データベースシステム 128－52
（２００２．７．１９）

ウェブデータを利用した概念ネットワーク構築

李 龍　　上林弥彦
京都大学情報学研究科社会情報学専攻
〒606-8501 京都市左京区吉田本町
{ryong, yahiko}@db.soc.i.kyoto-u.ac.jp

あらまし ウェブ情報間のリンクによる繋がりはウェブページの作成者の主観的な観点に依存しており、ウェブ全体の構造が非常に複雑になっている。そのため、ユーザがリンクを辿りながら連続的に情報を求めるようなウェブナビゲーションを行うときは、各ページの内容の把握、ユーザ目的に応じた知識抽出、適切なリンク（経路）の選択という作業を繰り返さないといけないため大きな負担がかかるという問題がある。本稿では、このようなウェブ情報ナビゲーションの問題点を解決するためにウェブ情報中に含まれている知識のマイニングによる概念ネットワーク構築手法と、その利用としてウェブサイトのイメージ要約、ウェブサイトの概念度評価の手法を提案する。

Web-based Conceptual Network: Construction and Utilization

Ryong Lee　　Yahiko Kambayashi
Department of Social Informatics, Kyoto University
Yoshidahonmachi, Sakyo, Kyoto 606-8501, JAPAN
{ryong, yahiko}@db.soc.i.kyoto-u.ac.jp

Abstract. The web has very complex structures, since web page makers have created a lot of links depending on their subjective opinion. It makes web searchers suffer from the requirements to repeat the following steps when they need to perform continuous search; (1) knowing the web pages by reading the full contents, (2) extracting interesting knowledge from them, and (3) selecting links to get more related information. In this paper, in order to handle these problems, we propose a method to construct conceptual network by mining hidden knowledge in the web, and its utilization functions for i) summarizing web sites ' image and ii) evaluating web sites ' conceptual levels.

1. はじめに

ウェブ上の膨大なデータがお互いにリンクで繋がっているが、その繋がりは情報間の意味的な関連よりも人それぞれの主観的な観点からの異質的なデータ間の関連を表している。Google[4]のようなリンク間の参照関係の強さでウェブページのランキングを計るようなウェブ情報検索システムは多くの人々から人気がある個々の情報を導くことには優れているが、情報間の意味的な関連を連続的に辿って総合的な情報検索を行う必要がある場合の検索コストは変わらないと問題がある。そのため、Yahoo![10] や Open Directory Project[3]のようなウェブディレクトリは 概念の階層を辿りながら適切なページを見つける機会を提供するがウェブ上で主に使われているキーワードとそのキーワード間の関係は言葉間の階層や類似語などのシソーラスに依存した固定的なもので、情報検索者の興味のある概念を動的に提供することは困難である。

本研究の目的は、大量のウェブ情報の中から有用な知識を意味論的なルールとして抽出して情報検索やウェブサイトのページ間の意味構造の評価に利用することである。一般に、ウェブの中に含まれている大量の情報は、時々刻々と変化してゆくために最新のものや、利用者の感覚といった他の情報源では得られないものを含んでいる。この大量の情報の中から有用な知識を抽出することは非常に重要であるが、元の情報量が多く、信頼性に欠けるデータもあるため一般に

は困難な問題である。そのため、ページごとの特徴的な言葉集合を計算して共起関係から意味論的なルールを作る方法が有用である。例えば、京都の「四条」に関する多くのページからは「フランス料理」、「和食」、「韓国料理」という言葉よく現れて、「四条には料理屋が多い」といった結果が得られることになり、情報検索に利用できるだろう。そして、このようなウェブマイニングの結果は、そのまま人々が考えている地域に関する知識構造として捕らえることが可能である。そのような個々の意味論的ルールを結合させると、一つの大きなネットワークとなる。そのネットワークには、ノードとして概念的な言葉と、ノード間にはその概念間の関係を表すと考えられる（以降、概念ネットワークと呼ぶ）。このようなウェブから構成される概念ネットワークは、次のような観点からその必要性が強く求められる。

あるウェブページ集合の全体の要約：

あるウェブサイトを辿る前に、そのウェブサイトの中にはどういう内容が含まれているかの要約は連続的なウェブ情報検索に非常に役に立つものである。しかし、ウェブサイト管理者にとってはサイト中のユーザの数や全般的なウェブの規模が大きくなるにつれ、そのような要約を手で作ることはコストがとても高い作業となる。また、その要約も一人の個人が作成すると客観性やその記述のレベルを維持することは難しくなる。したがって、サイト中の要約を訪問したユーザの目的に応じて動的に生成する必要がある。特に、サイト全体に頻繁に現れる概念ネットワークの提供は、サイト訪問者にとってサイト全体内容のイメージを掴むのに非常に有効な方法である。

ウェブサイトの意味的なリンク構造管理の評価方法：

一つのウェブサイト中の情報が連続的な情報検索にどの程度向いているといった評価は、そのサイトの個々のページが含めんでいる知識とその知識間の関係に基づいてウェブページがよくリンクされているかによって決まられる。お互いに関係ある二つのページはリンクされるべきであるが、あまりにもリンクが多すぎるとユーザにとって情報選択の大きな負担になってしまうので、適切なリンク構造の自動的計算や評価手法が要求される。

以降は、上記の二つの課題を実現するために、2章でウェブページ分析によってページ中に含まれている知識構造を代表的な概念集合とその間の関係とする概念ネットワークを構成する手法を提案する。3章ではウェブサイト管理の観点から上記の二つの課題を概念ネットワークを利用してどう解決するかについて述べる。

2 概念ネットワークの構築

ウェブページから知識構造を作るために従来のデータマイニングの手法を適用している。本稿で目的としている概念ネットワークは、図 1のように現在のウェブページ内容からよく現れる連想ルールを導いてそれによる結合構造をしている。特に、その連想ルールは、A→B の型で A が現れるページから B が現れる比率を計算する。ここでは、A,B が一つの語について考える。そして、ターゲットするデータセットからA とB が一緒に現れる比率をこのルールの Support、そして A が

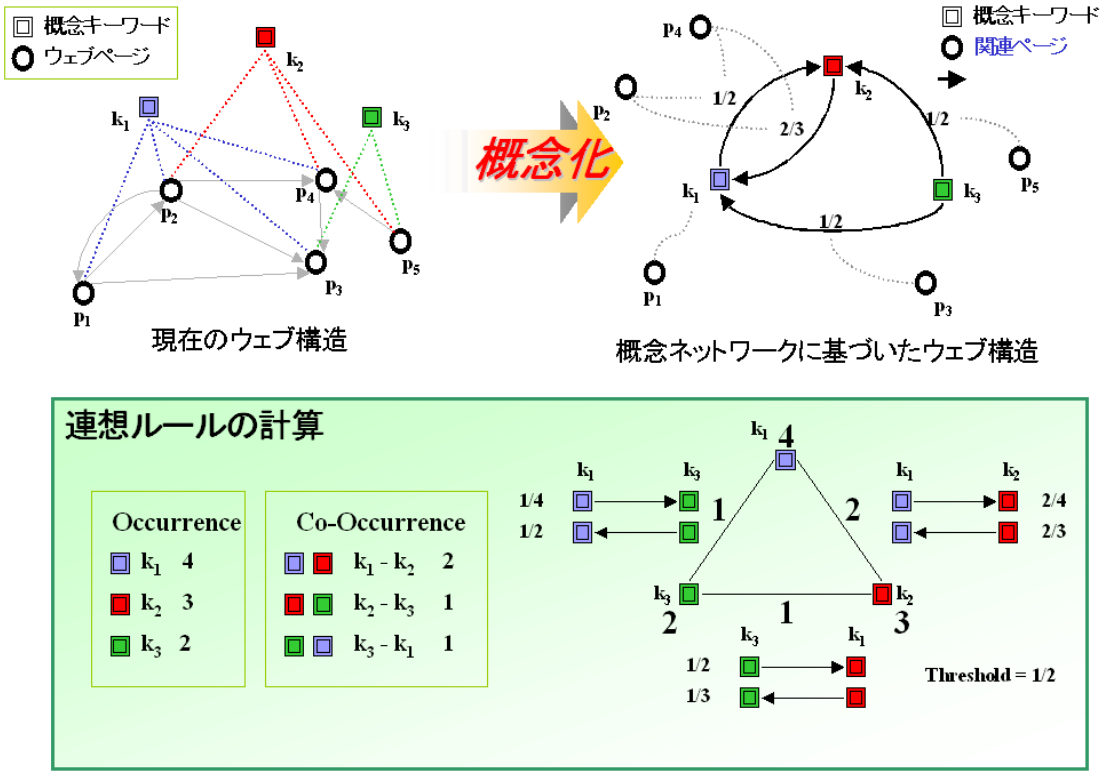

図１．ウェブデータを利用した概念ネットワークの構築

現れた時に B が現れる比率をこのルールの Confidence と定義する。その概念構築はいくつかのステップに分けて行っている。

① 概念抽出対象ページの収集 ： ウェブは一般的に非常に異質な内容を含めているので、連想ルール抽出での Support をあげるためには、類似性が高いページ集合を絞る必要がある。

② 概念キーワードの抽出 ： 上記のデータセットのようにウェブページからの一つページ中で現れる単語の数が予測できないほど大きくなる場合があるなど、ページごとの中心的で代表的なキーワードを絞って計算しないと計算時間と効率の問題が起こる。そのため、一つのウェブページに対応する概念キーワードを、TF/IDF 手法による特徴ベクトルを計算して各ページでの tf・idf の値が高い10個まで単語を選択するようにした。すなわち、以降の連想ルール計算においては、この各ページ当り10個までのキーワードのみを扱っている（PageID = {K1, K2, …,K10}）。

③ 連想ルールの計算 ： ターゲットするウェブページ全体のセットからよく現れるルールを導くために、Apriori アルゴリズムを適用する。それによって、各ルールでの Confidence は、次のように計算される。

$$\text{Confidence}(A \rightarrow B) = P(B|A) = \text{support\_count}(A \cup B)/\text{support\_count}(A)$$

なお、Apriori 計算の結果は、{ '京都 ', '観光 '}→{ '銀閣寺 '}のような、ルールが三つ以上の要素から構成されることまで計算される。これらの結果も概念ネットワーク構造の一部として使えるが本稿では簡単に二つの要素からもの構成されるルールについて説明する。

④ 要素ルールから概念ネットワークの構成 ： 上記までのステップで、{ '京都 '}→{ '観光 '}、{ '観光 '}→{ 'お寺 '}のような要素ルールが得られる。単純にこの二つのルールを結合すれば、{'京都 '}→{ '観光 '}→{ 'お寺 '}の三つのノードからなる簡単なネットワークが作られる。しかし、これらの要素がそれぞれ異なる Support と Confidence を持っているので、概念ネットワークのおけるこれらの結合の制約に対する考慮が必要となる。要素ルールに対する Support が高いほど全般のデータセットに対するそのルールの出現度が高いということで、この値が高いルールから構成される概念ネットワークはデータセット全体に対するおおまかなイメージを提供する。Confidence はルール間の関連の強さを表すので概念ネットワークでの各概念における関連概念の数をどの程度にするかを決める尺度となる。従って、概念ネットワークの構成には、要素ルールからどの程度のイメージと関連を求めるかを決める必要がある。

上記のように各ページで表れる概念キーワードからそれらの関係を概念ネットワークとして定義した。概念ネットワークでの概念間の関係は、方向性を持つようになる。それにより、「京都」に関しては「観光」という強い関係が意味を持つようになるが、「観光」から「京都」への関係は状況によってその強さが変わることになる。また、連想ルールの抽出は対象となるページの集合が大きくなるほど、その Support が低くなる問題点がある。しかし、その Support という尺度を逆に利用することによって、ウェブページ集合から必要なルールの数を決めることが可能であり、ルールの数によってそのウェブ集合の要約度を決めることになる。ルール抽出における Confidence という尺度は、特定の概念から関連する概念の数を決めることが可能となる。すなわち、Confidence を減らすことによってより関連する概念を多く求めることが可能となる。

2.1 概念抽出の実験

上記の概念ネットワーク構築のステップにおいて、実験のデータセットとして朝日新聞の英語ページサイト[2]からページを収集した（2001 年 6 月 26 日から 2002 年 6 月 10 日までの 9,766 ページを対象）。そして、ステップ②による各ページの概念キーワードのインデックスを生成して、ステップ③における連想ルール抽出を可能にした（ストップワード処理はこの段階で行った）。実際の連想ルール抽出では、全データセットから任意の 100ページを選び最小 Support は、1.0%から5.0%の範囲で、最小 Confidence は 80%に固定した条件で行った。これにより、80％のルールに対する確信度で、ルールの数を調整できるようになった。当然ながらSupport が低いほど多くのルールが現れるが、結果としては次の三つにグループとして分けられた。

Group A:
約 1%以下の Support で数千個のルールが出現
Group B:
約 1.5-2.5%までは 25～80個程度のルールが出現
Group C:
その以上に対しては、3個以下のルールが出現

上記の分類で、特に Group B の場合に結果となるルールの数は大きく変わるが、ルールを結合してグラフのコンポーネントとして表すと、約 10個程度になった。それぞれのコンポーネントは、図 3のように概念ノード間の意味的な関係を表すようになっている。個々の連想ルールは知識を与えるには十分でないが、このようにルールの結合によって概念の全般的な関連がわかるようになる。しかし、そのコンポーネント中には、ノードが二つしかないようなものがあり、例えば、{ 'yasukuni '}→{ 'visit '}のような関連は新聞記事と関連した何かの知識を連想させるが情報としての価値が低くなるという問題があった。そのため、少なくとも三つ以上のノードからなるコンポーネント(Large Component)が概念ネットワークとして意味を持つようになると仮定して、図 2の#(Large Component)/#(Component)の比率を比較したら、約 60%程度のコンポーネントが概念ネットワークを構成できるという結果になった。

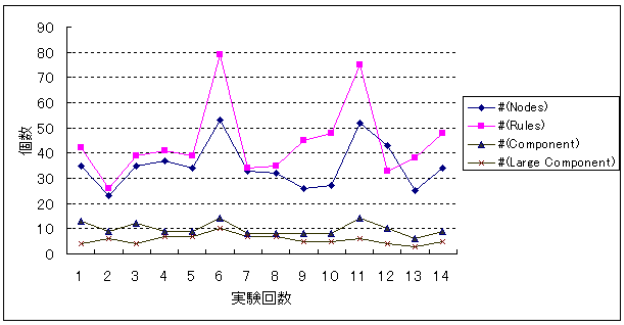

図 2. Group B (最小 Support: 1.5～2.5%)の実験結果

この実験では、類似語や上位・下位語、複数の単語からなる複合名詞などのようにシソーラスを使わずに、単語間の連想ルールのみによって概念ネットワークを構成した。そのため、図 3での左コンポーネントでのcup→worldのように、確かにworld cupを表しているにも関わらずノードとして分かられてしまう場合もあった。そのため、シソーラスの適用が望ましいが、計算上の効率面で上記の結果の場合には、ウェブページから名詞を抜き出す段階よりも最終的に図 3のコンポーネントまで抽出してから結合した方が有効である可能性がある。また、任意の 100ページに対して実験を数回繰り返して行ったが、ページ数を増やすにつれ、コンポーネントを構成できるルールの数を満たすためには、最小Supportを徐々に低くしないといけなくなり、Supportがどんどん無意味になってしまう場合がある。しかし、本実験のように、TF/IDFがその代わりとして十分に特徴的な言葉にフォーカスして、かつConfidenceで十分にルールの数をコントロールすることで適切なコンポーネット集合を導くことが可能である。

3. 概念ネットワークを利用したウェブサイト管理

3. 1 ウェブサイトが含めている知識のイメージ

一章で議論した「あるウェブページ集合の全体の要約」という課題に対して、概念ネットワークを自動で動的に生成する仕組みが要求される。あるウェブサイトを訪問したユーザがそのウェブサイトに含まれる情報を検索するために概念ネットワークを利用することを想定しよう。ユーザは最初からウェブサイトに対する詳しい概念ネットワークよりも、最初は簡単なものから次のステップでは興味のある部分に対して詳しいネットワークを知ろうとするだろう。その行動は次のようにまとめられる。

① 概念ネットワークの把握：ターゲットにしているページ集合にどういう概念ネットワークがあるか知りたい：全般の知識のイメージだけでいい（特に、情報検索の最初の段階には）

② **概念ネットワークの拡張：**知識ネットワーク中で特定の部分に関して詳しいネットワークが知りたい

③ **概念ネットワークの詳細化：**特定の概念に対して、もっと他の概念との関連を詳しく知りたい

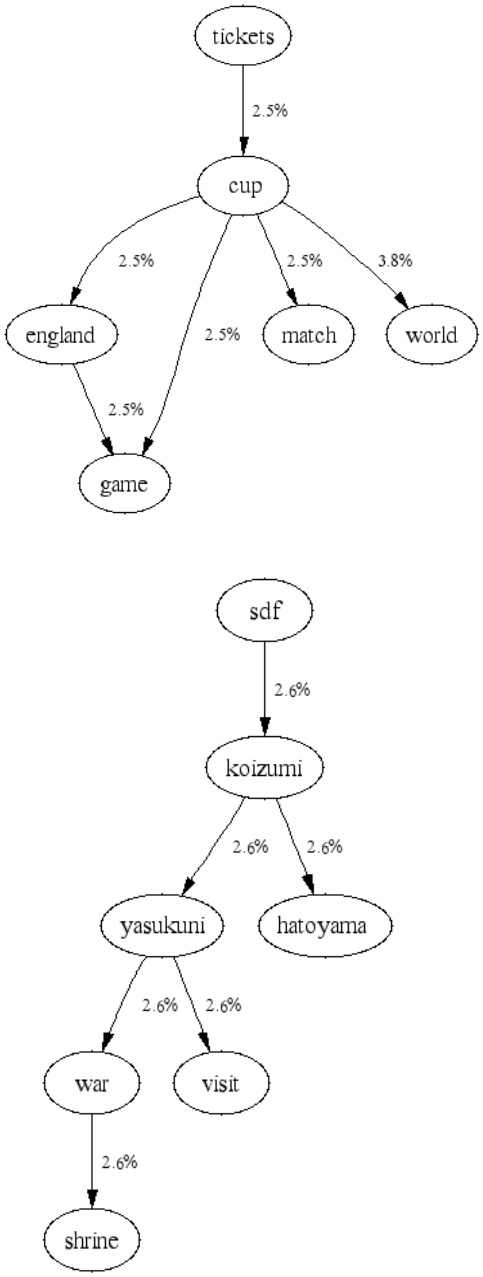

図 3. 連想ルールの結合によるコンポーネント生成の結果（数字は各ルールの Support であり、最小 Confidence は 80%である。）

上記のステップ①では、そのサイトのもっとも簡単な概念ネットワークのみを提供する。一般的なウェブディレクトリと異なって、ここではウェブサイトが我々の概念ネットワークを利用して情報検索を支援していると想定している。そのため、そのようなネットワークの周期的な生成により常にウェブサイト全体に対する新しいイメージを提供することが可能である。たとえば、新聞サイトで最近のニュース中でサッカーに関する記事が

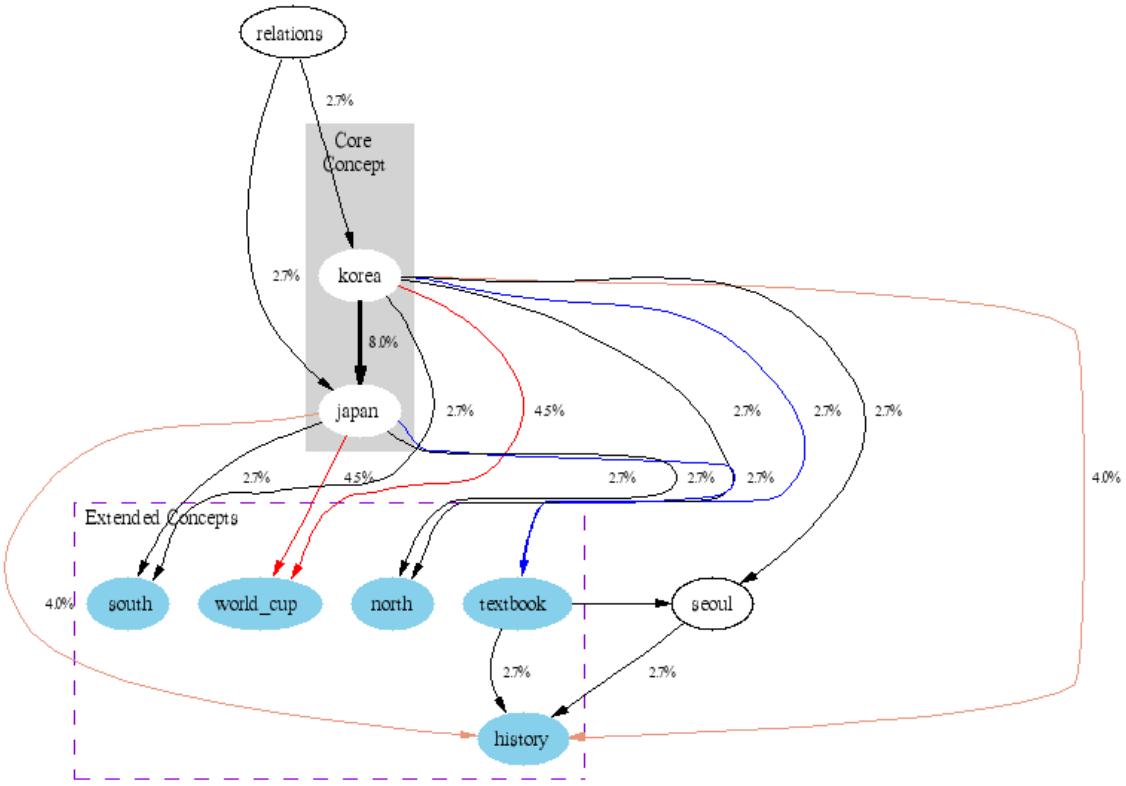

図 4. 概念ネットワークの拡張（最小 Support が 8.0%の場合を 2.7%まで減らしてコンポーネントを拡張した。）

多ければ、「記事→サッカー」のようなイメージを提供することが可能であろう。

次のステップ②では、図4のように、最初にユーザにウェブサイト中でもっとも特徴で簡単な概念ネットワークの一つのルール「Korea→Japan」を提供するとしよう（ここでは、2つのコアノードから始まるとする）。しかし、ユーザにとってはこのサイトが韓国と日本に関してどういう情報を提供しているかよく分からないので、概念ネットワークを拡張するために、現在の最小Supportを下げてサイト全体のイメージを拡張した。それによって、textbook, world cup, history, south, northと関わる内容を含めていることが分かる（実際に、図 4は朝日新聞データセットからの結果である）。特に、最初のコア概念の Korea と Japan から共通的に関わる Extended Conceptsがもっとも有効な情報であるという結果になっていた。

最終のステップ③は、例えばユーザは図 4のworld cupに関わる他の概念を知りたい時に、その概念を選択して、world cupからのルールの最小Confidenceを下げて特によく現れない内容でなくても広範囲の観点から情報を検索することを支援することが可能である。

上記のように、ウェブサイトの要約は、概念ネットワークをどの程度の拡張レベルで提供するかと、ユーザにどの概念にフォーカスをおいて概念ネットワークを提供する機能によって実現できる。また、これらは、2章の最小Supportと最小Confidenceの二つの尺度をコントロールすることによって計算できる。

3. 2 ウェブサイトのページリンク構造の概念度評価

新聞サイトなどではお互いの関連ページの間にリンクを貼ってユーザがもっと関係する知識を求めることを支援している。このようにウェブサイト中のお互いの関連を管理することは、ウェブサイト管理者の重要な役割の一つである。これに対して二つの観点から考えられる。まず、ウェブサイト中の全体のページがお互いの情報の関連性があるにも関わらずリンクがうまく貼られていないとか、あるいは一つのウェブページからの他のページへ到達するパスが長くなるとそのサイトの訪問者にとって情報を探しにくくなってしまうことになる。逆に、そのようなリンクの管理が面倒なウェブサイト管理者は、少しでもページの類似性があるページをお互いにリンクするようなプログラムを作ってしまうと楽になるかもしれないが、訪問者にとって今度はあまりにも多くのリンク中でどのページへ行くべきかを決めないといけなくなる。そのため、ページ中に含まれている知識のお互いのもっとも強い関係によって、各ページに関連するページへのリンクを最低限に貼る必要がある。

ウェブサイトのその内部にあるウェブがお互いにどの程度概念的に繋がっているか計るために、本稿ではウェブサイトの概念度を次のように提案する。

まず、ルールの各概念に関わるページがどの程度の繋がりになっているかを計算する。図 5のように、ルールA→Bにおけるそれぞれの対応するウェブページ集合をW(A)とW(B)としよう。そして、a_iはW(A)の要素であり、ab_iはW(A)∩W(B)の要素である。そして、この

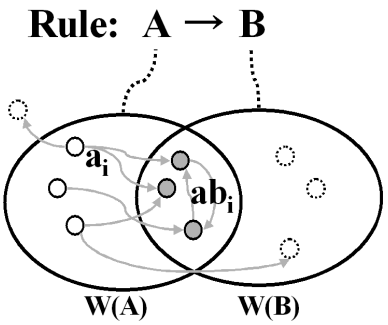

図５．ルールと対応するウェブページ

ルールに対して、ウェブサイトがどの程度リンクされているかを次のように計算する。

$$concept\_support(A \rightarrow B) =$$

$$\frac{1}{|w(A)|} \sum_{w(A)} \frac{\#(links\ to\ ab_i)}{\#(total\ output\ links\ from\ a_i)}$$

$$website\_concept\_support =$$

$$\frac{1}{|rule\_set|} \sum_{rule\_set} concept\_support(each\_rule_i)$$

そしてwebsite_concept_supportが1に近いほど、概念ネットワーク的なリンク構造をしていうことを意味する。この評価方法により、ウェブサイトにおけるリンク構造の全体または部分に対して連続的な情報検索に向いているかを評価することができ、その概念度が低いときには、そのウェブ集合からのページから導いたルールに基づいてリンクを自動的につけることも可能である。

４．関連研究

ウェブ情報から連想ルールに基づいた情報検索を行う手法は非常に直感的で有効な方法として知られている。Kawano[5]らによる Mondou というウェブ検索システムでは、ユーザが出した質疑語に対する関連語をウェブページ対するマイニング結果から得て、ユーザにフィドバックとして提供している。そして、Oyama[7]らによる研究では、ウェブから抽出したルールを利用してユーザの質疑を自動的に修正するエージェントシステムを提案した。また、ウェブネットワークと単語ネットワーク間をナビゲーションについても多く研究が知られている。Takano[8]らによる DualNAVI という情報検索システムでは、ウェブページに現れる単語を TF/IDF 手法を中心に特徴化して、あるページにもっとも関連する単語と、ある単語にもっとも関連するページの対応による情報ナビゲーションを支援している。また、WebBrain[9]というウェブ検索エンジンでは、単語間の階層的な関連を辿ることによってウェブページが検索できる優れたユーザインタフェースを提供している。我々の研究では[6]、KyotoSEARCH という地域ウェブ情報検索システムを開発した。このシステムは、地図、キーワード関連、ウェブ検索結果リストという三つの部分インタフェースとして構成されている。このシステムは、地域に関する知識をウェブ情報から連想ルールとして抽出してキーワード関連という部分で、関連あるキーワード間のナビゲーションができるようにしている。その関連は、連想ルールに現れるキーワードが地名(G)と非地名(N)に分かられていて、G→N，G→G，N→G，N→N の四つのタイプになっている。そして、それぞれのルールは、地域に関する知識を表現していると考えている。たとえば、ルール G→N の中には、'京都 '→ '観光 'というものがあり、これらは実際にウェブから '京都 'に関する多くのページから '観光 'に関するページが見つかるという結果から得られたものである。そして、キーワード関連インタフェースよってこれら四つのルールを移動することを実現している (たとえば、G1→N1→N2→G 2の経路では、'京都 '→ '観光 '→ 'お寺 '→ '銀閣寺 'ようのな例がある)。

5. 終わりに

本稿では、ウェブ上での連続的な情報検索を支援するために、ウェブページから概念ネットワークの構成とそれに基盤した次の二つの方法でその具体的な支援ついて述べた。

?? あるウェブページ集合の全体の要約
?? ウェブサイトの意味的なリンク構造評価

これらを実現するためには適切な概念ネットワークの構築が重要であり、そのために 2 章で述べたマイニング手法における Support と Confidence のコントロールとその生成されたルールの結合からの概念ネットワークの構築手法についても説明した。

なお、本研究は科学技術振興事業団（JST）戦略的基礎研究推進事業(CREST)における「デジタルシティのユニバーサル」プロジェクトの支援によって行われた。

参考文献

[1] R. Agrawal and R. Srikant. “Fast algorithm for mining association rules,” In Proc. of the 20th VLDB Conference, pages 487-499, Santiago, Chile, 1994.

[2] **朝日新聞英語サイト**
http://www.asahi.com/english

[3] DMOZ - Open Directory Project
http://dmoz.org

[4] Google
http://www.google.com

[5] H. Kawano, T. Hasegawa: Mondou: Interface with Text Data Mining for Web Search Engine. HICSS (5) 1998: 275-283

[6] R. Lee, H. Takakura, and Y. Kambayashi, "Visual Query Processing for GIS with Web Contents,” Proc. of the 6th IFIP Working Conference on Visual Database Systems, pp. 171-185,May 29-31, 2002.

[7] S. Oyama, and T. Ishida, “Applying Association Rules to Information Navigation,” IEICE Transaction on Information and Systems, Vol.J84-D-I, No.8, pp.1266-1274, 2001, (in Japanese)

[8] A. Takano, Y. Niwa, S. Nishioka, M. Iwayama, T. Hisamitsu, O. Imaichi, H. Sakurai, “Associative Information Access Using DualNAVI,” 2000 Kyoto International Conference on Digital Libraries: Research and Practice, pp. 285-289. Nov. 13th-16th, 2000, Kyoto University, Kyoto, Japan.

[9] WEBBRAIN2.0
http://www.webbrain.com/html/default_win.html

[10] Yahoo!
http://www.yahoo.com