A2-3 リンク情報に基づく検索エンジンの比較

樺島 結城† 廣川 佐千男‡
†九州大学大学院システム情報科学府 ‡九州大学情報基盤センター

WWW 上で必要な情報を探すために、我々は一般に検索エンジンを用いている。このような検索エンジンは多数あり、我々はそれら検索エンジンの検索結果に対し漠然と良い、悪いという感想を持ち、それぞれの気に入った検索エンジンを利用している。検索エンジンはそれぞれ異なった検索手法を持ち、また、データベースに持っている情報の量なども違う。従ってその検索結果がそれぞれ異なることは当然であるが、検索結果の善し悪しは多分に主観的なものである。本発表では、ユーザが漠然と持つそれぞれの検索エンジンに対する印象を、検索結果のページについてのリンク情報を用いて数量化する方式を提案する。具体的には、検索結果のページ群について入次数と出次数、オーソリティ度とハブ度という二組の尺度を用いて検索エンジンの比較を行う。

1 はじめに

今日のインターネットの普及に伴って、WWW 上の情報量は飛躍的に増加し、この情報の氾濫の中で自らの求める情報を得るためには検索エンジンの利用が欠かせないものとなった。しかし、多数ある検索エンジンの中から、どの検索エンジンを利用するかに応じて、求める情報を得るのにかかる手間と得られる情報の質が大きく異なることもしばしばである。このような状況の中で、各検索エンジンの特徴を解明することは、検索エンジン選択のためには有用である。本論文では、検索結果から得られるリンク情報を分析することによって検索エンジンの特徴を比較する。

あるページから多くのリンクが出ていれば、そのページは多くの情報の中心的ページと考えられる。また、あるページが多数のページからリンクされていれば、そのページは他のページの作者から権威とみなされていると考えられる。

ところで、Web ページを節、ページ間のリンクを枝とみなせば、Web 空間は有向グラフと考えられる。あるページから出ているリンクの数と、そのページへのリンク数は、このグラフにおける出次数(outdegree) および、入次数(indegree)になる。本論文ではまず、入次数と出次数を尺度として Web 空間の特性を解析する。

分析の対象として、一様なリンクの分布を持つランダム空間、一般の Web 空間、検索結果で決まる空間（これを検索空間と呼ぶ）の３種類の空間について、入次数と出次数によって比較を行ない、検索空間の持つ特殊性を示した。次に、各検索エンジン毎に求まる検索空間を、入次数と出次数を用いた分析により比較を行なった。

入次数と出次数は、そのページの評価や有用さとも考えられるが、各ページの評価は互いに依存しあって決まる。第４章では Kleinberg の HITS アルゴリズム[4]で求まるオーソリティー度、ハブ度という尺度について述べ、第５章ではこの尺度を用いて検索エンジンが持つページ情報をとらえ、その分布により検索エンジンを比較した。

2 入次数と出次数を用いた Web 空間の解析

2.1 入次数と出次数

ページを節、リンクを枝と考えると、Web 空間は有向グラフとみなせる。その意味で、あるページから出ているリンクの数を出次数(outdegree)、

あるページを指しているリンクの数を入次数(indgree)と呼ぶ。

Web空間におけるリンクは人為的に関係付けたものであり、当然そこには意味がある。つまり、あるページにおけるリンクとは、その製作者の基準により選ばれたページへのリンクであると言える。よって、Web上において多数のページからリンクされているページは良質なものであると思える。特定のキーワードに関連付けられた集合においてはなおさらであろう。また、多数の有用なページに対してのリンクを持つページもまた、情報を収集する際の起点という意味で有用であると考えられる。

以上の理由により、あるWeb空間に含まれる各ページについて、入次数、出次数がどのように分布しているかを分析することにより、その空間の特徴がわかるのではないかと考えた。

2.2 Web空間と検索空間の解析

検索エンジンに対してキーワード検索を行った結果から上位50件のページを取り、初期集合と呼ぶ。初期集合中のページについて、それぞれのページからリンクしているページと、それぞれのページへとリンクしているページを加えた集合を求め、これを基底集合と呼ぶ。以降、本論文ではこのような集合を検索空間と呼ぶ。なお、今回の実験では逆リンクはAltaVistaの逆リンク検索を利用し、上限を50件として収集している。

Web空間におけるリンクはページの製作者の意思によって張られる。したがって、ランダムに生成されたグラフとは異なった特徴をもつと考えられる。また、検索空間においては、Web空間に対して検索という作業を行っているため、その空間はWeb空間と異なった特徴を持つと考えられる。そこで、Web空間、検索空間の特徴を調べるため、一様分布に従うランダムなグラフ、Web上からランダムに生成したWeb空間、検索空間のそれぞれについて、各ノードの入次数、出次数をもとめ、その分布をプロットし、解析を行った。

一様分布によるグラフについては、節数3000、枝数1000の一様分布に従う有向グラフをランダムに生成し、これについて入次数、出次数を求めた(図2.1)。具体的には、3000×3000次元の行列において、i、jを一様分布となるようにランダムに選択し、i番目の節からj番目の節への枝を作ることによりグラフを構成した。節数、枝数を変えても同様に、大きなばらつきがなく一様に入次数、出次数が左下半部に分布することが確認できた。

Web空間全体から、ランダムにある個数のページを選び出すことは容易にできることではなく、それ自体が一つの研究テーマとなっている。LawrenceとGiles[7]は、複数の検索エンジンの検索結果の重複を計算することにより、世界のWebページの総数の推定を行なった。日本国内を対象とする同様の実験は、来住等[8]によっても行なわれている。本稿では以前ロボットで収集した国内のドメイン名86687件のデータの中から100件ランダムに選び、そこから順リンク、逆リンクをたどった空間をランダムなWeb空間として入次数、出次数を求めた(図2.2)。実際のWebの空間においては、図2.2のように入次数の方が出次数よりも高くなる傾向にある。これは、リンクを張るのは主に人手で行うという理由のためであると考えられる。

Web空間におけるリンクは、それぞれのページの製作者の考えにより張られているので、その分布は一見ランダムなもののように思われる。しかし、最近の研究によりWeb空間はランダムなグラフではなく、その入次数、出次数がべき分布(Power Law)に従うことが一般に知られている[1,5,6]。ここで、出次数がべき分布に従うということは、出次数がxであるノードの個数をyとするとき、ある定数a、bについて、log(y)=a-b*log(x)となることである。x、yの代わりにlogをとった値X、Yで考えると、Y=a-b*Xとなる。(X,Y)をプロットすると、x切片がa,傾きが-bの直線となる。

しかし、検索空間においては一様分布ともべき分布とも違う傾向がある。図2.3はinfoseek

Japan に対して Artificial と Intelligence の and 検索を行った検索空間についての入次数、出次数のグラフである。このグラフから、検索空間においては入次数や出次数が高いページも一般の Web 空間に比べて多く存在することがわかる。

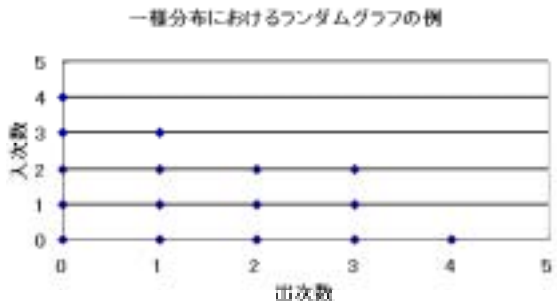


図2.1: 一様分布のグラフ(節数3000、枝数1000)。横軸に出次数、縦軸に入次数

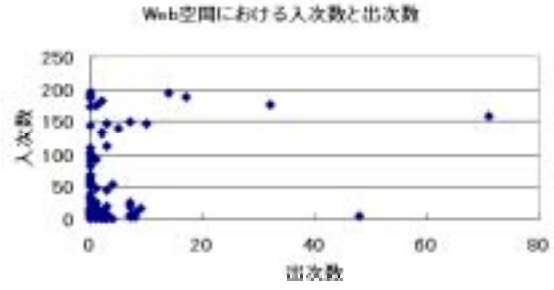


図2.2:ランダムに生成した Web 空間

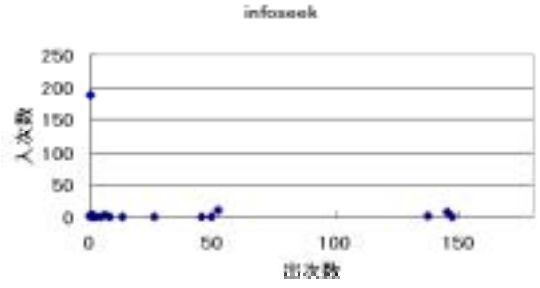


図 2.3：検索空間（infoseek Japan、キーワード：Artificial Intelligence）

3　入次数と出次数を用いた検索エンジンの比較

3.1　実験方法

Alta Vista[1]、goo[2]、google[3]、infoseek Japan[4]の四つの検索エンジンに対してキーワード検索を行い、その結果に対して検索空間を求める。この検索空間の各ノードに対してそれぞれ入次数、出次数を求め、プロットする。

実験に使用したキーワードは、citeseer[5]の Computer Science Directory と、yahoo![6]のカテゴリーの中から選んだ。

3.2　実験結果の分析

例としてキーワードに Artificial と Intelligence の and 検索を行った際の各検索エンジンの結果を図 3.1～図 3.4 に示す。

まず、第一に検索空間について気付くのは、その基底集合の広がりの違いである。検索結果の上位から同数のページを起点とした空間でありながら、ノードの数に大きな違いの出ることがある。goo や google においては検索対象が変化してもそれほどノード数の違いは大きくなく、平均して広がりのある検索空間を返す傾向にある。一方、Alta Vista や infoseek Japan では、検索対象によっては非常に狭い空間を返すことがある。

入次数,出次数については検索に使用するキーワードによって違いがあるのだが、検索エンジンごとの特徴をはっきりと示すような結果は得られなかった。ただ、infoseek においては出次数の多いノードを多く持つ検索空間を返すことが多い。

これらのグラフの特徴は、プロットされた点が主に軸線上に存在しており、入次数、出次数ともに高い値を持つノードが数少ないということである。この特徴は Web 空間の特徴でもある。Web 空間において各ページはリンクされているかリンクしているかのどちらかに偏っている場合が多いということになる。

---

[1] http://www.altavista.com/
[2] http://www.goo.ne.jp/
[3] http://www.google.com/
[4] http://japan.infoseek.com/
[5] http://citeseer.nj.nec.com/cs/
[6] http://www.yahoo.com/

また、一般のWeb空間では見られなかったような、出次数の高いページが各検索空間で見られる。このような出次数の高いページを実際に確認したところ、大学等のリンク集や検索サイト、用語集のインデックスなどであり、AIやArtificial Lifeについてのページであった。goo、googleの検索空間について見られる入次数、出次数がともに高いページについても調べたところ、googleの入次数194、出次数58のページはyahoo!のAIについてのディレクトリ(Science/computer_science/artificial_intelligence)であった。

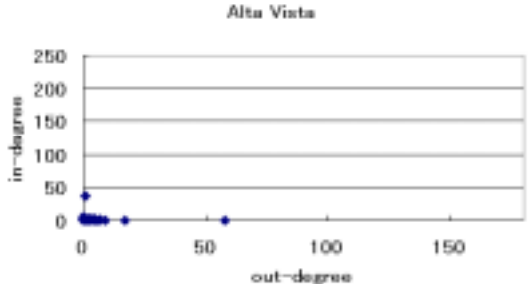


図3.1:AltaVistaにおける入次数,出次数の分布

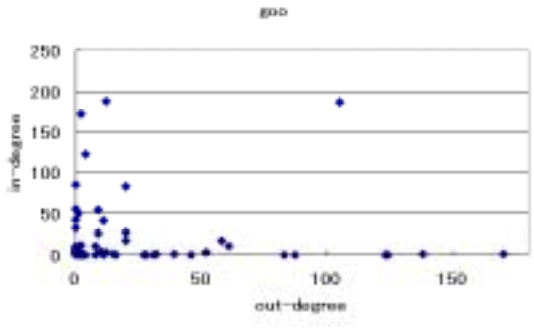


図3.2:gooにおける入次数,出次数の分布

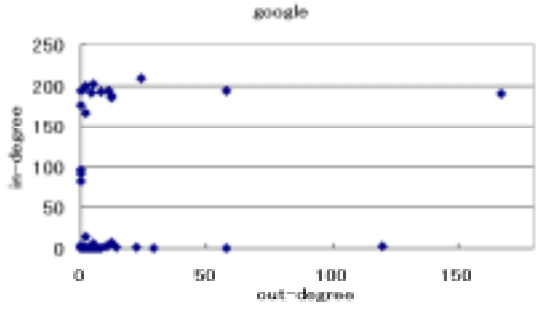


図3.3: googleにおける入次数,出次数の分布

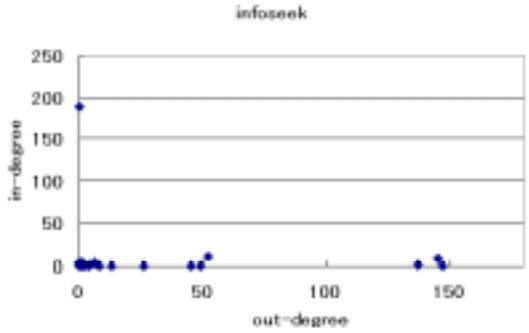


図3.4: infoseek Japanにおける入次数,出次数の分布

4　ハブ度とオーソリティー度

4.1　入次数と出次数の問題点

第3章では、入次数、出次数を用いた分析を行ったのだが、この手法には欠点が存在する。たとえばトップページやインデックスでは多くのリンクが存在する場合もあるが、そのような場合においてそこから参照されるWebページは、他からは参照されていない、キーワードとは無関係の内容であるかもしれない。このような時、出次数は大きくなるのだが、ユーザにとって不要なリンクも多く含み望む結果が得られない。

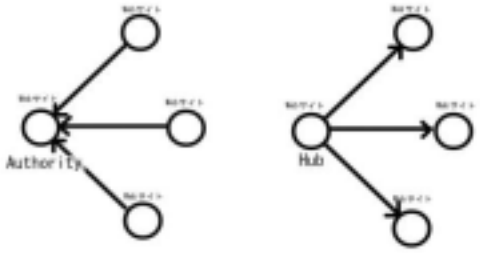
上記のような欠点を改善するためには、不要なリンクによる影響を取り除く必要がある。そこで、Kleinberg等はHITSアルゴリズムを導入した。これは、各ノードに値を与えることでリンクに重みをもたせ、不要なリンクの影響を取り除くものである。

4.2　ハブ度とオーソリティー度

KleinbergはWeb上のページの質を計る基準としてハブ度、オーソリティー度という尺度を定義している[4]。

現実の世界においては、あるページの管理者がリンクを張るときには当然その管理者が有用であると考えるページに対して張り、不要に思うページに対してリンクを張ることはまず無い。したがって、同一のキーワードを持つ数多くのページからリンクされるようなページは、その分野におけ

る権威(authority)である可能性が高いといえる(図4.1(a))。また、多くの有用なページをリンクしているようなページは検索の開始点(hub)の役割(図4.1(b))を果たしており、このようなページからリンクされているページは、たとえ広く知られていなくても、有用である可能性は高い。



(a) (b)

図4.1:オーソリティ、ハブの直感的概念

具体的には、ハブ度はそのノードがリンクしているノードのオーソリティー度の合計であり、オーソリティー度はそのページをリンクしているノードのハブ度の合計である。x_i, y_i をそれぞれノード i のオーソリティー度、ハブ度とすると、

$$x_i = \sum_{i \leftarrow j} y_j \quad , \quad y_i = \sum_{i \rightarrow j} x_j$$

となる。

つまり、ハブ度が高いということは質(オーソリティー度)の高いページを数多くリンクしていることになり、それだけ質の高い中心となるページであることをあらわす。同様に、オーソリティー度が高いということは中心としての質(ハブ度)の高いページから多くリンクされていることを表し、より権威あるページであることを表すのである。

図4.2に示すような小規模で具体的な例についてみてみると、このグラフにおけるハブ度、オーソリティー度は表4.1のようになる。①, ⑥のような、多くのノードからリンクされているノードは、表4.1のようにそれぞれ0.657192という高いオーソリティー度を持つ。一方、⑩は、そのノードをリンクするノードの数が少ないのだが、リンクしているノード②、⑦のハブ度がそれぞれ0.435162と高いので、オーソリティー度が0.369048と高い値となっている。

また、ハブ度についても、③、④、⑧、⑨、⑩はすべて一つのノードに対してのみリンクを持っているのだが、③、④、⑧、⑨のハブ度が0.278673であるのに対して⑩のハブ度は$8.07633\times10\text{-}11$という非常に低い値になっている。これは、ノード①、⑥が高いオーソリティー度を持つのに対し、⑩がリンクしているノード⑦のオーソリティー度が低いからである。

このように，オーソリティー度とハブ度はWebにおけるリンクの意味を利用することで、そのページの性質を導き出している。

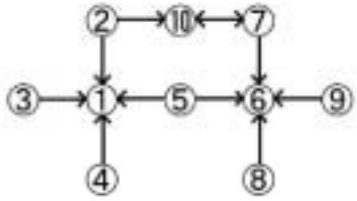


図4.2:ノード数10での具体例

ハブ度	オーソリティー度
0.000000	0.657192
0.435162	0.000000
0.278673	0.000000
0.278673	0.000000
0.557345	0.000000
0.000000	0.657192
0.435162	$1.90464E-10$
0.278673	0.000000
0.278673	0.000000
$8.07633E-11$	0.369048

表4.1：図4.2のグラフにおけるハブ度とオーソリティー度

5 ハブ度とオーソリティー度を用いた検索エンジンの比較

5.1 実験方法

第3章の実験で対象とした検索空間について、その各ノードのオーソリティー度、ハブ度をHITSアルゴリズムによって求め、その分布をプロットした。

5.2 実験結果の分析

第3章と同様、キーワードとしてArtficialとIntelligenceをand検索した検索空間について、各検索エンジンのハブ度、オーソリティー度の分布を図5.1～図5.4に示す。

一見して、Alta Vistaがハブ度の大きなノードを返しているのに気付く。オーソリティー度、ハブ度のベクトルはそれぞれ正規化されているため、値が1のものが存在するということは、すなわち中心としての値を持つノードが一つしかなかったということである。オーソリティー度に関しては0.1313程度の同じ値を持つノードが複数あったが、これらは全て、ハブ度が1のノードからのリンクがあるノードである。このことが示すのは、図5.1におけるAlta Vistaの検索空間は、一つのかたまった集合のみがオーソリティー度、ハブ度を持ち、多くのノードやリンクがオーソリティー度やハブ度には関与しない空間であったことを示す。

図5.4のinfoseek Japanについての図では、infoseekの特徴がよく出ている。入次数、出次数についての比較においても、出次数の高いノードを返す傾向があったが(図3.4)、ハブ度、オーソリティー度についての比較ではよりはっきりとその傾向が見て取れる。これはinfoseek Japanの大きな特徴といえる。

gooとgoogleについては、ハブ度、オーソリティー度ともに大きな偏り無く存在しており、特にgoogleではハブ度、オーソリティー度がともに高いノードが存在する場合が多い(図5.2、図5.3)。

次に、オーソリティー度と入次数、ハブ度と出次数の関係を図5.5、図5.6にしめす。各検索エンジンの違いが分かりやすいように4つの検索エンジンの結果を同一のグラフ上に示している。

第3章において入次数、出次数の高いページについて実際に調べたのと同様、オーソリティー度、ハブ度が高いものについて実際に調べてみた。Alta Vistaやinfoseekでは入次数、出次数の高いページがそのままオーソリティー度、ハブ度の高いページであった。一方、gooやgoogleにおいては入次数や出次数が高くないページが高いオーソリティー度やハブ度を持つ場合があった。図5.5、図5.6を参照すると、確かにそうであることがわかる。

また、googleにおいてオーソリティー度の高いページは、JAIR(Journal of Artificial Intelligence Reserch)と、IIT(The Institute for Information Technology)のAIに関するページであり、オーソリティー度が高いことのうなずけるページであった。これは、恐らくgoogleが検索のランキングにリンク情報を利用していることが関係していると思われる。

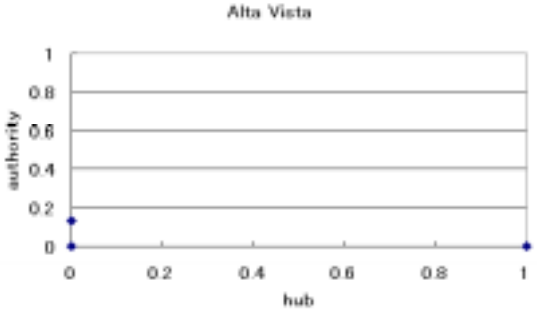


図5.1: Alta Vistaにおけるハブ度、オーソリティー度の分布

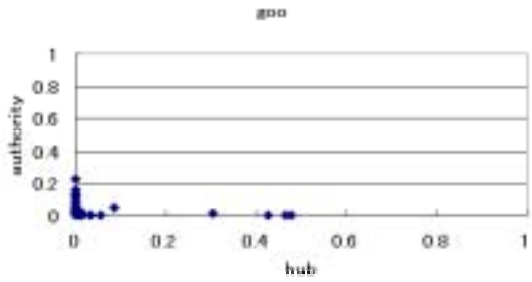


図5.2：gooにおけるハブ度、オーソリティー度の分布

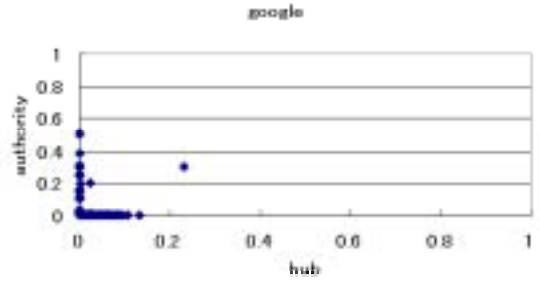


図5.3:google におけるハブ度、オーソリティー度の分布

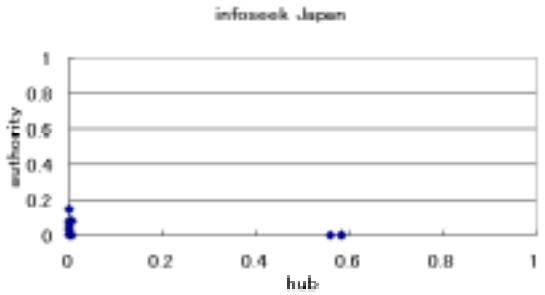


図5.4:infoseek Japan におけるハブ度、オーソリティー度の分布

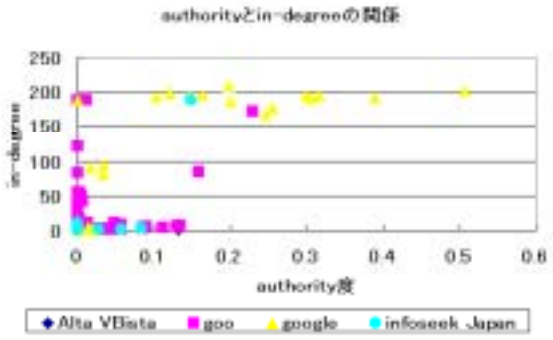


図5.5:検索空間でのオーソリティー度と入次数の関係

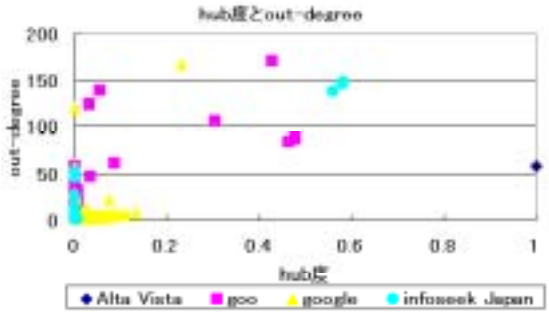


図5.6:検索空間におけるハブ度と出次数の関係

６　まとめと今後の課題

リンク情報を解析することで、一般の Web の空間と検索空間との違いを見出し、また、各検索エンジンの特徴や傾向をある程度比較することができた。これは、Web におけるリンクというものの特殊性があるからこそ可能なことである。しかしながら、リンク情報は Web の持つ性質の一側面であり、このことだけで検索エンジンによる検索空間や Web グラフの性質の全てを表現することはできない。

特異値分解における特異値の推移(スクリープロット)を見ることで、その空間の複雑さを推し量ることができる[9]。したがって、各検索エンジンについて特異値という面から評価することで、今回の実験とはまた違った側面が見えてくるであろう。また、最大の特異値以外の特異値に対応する特異ベクトルを用いることで、クラスタリングを行うことも期待できる。

特異値の推移という点では、絞込検索による空間の変化という点についての研究も考えられる。

参考文献

[1]R. Albert, H. Jeang and A. Barabasi, “Diameter of the World Wide Web,”Nature,401,130-131, 1999.

[2]M. W. Berry, S. T. Dumais and G. W. O’Brien, “Using Linear Algebra for Intelligent Information Retrieval,”Computer Science Department, CS-94-270, December 1994.

[3]M. R. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork, “Mesuring Index Quality using Random Walks on the Web,” Proceedings of the 8th WWW, 213-225, 1999.

[4]J. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” Technical Report RJ 10076,

IBM, May 1997.

[5]J. Kleinberg, S. R. Kumar, P. Raphavan, S.Rajagopalan and A. Tomkins, "The Web as a Graph:Measurments, Models and Methods," Proceedings of the International Conference on Combinatorics and Computing, 26-28, 1999.

[6]R. Kumar, P. Raphavan, S. Rajagopalan, D. Sivakumar, A. S. Ztomkins and E. Upfal, "The Web as a Graph," Proceedings of the 32nd ACM Symposium on Theory of Computing(STOC 2000),171-180, 2000.

[7]S. Lawrence and C. Lee Giles, "Accessibility of information on the Web," Nature, 400, 107-109, 1999

[8]来住伸子, 大森貴博,笹塚清二,近藤晶子,水谷正大,小川貴英, "統計的推定による日本語Webの調査," インターネットコンファレンス99論文集, 21-28, 1999.

[9]廣川佐千男,池田大輔, "Web グラフの構造解析," 人工知能学会誌16(4),525-529,2001