N-Grammen zu ermitteln und anhand dieser Fingerabdrücke eine Ähnlichkeitsfunktion zu berechnen, um Kandidaten für eine nähere Untersuchung zu bestimmen. N-Gramme entstehen durch überlappendes Zusammenfassen von jeweils N Buchstaben oder Wörtern. Dieser Ansatz ist erfolgversprechend, da unabhängige Dokumente eine geringe Menge an gemeinsamen N-Grammen besitzen, während bei Plagiaten davon ausgegangen werden kann, dass bestimmte Formulierungen oder sogar ganze Sätze direkt aus der Quelle übernommen worden sind [BCR09]. Aus den in Frage kommenden Quellen werden anhand der N-Gramme konkrete Fundstellen ausfindig gemacht, wobei aus Effizienzgründen häufig ein invertierter Index zum Einsatz kommt. Hilfreiche Techniken sind das Sortieren der Wörter, um N-Gramme abzugleichen, die sich lediglich in der Reihenfolge unterscheiden, das Entfernen von Stoppwörtern, sowie Lemmatisierung bzw. Stemming.

Mit Hilfe von Heuristiken werden diese Fundstellen anschließend zusammengefasst, um komplette Abschnitte als Plagiat identifizieren zu können. Ob es der im Rahmen dieser Arbeit betrachtete Textkorpus entgegen der Erwartungen zulassen wird, Fundstellen auf ganze Sätze oder gar Absätze auszuweiten, wird sich erst im Laufe der Implementierung zeigen. Existierende Plagiaterkennungssoftware würde die Schwelle an dieser Stelle womöglich von vornherein zu hoch ansetzen und im schlechtesten Fall keine Ergebnisse zurückliefern.

Lyon et al. [LMD01] erläutern Grundlagen zu N-Grammen, führen Containment als Maß für die Übereinstimmung ein und diskutieren ihre Ergebnisse auf Basis von Trigrammen anhand mehrerer Korpora. Barrón-Cedeño und Rosso [BCR09] analysieren Plagiatsuche mit N-Grammen für unterschiedliche Werte von N und kommen für ihre Testdaten zu dem Ergebnis, dass $N = \{2, 3\}$ die besten Ergebnisse liefere, wobei Bigramme den Recall und Trigramme die Precision begünstigen. Monogramme finden einen hohen Anteil an Plagiaten, erzeugen aber aufgrund der Tatsache, dass häufig das gesamte Vokabular eines Satzes im jeweiligen Quelldokument vorkomme, zu viele False Positives. Für $N \geq 4$ leide der Recall aufgrund umgestellter oder neu hinzugefügter Wörter. In einem späteren Artikel [BC+10] stellen die Autoren einen ebenfalls auf N-Grammen basierenden, aber

lediglich die Länge der Wörter berücksichtigenden Ansatz vor, der den Aufwand im Hinblick auf Zeit- und Platzkomplexität deutlich verringert. Stamatatos [Sta11] nutzt N-Gramme von Stoppwörtern, die sich besonders zum Aufdecken von Plagiaten eignen, bei denen die Satzstruktur übernommen, aber einzelne Wörter durch Synonyme ersetzt wurden.

Alternativ zur inhaltsbasierten Auswertung können Plagiate, speziell im Bereich akademischer Arbeiten, anhand einer Zitationsanalyse erkannt werden, die Dokumente auf Vorkommen und Reihenfolge referenzierter Werke hin untersucht [GB10].

4 Vorgehensweise

4.1 Textmaterial

Im Gegensatz zu Texten, bei denen im Verdachtsfall nicht angegebene Quellen zunächst recherchiert werden müssen, sind die in diesem Fall die in Frage kommenden Dokumente bekannt und zu einem Großteil öffentlich verfügbar. Die PDF-Dateien können auf den Seiten der jeweiligen Konsultationen heruntergeladen werden, die im ersten Abschnitt verlinkt sind. Im Rahmen dieser Arbeit beschränkt sich die Auswertung auf die Kommentare, die während der Konsultationen der EU-Kommission eingegangen sind. Nicht berücksichtigt werden hingegen Lobbyanträge an Parlamentarier, die sich meist direkt auf den Kommissionstext bezogen. Derartige Abschnitte durch die Software unterscheiden zu lassen ist demzufolge nicht erforderlich.

Zieldokument bildet der abschließende Gesetzesvorschlag der EU-Kommission.Obwohl dieser in allen 23 gegenwärtig innerhalb der EU anerkannten Amts- und Arbeitssprachen vorliegt [EC12], steht dabei die englischsprachige Fassung im Fokus, da ein Großteil der Stellungnahmen nach erster Sichtung ebenfalls auf Englisch eingereicht wurde. Bei der Implementierung hat die vorliegende Sprache Auswirkungen für die Wahl von Stoppwortlisten, sowie die Regeln, die, falls genutzt, bei Lemmatisierung, Stemming und Satzerkennung zum Einsatz kommen.

4.2 Einlesen und Indexieren

Die Dokumente liegen im Portable Document Format (PDF) vor und werden zunächst in ein einfaches Textformat konvertiert. PDF wird durch ISO/IEC 32000-1:2008 spezifiziert und für Publikationen auf EU-Plattformen in Version 1.5 verwendet [EC13]. Ob die spätere Darstellung innerhalb der Software mit den Formatierungen der ursprünglichen PDF-Dokumente erfolgen kann, wie Überschriften, Einrückungen oder Fußnoten, oder

sich auf den Text wird beschränken müssen, hängt von den Möglichkeiten frei verfügbarer PDF-Lösungen für Java ab, die im Vorfeld der Implementierung evaluiert werden sollen. Für das Zieldokument könnte alternativ auf existierende HTML-Version zurückgegriffen werden[2].

Nach dem Einlesen werden die Dokumente vorbereitend durch geeignete Datenstrukturen indexiert. Dabei kann eine Java-Library wie Apache Lucene[3] zum Einsatz kommen.

4.3 Suchvorgang

Zu einem gegebenen Zeitpunkt wird immer nur ein Quelldokument mit dem Zieldokument verglichen werden können. Der Vergleich erfolgt anhand von wortbasierten N-Grammen, die für das jeweilige Quelldokument ermittelt werden. Im Kontext von Lucene werden derartige Wort-N-Gramme auch als Shingles bezeichnet [Ing11]. Die N-Gramme werden als Queries an das Zieldokument gestellt und Fundstellen entsprechend erfasst. Bei diesen Anfragen soll eine gewisse Unschärfe gelten und Übereinstimmungen trotz im Satz hinzugekommener oder umgestellter Wörter erkannt werden können. Zur weiteren Verbesserung der Ergebnisse wird abhängig von den Möglichkeiten der eingesetzten Java-Libraries Lemmatisierung oder Stemmimg zum Einsatz kommen.

4.4 Gewichtung

Wichtiger Bestandteil ist eine Bewertung der Fundstellen. N-Gramme sollen eine stärkere Gewichtung erfahren, wenn sie ausschließlich im betrachteten Quelldokument oder nur in einer Handvoll von Texten vorkommen. Derartige Fundstellen sind entscheidend für die Zielstellung dieser Arbeit, da sie sich als Hinweis darauf interpretieren lassen, welche Dokumente stärkeren inhaltlichen Einfluss auf die Gesetzgebung gehabt haben könnten. Im Unterschied dazu sollen N-Gramme schwächer gewichtet werden, die in einem großen Prozentsatz von Dokumenten des Gesamtkorpus auftreten und daher aus beliebigen Quellen stammen können. Im hier verwendeten Material wären dies beispielsweise „processing of personal data“ oder „European Data Protection Board“. Gleiches gilt für gängige Idiome wie „in case of“ oder „in so far as“, wobei diese Beispiele durch Herausfiltern von Stoppwörtern von vornherein ausgeschlossen werden können.

[2] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:EN:HTML
[3] http://lucene.apache.org/

4.5 Visualisierung

Innerhalb der Software werden sämtliche verfügbaren Quelldokumente aufgelistet. Durch Auswahl eines Dokuments werden die Fundstellen entsprechend ihrer Gewichtung innerhalb des permanent sichtbaren Zieldokuments hervorgehoben. Angesichts dessen Länge von 119 Seiten erscheint es sinnvoll, für den aktuellen Vergleich irrelevante Seiten oder Absätze auszublenden, oder ein schnelles Scrollen zu Fundstellen etwa über eine zusätzliche Auswahlliste zu ermöglichen. Auch eine Gegenüberstellung beider Dokumente ist denkbar. Durch die Wahl von der Darstellung unabhängiger interner Datenstrukturen sind zu einem späteren Zeitpunkt Verbesserungen der Visualisierung möglich. Denkbar wäre zudem ein Export der Ergebnisse, etwa in Form von annotiertem XML.

Die Auflistung der Dokumente ließe sich optional um eine Funktion ergänzen, die anhand der Fundstellen ein Ranking vornimmt und die Liste entsprechend sortiert. Zusätzlich könnten untereinander ähnliche Dokumente durch Clustering basierend auf ihren N-Grammen gruppiert werden [MKM05].

5 Auswertung

Während Textkopien aus Stellungnahmen an das Europaparlament bereits durch LobbyPlag nachgewiesen wurden, kann das Ergebnis in diesem Fall auch negativ ausfallen. Für die in dieser Arbeit relevante Fragestellung, welche Konsultationen der EU-Komission eine Auswirkung auf den Gesetzestext hatten, liegen keine vorab annotierten Daten vor. Daher können die üblichen Maßzahlen des Information Retrieval wie Precision oder Recall nicht ermittelt werden.

Literatur

[Bao+04] Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, Hai-Yan Liu und Xiao-Di Zhang. „Semantic sequence kin: A method of document copy detection“. In: *Advances in Knowledge Discovery and Data Mining*. Springer, 2004, S. 529–538.

[BC+10] Alberto Barrón-Cedeño, Chiara Basile, Mirko Degli Esposti und Paolo Rosso. „Word length n-Grams for text re-use detection“. In: *Computational Linguistics and Intelligent Text Processing*. Springer, 2010, S. 687–699.

[BCR09] Alberto Barrón-Cedeño und Paolo Rosso. „On automatic plagiarism detection based on n-grams comparison“. In: *Advances in Information Retrieval*. Springer, 2009, S. 696–700.

[EC09a] EC (European Commission). *Consultation on the legal framework for the fundamental right to protection of personal data*. Dez. 2009. URL: http://ec.europa.eu/justice/newsroom/data-protection/opinion/090709_en.htm (besucht am 01. 05. 2013).

[EC09b] EC (European Commission). *Review of the data protection legal framework*. Mai 2009. URL: http://ec.europa.eu/justice/newsroom/data-protection/opinion/090501_en.htm (besucht am 01. 05. 2013).

[EC10a] EC (European Commission). *Consultation on the future European Union (EU) - United States of America (US) international agreement on personal data protection and information sharing for law enforcement purposes*. März 2010. URL: http://ec.europa.eu/justice/newsroom/data-protection/opinion/100128_en.htm (besucht am 01. 05. 2013).

[EC10b] EC (European Commission). *Stakeholder consultation: meeting on the review of the EU's data protection regulatory framework*. Juli 2010. URL: http://ec.europa.eu/justice/newsroom/data-protection/events/100701_en.htm (besucht am 01. 05. 2013).

[EC11] EC (European Commission). *Consultation on the Commission's comprehensive approach to personal data protection in the European Union*. Jan. 2011. URL: http://ec.europa.eu/justice/newsroom/data-protection/opinion/101104_en.htm (besucht am 01. 05. 2013).

[EC12] EC (European Commission). *Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). COM(2012) 11 final*. Jan. 2012. URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52012PC0011:en:NOT (besucht am 01. 05. 2013).

[EC13] EC (European Commission). *PDF (Portable File Format)*. Aug. 2013. URL: http://ec.europa.eu/ipg/standards/document/pdf/index_en.htm (besucht am 12. 08. 2013).

[GB10] Bela Gipp und Jöran Beel. „Citation based plagiarism detection: a new approach to identify plagiarized work language independently“. In: *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. ACM. 2010, S. 273–274.

[Ing11] Grant Ingersoll. *What's a shingle in Lucene parlance?* Dez. 2011. URL: http://searchhub.org/2010/12/17/whats-a-shingle-in-lucene-parlance/ (besucht am 02. 05. 2013).

[LMD01] Caroline Lyon, James Malcolm und Bob Dickerson. „Detecting short passages of similar text in large document collections“. In: *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. 2001, S. 118–125.

[Lob13a] LobbyPlag. *Comparison of Amendments ans Lobby Proposals*. 2013. URL: http://lobbyplag.eu/influence (besucht am 12. 05. 2013).

[Lob13b] LobbyPlag. *Documents*. 2013. URL: http://lobbyplag.eu/docs (besucht am 12. 05. 2013).

[MKM05] Yingbo Miao, Vlado Keselj und Evangelos Milios. „Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering“. In: *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM. 2005, S. 357–358.

[Pot+10] Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein und Paolo Rosso. „Overview of the 2nd international competition on plagiarism detection“. In: *Notebook Papers of CLEF* 10 (2010).

[Sch13] Max Schrems. *„Forum Shopping“ für die IT-Industrie?* Feb. 2013. URL: http://www.europe-v-facebook.org/IMCO_pub_de_ON.pdf (besucht am 01. 05. 2013).

[Sta11] Efstathios Stamatatos. „Plagiarism detection using stopword n-grams“. In: *Journal of the American Society for Information Science and Technology* 62.12 (2011), S. 2512–2527.

[ZES06] Sven Meyer Zu Eissen und Benno Stein. „Intrinsic plagiarism detection“. In: *Advances in Information Retrieval*. Springer, 2006, S. 565–569.