

Humboldt Universität zu Berlin  
Seminar Network Mining  
Betreuer: Isabel Drost

# LINK SPAM-ERKENNUNG UND -UNTERDRÜCKUNG

## **Zusammenfassung**

Aus meist kommerziellen Interessen wird durch sogenannten Linkspam versucht, den Pagerank [Brin and Page, 1998] bzw. HITS [Kleinberg, 1999]-Wert, der für populäres Ranking innerhalb einer Suchmaschine verantwortlich ist, zu erhöhen. Dieses Paper soll einen Überblick über Probleme und Lösungsansätze geben, um Suchmaschinenergebnisse von derartigem unerwünschtem Spam zu bereinigen.

Florian Holzhauer  
fh-hu@fholzhauer.de  
5. Februar 2007

# 1 Einleitung

Moderne Suchmaschinen haben Milliarden Webseiten in ihren Datenbanken indiziert, und stellen Suchwortabhängig die wahrscheinlich zutreffendsten Links zur Verfügung - die bekannteste Suchmaschine Google nutzt zur Bestimmung beispielsweise den sogenannten Pagerank-Algorithmus [Brin and Page, 1998]. Für einen kommerziellen Webseitenbetreiber ist gute Auffindbarkeit bei relevanten Suchworten, und damit ein hoher Pagerank, daher ausgesprochen wichtig - je populärer die Position in den Suchmaschinenergebnissen, desto wahrscheinlicher ist es, dass der Benutzer die Website besucht. Aus dieser Erkenntnis heraus entwickelte sich der sogenannte SEO (Search Engine Optimizers)-Markt, dessen Dienste darin bestehen, den Pagerank einer Seite möglichst hoch zu optimieren. Da eines der wichtigsten Elemente des Pagerank-Algorithmus die Anzahl der Links ist, die auf die zu optimierende Seite verweisen, werden häufig Seiten oder Links erschaffen, die nicht etwa sinnvollen Inhalt haben, sondern ausschliesslich dazu dienen, den Linkrank der beworbenen Seite<sup>1</sup> zu erhöhen.

## 1.1 Page-Rank

Der Pagerank-Algorithmus, in [Brin and Page, 1998] beschrieben, gilt als Grund des Erfolges der Suchmaschine Google. Er spezifiziert die "Wichtigkeit" einer Website, die mit der Formel

$$PR_i = \frac{1-d}{N} + d \sum_{\forall j \in \{(j,i)\}} \frac{PR_j}{C_j}$$

berechnet wird. Hier ist  $PR_i$  der zu berechnende Pagerank,  $N$  die Gesamtzahl der indizierten Webseiten, und  $d$  ein Dämpfungsfaktor.  $PR_j$  ist die verlinkende Seite,  $C_j$  ist die Anzahl der Seiten, die  $PR_j$  verlinkt.

Hinter dem Page-Rank versteckt sich die Idee des Random Surfer Modells - der Page-Rank will also das Verhalten eines zufälligen Websurfers modellieren, der nicht auf die Seiteninhalte achtet. Er klickt also zufällig auf einen Link einer Seite, und von dort aus wieder zufällig auf einen weiteren Link. Aus der Verlinkungshäufigkeit einer Seite ergibt sich so eine hohe Wahrscheinlichkeit, mit der der Surfer auf die Seite gelangt, und damit ein hoher Page-Rank.

---

<sup>1</sup>Wichtig ist auch der Linkrank der Seiten, die auf die zu optimierende Seite verweisen - da dieser Linkrank aber ebenfalls effektiv auf der Verlinkung dieser Quellseiten basiert, ist das Kernproblem das selbe.

Der Dämpfungsfaktor simuliert hier die Wahrscheinlichkeit, mit der ein Surfer zufällig eine Seite besucht, ohne einem Link zu folgen - etwa über ein Browser-bookmark. So wird vermieden, dass nicht verlinkte Webseiten vollständig durch den Pagerank-Algorithmus vernachlässigt werden. Der Wert von  $d$  bewegt sich meist um etwa 0.85.

Kurz gesagt entscheiden also drei Elemente einen guten Pagerank, und damit eine prominente Darstellung einer Website in einer Suchmaschine: Einerseits die absolute Zahl der eingehenden Links, andererseits der Pagerank der verlinkenden Seiten, da dieser Pagerank ja "weitervererbt" wird. Der dritte Faktor, die Anzahl der Links, die eine verlinkende Seite besitzt, wird bei Spamming-Techniken üblicherweise nicht berücksichtigt.

Ein Spammer will somit seine zu bewerbende Seite möglichst häufig verlinken, und das durch Webseiten, die einen möglichst hohen Pagerank haben.

## 1.2 HITS

Ein ähnlicher Ansatz zur Bewertung der Relevanz von Webseiten findet sich in [Kleinberg, 1999] - statt einem spezifischem Pagerank werden hier zwei unterschiedliche Werte berechnet: Während der sogenannte Authority-Wert bzw. das Authority-Gewicht  $a_i$  analog zum Pagerank angibt, wie häufig eine Seite von anderen verlinkt ist, wird mit dem Hub-Gewicht  $h_i$  angegeben, wieviel Seiten von der Seite verlinkt werden. Eine Bookmark-Sammlung etwa bekommt so ein hohes Hub-Gewicht durch die vielen ausgehenden Links, aber nicht zwangsläufig einen hohen Authority-Wert, da sie ja nicht unbedingt von anderen Seiten verlinkt wird. Hinzu kommt, dass HITS rekursiv definiert ist - eine Seite, die Links auf Seiten mit gutem Authority-Wert bietet, bekommt einen hohen Hub-Wert, und umgekehrt.

Berechnet werden die beiden Werte wie folgt:

$$h_i = \delta \sum_{j=1}^n A_{ij} a_j$$
$$a_i = \lambda \sum_{k=1}^n A_{ik}^T h_k$$

Hier ist  $A_{ij}$  eine Matrix, die angibt, ob es einen Link von  $i$  nach  $j$  gibt - in diesem Fall ist  $A_{ij} = 1$ , sonst 0.  $A_{ij}^T$  ist die transponierte Matrix von  $A$ , die "Gegenrichtung".

## 2 Link-Spam-Systeme

Um einer Seite zu einer guten Suchmaschinenposition zu verhelfen, ist somit auch die Linkdichte interessant, wie aus den beiden Bewertungsalgorithmen klar erkennbar ist. Ein Spammer, der eine bestimmte Seite propagieren will, ist also vor allem an Links auf eine Seite interessiert.

Möglichkeiten, Links auf andere Webseiten zu erzeugen, gibt es ausgesprochen viele - die im folgenden besprochenen Papers widmen sich vor allem den beiden momentan am häufigsten auftretenden Erscheinungen, "Blogspam" und "Linkfarmen".

### 2.1 Blogspam

Ein sogenanntes Weblog zeichnet sich unter anderem dadurch aus, dass es verschiedene Mechanismen der Kommentierung zulässt, meist sogenannte Trackbacks und Kommentare. Während Kommentare Leser-Annotationen zu einem in einem Weblog veröffentlichtem Text sind, handelt es sich bei Trackbacks um Rückverweise anderer Blogs auf einen Artikel, die darauf hinweisen wollen, dass der Blogbeitrag in dem rückverweisendem Weblog behandelt wurde.

Beiden Kommentierungsmöglichkeiten gemeinsam ist, dass sie auch Weblinks zu anderen Seiten beinhalten dürfen. Blogspam (z.B. erläutert in [Mishne et al., 2005], siehe auch Abbildung 1) nutzt diese Eigenschaft aus, um so auf eine beworbene Seite zu verlinken, die inhaltlich meist nichts mit dem eigentlichen Textinhalt zu tun hat. Erwünschter Nebeneffekt dieser Verlinkung ist hier, dass der Pagerank-Algorithmus auch den Rank der Quellseite mit berücksichtigt, der Spammer so also auch von der Popularität des bespamten Weblogs profitieren.

Diese Spamtechnik ist selbstverständlich nicht auf Weblogs beschränkt, sondern findet sich in ähnlichen Variationen in nahezu allen anderen Websystemen wieder, die Benutzerinhalte in ihrer Seite zeigen - so seien hier noch Foren, Gästebücher oder Wikis erwähnt. Die im Weiteren vorgestellten Spamererkennung-Techniken lassen sich meist analog auf derartige Systeme anwenden.

### 2.2 Linkfarmen

Eine Linkfarm ist eine Ansammlung von automatisch generierten Webseiten, die sich gegenseitig sowie eine zu bewerbende Seite verlinken, und mit suchmaschinen-

relevanten Stichworten gefüllt sind (siehe etwa [Fetterly et al., 2004]). Durch die starke gegenseitige Verlinkung wird der Pagerank sowohl der Linkfarm als auch der beworbenen Seite in die Höhe getrieben. Inhaltlich und technisch sind Linkfarmen üblicherweise sehr stark darauf optimiert, für bestimmte Stichworte möglichst populär auf einer Suchmaschine gezeigt zu werden - so taucht das Stichwort an vielen Stellen der Seite auf, etwa im Domainnamen, dem Seitennamen oder auch in der Seite (siehe Abb. Abbildung 2) selbst.

Auch verschiedene andere technische Charakteristika sind bei Linkfarmen auffällig. So werden die Suchanfragen, über die ein Besucher auf eine Linkfarm gelangt, mitprotokolliert und entsprechend bei der nächsten Seitengenerierung mit berücksichtigt. Der Inhalt der Seiten ist also oft nicht statisch, sondern dynamisch generiert. Dies hat zwar den Vorteil für einen Spammer, seine Seiten auf aktuell populäre Suchbegriffe zu optimieren zu haben, ist aber gleichzeitig ein Indikator, der zur Spamererkennung genutzt werden kann. Die meisten seriösen Seiten haben zumindest grössere Anteile, die nach einer erstmaligen Veröffentlichung statisch bleiben.

### 3 Filterung durch technische Attribute

Ein naheliegender Ansatz diesen Spam zu erkennen ist eine quantitative bzw. linguistische Methodik, da sich Spam oft durch bestimmte Charakteristiken und Wortfolgen auszeichnet, die der Suchwort-Optimierung geschuldet sind. Viele dieser Eigenschaften spiegeln sich in Eigenschaften des Servers bzw. ähnlichen "Meta-Attributen" oder in der Wortwahl und -frequenz wieder.

#### 3.1 Erkennung durch Servereigenschaften

In die Bewertung des Suchmaschinenposition einer Seite fließt neben vielen anderen Attributen auch mit ein, ob das gesuchte Wort Bestandteil des Domain- bzw. Seitennamens ist - basierend auf dem PageRank einer Seite wird zum Zeitpunkt der Suche zusätzlich ein Inhalts-Scoring durch verschiedene Seitenattribute im Hinblick auf die Suchanfrage erstellt. Linkfarmen erstrecken sich daher auf viele meist automatisch generierte Subdomains und Seiten, die dynamisch bei jedem Besuch neu erzeugt werden. Die daraus resultierenden Besonderheiten, wie etwa eine sehr hohe Anzahl Subdomains, die alle einer IP-Adresse und damit einem Webserver zugeordnet werden können, oder auch die sehr hohe Linkdichte innerhalb einer solchen Seitenansammlung lassen sich als deutliche Abweichungen vom Durchschnitt erkennen.

ID	Karma	How Long Ago	Author	Post Title	Comment	Type
176209	-0.48	2 hours, 4 minutes	hxdabinv qpxkw	DR. HANSEN	spjxteyjj aientqhv hyoi pxytycflug a [...]	Emt
176230	-107.28	1 minute	Viagra	Links 49	Really nice interesting site. thank [...]	Cmt
176229	-91.9	20 minutes	xanax	THE PIRATE BAY	Excellent site. It was pleasant to [...]	Cmt
176228	-106.13	21 minutes	Viagra	Links 49	Thank you very much, for this site! [...]	Cmt
176227	-109.92	22 minutes	Viagra	Links 49	Good site you done here,man.	Cmt
176226	-57.2	29 minutes	Viagra	Links 49	You can also visit my page.	Cmt
176225	-86.78	30 minutes	cialis	THE PIRATE BAY	Nice site! Thank you!	Cmt
176224	-93.42	36 minutes	tramadol	THE PIRATE BAY	If you have to do it, you might as [...]	Cmt
176223	-68.77	37 minutes	phentermine	THE PIRATE BAY	Very interesting site. Hope it will [...]	Cmt
176222	-74.03	37 minutes	mikkola	THE PIRATE BAY	http://tramadol11.bloggingmylife.co [...]	Cmt
176221	-74.39	40 minutes	hydros	THE PIRATE BAY	http://adipex11.bloggingmylife.com/ [...]	Cmt
176220	-68.09	1 hour, 2 minutes	Michael	GELD ZURÜCK VON JAMBA & CO.	this site rocks! http://slimurl.jp/ [...]	Cmt
176218	-28.17	1 hour, 14 minutes	Ron	GELD ZURÜCK VON JAMBA & CO.	medicine cabinet generic drug vic [...] odin online pharmacy viagra dosage oxycontin matricidaxola funny creator viagra price	Cmt

Abbildung 1: Blogspam am Beispiel lawblog.de

buy online viagra ... cheap viagra

high-octane fuel. And men with ED who want to get more out of their sex lives? They take VIAGRA. **viagra price** Help maintain an erection during sex!  
**female viagra cream** When it comes to sex, you want to perform. You just need to know how to get there. Men who want to get more out of their cars use high-octane fuel. And men with ED who want to get more out of their sex lives? They take VIAGRA. **female version of viagra** Helps most men with ED achieve harder erections **viagra research** The benefits of VIAGRA free viagra Discuss your general health status with your doctor to ensure that you are healthy enough to engage in sexual activity. If you experience chest pain, nausea, or any other discomforts during sex, seek immediate medical help. **viagra picture** cialis versus viagra Highly effective cialis vs viagra **viagra online us** When you want it

**generic viagra**

Why VIAGRA? Because that look she gives is only meant for you. Because an empty nest is the chance to fall in love all over again. Because reading the Sunday paper do take all day. **viagra dosage** viagra story cheap viagra **cheap viagra order viagra**

[viagra alternative](#)  
[100mg viagra](#)  
[prescription viagra](#)  
[viagra on line](#)  
[viagra sex](#)  
[viagra](#)  
[cheap viagra online](#)  
[viagra prescription](#)  
[viagra canada](#)  
[get viagra](#)  
[blog comment post viagra](#)  
[buying viagra online](#)  
[herbal viagra](#)  
[viagra cost](#)  
[viagra retail discount](#)  
[cialis viagra vs](#)  
[100mg viagra](#)  
[viagra substitute](#)  
[order viagra](#)  
[herbal alternative viagra](#)  
[free viagra](#)  
[viagra pill](#)  
[cialis generic viagra](#)

http://www10.asphost4free.com/onlinesearch/ Helps most men with ED maintain an erection during sex **viagra online** VIAGRA is prescribed to erectile dysfunction (ED).

**low cost viagra**

Abbildung 2: Beispiel einer durch eine Linkfarm generierte Spamseite

Verschiedene Charakteristiken dieser Art werden in [Fetterly et al., 2004] beschrieben, auch weitere Eigenheiten wie eine sehr hohe Ähnlichkeit der Seiten selbst, sowie eine auffällig hohe Aktualisierungsquote die dem automatischen Generieren solcher Spamseiten bei jedem Besuch geschuldet sind, werden betrachtet. Hervorzuheben ist hier vor allem, dass die meisten erläuterten Erkennungsmechanismen keine weitergehenden Informationen über die Linkstruktur zwischen den Seiten benötigen.

Hier wird ein grosses Set an Seiten-Daten (einmal 150 Millionen Urls, einmal 429 Millionen) auf einige dieser Charakteristiken untersucht - so etwa, wieviel verschiedene Hostnamen auf eine einzelne IP-Adresse zeigen, Anzahl von ausgehenden und eingehenden Links einer Seite, oder das häufige Vorkommen einzelner Worte bzw ähnlicher Seiten innerhalb einer Seite (siehe Abbildung 3). Auch die Änderungsfrequenz des Website-Inhaltes wird betrachtet - ein bei jedem Besuch anderer Inhalt ist ein Indikator für eine automatisch generierte Seite, die nicht statisch vorgehalten wird.

### 3.2 Erkennung durch Seiteneigenschaften

Mit ähnlichen technischen Attributen befasst sich auch [Drost and Scheffer, 2005] - hier wird allerdings auf Eigenheiten der einzelnen Website sowie der verlinkenden und verlinkten Seiten der zu bewertenden Website eingegangen. Faktoren wie der Länge der Domain, Anzahl der Subdomains, die Topleveldomain und Eigenschaften verschiedener HTML-Elemente, die pro Website mit einem "tfidf-Vektor" (term frequency, inverse document frequency) angegeben werden, ergeben zusammen einen Spamwahrscheinlichkeitswert.

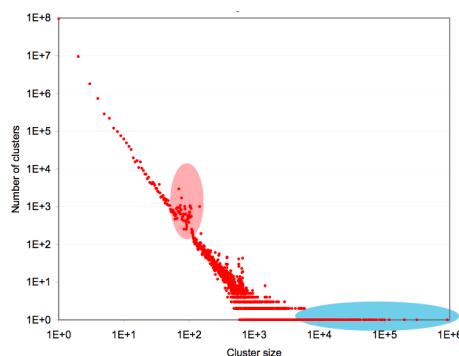


Abbildung 3: Stark ähnliche Seiten innerhalb einer Domain

Basierend auf einem manuell ausgewähltem Set von Spam- und Hamseiten wird anschliessend untersucht, welche Eigenheiten momentan die geeignetsten sind, um Spam zu klassifizieren. Wie anschliessend auch ausgeführt wird, ist dieses Ergebnis allerdings laufenden Änderungen unterzogen, da Spammer jedem neuen Erkennungsmechanismus mit entsprechenden Gegenmassnahmen begegnen.

### 3.3 Sprachliche Eigenschaften

Auch sprachliche Eigenschaften sind denkbare Ansatzpunkte einer Spamerken-  
nung. [Kolari et al., 2006] untersucht die Häufung von Wortteilen innerhalb einer Seite, wobei hier auch berücksichtigt wird, wo genau die Wortteile erscheinen, etwa innerhalb eines Metatags. Diese Eigenschaften werden hier zur Unterscheidung zwischen echten Weblogs und Weblogs mit reinen Spaminhalten untersucht - also eine Kombination aus Linkfarm und Blogspam.

Mittels einer "Support Vector Machine" (SVM), eines Klassifikationsmecha-  
nismus der zunächst mit manuell ausgewählten Seiten trainiert wurde, wird untersucht, welche Seitenelemente sich zur Unterscheidung eignen. Hier genutzte Elemente sind sogenannte Meta-Elemente, worunter der Seitentitel, die Seiten-Url und der "Generator"-Metatag verstanden wird, Link-Elemente der Seite sowie der eigentliche Seiten-Text selbst.

In der Ausführung wurden so Genauigkeiten von bis zu 88 Prozent erreicht, und untersucht, wo einige Probleme bei der fehlerhaften Klassifizierung liegen.

Vor allem für Blogspam geeignet scheint auch der Ansatz in [Mishne et al., 2005] - hier werden Unterschiede der Sprachmodelle zwischen einzelnen Kommentaren in einem Blog mit dem Blogposting selbst sowie der verlinkten Seiten betrachtet. Die hier genutzte Methodik macht sich zunutze, dass der Blogeintrag selbst sowie die verschiedenen Kommentare von unterschiedlichen Autoren mit entsprechend unterschiedlichem Stil geschrieben wurden.

Als Entscheidungskriterium wird hier die sogenannte "Kullback-Leibler-Divergenz" zwischen den einzelnen Elementen betrachtet, also die Sprachvariation zwischen den einzelnen Sprachmodellen der Kommentare und Blogeinträge. Jedes Sprachmodell bildet eine Wahrscheinlichkeitsfunktion in Form einer Gausskurve. Anhand der unterschiedlichen Ausprägungen, der Modellähnlichkeit zum ursprünglichen Blogpostings also, kann nun bestimmt werden, welche Kommentare wahrscheinlich zum Stil und Inhalt des Blogpostings passen. Je grösser die

Abweichung, desto höher die Spamwahrscheinlichkeit.

Es ist allerdings intuitiv klar, dass diese Lösung nicht grundsätzlich sinnvoll ist. So sind z.B. Weblogs denkbar, bei dem das Sprachmodell des ursprünglichen Eintrags bewusst anders als die Kommentare sind - ein Weblog mit Gedichten, die kommentiert werden, ist hier eine Idee. Auch ist naheliegend, dass ein Spammer sich die Idee zu eigen machen kann, in dem er das Sprachmodell zunächst analysiert und sein Kommentar automatisiert anpasst.

Zu letzterem Angriffs-Szenario machen sich die Autoren auch einige weitergehende Gedanken - ein derartiger Angriff bringt es dann mit sich, dass in verschiedenen Blogs sehr unterschiedliche Sprachmodelle mit den selben Links zu finden sind, so dass hier durch eine blogübergreifende Analyse Spam wiederum sehr schnell erkennbar ist.

## 4 Filterung durch Graphanalyse

Ein komplett anderer Ansatz findet sich in der Idee wieder, die zu bewertende Linkstruktur als gerichteten Graphen zu betrachten, und auf Unterschiede zwischen Ham und Spam zu untersuchen. Neben vielen Hintergrundinformationen zur Geschichte und Evolution von Linkspam finden sich grundlegende Überlegungen zu dieser Idee in [Metaxas and Destefano, 2005]:

Wir befinden uns heute in der dritten Suchmaschinengeneration. Während die erste Generation sich ausschliesslich mit Schlüsselwörtern auf der jeweiligen Seite im Text, in Meta-Tags oder ähnlichem beschäftigte, ging die zweite Generation bereits auf die Linkstruktur der jeweiligen Seiten ein - je häufiger eine Seite verlinkt wurde, desto besser die Suchmaschinenposition - eine Idee, die ausgesprochen trivial durch Linkspam angegriffen werden konnte.

Die dritte, heute aktuelle, Generation von Suchmaschinen hat diese Verlinkungsidee zwar aufgegriffen, aber durch den Pagerank-Algorithmus verfeinert, wie bereits in der Einleitung ausgeführt wurde.

Im Paper wird anschliessend ausgeführt, dass es noch weiterführende Ideen zu dieser Technik gibt - so ist etwa davon auszugehen, dass eine Seite, die Spam bewusst verlinkt, wahrscheinlich ebenfalls Spam ist. Gerade im Bereich des Blogspam ist allerdings offensichtlich, dass dieser Ansatz durchaus nicht immer zutrifft - solche Verlinkungen müssen nicht vom Seiten-Autor gewollt sein.



multipliziert mit einem Dämpfungswert ("trust damping")  $0 < \beta < 1$  - eine direkt von einer "guten" Seite verlinkte Website, bei der daher davon auszugehen ist, dass es sich hier ebenfalls nicht um Spam handelt, bekommt also den Trustrank  $1 * \beta$  - eine Linkebene weiter entsprechend  $1 * \beta * \beta$ . Je weiter eine Seite also vom Good Core entfernt ist, desto niedriger der Trustrank. Ein weiterer Mechanismus ist der des "trust splitting" - ausgehend von der Idee, dass eine Website mit wenigen Urls diese sorgfältiger prüft, wird der Trustrank der verlinkenden Seite entsprechend verteilt. Die verlinkte Seite erhält also nicht, wie im vorherigen Beispiel, einfach  $1 * \beta$ , sondern  $(1/Linkzahl) * \beta$ .

Eine Weiterentwicklung dieser Idee findet sich in [Gyongyi et al., 2006]:

Wenn eine Seite Links besitzt, die auf eine Element des Bad Cores, also der Gruppe als Spam bekannter Seiten, zeigt, ist sehr wahrscheinlich, dass diese Seite ebenfalls spammt. Ausgehend von dieser Idee werden zunächst zwei naive Ansätze zur Erkennung solcher Verlinkungen skizziert, zusammen mit einigen Gründen, warum ein naiver Ansatz nicht ausreichend ist.

Ausserdem muss ein Link auf eine Seite des Bad Cores nicht zwangsläufig bewusst erfolgen. Wie im Paper ausgeführt gibt es eine ganze Reihe von Gründen, warum eine solche automatisierte "Abstrafung" nicht sinnvoll ist. So ist es gängiges Verhalten von Spammern, alte und aufgegebene Domains zu kaufen, und sie mit Spam zu füllen. Ein Link kann allerdings bereits zu einem Zeitpunkt gesetzt worden sein, zu dem die Domain noch mit sinnvollen Inhalten vor der Löschung gefüllt war - also vor Übernahme durch den Spammer. Auch Blog-Spam ist hier ein Beispiel, der Link auf die Spamseite ist so in keinem Fall vom Seitenbetreiber erwünscht.

## 4.2 Graphanalyse ohne Good Core

Auf ein Good Core, also ein vordefiniertes Set "guter" Seiten, verzichtet hingegen [Wu and Davison, 2005]. Zunächst wird hier auf die Idee des sogenannten Bad Ranks eingegangen, der als

$$BR_i = E(A)(1 - d) + d \sum_{\forall j \in \{(j,i)\}} \frac{BR_j}{C_j}$$

definiert ist - die meisten Werte sind hier analog zum Pagerank [Brin and Page, 1998] definiert,  $E(A)$  gibt hier einen originären Bad-Rank-Wert an, der beispielsweise mittels Spamfiltern berechnet werden kann.

Die Autoren leiten anschliessend auf einen eigenen Ansatz, ParentPenalty genannt, hin. Interessant ist hier vor allem der Ansatz, einen Bad Core automatisch bestimmen zu können, eine manuelle Auswahl vordefinierter Spam-Seiten bzw. eines Good Cores wie in den vorherigen Arbeiten ist hier nicht zwingend erforderlich.

Hierfür gehen die Autoren davon aus, dass viele Linkfarmen sich dadurch auszeichnen, dass sie innerhalb der selben Domain viele Seiten gegenseitig verlinken - ein von Spammern gewollter Effekt, um so den PageRank zu erhöhen. Für jede Seite wird zusammengefasst, welche Domains von ihr verlinkt werden, und welche sie verlinken - es werden also zwei Domain-Mengen,  $INdomain(p)$  und  $OUTdomain(p)$ , definiert. Anschliessend wird die Schnittmenge der beiden Mengen gebildet - ist diese Menge höher als eine vorher definierte Schwelle, wird die Seite als Spam betrachtet, verlinkt sie doch auffallend oft andere Seiten innerhalb der selben Domain.

Ausgehend hiervon wird nun der Mechanismus der ParentPenalty definiert. Wenn eine Seite viele Spam-Seiten verlinkt, ist davon auszugehen, dass es sich bei der verlinkenden Seite ebenfalls um Spam handelt - auch über mehrere Linktiefen hinweg. Hierfür wird mit dem eben definierten Seed-Set, den potentiellen Spam-Seiten, eine Matrix  $A_n$  gebildet - ist die Seite  $n$  Spam, ist der Wert  $A_n$  1, sonst 0.

Nun werden die einzelnen Seiten ein weiteres Mal betrachtet: Wenn die Anzahl der Links auf Spam-Seiten einer Seite höher als ein vorab definierter Treshold ist, wird die Seite ebenfalls als Spam betrachtet,  $A_n$  wird also zu 1. Diese Betrachtung wird so lange wiederholt, bis sich  $A$  nicht mehr ändert.

Anschliessend wird auf mehreren Seiten evaluiert, wie Suchergebnisse basierend auf Pagerank im Vergleich zu bereinigten Ergebnissen mittels ParentPenalty verhalten - der Umfang dieser Untersuchung ist deutlich zu ausführlich für diese Zusammenfassung. Insgesamt zeichnet sich jedoch ab, dass der Ansatz sehr vielversprechend ist, aber einige Probleme mit sich bringt. So ist etwa die Auswahl eines geeigneten Tresholds nicht unbedingt trivial, ausserdem werden verschiedene Beispiele erwähnt, bei denen ParentPenalty zu Unrecht Spam vermutet.

## 5 Ausblick

Spam ist ein Wettkampf zwischen Suchmaschinenbetreibern und Spammern. Jeder neue Ansatz, jeder neue Algorithmus zu Erkennung und Beseitigung von Spam ist nur so lange effektiv, bis Spammer die Implementation verstanden haben, und ihre Mechanismen entsprechend angepasst haben - eine endgültige technische

Lösung für dieses soziale Problem ist nicht absehbar. Die meisten der in dieser Ausarbeitung beschriebenen Problemlösungen weisen auch deutlich darauf hin, dass ihr jeweiliger Ansatz nur vorübergehend hilfreich ist - sollte ein Mechanismus zu erfolgreich sein, ist es simpel, den Ansatz zu verstehen und entsprechend zu bekämpfen - so würden die meisten Ideen beispielsweise scheitern, sollte ein Spammer "unschuldige" Seiten mitverlinken, und so bewusst Kollateralschäden provozieren.

[Gori and Witten, 2005] gibt verschiedene Ideen, dieses Dilemma zumindest zu entschärfen, unter anderem, in dem ein Weg weg von einer globalen Suchmaschine zu personalisierteren Diensten skizziert wird. So gibt es keinen spezifischen Pagerank mehr, auf den maximal möglich optimiert werden kann. Die Verfasser fordern einen Paradigmenwechsel, eine "intellectually violent revolution": Verschiedene Benutzer suchen unterschiedliche Antworten auf die selbe Suchanfrage, so dass hier eine absolute Pagerank-Funktion nicht sinnvoll ist - eher eine auf den jeweiligen Besucher optimierte Funktionen, die je nach Personalisierung unterschiedliche Ergebnisse bieten. Gerade im Hinblick auf die Ansätze des Semantic Web gibt es sicher noch viele Ideen zu Suchalgorithmen, die weit über das hinausgehen, was heute als State of the Art gilt.

## Literatur

- [Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117.
- [Drost and Scheffer, 2005] Drost, I. and Scheffer, T. (2005). Thwarting the nigritude ultramarine: learning to identify link spam. In *Proceedings of the 16th European Conference on Machine Learning (ECML)*, volume 3720 of *Lecture Notes in Artificial Intelligence*, pages 233-243, Porto, Portugal.
- [Fetterly et al., 2004] Fetterly, D., Manasse, M., and Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages.
- [Gori and Witten, 2005] Gori, M. and Witten, I. (2005). The bubble of web visibility. *Commun. ACM*, 48(3):115-117.
- [Gyongyi et al., 2004] Gyongyi, Z., Berkhin, P., Garcia-Molina, H., and Pedersen, J. (2004). Combating web spam with trustrank. In *VLDB*, pages 576-587.
- [Gyongyi et al., 2006] Gyongyi, Z., Berkhin, P., Garcia-Molina, H., and Pedersen, J. (2006). Link spam detection based on mass estimation. In *VLDB'2006*:

- Proceedings of the 32nd international conference on Very large data bases*, pages 439–450. VLDB Endowment.
- [Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.
- [Kolari et al., 2006] Kolari, P., Finin, T., and Joshi, A. (2006). SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*. Computer Science and Electrical Engineering, University of Maryland, Baltimore County. Also available as technical report TR-CS-05-13.
- [Metaxas and Destefano, 2005] Metaxas, P. T. and Destefano, J. (2005). Web spam, propaganda and trust. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*.
- [Mishne et al., 2005] Mishne, G., Carmel, D., and Lempel, R. (2005). Blocking blog spam with language model disagreement.
- [Wu and Davison, 2005] Wu, B. and Davison, B. (2005). Identifying link farm spam pages. In *Proceedings of the 14th International World Wide Web Conference, Industrial Track*.