DEWS2006 3B-o2

リンク解析を用いたウェブ上のスパム発見手法に関する一考察

小野　拓史† 　豊田　正史†† 　喜連川　優††

† 東京大学大学院 情報理工学系研究科　〒 113–0033 東京都文京区本郷 7–3–1
†† 東京大学 生産技術研究所　〒 153–8505 目黒区駒場 4–6–1
E-mail: †{h-ono,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　ウェブスパムとは検索エンジンの検索結果において特定のサイトのランキングを不正に向上させることを目的とした行為のことを指す．ウェブスパムによって検索結果はページの質とは無関係になり精度が低下するため，検索エンジンのインデックスからスパムサイトを除去することが重要な課題となっている．ウェブスパムの手法のなかで，特に関連サイト同士で密にリンクを張ることによってランキングの向上を図るものをリンクスパムと呼ぶ．本論文ではリンクスパムの種類と分布について，日本のウェブアーカイブを用いて調べた結果について考察する．
キーワード　Web とインターネット，情報検索，テキスト DB

An Examination of Techniques for Identifying Web Spam by Link Analysis

Hiroshi ONO†, Masashi TOYODA††, and Masaru KITSUREGAWA††

† Graduate School of Information Science and Technology, The University of Tokyo　Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–0033 Japan
†† Institute of Industrial Science, The University of Tokyo　Komaba 4-6-1, Meguro-ku, Tokyo, 153–8505 Japan
E-mail: †{h-ono,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　Web spamming refers to actions intended to mislead search engines into ranking some pages higher than they deserve. Recently, the amount of web spam has increased dramatically, leading to a degradation of search results. Link spamming refers to the cases when spammers set up structures of interconnected pages in order to boost the link-based ranking, In this paper, we will analyze link spam types and distributions found in our archive of Japanese pages.
Key words　Web and the Internet, Information Retrieval, Text DB

1. はじめに

今日のウェブにおいて検索は日々の情報収集のために必要不可欠なツールとなっている．このため情報提供者にとって検索エンジンにおける結果の上位に自身のページがランキングされることは重要な意味を持つ．特に通信販売などの商用サイトにおいては，閲覧者を多数獲得する必要があることから，ページの記述方法やサイトの構成方法などを工夫してランキングを改善する SEO (Search Engine Optimization) と呼ばれる手法が多く用いられている．現在のウェブにおいては SEO 手法を悪用して検索エンジンに意図的に誤った評価をさせ，実際よりも高いランキングを得ることを主目的としてコンテンツを構成する行為が頻繁に行われている．この行為はウェブスパムと呼ばれ，これを行う主体をスパマーと呼ぶ．ウェブスパムが行われると，検索結果に関連のないページが多く含まれるようになり，検索結果に偏りをもたらしたり，得られる情報の品質の劣化を招いたりすることになる．ウェブスパムの手法は，主として二種類の方法に大別できる．一つはページのテキストを検索エンジンのクエリに適合するように調整する手法であり，関連するキーワードを多数ページに付加するなどの手法が用いられる．もう一方はリンク解析を用いたランキングを行う検索エンジンをターゲットとして，自身のサイトの周辺におけるリンク構造を操作してランキングを上げることを目的とするリンクスパム手法である．多数のサイトを作成しその間に密にリンクを張ることによって，参照数を基にしたランキングを欺くなどの手法が用いられている．現在，主要な検索エンジンではリンク解析が重要な要素としてランキングに用いられていることから，リンクスパムは頻繁に行われており，これへの対処が重要な課題

となっている．我々は，ウェブ上においてどのようなリンクスパムがどの程度行われているかを調査し，スパム対策手法を開発することを目標としている． その第一歩として， 本論文ではリンクスパムを行う際にウェブのグラフ構造上に現れる大きな極大クリークに着目し，それらの分布や種類について日本のウェブアーカイブを用いて調査した結果について報告する．本論文の構成は以下のとおりである．第 2. 節では，本研究に関連する既存の研究について述べる．第 3. 節では本研究において用いたスパム抽出手法について説明する．第 4. 節では本研究に使用したウェブアーカイブの詳細と実験に使用したサイトグラフの構成法について述べる．第 5. 節ではウェブアーカイブからの極大クリーク抽出についての実験とその結果について述べる．第 6. 節では共有ノード数によるクラスタリング実験とその結果について述べる．第 7. 節では得られた極大クリークに対する完全ハブ抽出実験の結果を示した．最後に第 8. 節では結論を記す.

2. 既存の研究

リンクスパムへの対処に関しては，まず単純な統計を用いたものがあり，Fetterly らはリンクの次数などの統計を用いたウェブスパムの調査を行っている [1]．リンクスパム手法がターゲットとするリンク解析手法は主に，Page，Brin らによる PageRank [5]，および，Kleinberg による HITS [7] の 2 つである [8]．PageRank はあるページに張られているリンクの数は一般的なウェブユーザーにとっての重要度をあらわしているという前提に基づいており，あるページの重要度はそれにリンクを張っているページの重要度によって計算される．このため，特定のページへリンクを集中させるリンクスパム手法が多く用いられている．リンクスパムが PageRank アルゴリズムに及ぼす影響は [2], [3] に述べられている．PageRank の改善手法については，Gyongyi らによって TrustRank [9] が提案されている．TrustRank はスパムでないことが判明しているページからスコアを伝播することでスパムサイトへ高いスコアを割り当て難くしている．HITS はウェブ上のすべてのページについてハブスコアとオーソリティスコアを割り当てる．HITS の定義によると，重要なハブページは多くの重要なオーソリティページにリンクしているものであり，また，重要なオーソリティページは多くのハブページからリンクされているものである．HITS アルゴリズムを用いる検索エンジンはもっとも高いハブスコアとオーソリティスコアを持つページをあわせたものを検索結果として返す．ハブスコアは知名度が高いページの多くにリンクをはることによって容易に上げることができる．高いオーソリティスコアを得るのは比較的難しく，重要だと思われるハブから多くのリンクを得なければならない．リンクスパムの手法は Gyongyi らによって [8] にまとめられており，以下のような手法が述べられている．

- 複数のサイトで協力して相互にリンクを交換する，または，自分で複数のドメインを取得しその間に密にリンクを張ることで相互に PageRank およびハブ・オーソリティスコアを上げることができる．リンクファームと呼ばれる．Wu らは密な相互リンクを抽出することでリンクファームを自動的に検出する手法を提案している [6].
- 一部のウェブディレクトリサービスには，誰もがリンクを登録できるものが存在する．またブログ，Wiki にはコメントなどにリンクを付加して登録できる．これらを利用するとスパマーはターゲットページへ外部からのリンクを加えることができる．ウェブディレクトリは高い PageRank スコアとハブスコアを持っていることが多いので，この手法はターゲットページの PageRank とともにオーソリティスコアもあげることができる.
- 一般的に役に立つ情報を持つページのコピーを用意し，それらがランキングを上げたいターゲットページを指すようにする．コピーしたページが他のページからリンクされると，ターゲットページのランキングを上げることができる.

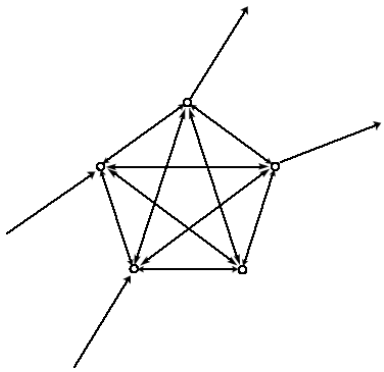

図 1　大きさ 5 のクリーク

3. スパム抽出手法

本研究では，リンクファーム，およびウェブディレクトリ等へのリンク登録を利用したリンクスパムを行った際にウェブのグラフ上に現れるクリークに着目してスパム構造の調査を行う．ここで扱うウェブのグラフは，各ウェブサイトをノード，サイト間に張られたリンクをエッジとした有向グラフである．これをサイトグラフと呼ぶ．ページ単位のグラフでは，複数のページにまたがるリンク交換を検出し難くなるため，サイト単位のグラフを用いている．以下に本研究で用いたスパム抽出法を述べる.

3.1 リンクファーム型スパムの抽出

リンクファームを用いたサイト同士はサイトグラフ上で密に結合されることになる．サイトグラフから，2 つのサイトの間に相互にエッジが張られている場合にのみ方向無しエッジが存在する無向グラフを抽出すると，ほぼすべてのリンクファームはクリークを含むことになる．クリークとはすべてのノードが互いにエッジによって相互連結されている部分グラフを指す．例として図 1 に大きさ 5 のクリークを示す．リンクファームを用いたスパムをリンクファーム型スパムと呼ぶ.

3.1.1 極大クリーク抽出法を利用したリンクスパム抽出

クリークのうち，他のクリークに包含されない極大クリークを抽出することでリンクファームの中心構造を捉えることができる．極大クリーク列挙には，牧野，宇野ら [4] によって提案

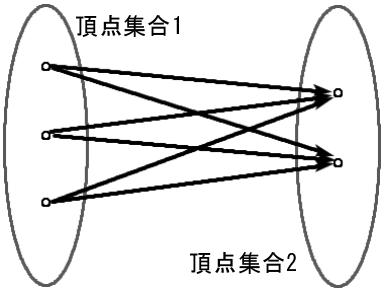

図 2　完全 2 部グラフ

されたアルゴリズムを使用した．このアルゴリズムはグラフのノード数 n，エッジ数 m，最大次数を Δ とすると，極大クリークを一つあたり $O(\Delta^4)$ の計算時間で列挙でき，メモリの使用量は $O(n+m)$ である．このアルゴリズムでは，次数が 80 以上では計算が困難であったので，次数を 80 以下としたサイトグラフにおいて，この実験を行った．第 5. 節ではサイトグラフから極大クリークを抽出することでリンクファームを抽出し，その種類について調査する．

3.1.2　近似的極大クリーク抽出法の提案

次に，極大クリークに近い構造を持つ部分を抽出する方法として，共有ノード数によるサイトのクラスタリングを行った．これは以下の手法により行われる．まず，無向グラフの各エッジをソートした状態で読み込む．次に各エッジの両端ノードに対して，双方のノードがリンクしているノードを読み出し，これらのノードの共通要素を調べる．これは最大次数のオーダで計算可能である．もしこれがある閾値 N 以上なら同じ集合としてクラスタリングする．クラスタリングのアルゴリズムとしては union-find アルゴリズムを使用する．これは，互いに素な集合を与えられた条件に従ってマージしていくものである．マージされた集合のサイズが N 以上ならばサイトグラフから取り出す．このようにして共通リンク先ノード数が N 以上，サイズ N 以上のクラスタが抽出できる．N の値を段階的に減少させることで，より共通リンク先ノード数が小さなクラスタが抽出可能である．また，全体の計算量はエッジ数と最大次数の積のオーダである．第 6. 節では，この手法により得られる極大クリークに類似した形のリンクファームの抽出を行い，その種類について調査する．

3.2　リンクファーム型スパムとハブ・オーソリティ型スパムを併用したスパムの抽出

リンク登録を利用したリンクスパムでは，登録サイトの集まりから，ターゲットサイトへの一方向のエッジが密に張られることになる．この構造は，2 部クリークを用いて捉えることができる．2 つのノード集合を持ち，任意のエッジが 2 つのノード集合間を結んでいるグラフを 2 部グラフと呼び，この部分グラフのうち 2 つのノード集合に含まれるノード同士が全て結合されているものを 2 部クリークと呼ぶ（図 2）．リンク登録を用いたリンクスパムは，登録サイトからターゲットサイトへ完全にエッジが張られた 2 部クリークを形成することになる．このようなスパムをハブ・オーソリティ型スパムと呼ぶ．また，図 2 において頂点集合 1 の各ノードは頂点集合 2 のノードの完全ハブとなっているという．第 7. 節では，第 5. 節で得られた極大クリークに対して完全ハブを抽出することでハブ・オーソリティ型のスパムを併用しているか調査する．

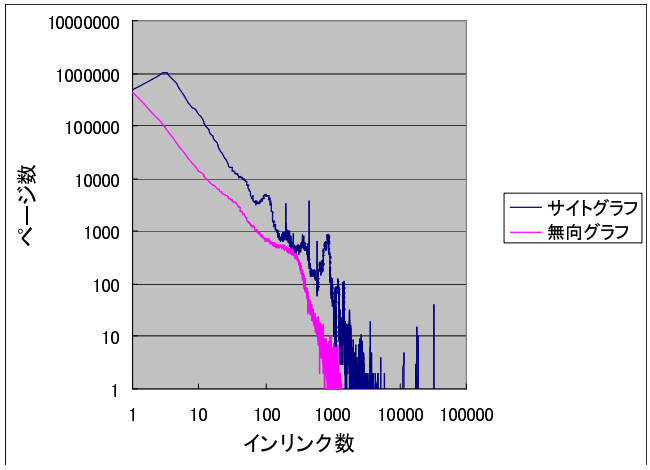

図 3　サイトグラフ，無向サイトグラフのインリンク数分布

4.　データセット

4.1　ウェブアーカイブ

本実験に使用したデータセットは 2004 年 5 月に日本語のウェブページを大規模にクロールをしたものである．クロールの方法としては，jp ドメインのページと日本語で書かれた海外ドメインのページを収集する形をとる．jp ドメイン以外のサイトについては数ページクロールして日本語のページを発見できない場合，そのサイトのクロールを停止する．アーカイブデータは 9600 万ページ，45 億のリンクから構成されている．

4.2　サイトグラフの作成

データセットからサイトグラフを作成する．このサイトのトップページとしてアーカイブ内にあるインリンク数が 3 以上のページを集めた．このようなページをシードページと呼び，シードページをサイトのトップページとしたサイトグラフを構成した．各シードページは URL が http://A/B/C/の形をしているものに階層を限定した．これは同一の団体によるページを一つにまとめる目的で行った．このグラフはノード数 680 万，エッジ数 2 億 8000 万である．このサイトグラフから，相互にリンクが張られている場合のみ方向無しエッジが存在する無向グラフを抽出した．この無向グラフはノード数 160 万，エッジ数 3900 万である．以下に元のサイトグラフ，無向グラフ双方のインリンク数分布を示す．

5.　極大クリークの抽出

実験は Itanium 2 1.6GHz x 8, メモリー 128GB のマシンで行った．この計算のために所要した時間を表 1 に示した．

表 1　極大クリーク計算のための所要時間

ノードの最大次数	計算時間
50	1 分
70	6 分
80	18 時間

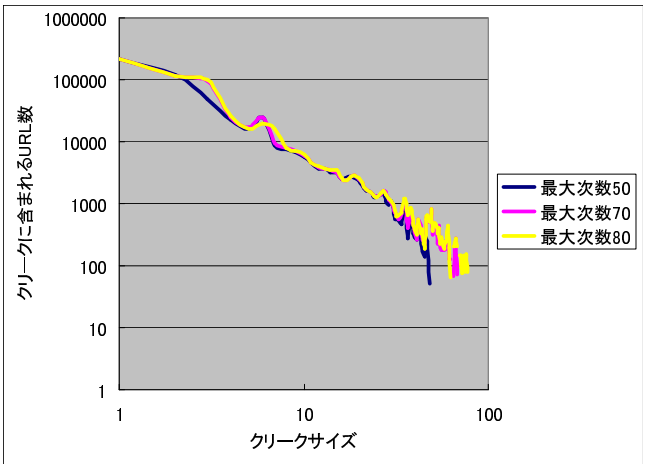

図 4　極大クリークのサイズごとに出現するサイト数の分布

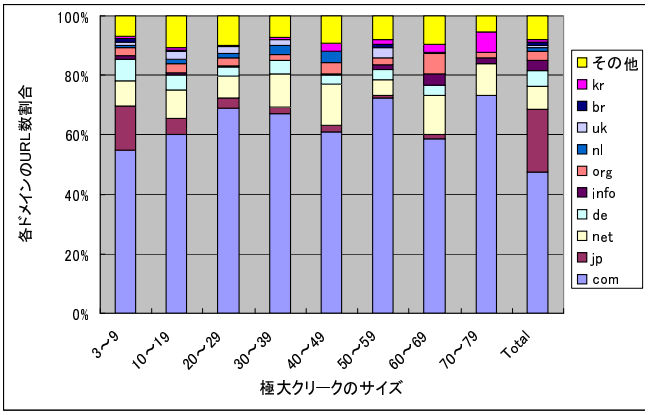

図 5　各サイズの極大クリークを構成するドメイン分布

極大クリークを列挙するとノードに重複のあるクリークが多数生成される．したがって極大クリークの数を集計するかわりに，各サイトが含まれる最大の極大クリークのサイズ (ノード数) を求め，極大クリークのサイズ毎にサイト数の分布を取った．その結果を図 4 に示す．

なお，元のサイトグラフとして，各シードページはインリンク数 3 以上のものとしたため，サイズ 3 以上のデータを示した．極大クリークのサイズ分布がべき乗則に従うことが分かる．最大次数を 80 としたとき，サイズ 3 以上の極大クリークに含まれるノードの総数は 60 万であり，元のサイトグラフに含まれるノード数の 37.5% であった．このときのグラフの極大極大クリークに含まれるサイトの主なドメインの内訳を図 5 に示す．全体的に国外のサイトが多く，特に.com ドメインのものが大半であることがわかる．jp ドメインのサイトは全体の 16% 程度であった．また，サイズ 30 以下の極大クリークには jp ドメインのものが数パーセントあるが，それ以上の大きさにおいてはほとんど見られない．

次に，極大クリークをランダムに選び，その内容を確認したところ，表 2 のようであった．「リンク集」とは，サイト内にリンク情報が中心的になっており，コンテンツがわずかであるようなサイトのことをさしている．「販売促進」サイトはある商品の宣伝を目的として作られたサイトである．オンラインカジノ，

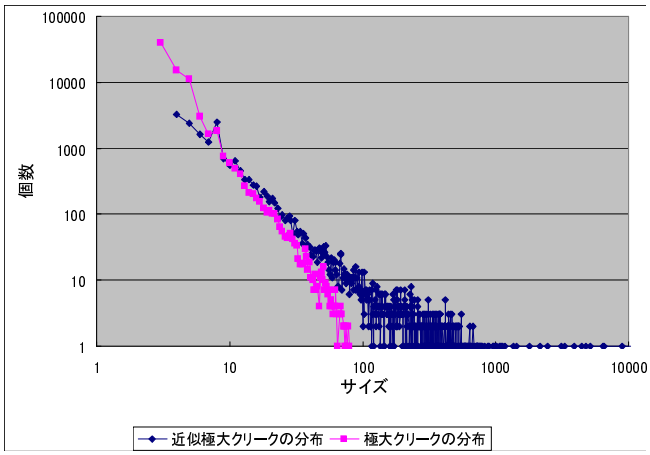

図 6　極大クリーク，近似的極大クリークのサイズ分布

懸賞サービスの勧誘サイトなどもこの中に分類した．主に医薬品などの販売をテーマとしているサイトの集合である．「一般」サイトとはリンク集，販売促進サイト，アダルトサイトのいずれにも該当しないものである．これには個人のサイト，一般企業のサイト，そして公的機関のサイトなどが含まれる．全体の 83% がスパムサイトと考えられ，アダルトサイトと販売促進サイトの割合が多い．極大クリークのサイズが大きくなるに従い，一般サイトの割合が減少している．

表 2　極大クリークのスパム分類

サイズ	一般サイト	販売促進	リンク集サイト	アダルト	総計
3-10	38	50	18	34	140
11-20	6	30	20	44	100
21-29	3	4	2	11	20
30-39	1	7	4	8	20
40-49	9	2	2	7	20
50-59	2	10	1	7	20
60-69	1	13	4	2	20
70-79	0	16	4	0	20
総計	60	132	55	113	360
割合	17%	37%	15%	31%	100%

6.　近似的極大クリークの抽出

クラスタリングにより得られた集合を近似的極大クリークと呼ぶ．図 6 では得られた近似的極大クリークのサイズ分布を極大クリークのものと比べて示す．この結果より極大クリークと比較して近似的極大クリークはサイズが大きいことがわかる．その理由として近似的極大クリークは極大クリークと比較して抽出される集合の制約が緩和されているためと考えられる．

このときのグラフの近似的極大クリークに含まれるサイトの主なドメインの内訳を図 7 に示す．極大クリーク抽出のときに得られた結果と同じ傾向にあると言える．

近似的極大クリークを極大クリークと同じ大きさの領域においてランダムに選び，その内容を確認したところ，表 3 のようであった．スパムの比率は 68% 程度であった．

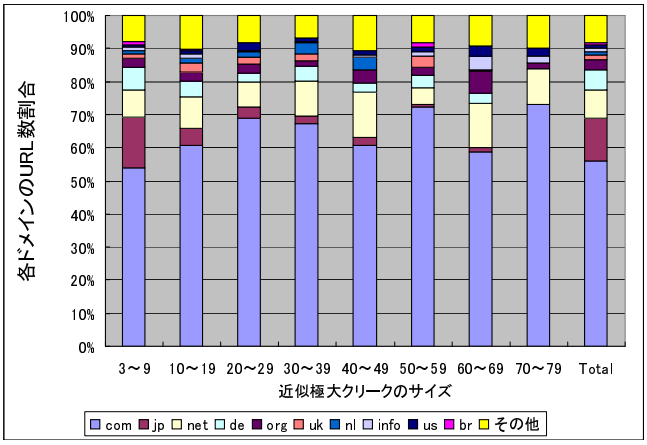

図 7　近似的極大クリークのサイズごとに出現するサイトのドメイン分布

表 3　近似的極大クリークのスパム分類 (小さいサイズ)

サイズ	一般サイト	販売促進	リンク集サイト	アダルト	総計
3-10	78	61	6	92	237
11-20	15	11	3	9	38
21-29	2	2	0	2	6
30-39	2	0	1	3	6
40-49	0	3	0	3	6
50-59	1	0	0	5	6
60-69	2	4	0	0	6
70-79	0	2	1	3	6
総計	100	83	11	117	311
割合	32%	27%	3.5%	38%	100%

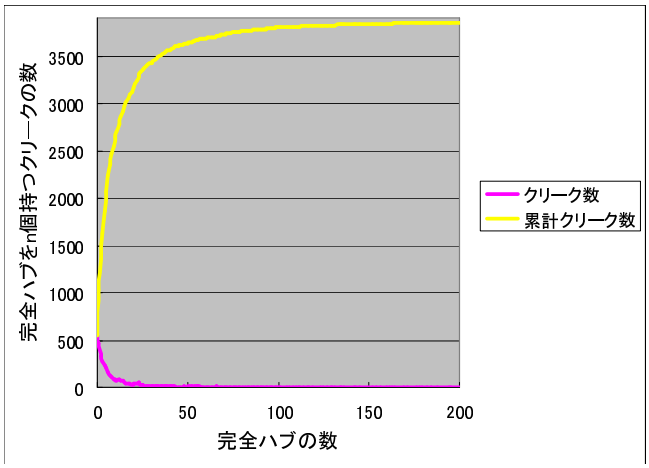

図 8　完全ハブを n 個持つ極大クリークの数

また，より大きなサイズの近似的極大クリークについてもランダムに選び，その内容を確認したところ，表 4 が得られた．スパムの比率は 99% 程度であり，そのうちアダルトが大きな割合を占めていることがわかる．

表 4　近似的極大クリークのスパム分類 (大きいサイズ)

サイズ	一般サイト	販売促進	リンク集サイト	アダルト	総計
100-199	0	5	1	28	34
200-299	1	6	2	17	26
300-	0	0	0	9	9
総計	1	11	3	54	69
割合	1%	16%	4%	78%	100%

7.　リンクファーム型スパムとハブ・オーソリティ型スパムを併用したスパムの抽出

第 5. 節で得られた極大クリークの内，サイズが 10 以上のものについて完全ハブを持つものの数の分布を示したものが図 8 である．これより，完全ハブを持たない極大クリークは全体の 14% 程度であり，半数の極大クリークは 4 以上の完全ハブを持っていることが分かる．従ってリンクファーム型スパムのほとんどがハブ・オーソリティ型スパムを併用しているといえる．

極大クリークの完全ハブサイトをランダムに選び，その内容

を確認したところ，表 5 が得られた．これより極大クリーク抽出と比較して，スパムサイトの比率が多くなっていることが分かる．

表 5　極大クリークの完全ハブサイトのサンプリング結果

サイズ	一般サイト	販売促進	リンク集サイト	アダルト	総計
10-19	6	1	2	1	10
20-29	6	1	3	0	10
30-39	3	3	4	0	10
40-49	3	2	5	0	10
50-59	3	1	0	1	5
60-69	3	5	1	1	10
70-79	1	9	0	0	10
総計	25	22	15	3	65
割合	38%	34%	23%	5%	100%

8. ま と め

本研究では大規模アーカイブを用いてスパムサイトを抽出しその分布を調べた．そして近似的極大クリークを抽出することにより，巨大スパムを抽出することができた．検出されたスパムの統計を取ったところ，.com スパムが大半であることが分かった．jp ドメインにおいては 2004 年 5 月の段階ではリンクスパムがあまり行われていないことも判明した．また，リンクファーム型とハブ・オーソリティ型の両構造を併用しているサイトは多く，それらはほぼすべてスパムであることも分かった．

文　献

[1] Fetterly, D., Manasse, M., and Najork, M.,“ Spam, damn spam, and statistics, ”Proc. 7th International Workshop on the Web and Databases (WebDB) Paris, France, June 2004.

[2] Baeza-Yates, R., Castillo, C., and Lopez, V.“ PageRank increase under different collusion topologies, ”Proc. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb) Tokyo, Japan, May 2005.

[3] Gyongyi, Z. and Garcia-Molina, H.“ Link spam alliances, ” Proc. 31st International Conference on Very Large Data Bases (VLDB) Trondheim, Norway, August 2005.

[4] Makino, K., and Uno, T.,“ New Algorithms for Enumerating All Maximal Cliques, ” SWAT 2004, LNCS 3111, pp. 260-272, 2004.

[5] Page, L., Brin, S., Motwani, R., and Winograd, T.,“ The PageRank citation ranking: Bringing order to the web, ” Tech. rep., Stanford University, 1998.

[6] Wu, B., and Davison, B.,“ Identifying link farm pages, ” Proc. 14th International World Wide Web Conference (WWW), Tokyo, Japan, May 2005.

[7] Kleinberg, J.,“ Authoritative sources in a hyperlinked environment, ”Journal of the ACM, 46(5),

[8] Gyongyi, Z. and Garcia-Molina, H.“ Web Spam Taxonomy, ” Proc. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb) Tokyo, Japan, May 2005.

[9] Gyongyi, Z., Garcia-Molina, H., and Pedersen, J.“ Combating web spam with TrustRank, ”Proc. 30th International Conference on Very Large Data Bases (VLDB) Toronto, Canada, August 2004.

[10] Benczur, A., Csalogany, K., Sarlos, T., and Uher, M.,“ SpamRank - Fully Automatic Link Spam Detection, ”Proc. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb) Tokyo, Japan, May 2005.

[11] Bifet, A., Castillo, C., Chirita, P., and Weber, I.,“ An Analysis of Factors Used in Search Engine Ranking, ”Proc. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb) Tokyo, Japan, May 2005.

[12] Henzinger M., Motwani, R., and Silverstein, C. ”Challenges in Web Search Engines,” SIGIR Forum 36(2), 2002.