Automatisierte Recherche von Markenpiraterie im Internet

Peter Ebinger · Ulrich Pinsdorf

{peter.ebinger|ulrich.pinsdorf}@igd.fraunhofer.de
Fraunhofer-IGD
Abteilung Sicherheitstechnologie
Fraunhoferstr. 5
64283 Darmstadt

Zusammenfassung

Das Internet hat sich zu einem wichtigen Instrument beim Auf- und Ausbau von Produktmarken entwickelt. Um die Reputation einer Marke zu schützen, haben Unternehmen ein Interesse am Aufspüren von missbräuchlicher Verwendung von Logos, Plagiaten, Grauimporten und Rufmordkampagnen. Die sog. *Brand Protection* ist sehr aufwendig und in elektronischen Medien bislang nur mit hohem personellem Aufwand möglich.
Dieser Aufsatz beschreibt eine Software, die ein flexibles Überwachen von Internetinhalten auf die unerlaubte Verwendung oder Beschädigung von Markennamen, Logos, unseriösen Verkaufsangeboten, etc. hin erlaubt. Der verwendete Algorithmus untersucht Web-Inhalte in mehreren Stufen bis hin zu einer semantischen Analyse. Das beschriebene System lässt sich flexibel auf bestimmte Überwachungsaufgaben anpassen und dient damit einem "Internet-Detektiv" als effektives Werkzeug bei der Recherche.

1 Motivation

Das Internet ist ein wichtiges Instrument beim Auf- und Ausbau von Marken. Unternehmen wollen ihre Präsenz im Internet verstärken und dabei die Entwicklung ihres Markenkapitals aktiv steuern und schützen. Dazu muss die Meinungsbildung im Internet, die Verwendung eines Markennamens, der zugehörigen Logos und der Vertrieb von Plagiaten oder Grauimporten im Internet überwacht werden. Der DIHK[1] schätzt den volkswirtschaftlichen Schaden durch Produkt- und Markenpiraterie allein in Deutschland auf 29 Milliarden Euro jährlich [7]. Dazu kommt die Vernichtung von geschätzten 70.000 Arbeitsplätzen in den letzten Jahren.

So haben auch das US-Juwelierhaus TIFFANY[2] und andere Luxusartikel-Hersteller das Online-Auktionshaus EBAY[3] verklagt, da dort Fälschungen von Luxusartikeln aufgetaucht seien. Die Klage ist Teil der Anstrengungen, mit der Anbietern von Fake-Artikeln das Handwerk gelegt werden soll. Nach den Untersuchungen von TIFFANY-Spezialisten seien

[1] Deutscher Industrie- und Handelstag
[2] siehe http://www.tiffany.com/
[3] siehe http://www.ebay.de/

rund 73 Prozent der bei EBAY angebotenen Stücke Fälschungen [10]. Das Internet ist nach China und Italien mittlerweile der weltweit drittgrößte Umschlagplatz für gefälschte Luxusartikel geworden, berichtet das Wall Street Journal.

Um der Marken- und Produktpiraterie zu begegnen, haben sich verschiedene Aktionsbündnisse zusammengeschlossen. Sie alle bauen mehr oder weniger auf Aufklärung der Kunden, Abschreckung der Nachahmer durch konsequente Strafvervolgung und Kooperation der betroffenen Firmen. Beispiele für solche Bündnisse sind der Aktionskreis APM[4], EBAYS VerI Programm[5] oder die Aktion Plagiarius[6], die jährlich einen Negativpreis für besonders dreiste Produktfälschungen vergibt.

Vor dem Verhängen von Konsequenzen gegen Markenpiraten – z.B. Abmahnung, Unterlassungsverfügung, Schadensersatzforderung, Strafanzeige – ist es notwendig, konkrete Verdachtsfälle zu ermitteln. Im Umfeld des Internets wird dieses Vorgehen als *Online Brand Monitoring* bezeichnet. Beim *Product Monitoring* hingegen geht es darum, den Vertrieb von Produkten in einem globalisierten Markt, bei dem sich die regionalen Marktsegmente auflösen und verschwimmen, zu schützen. Aufgrund der hohen Kosten für eine effektive Suche im Internet werden effiziente Prüfwerkzeuge benötigt. Diese sollten es erlauben, eine Vielzahl von Websites im Internet darauf zu überprüfen, ob Namen, Inhalte und/oder das Ansehen eines Unternehmens beschädigt werden.

Spezialisierte Dienstleister bieten Brand und Product Monitoring für Kunden an, welche die Nutzung ihrer Marken überwachen und schützen wollen. Dazu durchsuchen sog. *Internet-Detektive* das World Wide Web nach Webseiten, die den Markennamen oder das Logo des Kunden enthalten oder Plagiate über das Internet vertreiben. Diese Suche selbst erfolgt manuell bzw. mit Hilfe der gängigen Suchmaschinen wie GOOGLE, YAHOO, etc. Das Ergebnis der Recherche ist meist eine detaillierte Übersicht über Markenverletzungen und die damit verbundenen Personen. Diese wird dem Kunden – meist dem geschädigten Unternehmen – ausgehändigt, der dann ggf. weitere Schritte veranlassen kann. In bestimmten Fällen tätigen die Internet-Detektive auch verdeckte Testkäufe, um die gehandelte Ware eingehend untersuchen zu können und ggf. über Beweismaterial zu verfügen.

Diese Recherchen sind mit einem hohen personellen Aufwand verbunden. Es gibt keine spezialisierten Werkzeuge, die einen Internet-Detektiv bei seiner Arbeit wirkungsvoll unterstützen. Trotzdem ist die Recherchearbeit oft gleichförmig und mit einer geringen Trefferwahrscheinlichkeit verbunden. Eine Steigerung der Effizienz des Recherchevorgangs ist also sehr wünschenswert. Die eingesetzten Überwachungs- und Analysewerkzeuge sollen es den Internet-Detektiven ermöglichen, die Webinhalte effizient zu erfassen und anhand von kundenspezifischen Kriterien zu analysieren.

Dieser Artikel macht einen Vorschlag für ein flexibles Design, das Internet-Detektive bei der Recherche effizient unterstützt; er gliedert sich wie folgt. Der folgende Abschnitt 2 definiert die Anforderungen an ein solches Analysewerkzeug. Verwandte Arbeiten werden

[4] Aktionskreis Deutsche Wirtschaft gegen Produkt- und Markenpiraterie e.V. (APM), http://www.markenpiraterie-apm.de

[5] Verifizierte Rechte Inhaber Programm (VeRI) von EBAY zum Schutz von immateriellen Rechtsgütern, http://pages.ebay.de/help/community/vero-program.html

[6] Aktion Plagiarius e.V., http://www.plagiarius.com

in Abschnitt 3 gewürdigt. In Abschnitt 4 wird der Lösungsansatz skizziert, der danach in Abschnitt 5 detailliert beschrieben wird. Die Umsetzung der Architektur und die Einbindung in den Arbeitsablauf eines Internet-Detektivs erläutert Abschnitt 6. Schließlich fasst Abschnitt 7 den Artikel zusammen und gibt einen Ausblick.

2 Zielsetzung und Anforderungen

Das Aufspüren von Markenpiraterie, Grauimporten oder auch nur unberechtigt verwendeten Markenlogos erfordert die systematische semantische Analyse von fremden Webinhalten. Unter *fremden Webinhalten* werden Portale, einzelne Webseiten, Internet-Shops und Internet-Auktionen verstanden, die von Dritten betrieben werden.

Anhand von drei Beispielszenarien soll die Bandbreite der Suchaufträge illustriert werden, die bei der Recherche vorkommen.

Beispielszenario 1: Ein kommerzieller Anbieter von Bildern im Internet möchte herausfinden, welche Bilder auf fremden Webseiten auftauchen. Dazu hat er diese vorher mit einem digitalen Wasserzeichen versehen.

Beispielszenario 2: Der Hersteller eines Produktes sucht Angebote in einem Online-Shop, deren Preis so niedrig ist, dass es sich mit einer gewissen Wahrscheinlichkeit um einen Grauimport oder gar ein Plagiat handelt.

Beispielszenario 3: In einem Online-Auktionshaus sollen Angebote gefunden werden, die ein bestimmtes Markenlogo enthalten und deren Anbieter kein Vertragshändler ist.

Diese Suchekriterien sind – verglichen mit herkömmlichen Suchmaschinen – sehr "weich", lassen sich also algorithmisch schwer fassen. Gleichzeitig ist die Recherche sehr genau auf die Bedürfnissen und Rahmenbedingungen des Auftraggebers ausgerichtet.

Aus dem Umstand, dass die Recherche verdeckt erfolgt, ergeben sich weitere Randbedingungen für die automatisierte Suche. Zum einen darf sich der Rechercheautomatismus nicht durch auffällige Signaturen als solcher verraten. Die Abfragen von Produkt-Webseiten soll auf den Betreiber wie normales Nutzerverhalten wirken. Dies ist gerade bei einem systematischen Durchsuchen nicht gegeben und verrät die Recherche. Auch die Absenderadresse ist bei häufigen Abfragen verräterisch. Aus diesem Grund wird eine stark *verteilte Architektur* angestrebt, die von einem zentralen Arbeitsplatz kontrolliert und gesteuert wird. Die Verteilung ahmt reguläres Nutzerverhalten nach und sorgt zudem für Skalierung und Lastverteilung.

Bei einem verteilten System, das eine flexible Anpassbarkeit des Programmcodes erlaubt, besteht die Gefahr eines Missbrauchs durch Dritte. Eine Kompromittierung hätte nicht nur die Sabotage der Recherche zur Folge; mit einem kompromittierten System ließen sich auch DDoS-Angriffe[7] auf beliebige Rechner durchführen. Darum ist die *Sicherheit der verteilten Subsysteme* in einem besonderen Maße zu berücksichtigen.

Zusammenfassend ergeben sich die Anforderungen an ein sicheres, verteiltes System, mit dem sich fremde Webinhalte semantisch analysieren lassen. Das Gesamtsystem soll autonom arbeiten und sich flexibel an neue Aufgabenstellungen anpassen lassen. Die Arbeit

[7] Distributed Denial of Service

des Internet-Detektivs besteht darin, aus den automatisch recherchierten Hinweisen diejenigen auszuwählen, die auch nach menschlichem Ermessen den Verdacht auf Markenpiraterie nahelegen.

3 Verwandte Arbeiten

Wesentliche Vorarbeiten zur Realisierung von sog. Web-Crawlern wurden im Forschungsbereich *Information Retrieval* geleistet [16, 17]. Suchmaschinen verwenden üblicherweise eine lexikalische Analyse beim Archivieren und der Verschlagwortung von Internetinhalten [3, 6]. Üblicherweise arbeitet man bei der inhaltsbasierten Analyse mit einer Bewertungsfunktion P, die einen Inhalt I in Bezug auf eine Anfrage A auf einen skalaren Wert $P(I, A)$ abbildet.

Beim *Focused Crawling* [5] wird versucht, gezielt die relevanten Seiten zu einem definierten Menge von Themen zu finden. Diese Themen werden dabei nicht durch Schlüsselwörter, sondern durch Beispieldokumente spezifiziert. Anstatt alle verfügbaren Webinhalte zu sammeln und zu indizieren, um jede mögliche Anfrage beantworten zu können, werden nur die Referenzen auf andere Webseiten verfolgt, die wahrscheinlich die höchste Relevanz haben und damit irrelevante Regionen des Webs vermieden. Focused Crawling reduziert dadurch den Indizierungsaufwand von Suchmaschinen, da nur als wichtig erachtete Webseiten untersucht werden.

In [11] werden *autonome, aber kooperierende Web-Crawler* untersucht. Es wird ein Algorithmus für kooperative Crawler und ein Verteilungsprotokol vorgestellt. Kooperierende Web-Crawler unterrichten sich dabei gegenseitig über Änderungen und Neuigkeiten im Internet, um ein besseres und aktuelleres Abbild des Webs zu erhalten als es mit traditionellen Abfrage-Crawlern möglich ist. Die Web-Crawler können dabei wechselseitig ihre Ergebnisse an die anderen Crawler weitergeben und von den Ergebnissen der Anderen profitieren. Der Gewinn an Leistungsfähigkeit durch Kooperation ist dabei um so größer, je mehr Crawler der Gruppe beitreten.

Inhaltsbezogene Recherche ist aus den Bereichen *Data Mining* und *Image Retrieval* bekannt. Beim Data Minig werden unstrukturiert abgelegte Daten verknüpft und Informationen abgeleitet, die in dieser Form nicht vorlagen und für den Benutzr einen Mehrwert darstellen. Systeme zum automatischen Image Retrieval [14] analysieren Bilder nach den dargestellten Inhalten. Sie abstrahieren also von der exakten digitalen Repräsentation und machen unscharfe Aussagen über den möglichen Inhalt eines Bildes.

Tim Berners-Lee, einer der Erfinder und Vordenker des Internets, schlägt eine semantische Verknüpfung von Webinhalten vor [1, 2]. Dies würde die heute gängige syntaktische Suche ablösen und zu einem semantisch durchsuchbaren Informationsnetz, einem *Semantic Web*, führen.

4 Lösungsansatz

Dieser Abschnitt gibt einen Überblick über den gewählten Lösungsansatz. In Abschnitt 2 wurden aus der Problembeschreibung bereits Anforderungen an die Lösung abgeleitet. So soll das System stark verteilt, flexibel anpassbar und sicher gestaltet sein.

Diese Anforderungen passen genau zu den Leistungsmerkmalen von *Mobile Code* Technologie. Diese erlaubt das Verteilen von autonom arbeitenden Algorithmen in einem Netzwerk. Es gibt eine Reihe von Projekten, die diese Vorteile mit den notwendigen Sicherheitsmechanismen verbinden.

Ein Kernproblem ist das Verstehen der unbekannten Webinhalte, wenigstens in soweit, dass man eine Aussage in Bezug auf die Suchanfrage machen kann. Um das strukturierte Analysieren von unstrukturierten Daten zu ermöglichen wird eine stufenweise Bearbeitung gewählt. Dazu wird der HTML-Code in drei Stufen analysiert und gefiltert: syntaktisch, semantisch und logisch.

Jeder der spezifischen Filter in den drei Filterstufen – eine Stufe wird typischerweise mehr als einen Filter enthalten – ist ein eigenständiger Algorithmus. Ein Netzwerk aus Filter- und Analyseservern kann mit einer beliebigen Zahl von Filtern ausgestattet werden. Diese lassen sich auch nachträglich in ein laufendes System integrieren.

Zum Analysieren werden die Filter der einzelnen Stufen miteinader kombiniert. Diese Kombination und die zugehörige Parametrisierung wird als *Rechercheauftrag* bezeichnet. Der Rechercheauftrag selbst ist, wie die Filter, autonomer mobiler Code; in Abschnitt 6.3 sprechen wir daher von Auftragsagenten. Abbildung 1 zeigt die einzelnen Filterstufen und deren Koordination durch einen Suchauftrag.

5 Extraktion und Aggregation

Eine wichtige Randbedingung ist, dass die Recherche üblicherweise verdeckt erfolgt. Das wiederum bedeutet, dass von nicht-kooperativen Datenquellen ausgegangen werden muss. Die Suche stützt sich also auf öffentlich zugängliche Webseiten und Online-Portale.

Serverseitig werden diese Webseiten i.d.R. dynamisch aus Datenbanken generiert. Aus Geschäftsdatenbanken für Angebote, Kundendaten, Workflow-Management-Systeme, etc. werden Daten mittels PHP, SHTML oder JSP verknüpft und dem Benutzer optisch ansprechend präsentiert. Die HTML-Dateien werden üblicherweise über die Protokolle HTTP bzw. HTTPS zum Benutzer übertragen. Einen direkten Zugriff auf die dahinter liegenden Datenbanken erhält er i.A. nicht.

Durch die Erzeugung der Webseite geht die Strukturierung der Daten verloren. Die Extraktion und Analyse der Daten muss die präsentierten Daten also wieder auf eine strukturierte Form zurückführen, damit man Schlüsse aus dem Inhalt der Webseiten ziehen kann.

5.1 Ablauf und Datenfluss

Die Akquisition von Daten und die Suche nach relevanten Ergebnissen gliedert sich in die folgenden Schritte:

1. Auswahl und Generierung der Webadressen
2. Automatisches Abfragen der Webangebote per HTTP bzw. HTTPS
3. Syntaktische Analyse der Daten
4. Semantische Analyse der Daten

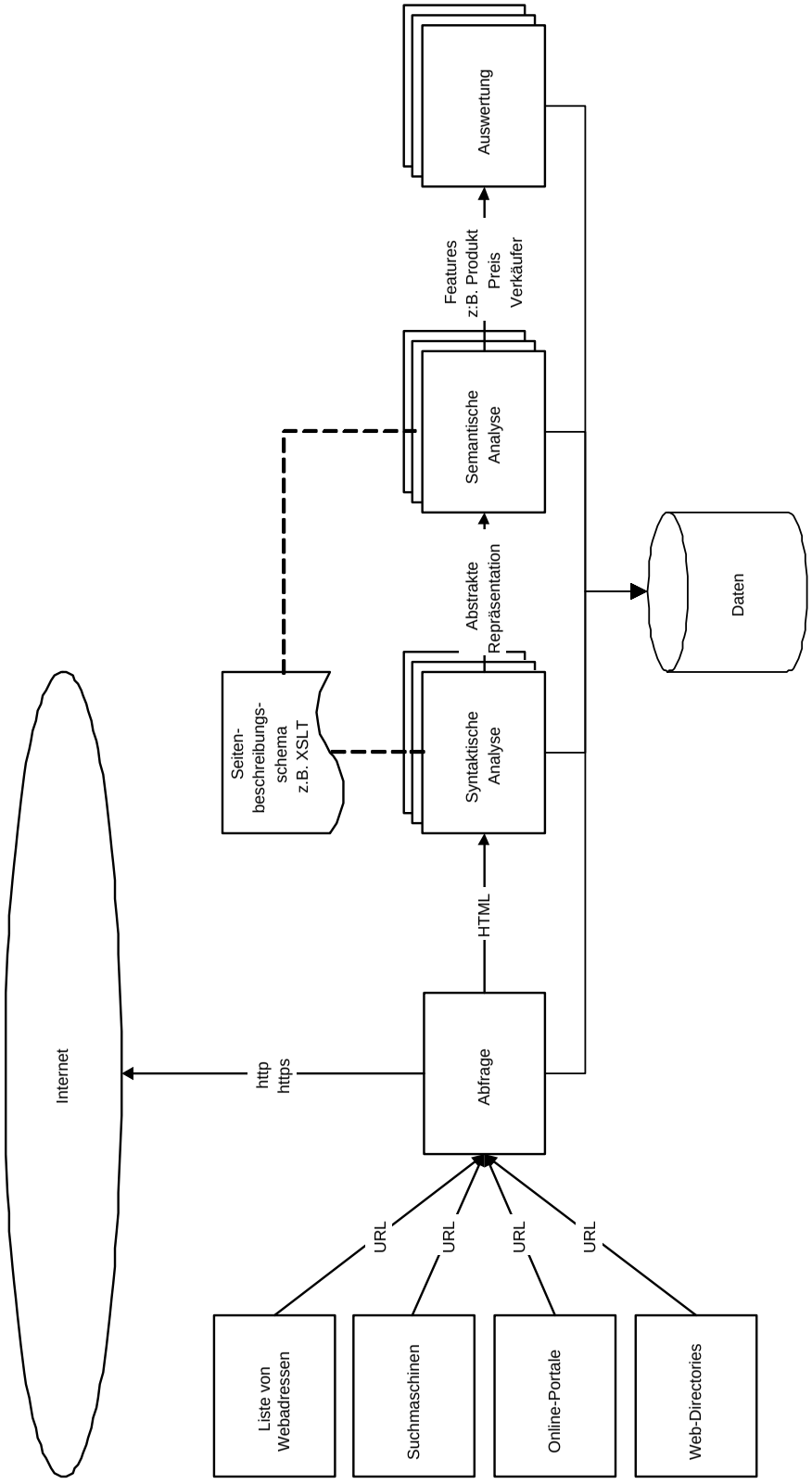

Abbildung 1: Datenfluss im Filter & Analyse-Server

5. Logische Analyse der Daten
6. Analyse, Recherche und Bewertung der relevanten Treffer durch den Internet-Detektiv

Abbildung 1 zeigt den schematischen Aufbau der Datenflüsse. Zunächst werden die zu untersuchenden Webadressen generiert und an das Abfragemodul übergeben. Die Webseiten werden dann inkl. aller Bilder, etc. aus dem Internet abgerufen und in der Datenbank abgelegt. Danach werden die Daten in drei Stufen zunächst syntaktisch und semantisch analysiert und dann ausgewertet.

Das Analysesystem ist modular aufgebaut und kann flexibel um neue Filteralgorithmen erweitert werden. In jedem Schritt werden die Daten in mehreren Modulen parallel unter verschieden Gesichtspunkten analysiert. Die Ergebnisse werden wiederum in der Datenbank abgelegt, sie stehen so für den nächsten Verarbeitungsschritt zur Verfügung. Im folgenden werden die sechs Einzelschritte genauer beschrieben.

5.2 Auswahl und Generierung der Webadressen

Im ersten Schritt werden die zu untersuchenden Webadressen ausgewählt oder generiert. Die Adressen können nach verschiedenen Kriterien generiert werden. Im einfachsten Fall werden die Webadressen einer vorgegebenen Liste entnommen, sie können aber auch dynamisch mit Hilfe einer Suchmaschine, eines Internet-Portals oder eines Web-Directories erzeugt werden.

Bei der manuell eingegebenen *Liste von Webadressen* kann es sich z.B. um URLs zu Angeboten handeln, die in einem Chat-Room erwähnt wurden. Alternativ werden die Webseiten mit Hilfe von *Suchmaschinen*, z.B. GOOGLE[8], zu bestimmten Schlagwörtern ermittelt. Auf diese Weise ist es möglich, den Suchhorizont auf eine unscharfe, aber doch zielorientierte Weise zu erweitern und evtl. auch zufällig auf neue Webseiten mit Plagiaten oder Markenschutzverletzungen zu stoßen. Weiterhin können bekannte *Internet-Portale* zu einem bestimmten Themenbereich durchsucht werden, z.B. zu Angeboten in einer bestimmten Produktkategorie bei EBAY. Im letzten Fall wird in ein *Web-Directory*, z.B. YAHOO[9], genutzt, in dem Webseiten kategorisiert einem bestimmten Themengebiet zugeordnet sind. Es können in diesem Fall z.B. alle Internetadressen ausgewählt werden, die in einer bestimmten Kategorie aufgelistet sind.

Das Ergebnis des Verarbeitungsschritts ist in jedem Fall eine Liste von URLs.

5.3 Syntaktische Analyse der Daten

Die zuvor generierten Webadressen werden via HTTP oder HTTPS angefragt und die zurückgelieferten Daten vorgefiltert und formatiert. Das Abfragemodul verhält sich wie ein üblicher Webbrowser.

Sofern die Struktur der Webseiten bekannt ist (z.B. der prinzipielle Aufbau der Produktpräsentation eines bestimmten Portals), lassen sich angepasste syntaktische Beschreibungsschemata (*Syntactic Site Description Schemes*) definieren, die bei der Analyse ver-

[8] siehe http://www.google.de/
[9] siehe http://de.yahoo.com/

wendet werden können. Diese ermöglichen es, automatisiert, ausschließlich relevante Informationen und Objekte zu extrahieren. Andernfalls wird ein generisches Standardschema verwendet, das für die Recherche allgemein interessante Objekte und Informationen extrahieren soll. Dabei handelt es sich besonders um die Elemente: Struktur der Webseite, Bild-Objekte, Multimedia-Dateien oder Text.

Die *Site Description Schemes* können z.B. in der Form von XSLT-Dateien formuliert werden, welche zusammen mit den vorgefilterten und formatierten HTML-Dateien als Eingabe für einen entsprechenden Transformer verwendet werden.

Mit den gefundenen Objekten sollten dabei die folgenden Parameter für die spätere Einordnung verknüpft werden:

- URL der Seite auf der sich das gesuchte Objekt befindet,
- URL des Objektes,
- Name des gefundenen Objektes,
- Datum und Uhrzeit,
- das gefundene Objekt selbst (bzw. ein Verweis darauf),
- Bildeigenschaften: Größe, Höhe, Breite, Format,
- Alternativtext für das Bild und
- die Position innerhalb der Seitenstruktur.

Das Resultat dieses Verarbeitungsschrittes sind Datenbankeinträge zu bestimmten syntaktischen Strukturobjekten einer Webseite.

5.4 Semantische Analyse der Daten

Nach der syntaktischen Extraktion von Datenelementen folgt im nächsten Schritt die semantische Analyse der Daten. Dabei können Datenelemente in Beziehung zueinander gesetzt werden und deren Bedeutung für die weitere Verarbeitung durch ergänzende Parameter angegeben werden. In einem semantischen Beschreibungsschema (*Semantic Site Description Scheme*) lassen sich die Bedeutungen der Datenelemente (im Fall von Bildern z.B. abhängig von Position oder Bildgröße) beschreiben, wodurch dann bestimmte Elemente z.B. als Produkt-Logos identifiziert werden können. Für die semantische Analyse werden spezielle inhaltsbasierte Algorithmen eingesetzt, welche auf Basis der folgenden Aspekte, eine genauere Einordnung der Datenelemente ermöglichen:

- Wasserzeichen-Verfahren, z.B. zur Extraktion von Urheberinformation,
- Fingerprinting-Verfahren, z.B. zum Ähnlichkeitsvergleich zu gegebener Vorlage,
- Dateieigenschaften, z.B. Typ, Name, URL,
- Bildeigenschaften, z.B. Größe, Auflösung,
- Meta-Information, z.B. Datum, Bildbearbeitungsprogramm, oder
- Kontext-Bezug, z.B. Textinhalt der Webseite.

Für die semantische Analyse sind die Wasserzeichen- und Fingerprinting-Verfahren als Basistechnologie von besonderer Bedeutung. Durch Wasserzeichen-Verfahren [12] ist es

möglich, in Multimedia-Daten versteckt Information einzubringen ohne die Qualität merklich zu beeinträchtigen. In dem hier beschriebenen Anwendungsszenario sind vor allem Wasserzeichen-Verfahren für Bilder [4] interessant, da es damit möglich ist, Urheberrechtsinformationen in Firmen-Logos, Produktbeschreibungen oder kommerziell angebotene Bilddaten einzubetten.

Fingerprinting-Verfahren bieten die Möglichkeit gleiche oder ähnliche Multimediadaten mit einem kurzen "Fingerabdruck" zu identifizieren. Sie sind vor allem dann von Vorteil, wenn die Inhalte bereits ohne Wasserzeichen im Internet kursieren oder man aus anderen Gründen keine Veränderungen an den Daten vornehmen kann. Auch hier haben Verfahren für Bilder [18, 19] eine besondere Bedeutung, da diese neben Text die wesentlichen Bestandteile von Webseiten sind. Andere Multimedia-Formate, wie Video und Audio, gewinnen aber auch an Bedeutung und lassen sich mit entsprechenden Fingerprinting-Verfahren [8] identifizieren.

Das Ergebnis dieses Analyseschrittes sind Datenbankeinträge zu semantischen Inhalten.

5.5 Logische Analyse der Daten

Im letzten Schritt werden die Daten in Bezug auf den Suchauftrag ausgewertet und ein Bericht erstellt. Dazu werden die verschiedenen Datenelemente aus der semantischen Analyse anhand eines spezifischen Rechercheauftrags verknüpft und die Ergebnisse in einem Bericht zusammengestellt. So werden zum Beispiel als Verkaufspreis erkannte Texte mit einem vorgegebenen Minimalpreis oder Wasserzeichen mit der URL des Logos verglichen.

Die Ergebnisse der einzelnen Analyse-Schritte werden jeweils in die SQL-Datenbank geschrieben.

6 Systemarchitektur und Anwendung

In den folgenden Abschnitten werden die Gesamtarchitektur und die einzelnen Komponenten des Systems näher beschrieben. Danach wird anhand eines Recherchevorgangs das Zusammenspiel der Teilkomponenten näher erklärt, ausgehend vom Arbeitsplatz des Internet-Detektivs, über die Extraktion und Aggregation der relevanten Daten im Filter & Analyse-Server bis zum Zurückliefern der Ergebnisse.

6.1 Architektur

Das vorgestellte Verfahren baut auf einer flexiblen Verarbeitungspipeline auf. Die Suchaufträge der Internet-Detektive manifestieren sich in den Verarbeitungsmodulen und dem Datenpfad zwischen ihnen. Diese Konfiguration muss sich leicht kontrollieren und an neue Suchstrategien anpassen lassen. Aus diesem Grund wird mobile Agententechnologie eingesetzt. Sie erlaubt das Verteilen spezifischer Algorithmen auf vielen Wirtssystemen, was an dieser Stelle notwendig ist. Dem Internet-Detektiv stellt sich das System wie ein gewohnter Client/Server-basierter Arbeitsplatz dar.

Das Gesamtsystem besteht aus drei Hauptkomponenten: dem Arbeitsplatz des Internet-Detektivs, den Filter & Analyse-Servern und den nicht-kooperativen Websites im Internet. Eine Übersicht ist im Architekturbild in Abbildung 2 dargestellt.

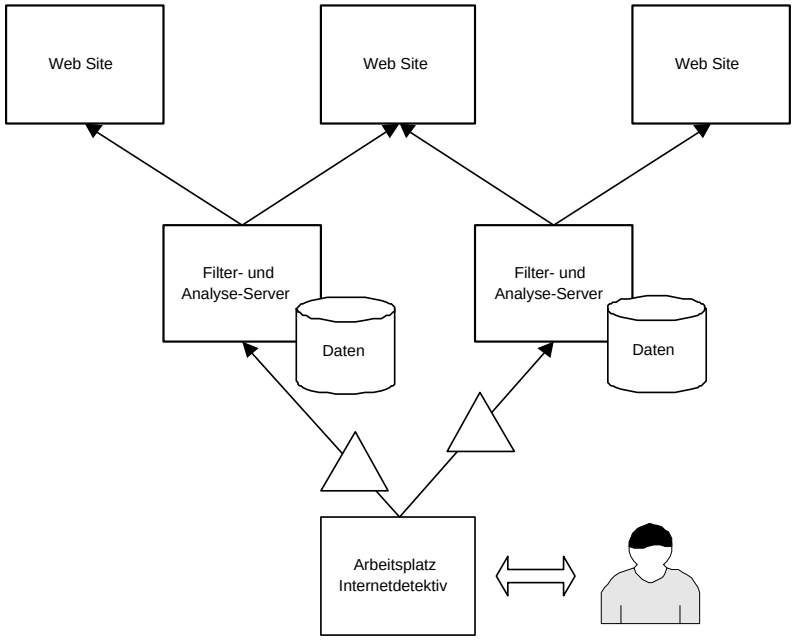

Abbildung 2: Gesamtarchitektur

Bei der Umsetzung des Systems kann man sich gängiger Software-Pakete bedienen. So bietet die freie Agentenplattform SeMoA [13, 15] als komponetenbasiertes Framework sowohl die notwendige Mobilität des Agentencodes, als auch ausgefeilte Sicherheitsmechanismen, um sowohl den Auftragsagenten als auch die verteilten Server zu schützen. Die Agentenplattform gibt als Programmierumgebung Java vor. Sowohl am Arbeitsplatz des Internet-Detektivs als auch auf dem Filter & Analyse-Server wird ein herkömmliches Betriebssystem, z.B. Windows, Solaris oder Linux vorausgesetzt. Zusätzlich muss auf dem Server noch eine Datenbank installiert werden.

6.2 Arbeitsplatz Internet-Detektiv

Der Arbeitsplatz des Internet-Detektivs bildet die Benutzerschnittstelle des Systems, an der die Suchagenten anhand der Kundenaufträge für eine spezifische Recherche erstellt werden. Die Bedienung, Einrichtung und Anpassungen der Suchagenten soll ohne Programmierkenntnisse mit einer grafischen Oberfläche möglich sein.

Der Anwender kann flexibel konfigurieren, welche Webseiten durchsucht werden sollen, und dabei verschiedene Zeitintervalle für die Suche angeben. Je nach Art der Seite, können bereits besuchte Seiten häufiger oder nicht so oft erneut angefragt und analysiert werden.

6.3 Auftragsagenten

Auftragsagenten werden auf dem Arbeitsplatzrechner des Internet-Detektivs generiert und konfiguriert. Danach migrieren sie zu einem Filter & Analyse-Server, um dort die Ausführung des Rechercheauftrags zu initiieren und den Ablauf zu überwachen. Die Auftragsagenten übernehmen das Workflow-Management und sorgen dafür, dass regelmäßig oder nach Beendigung eines Auftrags Berichte erstellt werden und an den Internet-Detektiv gesendet werden.

6.4 Filter & Analyse-Server

Die Filter & Analyse-Server bilden den eigentlichen Kern des Systems, hier erfolgt die Suche nach relevanten Webseiten im Internet, die Analyse und Vorfilterung der Daten und die Auswertung der Ergebnisse. Dazu werden zunächst Webadressen generiert und abgefragt und dann in mehreren Schritten analysiert (vgl. Abschnitt 5.1). Die Filter & Analyse-Server sind hochgradig dezentral lokalisiert, damit sich die Rechercheure nicht anhand der Routingwege verraten. Gleichzeitig sorgt dies für eine Lastverteilung.

6.5 Interaktion der Komponenten

In diesem Abschnitt wird beispielhaft die Abarbeitung eines Rechercheauftrags beschrieben, um das Zusammenspiel der einzelnen Komponenten zu verdeutlichen. In Abbildung 3 ist dazu der Workflow im Filter & Analyse-Server und die Verknüpfungen der verschiedenen Instanzen der einzelnen Analysestufen detailliert dargestellt.

1. Der Internet-Detektiv erstellt eine Anfrage basierend auf einem Kundenauftrag, d.h. er konfiguriert einen Auftragsagenten mit Hilfe einer grafischen Benutzeroberfläche. Dazu wählt er die zu durchsuchenden Webadressen aus, z.B. die Webseiten eines Internet-Portal zu einem bestimmten Thema oder die GOOGLE-Treffer für bestimmte Suchbegriffe. Weiterhin spezifiziert er die relevanten Elemente, Filter und Analysealgorithmen für die syntaktische und semantische Analyse, z.B. die Extraktion von Bildern und die Parametrisierung mit einem Firmenlogo. Er bestimmt die Bewertungsfunktion, konfiguriert und verknüpft die verschiedenen Elemente. Schließlich wählt er einen oder mehrere Filter- und Analyse-Server für seinen Rechercheauftrag aus.

2. Dann wird automatisch der Agent erstellt und migriert zum entsprechenden Filter & Analyse-Server. Soll die Abfrage von mehreren F & A-Servern aus durchgeführt werden, so werden entsprechend viele Agenten erstellt, unter denen sich die Suche verteilt.[10]

3. Dort wird der Rechercheauftrag ausgeführt. Die Webadressen werden generiert, abgefragt und die erhaltenen Daten in der Datenbank abgespeichert. Dann werden die relevanten Elemente in der syntaktischen Analyse extrahiert und in der semantischen Analyse z.B. Wasserzeichen gesucht oder Ähnlichkeitsvergleiche mit Hilfe von Fingerprinting-Verfahren durchgeführt. Schließlich werden die Ergebnisse anhand der Anfragevorgaben bewertet und ein Bericht erstellt.

4. Der Auftragsagent kehrt zum Internet-Detektiv zurück und präsentiert ihm das Ergebnis.

Besonders herauszuheben ist dabei, dass es durch das modulare Konzept möglich ist, bereits existierende Auswertungsmodule flexibel zu kombinieren und das System durch neue Module zu erweitern (siehe Abbildung 3). Der Internet-Detektiv kann dazu bequem per grafischer Oberfläche die entsprechenden Algorithmen und Verfahren auswählen und kombinieren. Er konfiguriert und parametrisiert damit implizit einen Auftragsagenten.

[10] Mittels Agentenkommunikation kann die Suche auch bei dynamischen URL-Listen zwischen den Agenten koordiniert werden.

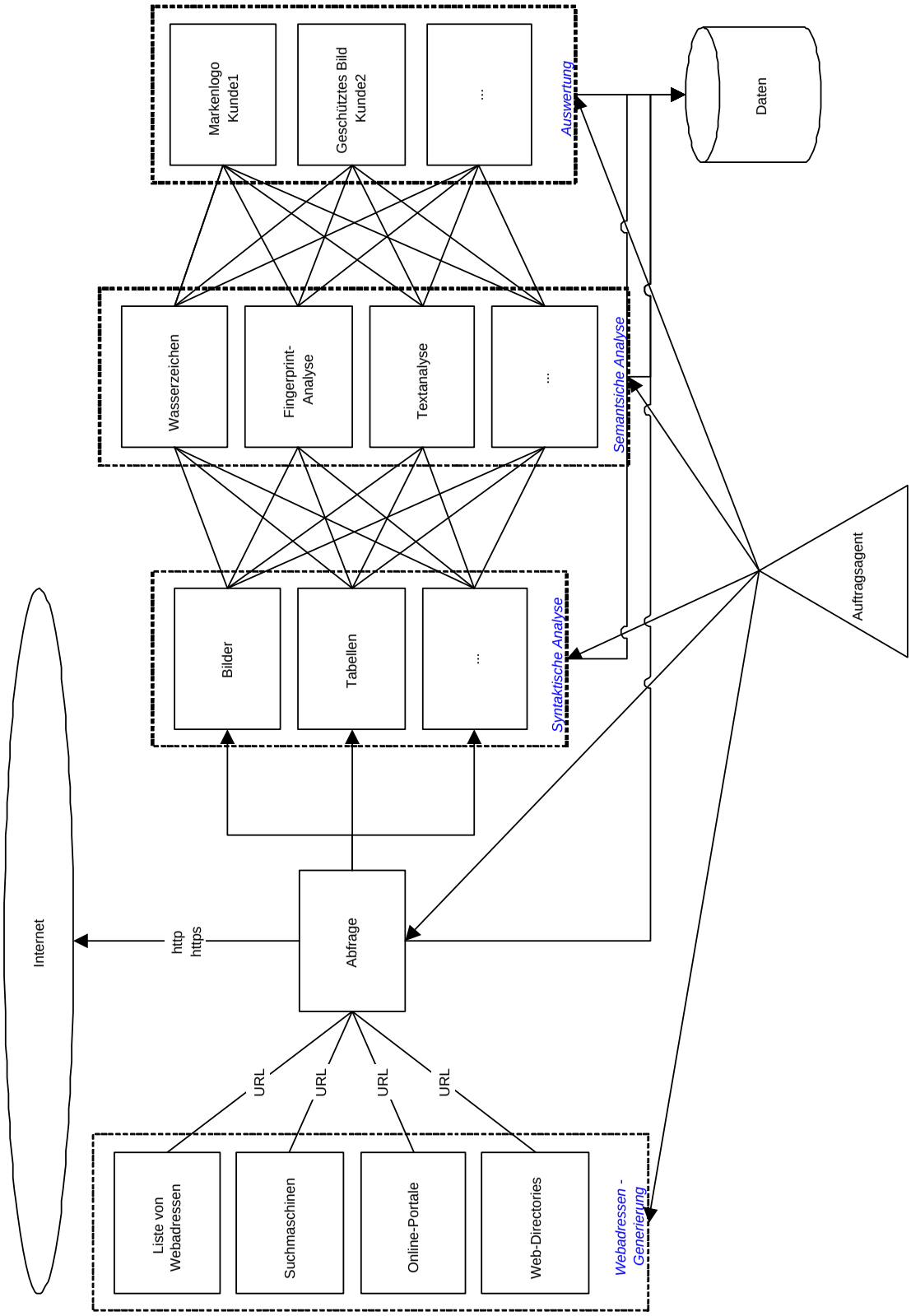

Abbildung 3: Workflow Filter- und Analyse-Server

Dieser Agent sorgt dafür, dass der Rechercheauftrag korrekt ausgeführt wird, und installiert gegebenenfalls die entsprechenden Algorithmen und Software-Komponenten auf dem Filter & Analyse-Server.

7 Zusammenfassung und Bewertung

Der Artikel beschreibt ein System, das die automatisierte Recherche nach Markenpiraterie im Internet ermöglicht. Es soll Internet-Detektive beim Aufspüren verdächtiger Seiten unterstützen. Die vorgeschlagene Lösung vereinigt eine Reihe von Vorteilen, die in dieser Form bei existierenden Werkzeugen nicht gegeben sind.

Die Suche lässt sich auf beliebige Fragestellungen anwenden. Dafür lassen sich für die inhaltsbasierte Suche Algorithmen individuell wählen und beliebig miteinander kombinieren. Die mehrstufige Analyse erlaubt eine flexible Suche nach verschiedenen Kriterien, die je nach Strukturierung der Webseite und der Aufgabenstellung variieren. Das verteilte System aus Filter & Analyse-Servern erlaubt die Untersuchung eines Webangebotes von verschiedenen Servern aus und zu verschiedenen Zeitpunkten. Solche Recherchen sind weniger auffällig als das Abfragen durch einen Web-Crawler, was im Rahmen des Einsatzgebietes Markenschutz einen großen Vorteil darstellt.

Interessant für den Bereich der Schutzrechtsverletzung ist der Aspekt der sozialen Kontrolle im Netz. So setzt bspw. die freie Online-Enzyklopädie WIKIPEDIA[11] [9] auf die Kontrolle ihrer Leser um Urheberrechtverletzungen bei den verwendeten Bildern und Texten aufzuspüren. Würden die in diesem Artikel beschriebenen Analysewerkzeuge jedem Internetbenutzer kostenfrei zur Verfügung stehen, so ist durch die entstehende soziale Kontrolle mit einer deutlichen Verbesserung des Webangebotes im Bezug auf Marken- und Urheberrechtsverletzungen zu rechnen.

Literatur

[1] Tim Berners-Lee. Semantic web road map. Internal note, World Wide Web Consortium, September 1998. See http://www.w3.org/DesignIssues/Semantic.html.

[2] Tim Berners-Lee, Jim Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[4] S. Burgett, E. Koch, and J. Zhao. Copyright labeling of digitized image data. *IEEE Communications Magazine*, 36(3):94–100, March 1998.

[5] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11–16):1623–1640, 1999.

[6] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30:161–172, 1998.

[11] siehe http://de.wikipedia.org/

[7] *Keine Gnade für Plagiate. Gewerbliche Schutzrechte nutzen.* Deutscher Industrie- und Handelskammertag (DIHK), 2001.

[8] Ton Kalker. Applications and Challenges for Audio Fingerprinting. In *Proc. 111th AES Convention, in the Watermarking versus Fingerprinting Workshop*, december 2001.

[9] Michael Kurzidim. Wissenswettstreit. Die kostenlose Wikipedia tritt gegen die Marktführer Encarta und Brockhaus an. *c't Magazin für Computer und Technik*, page 132ff, 2004.

[10] Torge Löding. Luxusfirmen verstärken Anti-Fake-Kampagne im Internet. Heise Online, 26.08.2004. http://www.heise.de/newsticker/meldung/50390.

[11] Gregory Louis McLearn. Autonomous cooperating web crawlers. Master's thesis, University of Waterloo, Canada, 2002.

[12] Peter Meerwald and Andreas Uhl. A Survey Of Wavelet-Domain Watermarking Algorithms. In *Proceedings of SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents III*, 2001.

[13] Ulrich Pinsdorf and Volker Roth. Mobile Agent Interoperability Patterns and Practice. In *Proceedings of Ninth IEEE International Conference and Workshop on the Engineering of Computer-Based Systems (ECBS 2002)*, Computer Graphics Edition, pages 238–244, University of Lund, Lund, Sweden, April 2002. Institute of Electrical and Electronics Engineers, IEEE Computer Society Press. ISBN 0-7695-1549-5.

[14] Volker Roth. *Sichere verteilte Indexierung und Suche von digitalen Bildern.* Ph.D. thesis, Technische Universität Darmstadt, Wilhelminenstraße 7, 64283 Darmstadt, Germany, June 2001.

[15] Volker Roth and Mehrdad Jalali. Concepts and Architecture of a Security-centric Mobile Agent Server. In *Proc. Fifth International Symposium on Autonomous Decentralized Systems (ISADS 2001)*, Dallas, Texas, U.S.A., March 2001. IEEE Computer Society Press.

[16] G. Salton. *Information Retrieval: Data Structures and Algorithms.* Addison-Wesley, Reading, Massachussetts, 1989.

[17] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[18] Jin S. Seo, Jaap Haitsma, Ton Kalker, and Chang D. Yoo. A robust image fingerprinting system using the radon transform. *Signal Processing: Image Communication*, 19, 2004.

[19] R. Venkatesan, S. M. Koon, M. H. Jakubowski, and P. Moulin. Robust Image Hashing. In *Proc. IEEE Int. Conf. Image Processing 2000*, pages 325–339. IEEE, 2000.