検索エンジンにより検閲されたホームページの発見

諸 井 教 紀[†1] 吉 浦 紀 晃[†2]

現在，インターネットは私たちの生活に欠かせないものである．インターネットの普及に伴い，その情報量は増加し続けている．膨大な情報の中から自分が欲しい情報を得ることに役立つのが検索エンジンサイトである．検索エンジンサイトの中でも特に人気があるのは Google である．しかし，Google の検索結果が人為的に操作されているという事実がある．多くの人が検索エンジンサイトを使う現在において，人為的操作が行われていることは重大な問題である．本研究では日本におけるグーグル八分を発見するシステムを開発し，収集したデータ用いて，傾向の分析を行った．

Searching Web Pages Censored by Search Engine

Takanori Moroi [†1] and Noriaki Yoshiura[†2]

1. はじめに

現在でもインターネットの利用は拡大し，私達が生活する上で欠かせないものとなっている．そして，インターネットを利用する上で，欠かせないものとして，検索エンジンがある．数ある検索エンジンの内，世界中で人気があるのは Google である．的確であると言われている Google の検索結果には「PageRank」という技術が大きな役割を果たしている[5]．的確な検索結果が得られること，Google 自身が「複雑で自動化された検索方法には人為的介入がありません」と公表していることから[6]，Google の検索結果は客観的尺度に基づいて決定される順位で表示されると思われている．

しかし，その得られる検索結果が人為的に操作されている事実がある．企業や団体などが世間に知られては都合が悪い情報に関して，Google にクレームを寄せたり，圧力をかけ，本来ならば検索結果のランキングの上位に表示されるはずのページが，表示されないということが起こっている[16]．

Google の検索結果が人為的操作をされ，本来表示されるはずページが表示されないことを，日本では江戸時代の村社会で行われていた村八分に例えて，グーグル八分と呼ばれる．海外においても「Google Censorship」と呼ばれ問題視されている．

実際に，検索結果から除かれているのは名誉棄損・誹謗中傷を行っているページ，犯罪に関係する記述のあるページ，児童ポルノなどのページ，検索順位を向上させる操作である Search Engine Optimization(SEO) を過度に行っているページ，法人や企業などの告発を扱っているページなどである．検索結果から除かれているページの内容が正しい情報であったとすると，Google の利用者はその正しい情報が書かれたページを発見することができない．また，Google の検索結果から除かれているページが存在するとき，「検索結果に表示されていないページがあります」というメッセージが表示されるが，どのページがどのような理由で検索結果から除かれているか利用者は知ることができない．そのため，情報操作が行われても利用者が気付かない可能性がある．インターネットから情報を得るために，検索エンジンに頼らざるをえない現在において，重大な問題であると考えられる．利用者が正しい情報の取捨選択できるように，検索結果を監視し，どのような内容のページがなぜ検索結果から除かれているのかチェックする必要がある．

グーグル八分は全世界で共通して行われているものと各国ごとに行われているものが存在する．全世界で共通して行われているグーグル八分を回避する方法は現在までに見つかっていないが，各国ごとに行われているグーグル八分は他国の Google を検索に使用することで回避することができる．

本研究では国ごとにグーグル八分が行われていて各国で対象となるページが異なっていることを利用し，

†1 埼玉大学大学院理工学研究科数理電子情報系専攻
†2 埼玉大学大学院理工学研究科数理電子情報部門

日本におけるグーグル八分を発見するシステムを作成する．また，発見したグーグル八分のページのデータを用いて，日本におけるグーグル八分の傾向の分析を行う．

本論文では，2 章において関連研究を挙げ，3 章においてグーグル八分とその実例を挙げグーグル八分の説明を行う．4 章においてグーグル八分の発見方法について記述する．5 章ではシステムによって得られた結果を述べ，考察を行う．6 章で本論文をまとめる．

2. 関 連 研 究

グーグル八分は一種の情報操作とも考えられるが，情報操作の例としては，中国におけるインターネット版万里の長城があげられる[7),10)]．また，インターネット上の検閲に関する研究も行われている[9)]．これらの多くは，中国における検閲に関するものであり，検索エンジンにおける検閲についての研究は，著者が知る限り見受けられない．また，検索エンジンの技術や検索エンジンに関する研究は数多くあるが[11),15)]，検索エンジンにおける検閲についての研究はほとんど無い．グーグル八分の発見に関して，情報処理推進機構 2007 年度第 2 期未踏ソフトウェア創造事業[12)] において，グーグル八分発見システム「∞ Eyes」[13)] が採択され，開発された．このシステムはキーワードを入れ実行することで，数種類の検索エンジンの検索結果を比較したものを表示，検索したキーワードの検索結果を時系列で保存を行う．しかし，グーグル八分発見システムとしているが，実際にグーグル八分を発見するシステムでは無い．

3. グーグル八分について

3.1 グーグル八分とは

グーグル八分とは，単に検索エンジン Google の検索結果に表示されないことではなく，検索結果に表示されるページの一覧に対して，Google が人為的な操作を行い，特定のページを表示しないようにすることである．

単に検索結果に表示されないことがグーグル八分と間違えられ易いが，Google の検索結果に表示されない理由は人為的操作と技術的問題の 2 つが考えられる．前者は先ほど述べた通りグーグル八分であるが，後者は検索用のデータベースを構築するときの問題でグーグル八分には含まれない．具体的には，Google は検索用のデータを集める際，クローラーなどと呼ばれる自動ページ収集プログラムを使用しているが，このプログラムによるアクセスを拒否するような設定をしているページ，他のページとのリンクが全く無く，クローラーがアクセスできないページは，データベースに反映されないため，検索結果に表示されない．また，作成したばかりのページも，クローラーが情報を収集できていない場合があり，検索結果に表示されない場合がある．

3.2 グーグル八分の種類

グーグル八分は以下のように，大きく 2 つ分類することができる[13)]．

- **世界的グーグル八分**
  Google はアメリカの企業でり，アメリカ版 Google でグーグル八分になっているページは，どの国の Google でも検索結果に表示されない．このグーグル八分を世界的グーグル八分という．
- **国別グーグル八分**
  世界的グーグル八分とは別に，国ごとでグーグル八分が行われている．例としてアメリカ版 Google，日本版 Google で「悪徳商法」というキーワードで検索をすると，アメリカ版 Google では表示されるが，日本版 Google では表示されないページが存在する[16)]．日本版 Google ではグーグル八分が行われているのである．このグーグル八分を国別グーグル八分という．

3.3 どのようなページがグーグル八分になっているか

実際に Google は次のようなページを検索結果から除いている[14)]．

- Google が Web サイトの望ましい在り方として規定しているガイドライン[6)] に明らかに違反しているページ
- スパム的な手法により検索順位を向上させようとしているページ
- 児童ポルノ，麻薬販売などの犯罪に関係したり，法律に違反しているページ
- 個人や法人などの第三者が「自分の権利を侵害している」とクレームがあったページ

3.4 グーグル八分表示

グーグル八分は Google 自身が自主的に行う場合と第三者からの依頼により行う場合がある．グーグル八分が行われているとき，検索を行ったときに検索結果の一番下に，図 1 のような表示がある．

しかし，このような表示があるだけで，実際にどのページが表示されていないかを知ることはできない．表示には，ChillingEffect.org で苦情を確認できると記述されているが，実際に調べてみても得られる情報はほとんどない．

図 1　悪徳商法で検索を行ったときのグーグル八分表示例

図 2　info:の使用例

3.5　グーグル八分の影響

Google は世界的に大きなシェアを持っている．アメリカの comScore 社[3] が発表した，2007 年 12 月の世界の主要国における検索エンジンのシェアの調査[1] によると，1 位は Google でシェアは 62.4%という結果が出ている．この調査結果からは半数以上の人が検索エンジンとして Google を使っていることが分かり，その社会的影響が大きいと判断できる．グーグル八分が行われていると，Google を使っている人はそのページを見つるけることができない，つまり，その人にとってネット上に存在しないことと等しいのである．

世界の検索エンジンシェアの 6 割を占める Google に表示されないことにより，ある人物のネット上での社会的発言を封じることが可能になると危惧する意見，グーグル八分が行われないと個人や企業の誹謗中傷が行われたとき，多くの人がそれを目にすることになり，権利を侵害する可能性があると言う意見があり，一概に善悪を判断することはできない[16]．

4.　グーグル八分発見法

本研究では，日本における国別グーグル八分を発見することを目的とする．3.2 章で挙げたように，国によってグーグル八分の対象となるページは異なっていることを利用し，他国の Google と日本版 Google の検索結果と比較することで，日本における国別グーグル八分を見つけ出す．

4.1　発見方法の概要

本研究では，他国の Google としてアメリカ版 Google を利用する．なお，イギリス版 Google などアメリカ以外の国の Google と比較しても，日本におけるグーグル八分を発見できる．アメリカ版 Google との検索結果の比較を，以下のような手順で行った．

(1)　あるキーワードでアメリカ版 Google，日本版 Google で検索をする．

(2)　得られた検索結果を比較し，アメリカ版 Google にあり，日本版 Google にないページがあるか調べる．

(3)　日本版 Google のみで見つからないページを，グーグル八分の可能性のあるページとする．

しかし，これだけでは本当にグーグル八分であるかは判別することは不可能である．図 1 で示したようなメッセージ表示されるときグーグル八分であるといえる．そこで，グーグル八分であるかの確認を行うため，Google の検索オプションの「info:」を用いる．「info:」の後に URL を入れ検索を行うと，Google のインデックスに記載されている URL の情報を検索することができる．グーグル八分の可能性のあるページの URL で「info:」を用いた検索を行ったときに，図 1 で示したようなメッセージが表示されれば，そのページがグーグル八分であると判断することができる．

図 2 に，「info:」を用いた検索例を示す．グーグル八分でなければ，図 1 にあるようなメッセージは表示されない．

4.2　発見システム概要

4.1 章で挙げた操作をすることでグーグル八分を発見することができる．しかしながら，検索のために利用するキーワードをどのように取得し選択するかが問題となる．そこで，本論文では，最初にいくつかのキーワードを与えておき，そのキーワードにより発見されたグーグル八分になっているページからキーワードを新たに取得して，このキーワードをグーグル八分の検索に利用する．これを繰り返すことによりグーグル八分を発見する．

本論文では，この操作を行うプログラムを perl を

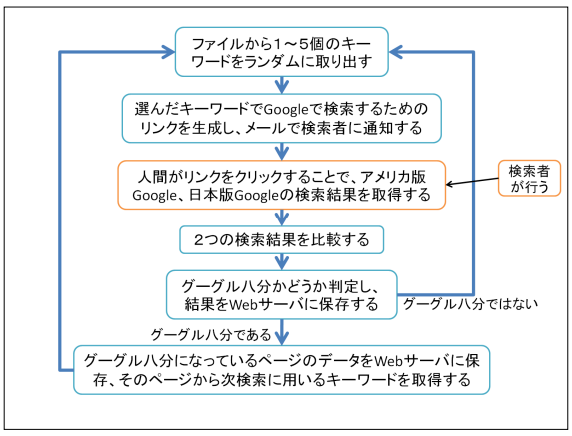
用いて実装した．この一連の操作を全てプログラムにより自動化することは可能ではあるが，Google の利用規約では，自動化された方法によりアクセスをすることを禁止していている[4]．本システムは，検索自体は人間の手によるものとして，その支援を行うシステムを実装した．

実際に，キーワードを取得するには，ページにある文章を単語ごとに分解する必要がある．単語ごとに分解する機能は，ChaSen を用いて実現した．ChaSen[2] は奈良先端科学技術大学院大学松本研究室が開発した日本語の形態素解析プログラムである．このプログラムは文字列を辞書ファイルと比較，単語ごとに分解し，それぞれの品詞を表示する．

4.2.1 プログラムの動作概要

図 3 はプログラムの動作概要図である．システムとしては，Web サーバプログラム，ChaSen，及び，本論文で作成したプログラムからなる．このプログラムは，以下の手順で出てくる一連の CGI の生成，キーワードの選択などの機能を持つ．以下では，Web サーバプログラムがグーグル八分を検索するために提供するホームページを検索用ページ，検索を行う人を検索者と呼ぶ．また，検索結果は，Web サーバプログラムにより保存され，この保存結果に対して，ChaSen を利用して，キーワードの抽出を行う．

図 3 プログラムの動作概要図



プログラムは初期設定としてキーワードを保存したファイルを用意する必要がある．その後，図 3 の動作を繰り返す．

次検索で用いるキーワードは新たに見つかったグーグル八分になっているページに対して ChaSen を使用し，取得する．しかし，取得する単語数が膨大になるため，以下の条件を設けた．

- 品詞が名詞，もしくは固有名詞の単語
- グーグル八分になっているページにおける出現回数が 10 回以下の単語

この条件を設けた理由は動詞などより名詞や固有名詞がグーグル八分発見につながると予想したこと，グーグル八分のページにだけ存在する珍しいキーワードを取得したいという 2 点である．

5. 実験結果と考察

5.1 実 験 環 境

実験には 3 台の PC を使用した．キーワードのデータは 3 台で共有するのではなく，それぞれに与え，PC ごとに実験を行った．同一の IP アドレスから短期間に多量のアクセスを行うと Google に迷惑をかけるため，各計算機にパブリックな IP アドレスを与えて，検索を行った．検索の頻度は，1 台の PC で，1 時間に 2 つのキーワードの組合せを選び，検索を行うように CGI を作成した．よって，検索者がメールを常時監視して，グーグル八分の検索を行うとすれば，1 台の PC あたりで，1 時間に 2 つの組のキーワードに関して検索を行うことが出来る．

5.2 実 験 結 果

2007 年 10 月 14 日日曜日から 2008 年 1 月 19 日土曜日までの 14 週間調査を行った．総検索回数が約 14,000 回であった．最初のキーワードとしてグーグル八分が観測されている悪徳商法関係で 2 個，宗教関係で 6 個，政治家関係で 50 個用意した．その結果，14 週間でグーグル八分になっているページを 169 ページ発見することができた．

5.2.1 グーグル八分の傾向

発見されたグーグル八分になっているページの内容を分析した．なお，発見されたグーグル八分のページの内容，つまりソースファイルを保存する機能は実験途中で追加したため，システムが保存しているページは 145 件である．これらのページを分析し，以下のように分類できた．

- **2 ちゃんねるや他の電子掲示板またそれらの過去のデータを保存している電子掲示板形式のページ**
  参加者が自由に文章などを投稿し，書き込みを連ねていくことができる Web ページのことである．投稿は時系列や記事の参照関係を元に並べられる．中でも，2 ちゃんねるは電子掲示板が集まって構成されているため，不特定多数の人からの情報が集まっている．削除されているのは，告発・企業・悪徳商法・性犯罪・麻薬・著作権・政治家・宗教などに関する記述があるページ，個人情報が公表

されているページ，誹謗中傷が行われているページなどであった．

- **wikipedia[8] などの利用者が特定の事柄に関する情報を共有するページ**
  Web ブラウザから簡単に Web ページの発行・編集などが行える Web コンテンツ管理システムである Wiki を利用している．複数人が共同で Web ページを構築していく利用法を想定していて，閲覧者が簡単にページを修正，追加することができる．電子掲示板に近いシステムである．削除されているのは，企業・裁判に関する記述があるページなどであった．
- **ブログ (ウェブログ)** 個人やグループで運営される，日記的なページである．削除されているのは，告発・犯罪・宗教・企業・裁判に関しての記述があるページであった．
- **マイナーなニュースサイト**
  地方紙や情報誌などのウェブページ向けのページである．削除されているのは，企業・不祥事・政治家・犯罪などに関しての記述があるページであった．
- **個人または団体が運営するページ**
  ある目的の下に構成されたページである．例えば悪徳商法の告発などを扱っているなどである．削除されているのは，悪徳商法・裁判・宗教・告発・医療ミス・犯罪に関して記述があるページであった．

なお，この分類は，筆者らの主観に基づいておこなわれている．表 1 は前述した分類ごとに分けたページ数，表 2 はその分類とグーグル八分の原因と考えられる内容で分けたページ数，表 3 はその分類とグーグル八分の原因と思われる理由の数で分けたページ数をまとめたものである．表 2 では，1 つのページでグーグル八分にされたと考えられる原因が複数ある場合があるので，ページ数が表 1 に比べて多くなっている．表 3 中の※は，グーグル八分にされた原因を見つけられなかったページである．表 3 における原因の数え方は，1 つの事に関しての記述が複数個ある場合でも 1 つと数えている．例えば，ある企業に対して告発，個人情報，裁判などの記述が複数個あっても 1 つと数えている．

5.2.2 システム稼働時間とグーグル八分発見数

図 4 は各週で新規に発見したグーグル八分になっているページの数をまとめたもの，図 5 は新規に発見したページだけでなく，過去に発見したページの重複を許したグーグル八分発見数を週ごとにまとめたものである．2 つの図からこのシステムによるグーグル八分

表 1 分類ごとに分けたページ数

掲示板	75 ページ
Wiki 等	17 ページ
ブログ	22 ページ
ニュースサイト	4 ページ
個人または団体のページ	27 ページ

発見数が時間とともに収束していることがわかる．

図 4 週経過に伴う新規発見数

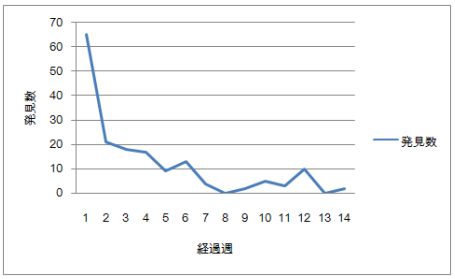
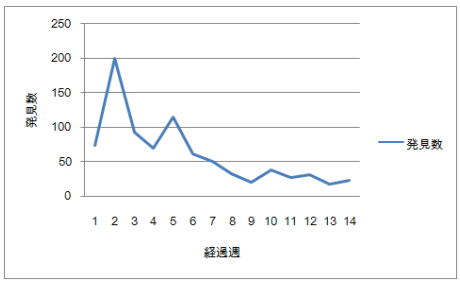


図 5 週経過に伴う重複を許した発見数



5.3 考　　察

5.3.1 グーグル八分が行われる範囲について

発見されたグーグル八分のページで URL が類似しているページは内容も類似していた．URL が類似しているページは特定のドメイン名の下にあるページの集まりであるサイトであることが多く，作者が同じであり，一つの目的に沿ってページを作成しているためだと考えられる．そのことからグーグル八分がサイト単位で行われているのではないかと推測される．そこで，グーグル八分がサイト単位で行われているかを Google の検索オプションの「site:」を用いて確認した．図 6 は実際に使用した例である．なお，実際のサイト名は伏せてある．

検索結果から同じサイト内のページは Google の検索結果に表示され，グーグル八分の対象がサイトではないということが判断できる．しかし，対象がページであると結論することはできない．例えば，ある URL

表 2　分類と原因で分けたページ数

分類 / 原因	掲示板	Wiki 等	ブログ	ニュースサイト	個人や団体のページ	合計
企業への告発	45	21	16	7	35	124
個人情報	55	0	12	4	3	74
誹謗中傷	59	0	9	0	4	72
犯罪	44	0	13	3	5	65
宗教	5	14	6	2	18	45
裁判	0	14	7	2	17	40

表 3　分類と原因の数で分けたページ数

分類 / 数	掲示板	Wiki 等	ブログ	ニュースサイト	個人や団体のページ
5 個以上	41	0	11	4	4
2〜4 個	12	0	2	0	0
1 個	14	17	9	0	23
0 個	2 ※	0	0	0	0
ページ合計	75	17	22	4	27

図 6　site:(医療関係のサイト) の結果

パターンにマッチするページ群の単位でグーグル八分を行っている可能性もある．収集したデータでは特定の URL パターンでグーグル八分を行っている事は見受けられなかったが，更なるデータを集めて分析する必要がある．

5.3.2　グーグル八分の対象について

表 1 からグーグル八分になっているページが多いのは掲示板形式のページである．誰でも簡単に匿名で書き込みが行えることから，名前，年齢，出身地などの個人情報を公表することや誹謗中傷に値するような内容が多いと考えられる．表 2 から掲示板形式のページの内，59 ページで個人情報や誹謗中傷に関するの書き込みがあったことがわかる．不特定多数の人が閲覧すると考えられる掲示板形式のページで，誹謗中傷に対してグーグル八分が行われている事は，人権侵害を防ぐことに効果的であると考えられる．

表 2 から Wikipedia などの利用者が情報を共有するページや個人などが運営しているページでは企業，宗教，裁判に関するの記述が多いことがわかる．誰かに訴えられたなどの情報は企業のマイナスイメージとつながるためグーグル八分になっているのでは無いかと考えられる．

ブログや地方紙などニュースサイトのページは企業や犯罪などのニュース記事を扱っているものが多く，グーグル八分になっている原因は告発や犯罪などで名前が公表されているからであると考えられる．

どの形式においても，過度の SEO 行っているページや児童ポルノに関連するページは発見することができなかった．

表 3 から分類によってグーグル八分の原因と思われるの数にも傾向があることが分かる．掲示板形式のページやブログ，ニュースサイトでは原因の数が多く，Wikipedia などのページや個人などが運営するページでは原因の数が少ない．この理由は，掲示板形式のページでは一つのテーマに沿って，多くの人々が書き込みを行うためであると考えられる．例えば，「悪徳商法について語ろう」というテーマで掲示板形式のページが作られたとき，多くの人々が悪徳商法に関する書き込みを行うため，どの記述がグーグル八分にされた特定が難しい．しかし，一見グーグル八分とは関係の無いテーマの掲示板に誹謗中傷の書き込みを行っている場合には，グーグル八分とされた原因が特定できる場合もある．一方，ブログやニュースサイトでは企業，不祥事，犯罪，など様々な事柄に関する記述があり特

定できないことが多い．Wikipedia などのページや個人などが運営するページでは 1 つのページで 1 つの企業，団体を扱っていて，原因を特定できる事が多い．

5.3.3 新たに発見したグーグル八分になっているページについて

以前より，悪徳商法や企業の告発に関係するグーグル八分は確認されていたが[16]，本研究において新たに医療関係のページや地方紙のページでグーグル八分になっていることを確認した．グーグル八分になっている医療関係のページと同じサイトに対して，Google で検索オプションの「site:」を用いて，サイト内にグーグル八分がどの程度含まれているか調べた．図 6 が実行例である．なお，実際の URL は伏せてある．

本システムでは 2 ページを発見していたが，この検索結果から同じサイト内で，他にもグーグル八分になっているページがあることがわかる．これらのページの主な内容は医療訴訟についてである．医療は誰にでも関係し，命に関わるものである，このような内容のページがグーグル八分になっていることは問題であると考えられる．

5.3.4 キーワード同士の関連性

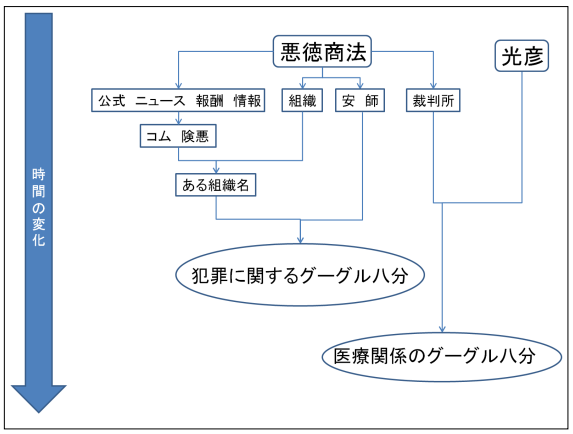
最初に与えたキーワードと，その後得られたキーワードには関連性があると考えられる．例えば「詐欺」「企業」「悪徳」「宗教」などのキーワードが多くのページで存在した．このことからシステムが発見するグーグル八分になっているページの傾向は，最初に与えたキーワードに依存すると考えられる．しかし，これらのキーワードとあまり関係が無いと思われる医療関係や別件の犯罪のグーグル八分が発見されていることから，最初に与えるキーワードとは違う分野のグーグル八分も発見できる可能性があることが明らかとなった．

図 7 はキーワードの関連性についてまとめたものである．矢印はキーワード同士の関係性を表している．例えば「悪徳商法」から「組織」への矢印は，「悪徳商法」というキーワードを利用した検索結果から発見されたグーグル八分のページから，「組織」が得られたことを表している．また，「悪徳商法」と「光彦」は実験開始時の初期キーワードとして用意したものである．図 7 から関係ないと考えられるキーワードにもつながりがあることが分かる．なお，図 7 にある「ある組織名」は本論文では伏せてある．

5.3.5 グーグル八分発見効率について

図 4 から新規のグーグル八分にされているページの発見数は時間とともに減少していることわかる．また，図 5 から重複を含んだグーグル八分にされているページの発見数も時間とともに減少していることがわかる．どちらも時間とともに収束し，グーグル八分の発見効率が下がっていくことが分かる．

図 7 キーワード同士の関連性



新規のページの発見数の合計は 169 ページ，重複を許したページの発見数の合計は 855 ページである．新規に発見したページは重複を許した場合の 2 割程度で，キーワードの取得元になったページが見つかることが多い．システムが使用したキーワードを分析してみると「アイスクリーム」「椅子」「その他」「10 月」などの一般的な名詞のキーワードを使用していた．このようなキーワードはインターネット上の多くのページに存在し，グーグル八分とは関連性が低く，グーグル八分発見につながらなかったと考えられる．グーグル八分になっているページ内には，人名や法人格を持つ団体名などの固有名詞が多くのページに存在していた．そのため「アイスクリーム」などの一般的な名詞ではなく，団体名などのキーワードで検索を行ったほうがグーグル八分発見の効率が良いと考えられる．

本システムではキーワード取得に形態素解析システム ChaSen を利用したが，ChaSen は自身の辞書ファイルに登録していない単語を認識することができない．例えば「グーグル八分」は「グーグル」と「八分」で認識される．実際にグーグル八分になっているページ内で使用されている，ChaSen に認識できないできない固有名詞を用いて検索を行ったところ，新たにグーグル八分になっているページを発見することができた．よって，固有名詞を用いる方が発見の効率がよいと思われる．

また，キーワードの取得元をグーグル八分になっているページに限定せず，ニュースサイトやブログから最近起こったネガティブな事件や話題になっているキーワード等を取得する機能が効果的であると考えられる．

5.4 Google の検索結果の不安定性について

本研究において，稀にグーグル八分になっているページが表示される事が観測された．これは Google がデータベースを更新する際に検索結果が不安定になることが原因であると考えられる．グーグル八分表示がある検索結果を時系列で観測し，稀に表示されるグーグル八分になっているページを見つけ出す事で，世界的グーグル八分発見につながる可能性がある．

6. おわりに

本論文でグーグル八分を発見し，分析を行ったが，収集できたデータが少ないため，まだ不明な点が多い．さらなる分析を進めるためにもより多くのデータを収集する必要がありシステムの改善が求められる．また，世界的グーグル八分の発見，グーグル八分の表示が無くなった場合，検索結果から削除されるのでは無く，ランキングを下げられ，結果として検索結果に表示されなくなる場合，それぞれに対応する機能を実現する必要がある．具体的には以下のような点が課題として挙げられる．

- 「アイスクリーム」などの一般的な名詞などではなく，企業名や団体名などのキーワードを認識し，取得する機能が必要である．
- ニュースサイトやブログから企業名などの固有名詞，最近知られるようになった未知語を取得する機能が必要である．
- グーグル八分になっているページの内容には関連する内容が多い．例えば企業と裁判などである．そのため，特定の分野でのキーワードの組合せ方法を改善する必要がある．
- システムが発見した医療関係のグーグル八分において「site:」を用いた検索を行ったところ，新たに 17 ページのグーグル八分を発見することができた，そのことからグーグル八分になっているページが存在するサイトはグーグル八分と関連性が高いと考えられるので，サイト単位で調べる機能を組込むことでグーグル八分の発見ができると考えられる．
- 時系列における検索結果の監視を行い，グーグル八分を発見する機能を実装する機能が必要である．
- 本研究ではリンク構造に関しての分析を行わなかったが，ページ同士の関係やリンク構造を分析をすることにより，グーグル八分を発見できるかどうかを検討する必要がある．

参 考 文 献

1) Baidu Ranked Third Largest Worldwide Search Property by comScore in December 2007 [comScore] http://www.comscore.com/press/release.asp?press=2018
2) ChaSen's wiki http://chasen.naist.jp/hiki/ChaSen/
3) comScore http://www.comscore.com/
4) Google 利用規程, http://www.google.com/accounts/TOS
5) 「Google の秘密-PageRank 徹底解説」, 馬場肇 http://www.kusastro.kyoto-u.ac.jp/baba/wais/pagerank.html
6) Google について http://www.google.co.jp/intl/ja/about.html
7) The OpenNet Initiative: Internet Filtering in China in 2004-2005: A Country Study, (June 2004), http://www.opennetinitiative.net/studies/china/ONI_China_Country_study.pdf
8) Wikipedia http://ja.wikipedia.org/wiki/Main_Page
9) J.R. Crandall, D.Zinn, and M.Byrd, ConceptDoppler: A Weather Tracker for Internet Censorship, Proceedings of the 14th ACM Conference on Computer and Communications Security, pp.352–365, 2007
10) R.Clayton, S.J.Murdoch, and R.N.M.Watson, Ignoring the Great Firewall of China, Proceedings of 6th International Workshop of Privacy Enhancing Technologies, LNCS 4258, pp.20–35, 2006
11) T.Tashiro, T.Ueda, T.Hori, Y.Hirata, H.Yamana, EPCI:Extracting Potentially Copyright Infringement Texts from the Web, Proceedings of the 16th Internatioanl Wolrd Wide Web Conference, pp.1151–1152, 2007
12) 独立行政法人情報処理推進機構 http://www.ipa.go.jp/
13) グーグル八分対策センター http://www.google8bu.com/
14) グーグル村上社長“ Google 八分 ”を語る：ITpro http://itpro.nikkeibp.co.jp/article/NEWS/20060630/242220/
15) 吉田泰明，上田高徳，田代崇，平手勇宇，山名早人，商用検索エンジンのランキングに関する定量評価と特徴解析，情報研報 (DBS)，Vol.2007, No.65, pp.441–446, 2007
16) 吉本敏洋，『グーグル八分とは何か』九天社，2006 年