社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＦＩ－74　(13)
2004－ＤＤ－43　(13)
2004／3／26

リンク構造と共起関係を用いたWeb空間の視覚化

山本仁志[1)] 石田和成[2)] 岡田勇[3)] 太田敏澄[1)]

hitoshi@is.uec.ac.jp

電気通信大学大学院 情報システム学研究科[1)]
東京農業大学 国際食料情報学部[2)]
創価大学 経営学部[3)]

あらまし：
本研究では、リンク構造において結びつきの強いサイトと、用語間関係において結びつきの強いサイトを２次元平面上で視覚化することで情報検索を支援する。Webで情報検索をする際には、検索対象に対するアンカーとなるサイトを見つけたら、その周辺にどのような情報があるのかを辿ることが多い。しかし、リンクを辿るだけでは、直接のリンク関係はないが、関連する情報を見つけることができない。検索エンジンだけでは、アンカーサイトの周辺情報を集めることが困難である。更に、同じハイパーリンクでも、アンカーサイトと関連の深いサイトへのリンクとポータルなどへのリンクでは重みが違うと考えられる。リンク構造と共起関係の平面上に視覚的に表示することで、情報検索支援が可能になると期待される。本研究で提案する手法により分析した結果、「関連の深いサイト」「関連があると考えられるサイト」「無関係なサイト」という分類が可能になった。

Analysis of Web Space using Link Structure and CTR

Hitoshi YAMAMOTO[1)], Kazunari ISHIDA[2)]

Isamu OKADAand[3)] and Toshizumi OHTA[1)]

University of Electro-Communication[1)]
Tokyo University of Agriculture[2)]
Soka University[3)]

Abstract: We have analyzed internal relations among web sites using link structure and click through rate (CTR). Information retrieval tends to begin with an anchor and related information is searched for using hyperlinks. On the one hand, it is difficult to find related information simply by tracing a link. On the other hand, it is difficult to collect information associated with an anchor simply by using a search engine. Furthermore, even with the same hyperlink from the anchor, the weight of the link is different from that to a related site in a link to a portal. Based on this analysis, we propose a two-dimensional map that has two axes (link structure and CTR) to support information retrieval. Using this map, we can classify web sites into "closely related to anchor", "related to anchor", and "not-related to anchor".

1. はじめに

ハイパーテキストのリンク構造を分析することで、Web上のサイトの重要度を定義し、情報検索を支援する試みは多くの研究者によってなされている(Kleinberg,1998), (Brin and Page,1998)。高

橋・赤堀(1999)は、情報検索支援のために、相互に結びついたリンクを群として捉える手法を提案している。

本稿の提案では、相互のリンクにも結びつきの強いリンクと弱いリンクがあると考え、結びつきの強弱を提示することで、より効率的な情報検索がおこなえると考えた。更にリンク構造だけでなく意味的に近いサイトを提示することが情報検索支援に効果があると考えた。本稿では、情報検索行動の中で、検索者が注目したサイト（アンカーサイト）の周辺にどのようなサイトが存在しているのかを示すことで情報検索を支援することを試みる。ここでいう周辺とは、アンカーサイトと扱っている情報が意味的に近いサイトや、ハイパーリンクで辿れるリンク構造において近いサイトを指す。例えば、学術的な情報を検索している場合、アンカーサイトと意味的に近いサイトは、同様の研究テーマを扱っているサイトと考えられる。また、リンク構造において近いサイトは、そのサイトの運営者と学会や大学において近い関係にある研究者であると考えられる。こうした周辺情報を効率的に探索することで、検索者は検索対象とする情報を網羅的に検索することができると考えられる。本稿では、インターネット上の情報検索行動を支援するために、Web 空間を「リンク構造」および「共起関係」の平面で視覚化する分析手法を提案する。

2. 視覚化システムの開発

本節では、リンク構造と共起関係による視覚化システムの概要を述べる。サイト間の関連性を議論するためには、ひとつのサイトの範囲を規定する必要がある。本研究では、以下の規則にしたがってサイトの範囲を規定している。URL がチルダ（~）より下層のディレクトリは一人のユーザによって管理されていることが多いと考えられるため、チルダ以下はひとつのサイトとする。また、チルダが使われていない場合、最下層のディレクトリをひとつのサイトとすることとした。

以下に、リンク構造と共起関係による視覚化システムの概要を述べる。

ハイパーテキストのリンク構造は有向グラフとして捉えることができる。サイト間におけるハイパーリンクの有無を隣接行列として表現する。例として図 1に、リンク構造とそれに対応した隣接行列の例を示す。これを正規化することで、推移確率行列 A が得られる。これを用いて、アンカーサイト i からのリンク近接度 $Dist_i$ を式（1）のように定義する。

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

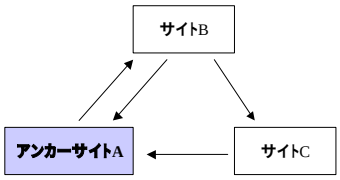

図 1：リンク構造と隣接行列

図 1の行列の各行を正規化することで、正規化された隣接行列 A が得られる。これは、あるサイトからハイパーリンクをひとつ辿ったときに他のサイトに到達する推移確率を現している。また、この行列を累乗することで、n クリック後に到達している確率となる。また、の転置行列は、図 1の転置行列を正規化することで、ひとつのハイパーリンクを辿ったときに自サイトに到達する被推移確率となる。これらを用いて、あるサイトから他のサイトへのリンク構造における近接度（リンク近接度）を $Dist_i$ として定義する。

$$Dist_i = -\log \delta_i \left(\sum_{j=1}^{k} w_j (A^j + A'^j)\right) \quad (1)$$

$Dist_i$：サイト i を起点とした、他の各サイトの i とのリンク距離

δ_i：i 番目の要素のみ 1、それ以外の要素は 0 の（$1 \times n$）ベクトル

k：推移確率計算上の推移回数の上限（現在は k=9 を利用（経験的な収束値））

A：正規化された隣接行列

A'：正規化された隣接行列の転置行列

w_j：重み係数。本稿では $w_j = 2^{-j}$

また､サイト間の意味的な近接関係は、用語間の共起関係を用いて定義する。文書間の意味的な関係を CTR によって視覚化する研究には、Ishia and Ohta(2002)がある｡アンカーサイト i と他のサイト j 間の用語近接度は、CTR を用いて定義する。CTR 行列 T の要素 t_{ij} は次のようにあらわせる。

$$t_{ij} = 1 \Big/ \left(\frac{CTR(i,j)}{|TR(i) \cup TR(j)|} \right) \qquad (2)$$

アンカーサイトから、リンク構造、用語間関係、の平面に展開したマップを作成し平面を図 2のように４つに分類することができる。第一は「アンカーサイト周辺かつ関連領域」である。第二は「アンカーサイトとは繋がりがないが関連領域」、第三は「繋がりはあるが、特に関連ではない」、そして最後に「つながりも関連も薄い」である。上記の分類によって１，２を重点的に検索すればよいという指針が得られると期待される。

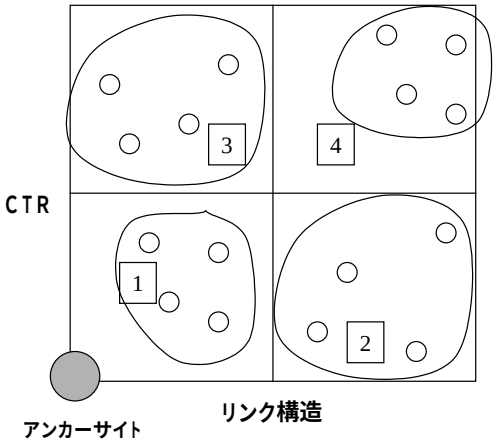

図 2：リンク構造と CTR による Web サイトの分布

3. Web 空間の分析

本稿では試験的にいくつかのアンカーサイトを起点に分析をおこなった。アンカーサイトは、マーケティングサイエンスを専門とする研究者のページ[1]、および経営情報を専門とする研究者のページ[2]を用いた。アンカーサイトは、筆者らが分析結果を観察して、アンカーサイトとの関連を考察するために、筆者らの専門領域と関連あるサイトを選択した。

図 3は、[1]をアンカーサイトに分析した結果である。この結果から関連サイトは A,B,C の３領域に分類できる。A の領域は、アンカーサイトの所属する学科の公式サイトと、フィールドリサーチを専門とするの研究者のサイトである。このことから、リンク構造における分析によって関連の深いサイトが抽出できていることがわかる。B の領域には、統計学や数学を専門とする研究者など関連があると考えられるサイトのほかに、掲示板サービスのトップページなどが含まれている。これは、サイト集約の際に最下層のディレクトリをひとつのサイトとしたことによるものと考えられる。また、CTR の軸では、非常に近接度が近い領域に掲示板サービスなどが抽出された｡これは、CTR による近接度を計算する際に、用語数で正規化した影響であると考えられる。領域 C には、やはり掲示板サービスやポータルサイトなどが抽出されている。これらはアンカーサイトとは関連がないと考えられるサイト群である。

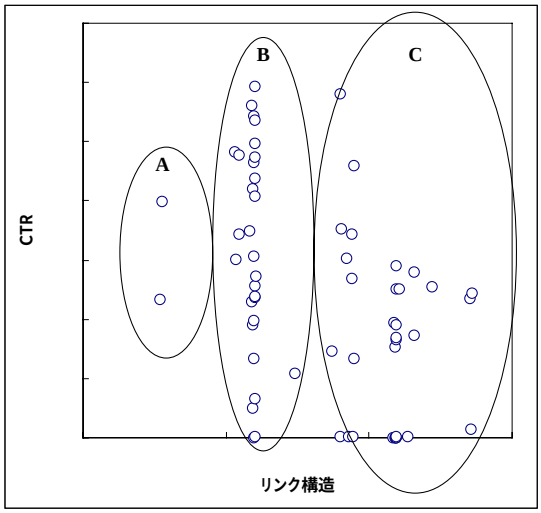

図 3：アンカーサイト[1]のリンク-CTR 平面による図示

次に領域 B にサイトが集中しているた

め、この領域を拡大したものを図 4に示す。B-1 には、[1]の研究者のゼミの連絡掲示板や、論文記述の手法に関するサイトが抽出された。これは、[1]のコンテンツに「論文作成に役立つリンク集」があるためである。B-2 には、周辺領域の研究者のサイトが多く観察されている。

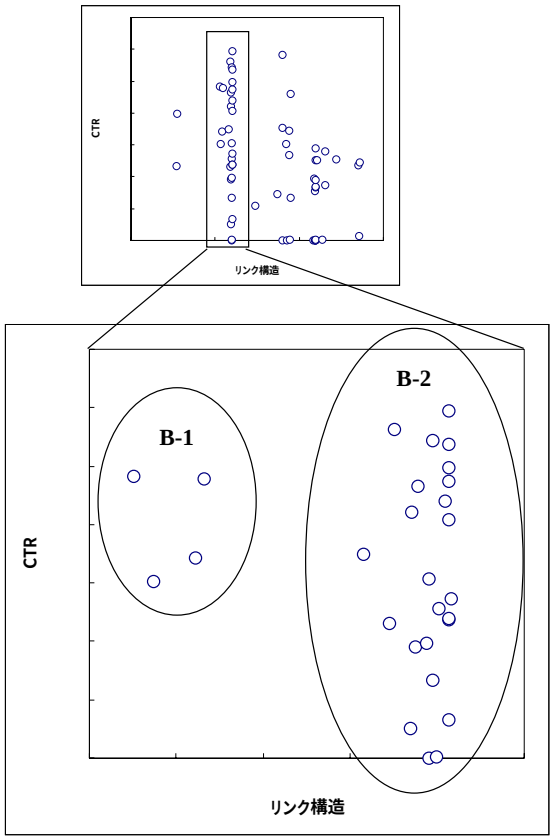

図 4：アンカーサイト[1]の領域 B の拡大

表 1は、アンカーサイト[1]に関して、領域Bのサイト群をリンク近接度が近い順にまとめたものである。ただし関連性の項の「○」は、本人が関連しているサイトであり、「△」がマーケティングや論文作成手法に関するサイト、「×」は無関係と思われるサイトである。「？」は、ディレクトリで集約したためにこの URL では閲覧できないものである。

表 1からわかるように、リンク構造が近いほど本人に関連したサイトが多く、遠くなるに従って領域的にも関係のないサイトが増えていることがわかる。

表 1：アンカーサイト[1]の領域 B に含まれる URL

関連	URL
○	www.hit-u.ac.jp/commerce
○*	6033.teacup.com/matsuizemi2001
△	hostgk3.biology.tohoku.ac.jp/sakai
△	sc1.cc.kochi-u.ac.jp/~yoshikur
○	marketing.cm.hit-u.ac.jp/~matsui
○	finito-web.com/doctormatsui
△	web.cc.osaka-kyoiku.ac.jp/~shakai
△	grape.c.u-tokyo.ac.jp/~makino
×	www.kanzaki.com
△	www.sal.tohoku.ac.jp/~gothit
△	meta.tutkie.tut.ac.jp/~ichikawa
×	fc2.com
△	www2.tokai.or.jp/kimijima
△	www.hyuki.com/writing
△	www.ceser.hyogo-u.ac.jp/naritas
△	www.gakushuin.ac.jp/~881791
×	www.kanzaki.com/docs
△	www.econ.tamacc.chuo-u.ac.jp
△	www.itojun.org/paper
△	syajyo.tamacc.chuo-u.ac.jp/~miyaken
×	www.hit-u.ac.jp
△	www.naruto-u.ac.jp/~rcse
△	base.econ.osaka-u.ac.jp/~nakajima
?	www.s.soka.ac.jp/~satomac
△	orion.mt.tama.hosei.ac.jp/hideaki
?	www.senshu-u.ac.jp/~thc0597
△	www.hyuki.com/wl
?	133.46.221.167/servlet

(*)表示はできなかったが、関連のサイトと考えられる。

図 5は[2]をアンカーサイトに分析した結果である。同様にこの結果から関連サイトは A,B,C の 3 領域に分類できる。A の領域には、アンカーサイトの所属する学科の公式サイトと、情報学に関する研究プロジェクトのサイトが含まれる。これらは[2]の領域と関連が深い。B の領域には、やはり関連領域である経営学や経営情報学の研究者が多く含まれている。また、図中の b は、工業部品の製造流通

企業であり、直接的な関係はないと考えられるが、この企業は経営情報学の領域では、情報仲介業（Hagel, 2001）の事例として非常に良く取り上げられる企業である。b のサイト自体は工業部品に関する情報が主体のため、CTR 近接度で遠い位置にプロットされている。C の領域には、[2]の所属する大学の総合案内などが多い。この大学は総合大学であるため、この領域の情報は[2]とは関連が薄いと考えられる。

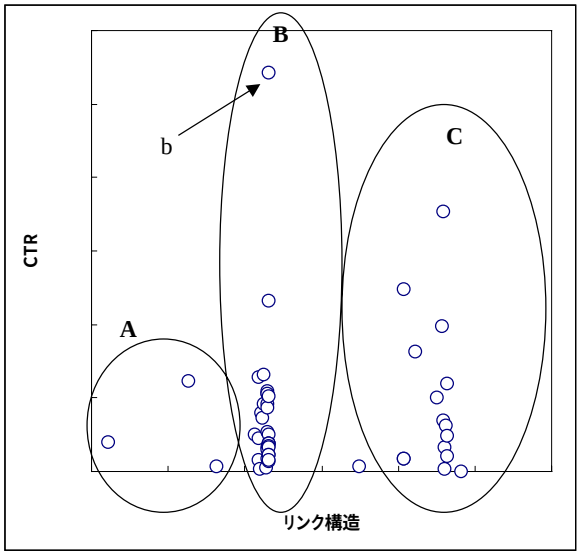

図 5：アンカーサイト[2]のリンク-CTR 平面による図示

表 2は、アンカーサイト[2]に関して、領域Bのサイト群をリンク近接度が近い順にまとめたものである。関連性の項の凡例は表 1と同様にリンク近接度が近いものほど関連のあるサイトであることがわかる。関連性のないポータルサイトや書店のサイトが多く含まれている。これは、著作などの紹介リンクで書店へのリンクが多いためと考えられる。そのため、書店関連のサイトも、経営や情報技術関係のサイトが抽出されている。しかし、領域Cにアンカーサイトの研究者が所属する大学の総合案内ページが含まれることから、本来これらとポータルの位置関係は逆転しているべきである。この原因の解明は、ディレクトリ集約のルールとともに今後の課題である。

表 2：アンカーサイト[2]の領域 B に含まれる URL

関連	URL
○	www.yokohama-mot.jp
○	www.yokotakeda.com/3dtrip
△	www.oscar.gr.jp
×	www.incs.co.jp/tsurezure
?	www.yokotakeda.com/basic
△	www.lib.ynu.ac.jp
×	www.jal.co.jp
△	www.iir.hit-u.ac.jp/reserch
×	www.diamond.bookpark.ne.jp
△	www.nc-net.or.jp
△	www.ecrp.org
△	www.bookpark.ne.jp/sosiki
△	www.jkokuryo.com
?	jmall.joshin.jp/servlet
×	www.mytrip.net
×	www.mapion.co.jp
×	www.forest.impress.co.jp
×	www.mapfan.com
×	www.amazon.co.jp
×	www.infoseek.co.jp
×	www.goo.ne.jp
△*	www.misumi.co.jp
△	www.glocom.ac.jp/odp
△	www.rieb.kobe-u.ac.jp
?	www.johogaku-a06.isics.u-tokyo.ac.jp
?	www.yokotakeda.com/incs
×	www.mbn.or.jp
△	www.commerce.or.jp
△	www.bookpark.ne.jp/hbr
△	www.e.u-tokyo.ac.jp/itme
×	www.nifty.com
?	www.yokotakeda.com/t-class

*：図 5 における点 b のサイト

続いて、図 3および図 5における領域A,B,C に含まれるサイトとアンカーサイトの関連性を図 6および図 7にまとめる。「○」「△」「×」「？」は表 1と同様である。図 6と図 7からわかるように、領域A には、関連の深いサイトが含まれてい

る。しかし、領域 A には、本人が関連するサイトが含まれているため、相対的に他の関連のあるサイトが領域 B に含まれてしまったことが考えられる。領域 B には、関連のあるサイトが多く含まれている。この領域のリンク構造が遠い位置には、無関係なサイトが多く含まれることから、領域 B の中でポータルサイトや掲示板サービスをより明確に差別化するアルゴリズムが求められる。これは今後の課題である。領域 C には、ほぼ無関係なサイトが含まれている。しかし 4 節で考察するように、CTR の軸による分類は今回の分析では有意に得られなかった。

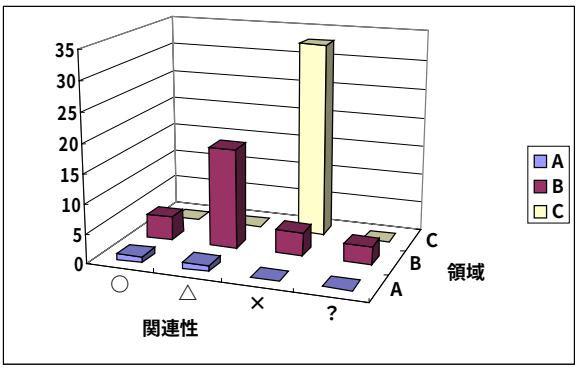

図 6：アンカーサイト[1]と他サイトの関連性

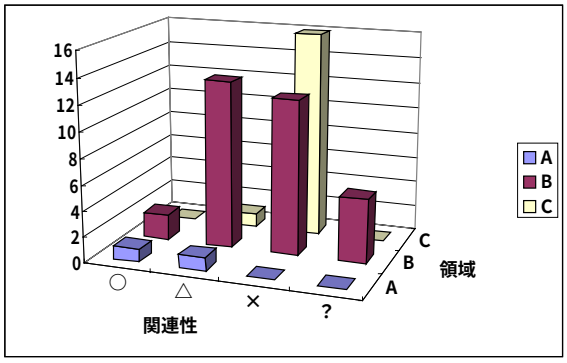

図 7：アンカーサイト[2]と他サイトの関連性

4. 考察

3 節の結果からわかるように、アンカーサイトの周辺に存在する情報を視覚的に捉えることができた。本研究で提案した手法により、情報検索の際、アンカーサイトを起点に周辺の情報を効率的に網羅的に探索することができる。特にリンク構造の軸では、アンカーサイトのごく周辺にあるサイト、関連のあるサイト、無関係なサイトという 3 段階に視覚的に分類できるため、情報検索の支援として適切であると考えられる。しかし、領域 B において顕著な結果として、関連情報のサイト、書店のサイト、掲示板サービスのサイトなどが CTR の軸で混在している。また今回、CTR が 0 に近い（非常に用語間関係が近い）サイトが多くあるが、むしろこれらは実際には関係の薄いサイトが多く、CTR が中程度のところに関係サイトがプロットされた。これは、書店サイトや掲示板サイトでは、用語が幅広く網羅的に使用されると考えられるために、結果として CTR が大きくなるためであると考えられる。また、サイトの範囲を定義するために 2 節で用いた集約ルールを用いているが、ひとつのディレクトリに多くの情報が存在する場合、やはり CTR 近接度やリンク近接度が近くなると考えられる。また、非常に用語量が少なくかつその用語がアンカーサイトで使われているサイトなどは、正規化の処理によって CTR 近接度が非常に大きくなると考えられる。例えば、トップページにほとんど具体的な情報を置かずに、典型的なホームページの用語だけが存在している場合などである。これに対する対策としては、正規化をせずに CTR の量だけによって分析を試み、結果を比較することが考えられる。また、領域 B にサイトが集中しているが、これらのサイトの中で検索の優先度を示すために、PageRank(Brin and Page,1998)などのアルゴリズムを併用し、各点のカラーリングなどで示すことでより有用になると考えられる。

本研究で提案した隣接行列を用いてネットワークを分析する手法は多くの応用が可能である。例えば、オンラインコミュニティにおける発言チャネルをリンクとして捉えることで、コミュニティの構造を理解する助けになると考えられる。

5. まとめ

本稿では、リンク構造と CTR を用いて Web 空間を視覚的に表示する手法を提案した。これにより、情報検索者が重要と考えて注目したアンカーサイトの周辺の情報を効率的に探索することが可能になる。本研究の結果、リンク構造による表示はサイトを「ごく周辺のサイト」「関連サイト」「無関連サイト」に分類可能であったが、CTR による表示は改善の余地があることがわかった。今後の課題は、CTR によるサイト分類の改良をおこなうことである。

アンカーサイト URL

[1] marketing.misc.hit-u.ac.jp/~matsui/
[2] www.yokotakeda.com/

参考文献

(Brin,1998) Brin, S., L. Page,"The anatomy of a large-scale hypertextual web search engine", Comput. Networks ISDN Systems 30(1-7)107-117. 1998.

(Hagel,2001) Hagel,J.,"Net Worth : ネットの真価―インフォミディアリが市場を制する", 東洋経済新報社,2001.

(Ishida and Ohta,2002) Ishida K. and T. Ohta, “An approach for organizing knowledge according to terminology and represen ting it visually”IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol 32, No. 4,pp.366 - 373,2002.

(Kleinberg,1998) Kleinberg,J.,"Authoritative sources in a hyperlinked environment",Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

(高橋・赤堀,1999) 高橋弘行,赤堀侃司,"検索効率を支援するサーチエンジンのインターフェースの評価",電子情報通信学会技術研究報告 Vol.98,No.643,1999.