DEWS2007 B9-4

リンク元ページのアドレス情報に基づくWebページの地域的支持度の分析

近藤　浩之†　　手塚　太郎††　　田中　克己††

† 京都大学工学部情報学科
〒 606-8501 京都市左京区吉田本町
† 京都大学大学院情報学研究科社会情報学専攻 〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{kondo,tezuka,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　Web 上には特定の国や地域の利用者を念頭において発信された情報と，世界中の利用者に向けて発信された情報が混在している．Web ページの持つこのような属性は，そのページに対してリンクを張っているページの空間的分布から抽出できると考えられるが，ハイパーリンクの構造のみを利用する PageRank やハブオーソリティモデルなどの既存手法では，Web ページに対する地域的な支持の度合いを計測することはできない．本論文では，ある Web ページに対してリンクを張っているリンク元の Web ページのアドレス情報を用いて，Web ページが地理的に近接する地域から多くリンクされているか，地理的に離れていた地域からリンクされているか等を表す地域的支持度を提案する．この手法によって，Web ページの地域的支持度を示すとともに，どの地域からリンクされているかを示すための視覚化も行う．

キーワード　地理情報，地域的支持度, Web マイニング

Analysis of Regional Support Degrees of Web Pages by In-Link Page Address Information

Hiroyuki KONDO†, Taro TEZUKA††, and Katsumi TANAKA††

† Department of Informatics and Mathematical Science, Faculty of Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-8501 Japan
† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-8501 Japan
E-mail: †{kondo,tezuka,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　On the Web, there are pages that are intended for users in a specific region or country, and pages that are intended for global use. The geographic distribution of pages that links to a page would indicate such characteristics, yet existing hyperlink-based algorithms such as PageRank algorithm and hub-authority algorithm are not capable of extracting this type of feature.In this paper, we introduce “regional support degree” which indicates form how large area web page is linked from. A web page may be linked from a small region or a wider region. We show some experiments based on this technique, and a prototype system that visualizes where web pages are linked.

Key words　Geographic information, Regional support degree, Web mining

1. はじめに

Web 上の情報は急速に増加を続けており，日常生活に関わる情報の取得において欠かすことのできない手段になりつつある．Web 上の情報は世界中からアクセス可能であるが，想定されている利用者は地域的に限定されていることも多い．その一例として，ユーザが近隣の店舗の情報を調べようとした場合，別の都市の店舗の情報が混じってしまうという問題が頻繁に発生する．また，英語圏のユーザが検索を行った場合には，検索結果の中に国外のページまでもが混在し，地域的な情報に対するノイズはさらに顕著である．

Web ページの重要度を測る指標として，PageRank 値 [1] やハブオーソリティモデル [2] などが広く知られているが，ウェブページがどの範囲の地域からリンクされているかという地域

的な支持度を考慮していないため，現状では Web ページを地域的な視点から評価することができない．そこで本研究では，Web ページが地域においてどれだけ支持されているかを示す指標である地域的支持度を定義し，それによって Web ページを評価する手法を提案する．さらに，その根拠となるリンク元の分布の視覚化を行う．

ある Web ページがどのような地域からリンクされているか調べようとして，検索エンジンを用いたとしても，検索結果として返される情報はリンク元の情報としてタイトルやスニペット，URL だけであり，それぞれのページに対してどのような国や地域から参照されているかは調べらることはできない．また，リンク元となる検索結果を個別に調べたとしても，検索集合全体のどの程度の割合で，世界中の国からリンクされているのか判断したり，どの地域からリンクされているのか判断することは容易ではない．本研究では，どのような地域からリンクされているかわかりやすくするためにリンク元を視覚化した．

本研究で求めた地域的支持度を応用することによって，検索結果のリランキングを行い，地域的に支持されているページでランキングすることができる．例えば，あらかじめ各ページの地域的支持度を求めておくことによって，ユーザの所在地や検索クエリに含まれている地名に応じて，検索結果のランキングを変更することができる．地域的支持度が高い一方で，ユーザの所在地やクエリに含まれている地名から地理的に遠い場所に位置するページは，ユーザにとっての利用価値が低いと考えられる．また，地域的支持度が低いページの利用価値は，ユーザの所在地やクエリ中の地名に依存しないと考えられる．

本論文の目的は，ある Web ページがどのような特定の国や地域からリンクされているのか，世界からリンクされているのかを視覚化し，地域的支持度として定義することによって，Web ページの地域性を測ることである．本研究の実験では，検索エンジン用いて調べたあるウェブページに対するリンク元のアドレス情報からそれぞれの Web ページの存在するサーバの国・地域を特定し，地域的支持度を求めリンク元の分布の可視化を行った．

1.1 関 連 研 究

1.2 ロボット型施設検索システム

ロボット型施設検索システムは，Google や Yahoo!などのロボットを用いてウェブをクロールし，ユーザが指定した種別の施設に関して情報を収集し，Google Maps API [3] を用いてユーザに提示するという手法 [4] である．このシステムの特徴は，データベースに登録されていない施設情報を得ることができる．また，このシステムを用いて「水田　コシヒカリ」のキーワードで検索した結果が新潟県に集中しているなど，米の品種による分布の違いを知ることができる．

このシステムと本システムの共通点は検索エンジンの Yahoo!や Google などのロボットを用いて URL を収集し，URL を Google Maps 上にマッピングして表示するところである．相違点は，このシステムでは検索エンジンを用いて「水田　コシヒカリ」などのクエリを検索し，その検索結果に含まれるそれぞれ Web ページのコンテンツに対して自然言語処理を行い，その中から地名情報を抽出し，座標変換して施設の位置を取得し，可視化しているのに対し，本論文では検索エンジンを用いてある Web ページに対するリンク元情報を調べ，それぞれのリンク元のアドレス情報を IP アドレスを利用して座標に変換することによって，リンク元の分布を可視化し，さらに地域的支持度について求めている点が異なる．

1.3 ウェブページからの地理情報の抽出

ウェブページから地理情報を抽出する手法として，IP アドレスによるマッピングする手法がある．Buyukkokten らは，特定の米国の Web サイトにリンクしているページを IP アドレスを元にマッピングし，地図上に表示するという手法を提案している [5]．このシステムを用いると，米国国内で，特定のサイトにリンクしている地域がどの地域なのかを知ることができる．このシステムを用いると，サンフランシスコの地方紙 “SFGate [6]”とニューヨークの新聞 “New York Times [7]” の Web サイトでは，米国国内からのリンクの分布が異なることがわかる．

このシステムと本システムの共通点は IP アドレスによって，ページを位置情報に変換して，リンクを可視化している点である．相違点は，このシステムでは米国の国内を対象としてリンク元の情報を可視化しているだけなのに対し，本論文では地域的支持度を求めるための尺度を示し，同時に世界各国でのリンク元を分析している点で異なる．

1.4 そのほかの地域性に関する関連研究

ページの地域性を評価するものとしてローカル度 [8] がある．これはページ中に含まれる地理用語から位置情報を取得し，それに対する MBR(Minimum Bounding Rectangle) を作成し，さらにその MBR を使って，地理用語の密度について考慮している．また，複数のクローラによって，それぞれに特定の地域の Web サイトを取るための手法として，地域性を判断させる基準を比較する研究も行われている [9]．米国におけるブログ上での地域的な話題のマイニングの手法の研究や URL を空間にマッピングし，空間リンクを考慮した拡張リンク上での HITS の研究もなされている [10] [11]．井上らは独自のウェブクローラを実装して京都市に関連するキーワードを含むページを 12 万集め，そのリンク構造に対して Google の PageRank の計算を行い,「地域的なページ集合の中での PageRank」を求めるという研究を行った [12]．しかし，井上らの研究では,「京都市に関連するキーワードを含む」という条件だけが空間的な条件となっており，実際にどの地点と関連しているかという情報は使われていないため，リンク元のページがどのように分布しているかといった指標を見ることはできない点が本研究とは異なる．

2. 地域的支持度

Web ページは様々なページからリンクされているが，リンクされているということはリンク元の Web ページがリンク先の Web ページを支持していると解釈できる．このようなリンクの支持度を考慮した評価法にページランクや HITS などのアルゴリズムがある．しかし，この支持度の計算には地域性が

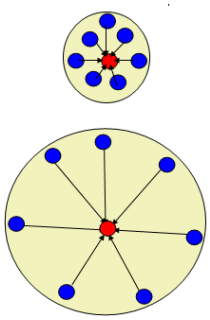

図 1　リンクの地域的支持度のイメージ
・狭い範囲からのリンクは地域的支持度が高い
・広い範囲からのリンクは地域的支持度が低い

考慮されていない．例えば，“google.com”のように海外から多く支持されているようなサイトや “e-kyoto” [17] といった京都のポータルサイトのように国内から多く支持されているようなページもある．このようなページの支持度の地域的な差が考慮されていない．

そこで，近くのサイトからリンクされているか遠くのサイトからリンクされているかを計算するために，地域的支持度を提案する．図 1 は地域的支持度のイメージ図である．

図 1 の上の円のように近くから多くリンクされていれば地域的支持度が高く，下の円のように離れたところから多くリンクされていれば，地域的支持度が低い．つまり地域的支持度とはどの程度近くのサイトからリンクされているかということを測るための指標となるための値である．地域的支持度が高いページは限られた地域からリンクされている Web ページであり，地域的支持度の低いページは様々な国や地域などからリンクされているページである．地域的支持度を求めるために次の 4 つの定義を述べる．

（1） 距離の相加平均による地域的支持度
（2） 距離の対数の平均による地域的支持度
（3） (1) に対する分散を用いた地域的支持度
（4） (2) に対する分散を用いた地域的支持度

次に 4 つの地域的支持度 (Regional Support Degree) を示す．4 つの手法における Web ページ p に対する地域的支持度を $RSD_1(p)$，$RSD_2(p)$，$RSD_3(p)$，$RSD_4(p)$ とする．

2.1　距離を用いた相加平均による地域的支持度

あるウェブページ p 対して，n 件のページがリンクされているとしてそれぞれのページを $p_1, p_2, p_3, \cdots, p_n$ とする．また，ウェブページ p の世界測地系における経度を λ_p，緯度を ϕ_p とする．二点間の緯度・経度から距離を計算するために次のような計算を行う．ウェブページ p_i に対する座標 $(\lambda_{p_i}, \phi_{p_i})$ は地理経緯度であるので，これを地心経緯度に変換する．ウェブページ p_i に対する地心経緯度を $(\lambda'_{p_i}, \phi'_{p_i})$ とすると，

$$\lambda'_{p_i} = \lambda_{p_i}$$

$$\phi'_{p_i} = \phi_{p_i} - 11.55' * sin(2 * \phi_{p_i})$$

となる．また，この変換した経緯度を用いて，二つのウェブページ $p_i(\lambda'_{p_i}\phi'_{p_i})$，$p_j(\lambda'_{p_j}\phi'_{p_j})$ 間の距離 $d(p_i, p_j)$ を求める．距離を求めるために，二点間における極からみたときの球の角度 θ をまず求める．球面三角法の余弦定理より

$$cos\theta = cos(\phi'_{p_i}) * cos(\phi'_{p_j}) * cos(\lambda'_{p_i} - \lambda'_{p_j}) + sin(\phi'_{p_i}) * sin(\phi'_{p_j})$$

$cos\theta$ より θ を求める．二点間の距離 $d(p_i, p_j)$ は極からみたときの角度に対して地球の半径である 6369km をかけた

$$d(p_i, p_j) = \frac{6369 * \theta}{10000}$$

で求められる．ただし，θ の単位はラジアン，距離の単位は万 km である．10,000 で割り，単位を万 km にしている．これは，地球の赤道の長さが約 4 万 km であり，距離が最大でも約 2 万 km となるため，単位を万 km にすることで，計算機上でオーバーフローを減らし，地域的支持度の値をある程度の大きさまでで抑えるためである．地域的支持度 $RSD_1(p)$ は

$$RSD_1(p) = \frac{n}{\sum_{i=1}^{n} d(p, p_i)} \tag{1}$$

この地域的支持度は物理的な距離の相加平均の逆数を意味する．つまり，リンク元 p_1，p_2，…，p_i と p との物理的距離が短いほど地域的支持度は大きくなる．

2.2　距離の対数による地域的支持度

この地域的支持度は物理的距離に対して，対数をとることによって，近くからのリンクの支持度をより大きく，遠くからのリンクの支持度をより小さく見せようとする手法である．地域的支持度 $RSD_2(p)$ は，

$$RSD_2(p) = \frac{n}{\sum_{i=1}^{n} ln(d(p, p_i) + 1)} \tag{2}$$

ここで，log の中で 1 を加えているのは，

$$ln(d(p, p_i) + 1) \geqq 0$$

として数値を正とするためである．

2.3　(1) に対する分散を用いた地域的支持度

この地域的支持度は物理的距離 $d(p, p_i)$ の平均に対する分散の逆数を表している．(1) の地域的支持度との違いは，物理的距離の平均の円上に近い地点からリンクされている Web ページは地域的支持度が高く，様々な距離の Web ページからリンクされている Web ページは低くなることである．リンク元からの距離の平均を $\overline{d_1}(p)$ を

$$\overline{d_1}(p) = 1/RSD_1(p)$$

とすると，地域的支持度 $RSD_3(p)$ は，

$$RSD_3(p) = \frac{n}{\sum_{i=1}^{n} \{d(p, p_i) - \overline{d_1}(p)\}^2} \tag{3}$$

2.4 (2) に対する分散を用いた地域的支持度

この地域的支持度は $ln(d(p,p_i)+1)$ の平均に対する分散の逆数を表している．(2) の地域的支持度と比べ，一定した距離からリンクされている Web ページは地域的支持度が高く，様々な距離からリンクされている Web ページは低くなる．リンク元からの距離の対数の平均を $\overline{d_2}(p)$ とする．

$$\overline{d_2}(p) = 1/RSD_2(p)$$

地域的支持度 $RSD_4(p)$ は，

$$RSD_4(p) = \frac{n}{\sum_{i=1}^{n}\{ln(d(p,p_i)+1) - \overline{d_2}(p)\}^2} \quad (4)$$

3. Web ページのマッピング

地域的支持度を求めるために Web ページを位置情報へと変換する必要がある．本論文では，Web ページの位置情報を求めるために IP アドレスを用いて位置情報を得た．Web ページを IP アドレスに変換する手法については次の通りである．

（1） Web ページの URL から独自の手法で IP アドレスを取得する．

（2） IP アドレスからデータベース [13] を利用して国名・位置情報を得る．

この二つの手法により Web ページから国名・位置情報を得た．

3.1 URL を IP アドレスに変換

URL からホスト名を調べ，IP アドレスに変換するために DNS を用いた．しかしこの手法だけではミラーサイトに関する問題がある．例えば，海外のサイト (“www.premier-ministre.gouv.fr”) の IP アドレスを調べるために，DNS を用いたとすると，ミラーサイトである (“a331.g.akamai.net”) のサーバの IP アドレスを返してしまう．このように地域的支持度を調べたい IP アドレスとは異なるサイトの IP アドレスが返されるため，データベースを利用してマッピングしたとしても，フランスの地域にマッピングされずに，ミラーサイトの位置のある日本にマッピングされることになる．このようなことを回避するために，ミラーサイトの影響を排除するために次のような判定を行った．

（1） URL と DNS 上のホスト名を比較し，文字列に関係性のない場合はミラーサイトである可能性が高いと判断する

（2） ミラーサイトと判断した場合，ルートサーバの情報を使用して，第一ドメインの IP アドレスを取得し，取得した IP アドレスを URL の IP アドレスとした．

一つ目の判断の例として，Yahoo!(“www.yahoo.co.jp”) と先ほどの海外サイト (“www.premier-ministre.gouv.fr”) の例を挙げる．DNS を用いたときに，URL の情報から IP アドレスと該当する DNS の名前を得ることができる．ミラーサイトではない場合，DNS の名前と URL の名前は一致または類似することが多い．URL が “www.yahoo.co.jp”である場合，DNS 名は “www.yahoo.co.jp”であり，一致しているので，ミラーサイトではない可能性が高い．しかし，“www.premier-ministre.gouv.fr”の場合，ホスト名は “a331.g.akamai.net”となってしまい，全く一致していない．このようなことから，URL とホスト名に”yahoo”などの同一のものが含まれている場合は DNS で得た IP アドレスをマッピングした．含まれていない場合は次の手法を用いて IP アドレスにマッピングした．

二つ目の手法として，ミラーサイトである場合は DNS の IP アドレスの情報から位置情報を得ることができない．したがって，さきほどの海外サイトの例では，ルートサーバに問い合わせることによって，”fr”が URL がどの国の IP アドレスか調べることができる．これを利用して IP アドレスを該当する国にマッピングした．

3.1.1 IP アドレスを座標に変換

IP アドレスを座標に変換するために，MaxMind 社 [13] の GeoLite City というデータベースを利用した．このサービスで変換された座標は IP アドレスに対応するサーバのある国や都市の座標である．このデータベースは IP アドレスを指定すると，そのアドレスの位置する国・地域・緯度・経度を返す．国単位での正確さは 97%，都市単位での正確さは 60%である．すべての IP アドレスに対して，都市の座標が登録されてないので，都市の座標が存在する場合は都市の座標を，存在しない場合は国の座標を使用した．

4. リンク元ページの偏り

前々章で述べた地域的支持度では，どのような国からリンクされているかといった国別の情報が含まれていない．Web ページがある特定の国から多くリンクされているのか，国内からだけリンクされているのかを調べるために，リンク元がどのような国からリンクされているかを可視化する必要がある．しかし，前章で示したような IP アドレスを用いた手法では IP アドレスの分布が国によって偏っていることが考えられる．IP アドレスの国別の偏り具合を考慮するため，次のような予備実験を行った．

4.1 予備実験-IP アドレスの分布-

（1） IP アドレスをランダムに 100 万件生成する

（2） 生成した IP アドレスから MaxMind 社 [13] のデータベースを利用して国名を得る

（3） 国名ごとに IP アドレスを集計し，国別の分布を調べる

このような実験を行い，IP アドレスの分布を調べた．ランダムに生成した IP アドレスは 100 万であり，データベース上に登録されていた IP アドレスは 581724 件であった．表 1 は IP アドレスのうちの国名の上位 10 件である．

表 1 からわかるように，IP アドレスのうちアメリカが約 58.17%，日本が約 6.06%，イギリスが約 3.98%，中国が約 3.87%を占めている．この結果から，ほとんどの IP アドレスがアメリカに割り振られていて，次に，日本，イギリス，中国が多いことがわかる．検索結果の多くがアメリカに分布する可能性があるため，この偏りを考慮して，IP アドレスの割り当ての少ない地域ほど，その地域からリンクされていることを強調しなければならない．

4.2 リンク元の国別での集計方法

どの国からどの程度の割合でリンクされているか調べるた

表 1　IP アドレスにおける国名の上位 10 件

国名	件数	割合 (%)
アメリカ	338374	58.17
日本	35259	6.06
イギリス	23143	3.98
中国	22505	3.87
ドイツ	20468	3.52
カナダ	16618	2.86
韓国	12089	2.08
フランス	11820	2.03
オランダ	9103	1.56
オーストラリア	7207	1.24

めに，次のような二種類の方法で国別リンクの割合を計算した．

- リンク元を国別に集計する
- リンク元をサーバの偏りを考慮して集計する

あるページ p に対するリンク元のページを国別に集計したものを $count_p(country)$ とする．また，予備実験の結果で得たそれぞれの国に対する IP アドレスの割合を $base(country)$ とする．

一つ目の手法では，それぞれの国の割合を調べるために，$count_p(country)$ を利用するが，二つ目の手法では，

$$count_p(country)/base(country)$$

を用いて，サーバの偏りを考慮した．

例えば，あるサイトにリンクしている国別の数がアメリカ 80 件，日本 15 件，韓国 5 件であったとする．このとき，一つ目の手法で示した割合はアメリカ 80%，日本 15%，韓国 5%と件数の割合のままである．二つ目の手法では表 1 で示したように，国別の IP アドレスの割合を考慮して，アメリカ 1.38，日本 3.30，韓国 2.40 となり，日本が高い割合となる．

5.　ページへのリンクの総数

リンク元から地域的支持度を求める前に，リンク元の数がどの程度であるのかを調べた．まず，Yahoo!API を利用し，”link:”構文を用いて，”google.com”へリンクしている上位 1000 件を取得した．このうち，インデックスページに対して，それぞれリンク元の数を調べた．有効なリンク数は 845 件であった．このうち，はずれ値として，リンク数で昇順に並べたときの 99%のデータである 837 件について分析した．またリンク数の最大値，最小値，平均値，メディアン，分散，標準偏差は次の通りである．

- 最大値　721
- 最小値　0
- 平均 46.9
- 分散 4112
- 標準偏差 64.1
- 中央値 40

図 2 はリンク元の数のヒストグラムである．リンク元の総数の平均値は約 47 件であり，標準偏差が約 64 であることと，ヒストグラムからリンク数として 100 件とれば妥当であると考えられる．

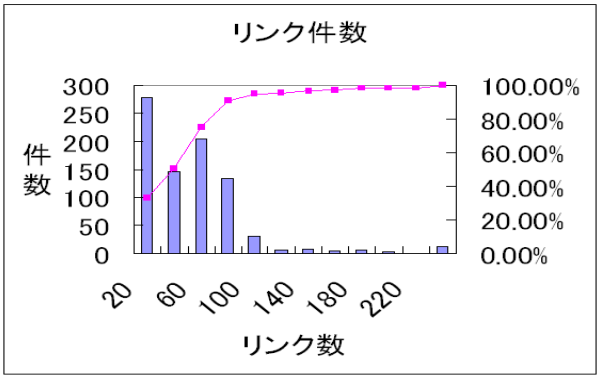

図 2　リンク元の割合

6.　実　　験

6.1　実験の概要

４つの地域的支持度がどのように評価されているかを調べ，国別の偏りを調べるために，次のようなシステムを作り，２つの実験を行った．

（1）　Yahoo!　API を利用して特定の Web ページにリンクするリンク元の上位 100 件を取得する

（2）　取得した 100 件の Web ページの URL を独自の手法で IP アドレスに変換する

（3）　それぞれリンク元の IP アドレスをデータベースを用いて座標に変換する

（4）　４つの地域的支持度を計算する

（5）　GoogleMaps 上にマッピングする

（6）　それぞれどの国からリンクされているか示す

一つ目の実験はこのシステムを用いて，様々な URL に対して４つの地域的支持度を求め，それぞれの URL ごとにリンク元の URL を GoogleMaps 上に表示し視覚化し，リンク元の国別の偏りを円グラフで示した．この実験でそれぞれの URL の事例に対する細かい分析をした．　二つ目の実験は同様の手法を用いて，数百件の URL に対して，地域的支持度を求め，それぞれの地域的支持度の数値の分布やリンク数との相関を求めた．本システムの細かい流れは次の通りである．

6.2　システムの流れ

6.2.1　リンク元ページの取得

実験対象となる URL からリンク元ページを取得するために，Yahoo!　API [18] を用いて，“link”構文を使いリンク元を求めた．予備実験よりリンク元の件数は 100 件以下であることが多いので，本実験では検索結果のうち 100 件の URL をリンク元ページとした．ただし，Yahoo！API では言語によるフィルタがかかり，特定の言語からのリンクしか取得することができない．そのため，ウェブページのうちで多くを占める英語のサイトから取得することとした．

6.2.2　URL を IP アドレスに変換

第４章で用いた変換を利用してリンク元ページを変換した．

表 2　実験サンプル

ページ	URL
Google	google.com/
FIFA	fifaworldcup.yahoo.com/
イギリス政府	www.direct.gov.uk/Homepage/fs/en
オーストラリア政府	www.australia.gov.au/
フロリダ地方紙	www.alachuatoday.com/
南アフリカ政府	www.gov.za/
フロリダマーリンズ	florida.marlins.mlb.com/NASApp/mlb/index.jsp?c_id=fla
バレーボール世界選手権	www.2006vball.jp/
京都ポータルサイト	www.e-kyoto.net/

6.2.3　IP アドレスを座標に変換

第 4 章で用いた変換を利用してリンク元ページを変換した．

6.2.4　地域的支持度の測定

第 3 章の 4 つの地域的支持度を用いて求めた．URL を座標変換する過程で，同一座標に複数の URL が割りあてらることがある．そのことを考慮し，URL が重複した地点に重みをつけ，定義と同じページ単位での地域的支持度を求めるようにした．

6.2.5　リンク元の可視化

URL の座標をすべて GoogleMaps 上にマッピングした．地域的支持度を求めたページはリンク元から線で結ばれている．ただし，同じ地点にマッピングされた URL 数をその地点でのマーカーの数とし，線の太さをマーカの地点の重複 URL 数とした．

6.2.6　国別円グラフ

リンク元を国別に示した円グラフを 2 つ表示する．1 つ目の円グラフはどの地域からのリンクが多いのかを示している．2 つ目の円グラフは国別の IP アドレスの分布を考慮した円グラフである．上位 5 件だけを示している．

6.3　実験結果 1

表 2 のサイトにおいてそれぞれ地域的支持度を求め，国別の偏りを示した．南アフリカをサイトで，システムの概要を示す．南アフリカ政府のサイトにリンクされている URL を調べるために

“http://www.gov.za/” を実験対象の URL とした．図 3，図 4，図 5 は実験結果である．

次に 4 つの地域的支持度を示す．

$RSD_1(p) = 1.210$　$RSD_2(p) = 2.427$

$RSD_3(p) = 1.860$　$RSD_4(p) = 6.383$

図 3 から，地域的支持度が低いことがわかる．これは，米国やイギリスからリンクされているということを示しているのではなく，南アフリカに自国のサーバが少ないので，リンクしているページの著者が南アフリカであっても，米国やイギリスのサーバが多くなり，地域的支持度が低下してしまうものであると考えられる．図 4 はリンク数の割合を示している．上位三件はアメリカ，南アフリカ，イギリスである．図 5 は IP アドレスの偏りを考慮した割合である．上位三件は南アフリカ，アラブ，イギリスである．この結果から，南アフリカからの支持が高いことがわかる．

図 3　“南アフリカ政府” における実験結果

図 4　“南アフリカ政府” における国別リンクの割合 (手法 1)

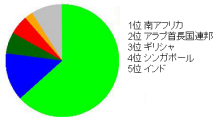

図 5　“南アフリカ政府” における国別リンクの割合 (手法 2)

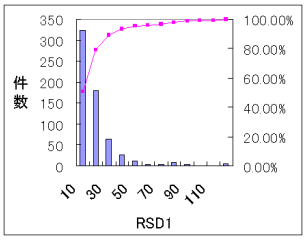

図 6　RSD_1 の分布

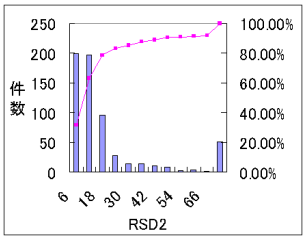

図 7　RSD_2 の分布

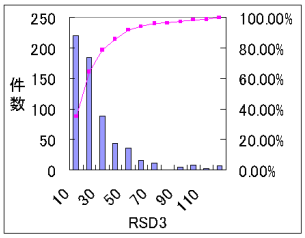

図 8　RSD_3 の分布

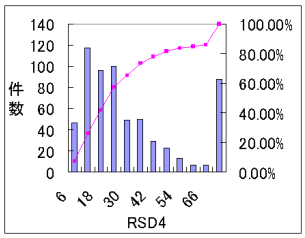

図 9　RSD_4 の分布

6.4　実　験　2

次に 4 つの地域的支持度の数値の分布とリンク元の件数と地域的支持度との関係を調べるための実験を行った．実験方法は次の通りである．

（1）　“google.com” にリンクしている 1,000 件をサンプルとする

（2）　実験 1 でのシステムを利用し，4 つの地域的支持度とリンク元の件数を求める

ここでのリンク元の件数とは地域的支持度に利用したリンク元 100 件ではなく，Yahoo! API に登録されているリンク元の総数である．Google.com でリンクを 1000 件取得しようとしたが，実際に URL が取得できたものは 845 件であった．これは，通信やプログラム，Yahoo! API の問題であると考えられる．845 件のうち，地域的支持度の計算ができた有効な URL は 636 件であった．計算できなかったものについては以下の原

表 3　地域的支持度の数値の範囲

手法	最大値	最小値	平均値	中央値	分散	標準偏差
RSD1	61.64	0.71	12.32	9.787	114.4	10.7
RSD2	164.65	0.77	15.79	8.61	602.1	24.54
RSD3	61.77	1.14	17.39	14.22	199.68	14.13
RSD4	225.08	2.22	30.86	20.59	1140.6	33.92

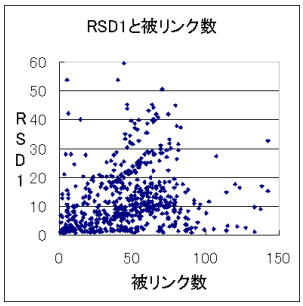

図 10　RSD_1 とリンク数

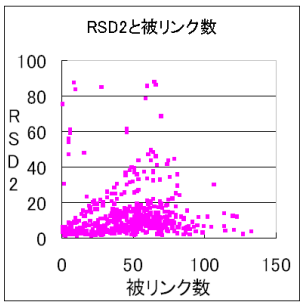

図 11　RSD_2 とリンク数

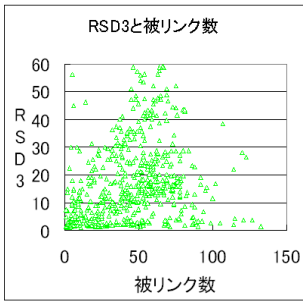

図 12　RSD_3 とリンク数

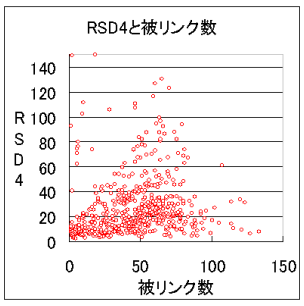

図 13　RSD_4 とリンク数

因が挙げられる．

- 被リンク数が 0 である URL
- 被リンクの位置がすべて，求めるべき URL と同一地点である

前者の場合，リンク数が 0 であるため地域的支持度の定義より地域的支持度が 0 となり有効ではない．後者の場合，同一の点となるとき，距離が 0 となる．これは地域的支持度の定義の分母が 0 となり，地域的支持度が無限大となってしまい，不適切である．このように，地域的支持度のうち，リンク数が非常に少ないものやマッピングされた地点が少ないものに関しては閾値を与えて排除することが有用であると考えられる．

被リンク数の分布については予備実験で示した図 2 の通りであり，多くの被リンク数は 100 件以下であることがわかる．

次に地域的支持度の数値の分布について調べた．このとき，さきほどの地域的支持度として有効な URL637 件のうち，昇順で 95%(604 件) のものを利用した．これは数値が大きすぎる上位 5%を省くことによって，正確な値を求めるためである．図 6，図 7，図 8，図 9 は RSD_1，RSD_2，RSD_3，RSD_4 の数値の分布図である．また，それぞれの RSD の最大値，最小値，平均値，中央値，分散，標準偏差は表 3 に示した．

表 3 より RSD_1，RSD_2，RSD_3，RSD_4 の数値の特徴をつかむことができる．RSD_1 と RSD_3 に関しては，最大値，最小値，平均値は変わらないが，RSD_3 のほうが分散が大きくなっている．このことから RSD_3 の指標のほうが数値のばらつきが大きく，定義のところで仮定した分布のばらつきを大きくすることができた．RSD_2 と RSD_4 は RSD_1，RSD_3 に比べて最大値と分散が大きくなっていることがわかる．

表 4　地域的支持度と被リンク数の相関関係

手法	相関係数
RSD1	0.186
RSD2	-0.049
RSD3	0.266
RSD4	-0.011

表 5　URL に対する地域的支持度

ページ	手法 1	手法 2	手法 3	手法 4
Google	2.201	2.939	5.384	11.164
FIFA	1.512	4.645	2.060	9.217
イギリス政府	**22.782**	**44.734**	**27.676**	**87.948**
オーストラリア政府	2.456	2.792	3.690	7.540
フロリダ地方紙	4.352	**42.240**	4.981	**61.203**
南アフリカ政府	1.210	2.427	1.860	6.383
フロリダマーリンズ	**4.696**	**30.527**	**5.419**	**46.287**
バレーボール世界選手権	1.163	5.799	1.656	12.205
京都ポータルサイト	**18.499**	20.471	**26.106**	43.273

太字：数値が平均値以上

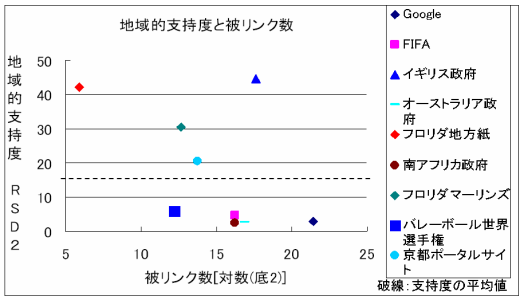

図 14　RSD_2 と被リンク数

また，リンク元と地域的支持度に関する関係を求めた．このときも先ほどと同じように被リンク数の下位 95%(604 件) と地域的支持度の下位 95%(604 件) について分析を行った．リンク数と地域的支持度の数値の両方ともが下位 95%以内に入っているリンク数のみを対象とした．リンク元と地域的支持度のうち，はずれ値を除いたデータを用いて，リンク数と RSD_1，RSD_2，RSD_3，RSD_4 との関係を図 10，図 11，図 12，図 13 の散布図に示した．それぞれの RSD とリンク数との相関を調べた結果が表 4 である．これらのことから，相関関係がほとんどないといえる．特に，RSD_2，RSD_4 に関しては無相関検定により，相関係数が 0 であるという帰無仮説は棄却されない．これらの指標は被リンク数とほぼ相関がないことから，既存の検索エンジンのランキングの指標となっている被リンク数にたいして独立な数値であると考えられる．よって，これらの数値を元に URL のリランキングをすることによって，地域的支持度を用いたランキングをすることが可能である．

6.5　実験考察

最初の実験では，9 件のサンプルの URL を用いて実験を

行った．表 5 にまとめる．実験では４種類の手法を比較した．最初の実験の９つのデータと実験２で行った RSD に関する平均値などのデータを比較することにより，地域的支持度が高いか低いかを比較することができる．これらのデータから，イギリス政府，フロリダ地方紙，フロリダマーリンズ，京都ポータルサイトの４つは地域的支持度が高いことがわかる．また，そのほかのサイトでは地域的支持度が低いことがわかる．図 14 は，これら９件の RSD_2 と被リンク数の対数との関係をグラフにしたものである．これより被リンク数とほぼ無関係に地域的支持度が評価されていることがわかる．

RSD_1, RSD_3 と RSD_2, RSD_4 の比較に関しては URL に対して距離の log をとり，地域的支持度の数値のオーバーフローを防ぐとともに，遠くからのリンクの影響をできるだけ軽減しようと試みた．表 3 から，RSD_1 と RSD_3 の両者においては数値の最大値，最小値は変わらないものの，分散が大きくなり，うまく分散していることがわかる．しかし，RSD_2 と RSD_4 の両者においては最大値と最小値の両方が分散とともに大きくなり，大きな違いはなかった．

次に，分散を取った RSD_3, RSD_4 と RSD_1, RSD_2 に関しては大きな違いが生じた．これは RSD_1, RSD_2 が距離の平均の逆数を地域的支持度としたため，近くからリンクされているページが多いほど地域的支持度が高くなったが，RSD_3, RSD_4 はそれぞれのリンク元と中心の Web ページとの距離の分布を見ていることになる．このため，リンクが平均距離の円状に均等に分布しているほど，支持度が高くなり，逆に，離れているほど低くなる．国内からが多いほど，平均距離の円は小さくなるので，分散も小さくなり，地域的支持度が高くなると考えられる．逆に，海外からのリンクが多いほど，平均距離の円が大きくなるので，分散も大きくなり，地域的支持度の値が小さくなると考えられる．

7. 結　　論

本研究では多様な手法で地域的支持度を求めたが，定義した数値のばらつきが存在する．今の数値では，それぞれの手法ごとに相対的に地域的支持度を測る方法しかない．しかし，数値の平均値，分散などを実験２によって求めているので，それぞれの値を正規化して，リランキングするときの指標として利用することができる．

実験２より４つの地域的支持度と被リンク数がそれぞれほとんど無相関であることがわかった．このことから，地域的な支持度を利用してランキングしたときに，既存のランキング手法と独立したランキングができることがわかった．

現在はリンク元を見つける手法として，Yahoo! API を使用して，言語を指定した．そのため，同じ Web ページに対して言語に依存した地域的支持度となった可能性があると考えられる．

謝辞　本研究は, 文部科学省 21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」(研究代表者:田中克己),　文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」, 異メディア・アーカイブの横断的検索・統合ソフトウェア開発 (研究代表者:田中克己),　文部科学省科学研究費補助金：特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」 計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者：田中克己,A01-00-02, 課題番号 18049041), 特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」(研究代表者：安達淳, Y00-01, 課題番号：18049073), 若手研究 (B)「ウェブ活用のための情報統合による信頼性判断支援」(研究代表者:手塚太郎, 課題番号: 18700086) によるものです.

ここに記して謝意を表すものとします.

文　　献

[1] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” WWW7 / Computer Networks, 1998
[2] J. Kleinberg, “Authoritative sources in a hyperlinked environment,” Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
[3] Google Maps API, http://www.google.com/apis/maps/
[4] 長屋務，森本泰貴，藤本典幸，出原博，萩原兼一，“Google Maps API を応用したロボット型施設検索型システムの試作,” DEWS2006 5B-i6 , 2006
[5] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar, “Exploiting Geographical Location Information of Web Pages,” WebDB, 1999
[6] SFGate, http://sfgate.com/
[7] The New York Times, http://www.nytimes.com/
[8] 松本知弥子, 馬強, 田中克己, “Web ページの地理情報と話題の日常性を考慮したローカル度検出とフィルタリング機構,” DBWeb2001, 2001
[9] Weizheng Gao, Hyun Chul Lee and Yingbo Miao, “Geographically Focused Collaborative Crawling,” Proceeding of the 15th International World Wide Web Conference, Edinburgh, Scotland, 2006, 2006
[10] Qiaozhu Mei, Chao Liu, Hang Su and ChengXiang Zhai, “A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs,” Proceeding of the 15th International World Wide Web Conference, Edinburgh, Scotland, 2006, 2006
[11] 張建偉, 石川佳治, 北川博之, “空間情報ハブ抽出のためのウェブリンク解析手法の開発,” DBSJ Letters, 2004
[12] 井上陽介, 李龍, 高倉弘喜, 上林弥彦, “地域情報検索のためのリンク構造分析によるウェブページと地域の関係抽出,” 電子情報通信学会データ工学ワークショップ, 2002
[13] MaxMind, http://www.maxmind.com/
[14] IP2Location, http://www.ip2location.com/
[15] hostip.info, http://www.hostip.info/
[16] cyberareasearch,http://www.arearesearch.co.jp/
[17] e-kyoto, http://www.e-kyoto.net/
[18] Yahoo! API, http://developer.yahoo.co.jp/