Datenschutzorientierte Analyse, Indizierung und Suche von Dokumenten in Sozialen Internetanwendungen[1]

Sergej Zerr[2]

Abstract: Obwohl es kaum denkbar ist, dass jemand private Informationen wie das Geburtsdatum oder eine private Fotosammlung einer unbekannten Person auf der Straße mitteilt, werden dennoch im Internet solche persönlichen Daten Tag für Tag von Benutzern öffentlich zugänglich gemacht. Nach Bekanntgabe solcher Informationen hat der Benutzer weder Einfluss darauf, wo und wie lange sie gespeichert werden, noch Kenntnis darüber, wer Zugang zu den Daten hat. Besonders viele Daten dieser Art werden in sozialen Netzen geteilt. In dieser Arbeit beschäftigen wir uns damit, Verfahren und Modelle zu entwickeln, die auf der einen Seite dem Benutzer erlauben, vertrauliche Daten sicher und effizient zu indizieren und danach zu suchen, auf der anderen Seite den Benutzer automatisch auf die Vertraulichkeit der Daten aufmerksam machen.

1 Einführung

Das heutige Web bietet Benutzern eine Vielzahl von Applikationen zur Veröffentlichung und zum Austausch von Bildern, Texten und anderen Informationen. Dies resultierte in der Entwicklung leistungsstarker Werkzeuge zur Veröffentlichung und gemeinsamen Nutzung von nutzergenerierten Inhalten und persönlichen Dokumenten. Um effektiven Informationsaustausch und Suche für die stark anwachsende Datenmenge zu gewährleisten, wurden Suchinfrastrukturen und Datenmanagementsysteme sowohl in dem Bereich Industrie, als auch im Web-Umfeld entwickelt. Zusätzlich verbinden Applikationen für reale oder virtuelle Arbeitsgemeinschaften Gruppen von gleichgesinnten Nutzern auf der ganzen Welt. Die Leistungsfähigkeit solcher Werkzeuge und die Einfachheit ihrer Anwendung führen jedoch häufig zu leichtfertigem Umgang mit vertraulichen Informationen. Dies macht eine Berücksichtigung von Datenschutzmechanismen bereits in der Systemdesignphase notwendig.

In dieser Arbeit beschreiben wir Lösungen zum Schutz der Privatsphäre der Benutzer in gemeinschaftlich genutzten Systemen. Insbesondere schlagen wir Lösungen für sichere Informationsspeicherung, Indizierung und den Austausch von Dokumenten vor. Dabei fokussieren wir uns auf: (1) Effiziente und sichere Indizierung von und Suche in gemeinsam verwendeten Dokumenten, (2) Skalierbare Analyse der inhaltlichen Vielfalt für große Dokumentenkorpora, sowie (3) Datenschutzorientierte Inhaltsanalyse von gemeinsam genutzten Inhalten mit automatischen Methoden zur Unterscheidung zwischen öffentlichen und potenziell vertraulichen Inhalten.

[1] Englischer Titel der Dissertation: Privacy Preserving Content Analysis, Indexing and Retrieval for Social Search Applications

[2] Forschungszentrum L3S, Leibniz Universität Hannover, Deutschland;
Web and Internet Science (WAIS) Research Group, University of Southampton, UK;
zerr@L3S.de, s.zerr@soton.ac.uk

Zunächst stellen wir *ZERBER+R* vor - ein System, das die Indizierung vertraulicher Dokumente ermöglicht und dabei den Informationsverlust in dem Index je nach Parametereinstellung unterschiedlich stark begrenzt und verhindert. Unser System ermöglicht effiziente und sichere Suchvorgänge, sowie die Auswahl der relevantesten Dokumente unter der Berücksichtigung der Zugriffsrechte für den Benutzer, ohne dass dabei Informationen über diese Dokumente für potentielle Angreifer sichtbar werden. Des Weiteren entwickelten wir effiziente Algorithmen zur Inhaltsanalyse von großen Dokumentensammlungen um ihre Themenvielfalt abschätzen zu können. Dies kann, unter anderem, zur Partitionierung einer Indexdatei hinsichtlich der Zugangsberechtigungen einzelner Benutzergruppen verwendet werden um die Leistungsfähigkeit des Systems weiter zu steigern. Schließlich präsentieren wir ein Framework, das den Benutzer automatisch bei der Auswahl adäquater Datenschutzeinstellungen für seine Dokumente und insbesondere Bilder unterstützt. Dabei untersuchen wir die Verwendbarkeit von ausgewählten Text- und Bildinformationen zur Bildklassifizierung im Kontext des Schutzes der Privatsphäre. Unsere Experimente zeigen, dass unsere Modelle, unter Benutzung sowohl textueller als auch visueller Merkmale, qualitativ hochwertige Klassifikationsergebnisse liefern.

Die in dieser Arbeit entwickelten Methoden und effizienten Algorithmen ermöglichen sowohl die direkte Anwendung in verschiedene datenschutzrelevante Nutzungsszenarien und Web-Anwendungen wie Flickr, Facebook oder Twitter, als auch die Integration in hochskalierbare öffentliche und industrielle Suchsysteme. Darüber hinaus eröffnen unsere theoretischen Erkenntnisse eine Reihe von möglichen Forschungsrichtungen in den Bereichen Indexeffizienz, Zugriffsrechteverwaltung und automatische Klassifikation von persönlichen Daten.

2 Effiziente und sichere Indizierung von und Suche in gemeinsam verwendeten Dokumenten

Dieses Kapitel beschäftigt sich mit der effizienten Indizierung und Suche vertraulicher unstrukturierter Informationen (Textdokumente) die sowohl innerhalb als auch zwischen Arbeitsgruppen ausgetauscht werden [Zea]. Solche Gruppen befinden sich oft innerhalb eines Unternehmens, können aber auch unternehmensübergreifend entstehen. Dabei gibt es typischerweise zur Überwachung und Verwaltung der Zugriffe keine zentrale Dienststelle, der alle beteiligten Unternehmen ihre Dokumente anvertrauen würden. Dagegen können die Mitarbeiter innerhalb eines Unternehmens den zur Verfügung stehenden Authentifizierungsmechanismen vertrauen. Solche Arbeitsumgebungen sind üblich für große private und staatliche Unternehmen sowie Universitäten und finden sich auch im sozialen Web. Eine ideale Lösung zur Indizierung vertraulicher Dokumente würde die Änderungen der Zugriffsrechte auf die Dokumente augenblicklich berücksichtigen und in den Suchresultaten widerspiegeln. Des weiteren würde eine ideale Lösung keine Informationen über die Dokumente liefern auch im Falle eines unautorisierten Lesens oder potentieller Übernahme durch einen Angreifer. Darüber hinaus würde eine ideale Lösung eine Suchanfrage eines autorisierten Benutzers genauso effizient und akkurat beantworten, wie ein herkömmlicher Index [MRS08], so wie er in den modernen Suchmachienen zum Einsatz kommt und in

der Abbildung 1 dargestellt ist. Hier wird für jedes in der Dokumentensammlung vorkommende Wort (Martha, Layoff, ImClone) eine *Postingliste* der Dokumente, die dieses Wort beinhalten angehängt. Diese Liste ist sortiert nach Relevanz der einzelnen Dokumente im Bezug auf das entsprechende Wort, sodass der Server einfach den Kopf der Postingliste an den Benutzer liefern muss. Leider lassen sich die Dokumenteninhalte dabei auch, je nach im Index gespeicherten Informationen, weitgehend wiederherstellen.

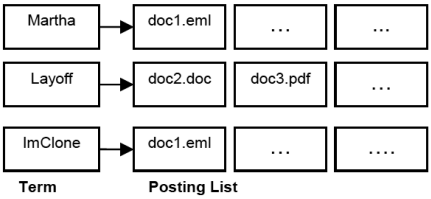


Abb. 1: Eine herkömmliche Indexdatei

Eine Implementierung der Idealen Lösung ist in der Praxis zwar nicht möglich, doch werden wir in diesem Kapitel eine weitgehende Annäherung vorstellen, bei der der Grad des Informationsverlustes in dem Index je nach Parametereinstellung unterschiedlich stark begrenzt werden kann und die sonst den obengenannten Zugriffsschutzanforderungen gerecht wird. Insbesondere soll die in diesem Kapitel entwickelte Indexingstruktur die Beantwortung folgender Fragen eines potentiellen Angreifers ihm Falle einer Indexübernahme erschweren: (a) Es soll weder möglich sein den Inhalt der Dokumente anhand der Indexdatei zu rekonstruieren, noch statistisch Hinweise über die Wortverteilung daraus ableiten zu können. (b) Es sollen keine Informationen über die aggregierte Worthäufigkeitsverteilung in einer Dokumentensammlung abgeleitet werden können. (c) Es soll nicht möglich sein festzustellen, ob ein bestimmtes Wort in einem Dokument oder einer Dokumentensammlung überhaupt vorkommt.

2.1 Vorgestellte Beiträge im Bereich Vertrauliche Suche

Die in diesem Kapitel vorgestellten Beiträge umfassen mehrere Aspekte der effizienten Indizierung und Suche vertraulicher Dokumente wie folgt:

i Als Erstes wird *r-confidentiality* als ein Maß für den Informationsverlust über die zugriffsgeschützten Dokumente aus einem Dokumentenindex entwickelt.

Definition 1 (*r*-Confidential Indexing) *Ein Index ist* r-*confidential genau dann, wenn*

$$\frac{P(X|B,I)}{P(X|B)} \leq r. \tag{1}$$

Hier ist r der Informationsgewinn (Wahrscheinlichkeitsfaktor) darüber, dass sich das Wort w in dem Dokument d befindet (X) nach der Betrachtung des Indexes (I), unter der Annahme, dass der Angreifer bereits ein Hintergrundwissen (B) abgeleitet aus einer ähnlichen Dokumentenmenge oder Untermenge besitzt. Diese Lösung bietet maximalen Schutz, wenn die Indexdatei keine zusätzlichen Informationen über X bietet, verglichen mit dem Hintergrundwissen des Angreifers.

ii Als Zweites wird *Zerber* vorgestellt - ein Dokumentenindex für vertrauliche Informationen. *Zerber* basiert auf Verteilung der mit **k** out of **n** verschlüsselten [Sh79] Teile des Indexes auf zentrale und nicht zwangsweise im vollen Umfang vertrauenswürdige Server. Dieses Verschlüsselungsverfahren ist sicher, solange nicht mehr als **k**-1 Server gleichzeitig durch Angreifer kompromittiert sind. Um statistischen Attacken Widerstand zu leisten, entwickelten wir in diesem Kapitel einen neuartigen Mechanismus zum Zusammenführen von Postinglisten (Postinglistmerging) innerhalb des Indexes. Dieser Mechanismus hat einen minimalen Einfluss auf die Effizienzeigenschaften des Indexes. Bei der Suche garantiert er eine schnelle Berücksichtigung dynamischer Änderungen der indizierten Dokumente, sparsame Benutzung der Bandbreite, benötigt keine Schlüsselverwaltung und liefert die Suchergebnisse vergleichbar effizient gegenüber dem herkömmlichen Index, der sich bei heutigen Suchmaschinen im Einsatz befindet. Postinglistmerging wird in der Abbildung 2 dargestellt. Dabei werden beispielhaft die Postinglisten für “Martha” und “Layoff” zusammengeführt und anschließend einzelne Elemente verschlüsselt, sodass jedes einzelne Element (und somit das Dokument) nicht mehr eindeutig einem Wort zugeordnet werden kann.

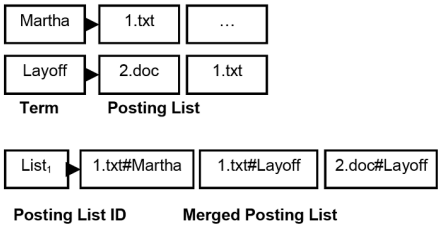


Abb. 2: Eine Indexdatei mit Zusammengeführten Postinglisten

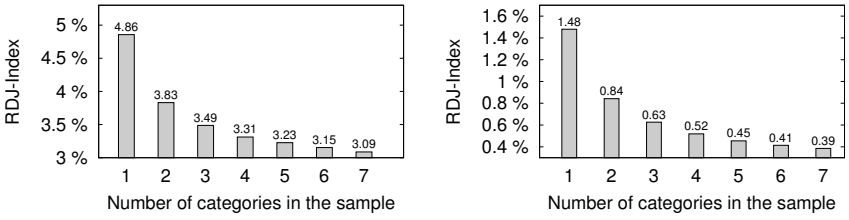
iii Der *Zerber* Ansatz wird zu *Zerber+R* erweitert, um die effiziente Auswahl der relevantesten Dokumenten von einem *r*-confidential Index zu ermöglichen. Es wird ein neuartiges Modell vorgestellt, das den Informationsabfluss an potentielle Angreifer bei der Auswahl der relevantesten Dokumente minimiert. Wir stellen dafür einen neuartigen Transformationansatz vor, bei dem die spezifische Verteilung der Relevanzwerte der einzelnen Worte nicht mehr erkennbar gemacht wird. Die damit transformierten Werte können auch auf einem nicht vertrauenswürdigen Server gespeichert werden, der dann in der Lage ist effizient und genau die Dokumentensuchanfragen zu beantworten und statistische Auswertungen der Inhalte der Dokumente verhindert.

iv Schließlich werden die vorgestellten Methoden auf zwei Datensätze aus der realen Welt angewendet und evaluiert.

Unsere Experimente belegen, dass die von uns entwickelten Suchmodelle auf der einen Seite wahrscheinlichkeitsbasierte Datenschutzgarantien liefern, auf der anderen Seite sparsam mit dem Datenverkehr umgehen, ein Minimum an Verwaltungsaufwand erfordern und sich dabei an Effizienz kaum von den herkömmlichen Suchmodellen unterscheiden.

3 Skalierbare Analyse der inhaltlichen Vielfalt für große Dokumentenkorpora

Generell haben sich die invertierten Indizes als sehr effizient für die Dokumentensuche erwiesen. Doch gilt dies nur unter der Annahme, dass alle indizierten Dokumente für alle Benutzer der Suchmaschine zugänglich sind. Falls jedoch Zugriffskontrolle auf einzelne Dokumentengruppen erforderlich ist, muss die Postingliste, nach für den Sucher zugänglichen Elementen, zeitaufwändig durchsucht werden. Eine Vorpartitionierung des Indexes unter Berücksichtigung der Zugriffsrechte kann die Anzahl der unnötigen Überprüfungen drastisch reduzieren. Dieses Modell hat zwei Extrema. Auf der einen Seite gibt es einen gemeinsamen Index, wo für jedes Dokument die Zugriffsberechtigung überprüft werden muss. Auf der anderen Seite gibt es Indizes für jede mögliche Zugriffsbeschränkung, deren Zahl wegen aller möglichen Kombinationen sehr hoch sein kann. Im ersten Fall benötigt der Index relativ wenig Speicherplatz und technische Pflege, die Ausführung dagegen benötigt sehr viel Zeit. Im zweiten Fall ist es umgekehrt, die Ausführungszeiten sind vernachlässigbar, der Speicherplatzbedarf jedoch steigt dramatisch an, da die Anzahl der einzelnen Indizes so hoch wie die Anzahl der Arbeitsgruppen sein muss. Clusteringverfahren können eine Lösung des Partitionierungsproblems bei sich stark überlappenden Arbeitsgruppen bieten. Viele Clusteringtechniken benötigen eine Vorabschätzung der Clusteranzahl um effektiv arbeiten zu können [Ja10]. In diesem Kapitel entwickeln wir ein Verfahren zur effizienten Analyse der Inhaltsdiversität in großen Dokumentenmengen [DSZ]. Der dabei berechnete Diversitätsgrad hat großes Potenzial für den Einsatz bei der Parameterabschätzung für die Indexpartitionierung wie die Abbildung 3 unserer Experimente zeigt. Hier haben wir für die Datensätze Reuters (Nachrichten) und Flickr (Fotoannotationen) die Anzahl der Kategorien (Abszisse) variiert und den Diversitätsgrad gemessen.



(a) Reuters RCV1 (b) Flickr Groups

Abb. 3: Diversitätsgrad korreliert mit der Anzahl der Kategorien (Cluster) in den Datensätzen

Die Bestimmung des Diversitätsgrades ist in mehreren Disziplinen seit Jahrzehnten ein wichtiges Thema. Dieser Wert gibt Einblicke in einige Eigenschaften wie die Religionszugehörigkeit oder die politische Ausrichtung verschiedener sozialer Gruppen oder Bevölkerungschichten eines Landes. In der Ökologie wird geringe Biodiversität als ein Maß für die Gesundheitsgefährdung eines biologischen Systems angesehen [Si49, St07]. Aber auch in der Informatik gewinnt die Diversitätsberechnung, zum Beispiel für Recommendersysteme, an Wichtigkeit, um dem Benutzer den Überblick über die vorhandene Datenmenge zu verschaffen. Leider sind die eingesetzten Berechnungsmethoden so komplex, dass sie nur auf kleine Objektmengen, wie Resultate einer Suchanfrage oder Lebensformen in einem Biosystem, angewendet werden können.

3.1 Vorgestellte Beiträge im Bereich Skalierbare Datenanalyse

In diesem Kapitel befassen wir uns mit der Diversitätsanalyse großer Dokumentenmengen. Um das Komplexitätsproblem dabei zu lösen, entwickeln wir zwei Algorithmen zur effizienten Berechnung des Diversitätsgrades.

i *SampleDJ* ist ein stichprobenbasiertes Verfahren, welches das Berechnungsproblem unabhängig von der Datensatzgröße in einer konstanten Zeit löst, die nur von dem Diversitätsgrad abhängt. Dabei werden Stichproben aus einem Datensatz gezogen und gemittelt bis die Stoppbedingung eintritt. Die Bedingung basiert auf der Tschebyscheff-Ungleichung, die eine obere Schranke der Wahrscheinlichkeit angibt, bei der die Stichprobe von dem Erwartungswert (Diversitätsgrad in unserem Falle) in einem bestimmten Intervall abweicht.

ii *TrackDJ* basiert auf "Min-wise hash" [Br00], einem Verfahren aus der Familie des "Locality Sensitive Hashing" und löst das Berechnungsproblem garantiert in linearer Zeit im Bezug auf die Datensatzgröße. Zunächst werden dabei die einzelnen Objekte mit Hilfe von "Min-wise hash" effizient in kleine Gruppen (Buckets) partitioniert. Dabei liegt der Wahrscheinlichkeitswert, dass sich zwei Objekte in einem Bucket ähnlich sind, sehr nah an deren Jaccard-Koeffizient, der als Basis für die Diversitätsgradberechnung in dieser Arbeit genommen wird.

iii Die vorgestellten Methoden wurden sowohl auf einem synthetischen Datensatz, als auch auf mehreren Datensätzen (zum Beispiel auf 1 Mio. Publikationstitel aus DBLP oder 144 Mio. Bilder aus Flickr) aus der realen Welt angewendet und evaluiert. So brauchen unsere Algorithmen einige Minuten (SampleDJ) bis wenige Tage (TrackDJ) für einen 20 Mio. Datensatz, während die herkömmliche Berechnungsmethode über ein Jahr lang dauern würde.

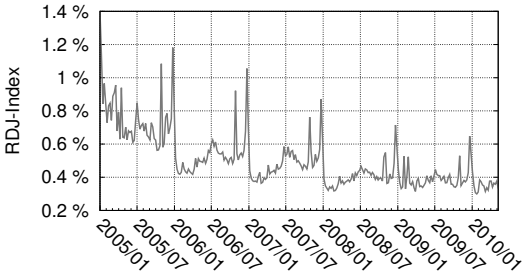


Abb. 4: Ähnlichkeitsgrad von Fotoannotationen über den Zeitraum 2005 - 2010

Die Experimente belegen die hohe Verwendbarkeit unserer Algorithmen. Im sozialen Web können unsere Methoden verwendet werden um zum Beispiel die Diversität von Benutzergruppen zu bestimmen. Darüber hinaus lassen sich unsere Methoden auf Datenströme anwenden, um deren zeitliche Entwicklung zu analysieren und interessante Informationen wie wiederkehrende Ereignisse und Trends zu entdecken. So zeigt die Abbildung 4 den von uns gemessenen mittleren Ähnlichkeitsgrad von Fotoannotationen über den Zeitraum 2005 - 2010, wo wiederkehrende Ereignisse wie "Weihnachten" als Ausschläge zu erkennen sind. Schließlich zeigten wir, dass unsere Methoden zum Abschätzen der Parameter der Partitionierung großer Datenmengen geeignet sind.

4 Datenschutzorientierte Inhaltsanalyse

In diesem Kapitel analysieren wir, wie Verfahren des maschinellen Lernens dazu benutzt werden können, den Vetraulichkeitsgrad von textuellen und visuellen Informationen zu bestimmen [Zeb]. Dabei können die in dieser Arbeit entwickelten Verfahren den Benutzer bei seinen Entscheidungen über die Veröffentlichung von potentiell privatsphäregefährdende Informationen unterstützen. Der Vetraulichkeitsgrad eines einzelnen Dokumentes kann von vielen Kriterien abhängen. Die Inhalte, die typischerweise von Webbenutzern veröffentlicht werden, sind sehr heterogen. Dazu gehören Texte, aber auch Bilder und Videos. Auch die Themen der Inhalte sind verschieden. Ideale Lösung bei der Bestimmung des Vetraulichkeitsgrades wäre (a) Bestimmen von Inhaltstyp und Kontext – ist das Foto auf dem Strand aufgenommen, oder in einem Gebäude – und (b) Bestimmung des Vetraulichkeitsgrades in diesem Kontext. Abbildung 5 zeigt Ergebnisse der Suche in Flickr mit der Suchanfrage “ronaldo”. Die Bilder in der linken Spalte stellen das professionelle Leben des Fußballstars vor, die in der rechten Spalte das private. Die Reihenfolge in der linken Spalte entspricht der in Originalergebnisliste. Die rechte Spalte wurde mit Hilfe unserer Methoden automatisch angeordnet.

(a) Top-3 public photos (b) Top-3 private photos

Abb. 5: Top-3 Suchergebnisse mit den Flickr IDs für die Anfrage “Ronaldo” (a) Originalreihenfolge, und (b) Anordnung mit unseren Suchmethoden (Stand: Okt. 2010).

Da die automatische Ausführung mehrerer Schritte zur Ermittlung des Vertraulichkeitsgrades fehlerbehaftet sein kann, stellen wir eine Methode vor, die auf Verfahren des maschinellen Lernens basiert, um den Vetraulichkeitsgrad direkt anhand der textuellen oder visuellen Eigenschaften des Objektes zu bestimmen. Solche Merkmale sind zum Beispiel das Vorkommen von Gesichtern in einem Bild oder die Farbverteilung. Lange und parallele Kanten deuten auf künstliche Umgebungen hin, kurze und chaotisch angeordnete Kanten dagegen auf Fotos der Natur. Andere effektive Merkmale sind sogenannte SIFT-Features [Lo04] - im Foto zu findende geometrischen Objekte. Auch textuelle Informationen, wie Tags, Titel oder Beschreibung der Bilder erwiesen sich als hilfreich in diesem Kontext. Solche Systeme können dem Benutzer helfen Gefahren für seine Privatsphäre rechtzeitig zu erkennen bevor er die Inhalte veröffentlicht.

4.1 Vorgestellte Beiträge im Bereich Vertrauliche Inhaltsanalyse

In diesem Kapitel stellen wir Verfahren vor, um den Vetraulichkeitsgrad von visuellen und textuellen Informationen zu bestimmen. Darüber hinaus entwickeln wir Algorithmen für privatsphäreorientierte Suche und Diversifizierung der Resultate.

i Unser erstes Ziel ist es, Techniken und Verfahren zu entwickeln, die es dem Benutzer ermöglichen, den Vetraulichkeitsgrad von eigenen Fotos anhand deren visuellen und textuellen Informationen zu bestimmen, um darauf basierend weitere Entscheidungen bezüglich deren Veröffentlichung zu treffen. Dafür entwickelten wir zunächst ein Spiel, bei dem Probanden die von uns zufällig ausgewählten Fotos aus der populären Fototauschplattform Flickr als ”privat” oder ”öffentlich” annotieren konnten. Der annotierte Datensatz wurde anschließend veröffentlicht und wird derzeit für Experimente in Forschungsprojekten an verschiedenen Instituten weltweit genutzt. Die Analyse der Annotationen auf Interrater-Reliabilität ergab, dass sich die Probanden im Großen einig über die Zugehörigkeit einzelner Fotos waren, gemessen mit Fleiss Kappa [Gw10] von 0,6. Wir benutzten dieses Feedback später beim Trainieren von Support Vector Maschinenmodellen [MRS08], basierend auf den dabei extrahierten visuellen und textuellen Merkmalen. Die dabei entstehenden Modelle können Informationen aus beliebigen Bildern evaluieren und dem Benutzer Empfehlungen zur Veröffentlichung generieren. Solche Warnsysteme können direkt in sozialen Applikationen eingesetzt werden, zum Beispiel als Browserplugins.

ii Als zweiten Aspekt entwickeln wir die privatsphäreorientierte Dokumentensuche. Heutige Suchsysteme erlauben keine gezielte Suche nach privaten Informationen. Zum einen soll es für den Benutzer möglich sein Dokumente, die über ihn im Web (von ihm, oder von Dritten) veröffentlicht sind frühzeitig zu identifizieren, um eventuell mit entsprechenden Providern Kontakt aufzunehmen. Zum anderen kann die Suchanfrage in Bezug auf den Privatanteil für eine Maschine mehrdeutig erscheinen. Bei der Suchanfrage “Ronaldo” zum Beispiel, könnte sich der Benutzer für Bilder interessieren, die das Privat- oder Berufsleben des Fußbalspielers zeigen. In dieser Arbeit benutzen wir die Ergebnisse der automatischen Bildklassifikation, um gezielt nach privaten Informationen über ein bestimmtes Thema zu suchen.

iv Schließlich stellen wir eine Methode der privatsphäreorientierten Diversifikation von Suchresultaten vor. Diversifikation von Suchresultaten wurde ausgiebig in der Literatur behandelt [GS, Cl] und wird eingesetzt um dem Benutzer einen Überblick über mögliche Facetten des Resultats zu geben. Mit ähnlicher Motivation versuchen wir mit unseren Methoden das Risiko der Unzufriedenheit des Suchenden in Bezug auf den Privatanteil der Suchergebnisse zu minimieren.

Unsere Experimente zeigen, dass es mit maschinellem Lernen möglich ist, das abstrakte Konzept der Privatsphäre in Bildern zu erfassen. Darüber hinaus beobachteten wir, dass verschiedene Aspekte dieses Konzepts bei Benutzern individuell verschieden gewichtet werden. Dies ergibt interessante Richtungen für zukünftige Forschung im Bereich Personalisierung.

5 Zusammenfassung

Die Entwicklungen in der Internet- und Netzwekinfrastruktur zusammen mit der Popularität von Web 2.0 Plattformen haben es sowohl für individuelle Nutzer, als auch für soziale und berufliche Benutzergruppen, möglich und einfach gemacht, große Datenmengen auszutauschen. Diese Entwicklung erfordert auch Mechanismen zur sicheren und gleichzeitig effizienten Veröffentlichung der Daten, wobei intelligente Warnsysteme ein Bestandteil bilden müssen. Sicherheit und Datenschutz in modernen Informationssystemen gewinnen an Bedeutung nicht nur in Unternehmen und sozialen Netzen, sondern auch im Bereich Cloudcomputing, in dem Datenverwaltung und Datenverteilung ohne Garantieen im Bezug auf Datenschutz nicht denkbar sind.

5.1 Ergebnisse

In dieser Dissertation haben wir uns drei wichtigen Herausforderungen gestellt: (1) Effiziente und sichere Indizierung und Suche auf gemeinsam verwendeten Dokumenten, (2) Skalierbare Analyse inhaltlicher Vielfalt für große Dokumentenkorpora, sowie (3) Datenschutzorientierte Inhaltsanalyse von gemeinsam genutzten Inhalten mit automatischen Methoden zur Unterscheidung von öffentlichen und potenziell vertraulichen Inhalten. Die Ergebnisse dieser Dissertation können in vielen Bereichen, wie Cloudcomputing oder Datenstromverarbeitung direkt angewendet werden und eröffnen gleichzeitig weitere Forschungsrichtungen.

5.2 Ausblick

Im Bereich Indizierung und Suche von vertraulichen Informationen, können Verfahren entwickelt und untersucht werden, die sowohl für strukturierte als auch teilstrukturierte Daten geeignet sind. Die Adaptierung von modernen Information Retrieval Mechanismen, wie Suche nach Synonymen, unscharfe Suche und Anfragevorschläge, bleibt ebenfalls eine Herausforderung für vertrauliche Daten.

Im Bereich der skalierbaren Inhaltsanalyse können unsere Verfahren eingesetzt werden, um einen Index durch effiziente Abschätzung der Clusteranzahl effektiv zu partitionieren. Eine andere vielversprechende Richtung ist die Analyse großer Datenströme, bei der unsere Samplingverfahren direkt eingesetzt werden können.

Schließlich, im Bereich privatsphäreorientierte Inhaltsanalyse, können Personalisierungstechniken eingesetzt werden, da das Empfinden von Privatsphäre stark subjektiv ist und sich zudem über die Zeit ändern kann. Obwohl unsere Experimente sich auf Bilder beschränkt haben, ist es denkbar die Methoden nach entsprechender Justierung auch auf Dokumente anderer Datentypen, wie Video, Audio oder Blogeinträge anzuwenden. Auch interdisziplinäre, kulturübergreifende Studien könnten unter Benutzung unserer Methoden zusätzliche Einblicke in das Empfinden von Privatsphäre in verschiedenen Personengruppen ermöglichen.

Literaturverzeichnis

[Br00] Broder, Andrei Zary: Identifying and Filtering Near-Duplicate Documents. In: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. COM '00, Springer-Verlag, London, UK, S. 1–10, 2000.

[Cl] Clarke, Charles L.A.; Kolla, Maheedhar; Cormack, Gordon V.; Vechtomova, Olga; Ashkan, Azin; Büttcher, Stefan; MacKinnon, Ian: Novelty and Diversity in Information Retrieval Evaluation. SIGIR '08, S. 659–666.

[DSZ] Deng, Fan; Siersdorfer, Stefan; Zerr, Sergej: Efficient Jaccard-based Diversity Analysis of Large Document Collections. CIKM '12, S. 1402–1411.

[GS] Gollapudi, Sreenivas; Sharma, Aneesh: An axiomatic approach for result diversification. WWW'09, S. 381–390.

[Gw10] Gwet, Kilem: Handbook of Inter-Rater Reliability. Advanced Analytics, LLC, second. Auflage, 2010.

[Ja10] Jain, Anil K.: Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8):651 – 666, 2010.

[Lo04] Lowe, David: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV), 60(2):91–110, January 2004.

[MRS08] Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich: Introduction to Information Retrieval. Cambridge University Press, New York, 2008.

[Sh79] Shamir, Adi: How to share a secret. Communications of the ACM, 22:612–613, November 1979.

[Si49] Simpson, E. H.: Measurement of Diversity. Nature, 163, 1949.

[St07] Stirling, Andy: A general framework for analysing diversity in science, technology and society. Journal of The Royal Society Interface, 4(15):707–719, 2007.

[Zea] Zerr, Sergej; Olmedilla, Daniel; Nejdl, Wolfgang; Siberski, Wolf: Zerber+R: Top-k Retrieval from a Confidential Index. EDBT '09, S. 439–449.

[Zeb] Zerr, Sergej; Siersdorfer, Stefan; Hare, Jonathon; Demidova, Elena: Privacy-aware Image Classification and Search. SIGIR '12, S. 35–44.

Sergej Zerr wurde 1977 in Karaganda (Kasachstan) geboren. Er studierte Medieninfomatik an der Fachhochschule Osnabrück (Diplom 2003) und ergänzte sein Studium durch Fokussierung auf Information Engineering an den Universitäten Osnabrück und Twente (Master of Science 2006). Inspiriert durch Fortschritte der Suchmaschinen und die Transformation des Internets zum Informationsaustauschmedium zwischen Benutzern, beschäftigte er sich bereits während seiner Masterarbeit mit dem Thema der autorisierten Suche in Zugangsbeschränkten Arbeitsumgebungen. Er vertiefte und erweiterte das Thema auf die Bereiche der sozialen Internetanwendungen als Doktorand am Forschungszentrum L3S an der Universität Hannover, wo er anschließend im Jahr 2015 promovierte. Zurzeit forscht Sergej Zerr an der Universität Southampton in Großbritannien im Forschungsteam Web and Internet Science (WAIS) in den Bereichen soziale Netzwerke und Human Computation im Eropäischen Projekt STARS4ALL (688135).