第6章　部分時系列グラフの抽出

第5章で行った例は比較的ページ数，すなわちノード数の少ないグラフ構造の例であった。一般にはキーワードにヒットするページ全体のグラフ構造の可視化は困難である．そこで，時系列グラフ Gt に対して，部分時系列グラフを抽出する．まず，全体のグラフ構造を大まかに捉えるようにページを選んで部分時系列グラフを作成する．そして必要に応じて，あるひとつのページ p に注目し，その周辺のグラフを取る．その際に，時系列ごとのつながり（Gt に対して，Gt+1 や Gt-1 がどのように変化するか）を考慮して部分グラフを抽出する．

6.1.　全体の構造を考慮した部分時系列グラフ

全体を大まかに捉えるための部分時系列グラフの抽出は次のように行う．

ページ p の重要度について，

1. ランキング上位 N 個のページを選ぶ．
2. その各ページにリンクしているページに対して，同様にランキング上位 N 個を取る

次に，そのページの重要度の計算手法について述べる．

6.1.1.　ある時間におけるページの重要度

本研究ではページの重要度 s(p,t)として，そのページのインリンクを用いる．インリンクにも様々な種類が考えられる．

1. Web 全体からページ p へのインリンク数：n(p,t)
2. キーワードにヒットしたページ集合からページ p へのインリンク数：i(p,t)
3. リンク元のアンカーテキストにキーワードを含むようなインリンク数：a(p,t)

後のものほど，キーワードへの関連の度合いは強い．本研究では，特に断りのない限り，a(p,t)を用いることする．

6.1.2.　時系列にわたってのページの重要度

6.1.1 で述べた 3 種類のページのスコア s(p,t)は各時間 t におけるページの重要度の尺度である．時系列に渡っての変化を見るために，これの時間積分

$$S(p) = \int s(p,t)$$

を時系列にわたってのページの重要度とする.

以上の S(p), s(p,t)について，状況に応じていずれかをランキングスコアとし，ランキングを行って，部分時系列グラフを得る．抽出の範囲 N は 10 とした.
さらにノード数を抑えるために，同じサイトのページを一つのノードにまとめてある.

6.2. あるページの周辺のグラフの抽出

ページ p の周辺のグラフの抽出は次のように行う.

1．すべての時系列において，p に隣接するノードを取り，その集合を U とする.
2. 各時間 t において，ut ∈ U となる ut を新たにノードとする.
3. 集合Uの中でつながるリンクだけをとり，エッジ（ut, vt| vt ∈ U）とする.

このように得られるグラフを p の周辺における部分時系列グラフとする.

ページ p については，ユーザが選んだページを得られるように実装した．これはページの URL を直接指定することで行う．ただし，あくまでキーワードに関するグラフ構造を想定しているため，その URL はいずれかの時間においてキーワードにヒットするページ集合に含まれている必要がある．現実的には，まず，6.1 の手法でおおまかなグラフ構造を捉えた後，その中で気になるページを直接指定することを想定している.

6.3. WebRelievo の改良

従来の WebRelievo は 3×２のパネルセットであったが，本研究の 8 回分の時系列にあわせて 4×２に変更した．また，ノードして選んだページにアクセスして内容を参照することを容易にするため，アーカイブ内の過去のページにもアクセスできるよう改良を行った.

第7章　部分時系列グラフの可視化

第6章のような手順で得た部分グラフを，改良したWebRelievoを用いて可視化する．第6章で述べたランキングスコアS(p), s(p,t) のいずれを使うか，ある特定のページの周辺のみを見た場合はどうか，といったそれぞれの場合について，時系列変化がどのように見えているか，その変化は実社会の事象を反映しているか，といった分析を，例を示しながら行う．

7.1. 全体的な時系列変化の例と分析

図 19 はキーワード"ベッカム"で得られたグラフから部分時系列グラフを抽出して描画したものである．ページのランキングは S(p)を用いている．図 19 を見ると，2002/02～2003/02 の間に新規ノードが多く現れていることが分かる．この部分をさらに詳しく見たものが，図20である．

図20では，2003/02 で大きく分けて 3 つのコンポーネントがそれぞれ現れていることが分かる．それぞれ中心のページにアクセスしたところ，右からそれぞれ，

- 個人による（サッカー選手）ベッカムのファンサイト
- 映画「ベッカムに恋して」の公式サイト
- メール件名に「ベッカム」と表示するウイルスを警告するニュース記事

であった．

これらのコンポーネントの次の時間についてみたものが図 21 である．まず，2003/07 には別個に現れたコンポーネントがそれぞれ結合している様子がわかる．これはウェブリンク構造の発展過程を良く表している．

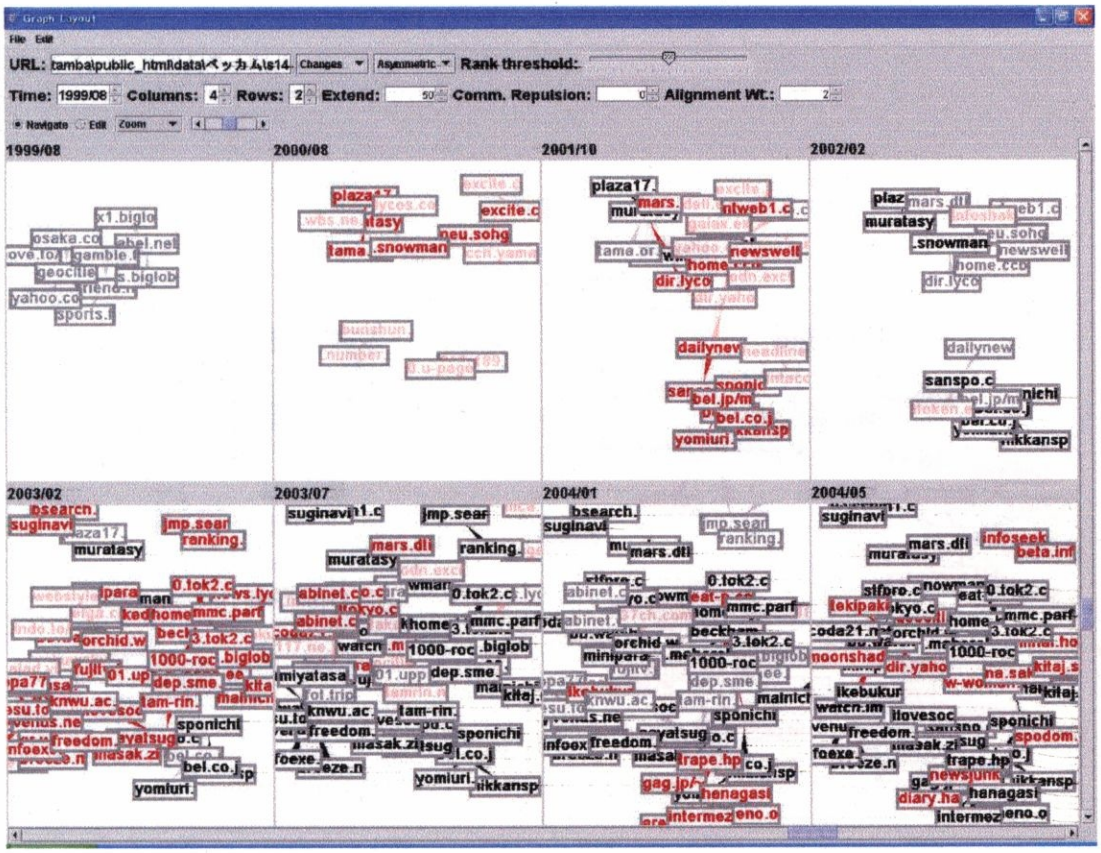

図 19：部分時系列グラフ（全時間，キーワード：ベッカム）

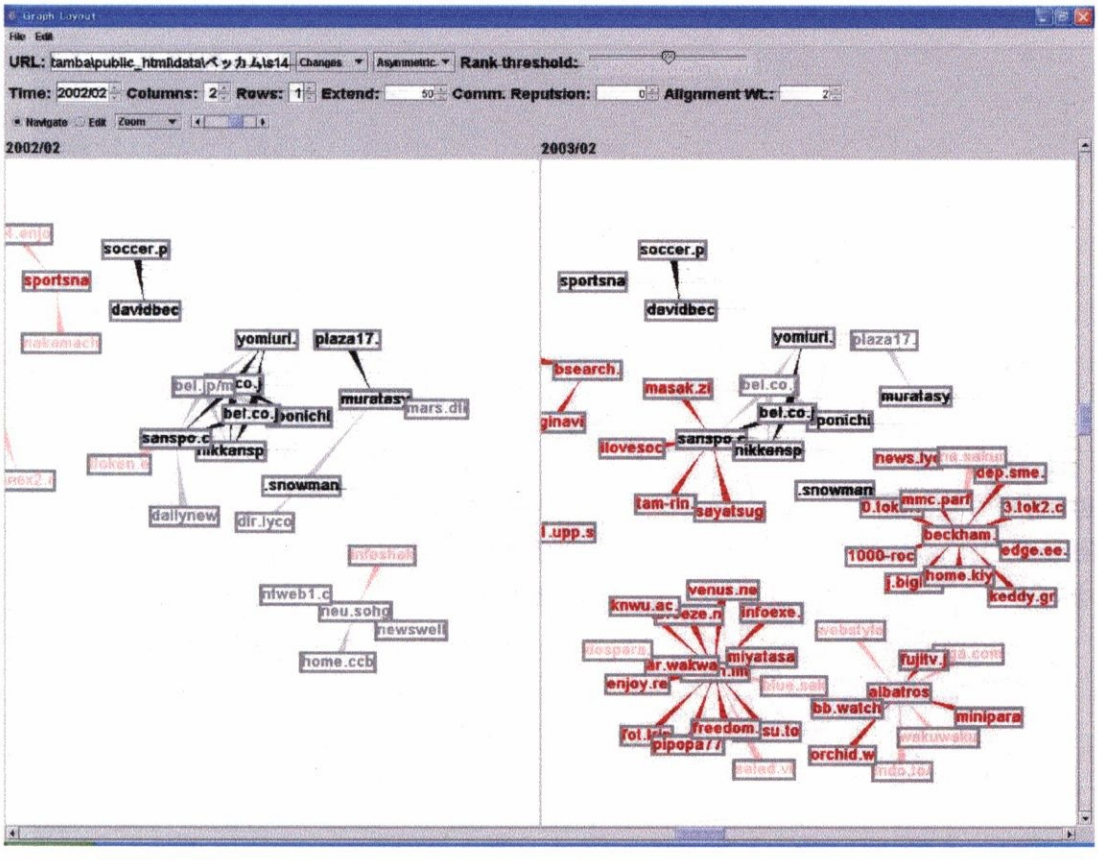

図 20：部分時系列グラフ（2002/02～2003/02，キーワード：ベッカム）

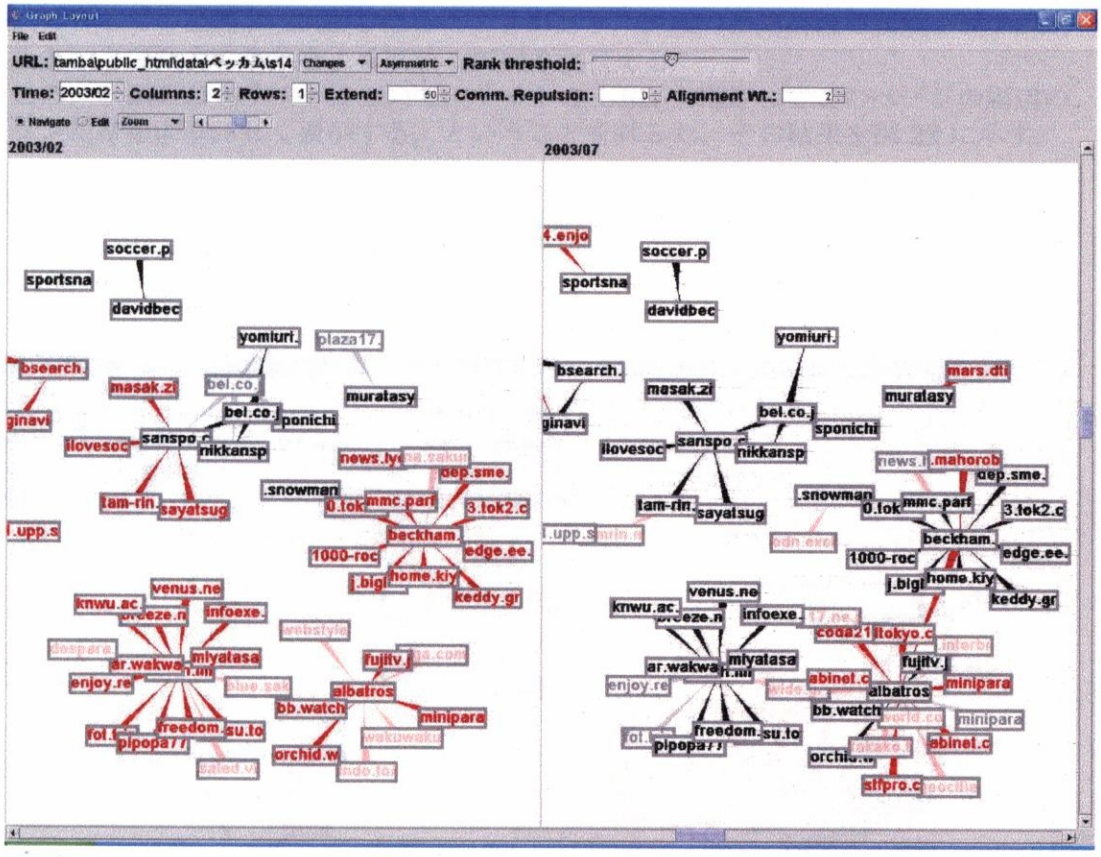

図 21：部分時系列グラフ（2003/02～2003/07，キーワード：ベッカム）

7.2. 一つのページに注目した例と分析

図 21 において，映画の公式サイト（図中右下に現れている塊）に関しては，2002/02～2003/07 にかけてさらに新規リンクが増えている．この変化の背景としては，この映画の日本での公開が 2003/4 だったことがあり，映画の公開前後で新規リンクが増えるという実社会の事象をよく反映していると言える

この，映画の公式サイトについて，サイトの URL を指定してこのページの周辺のグラフだけをより詳しく表示する，ということを試みた．その結果を図 22 に示す．

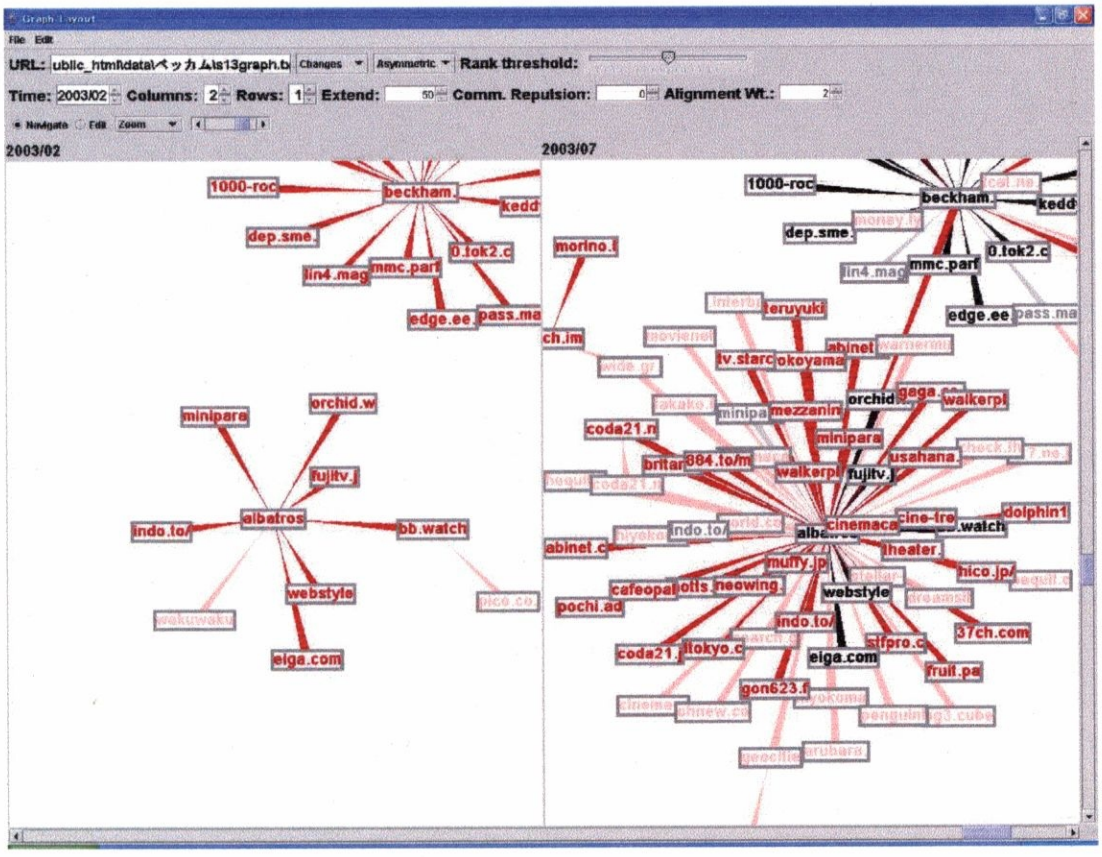

図 22：部分時系列グラフ（2003/02～2003/07，キーワード：ベッカム，url 指定:"http://www.albatros-film.com/movie/beckham/"）

図 22 を見ると，2003/02～2003/07 にかけて，図 21 よりも多くの新規リンクがあることが分かる．図21 ではキーワードについての時系列変化を大まかに知るために，ランキング上位のページについての部分グラフを可視化しているので，ランク外のページについては表示されない．そこで，この場合のように，改めて気になるページの周辺だけを切り取って部分グラフとすることで，そのページの周辺の変化がより詳細にわかる．

実社会の事象の反映としては，映画公開前はテレビ局や大手のレビューサイト等，いわば宣伝としての意味合いのリンクが多かったのが，公開後は大手ではない個人のレビューサイトや日記サイト等，映画を見た後の感想としての意味合いのリンクが増えていることがわかる．図 21 の段階の分析では，ランキング上位で切り取っている以上，公開後に増えたリンクも大手のレビューサイトが多く，それが宣伝の意味合いのリンクか乾燥の意味合いのリンクかはわかりにくいが，一つのページに注目して展開することで，実社会の事象の反映もより詳しく確認することができる．

7.3. ページのスコアを変えた場合の例と分析

ページのスコアを基準にランキングを取る際，場合によっては，そのスコアの取り方によってはランキングが変動し，表示されるグラフも異なったものになることがある．以下では，そのような場合の例を，実社会の事象との関わりも含めて分析する．

キーワード"BSE"における例について，全時間における可視化を行ったものを図 23，その中で大きな変化があった 2000/08～2001/10 のものを図 24 に示す．

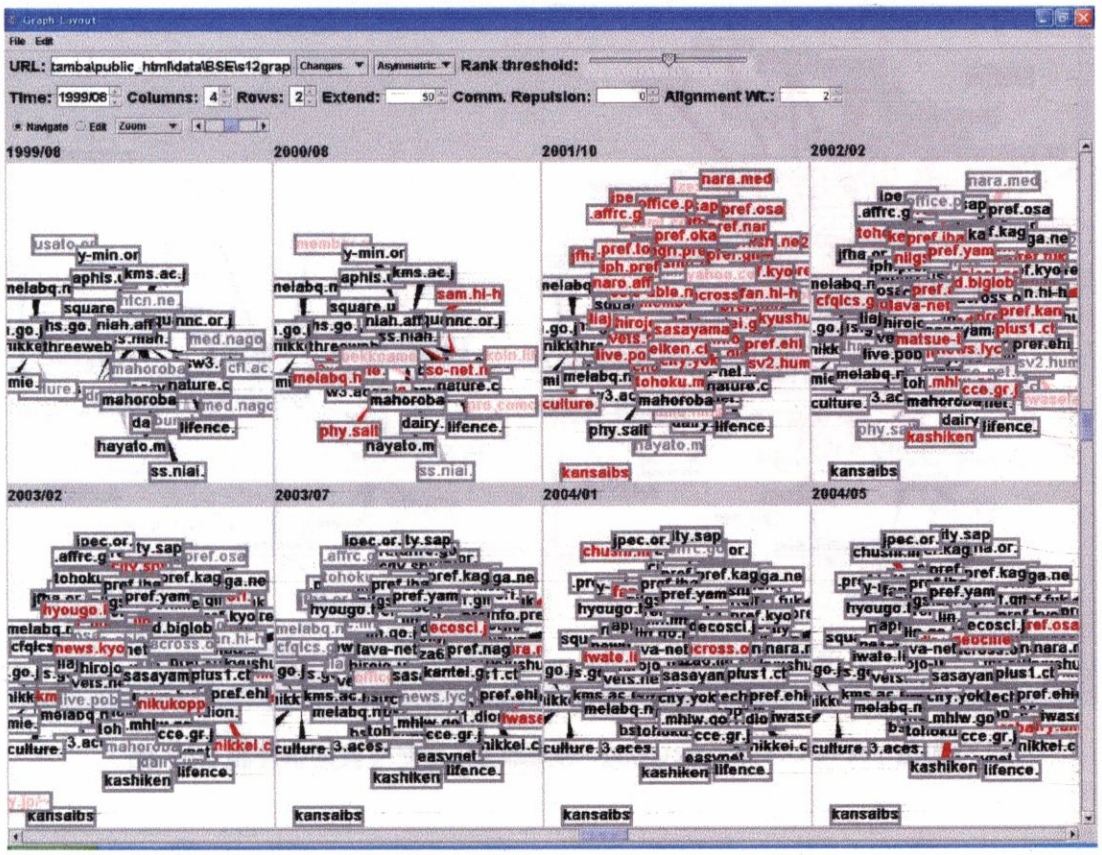

図 23：部分時系列グラフ（全時間，キーワード：BSE）

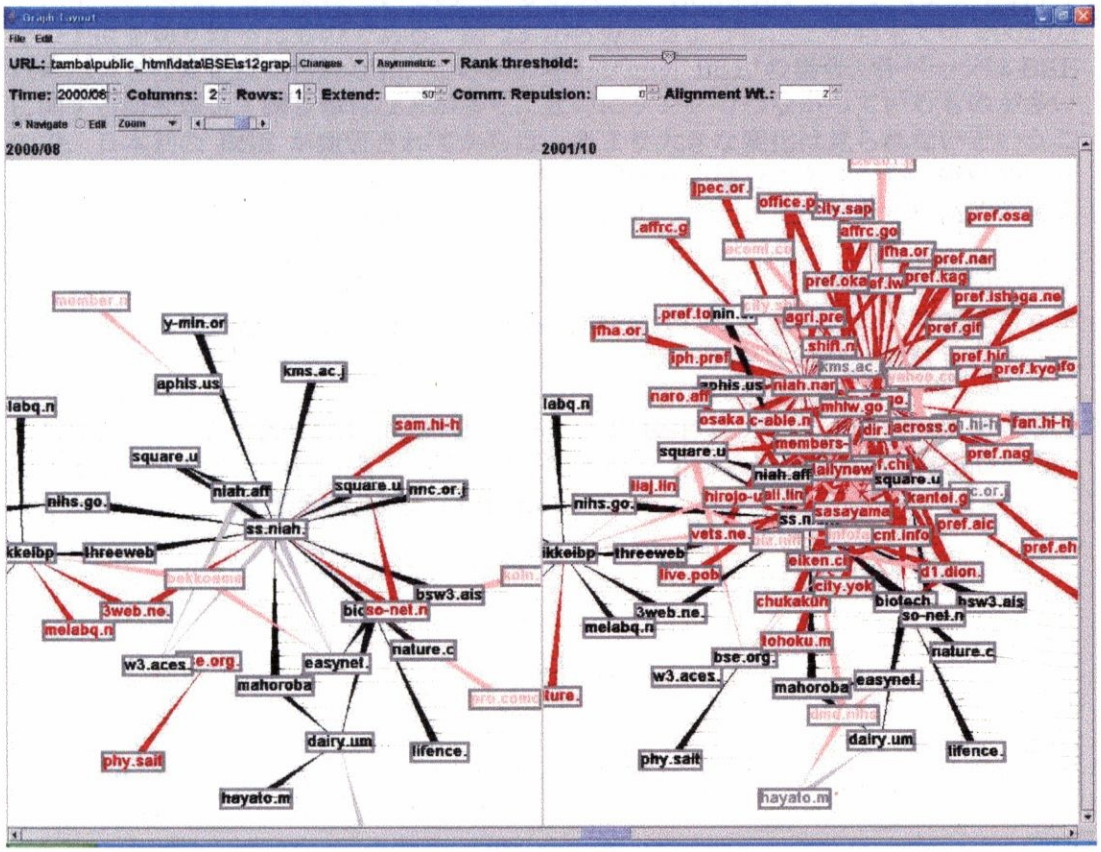

図 24：部分時系列グラフ（2000/08～2001/10，キーワード：BSE）

図 24 では大きな部分グラフが現れ，しかもそれ以前から存在していたページ群と密接につながっていることが分かる．これは，2001/09 に千葉県にて日本初の感染牛が確認された，という事象を反映している．2000/08 以前から存在するページ群は BSE そのものを解説したページなどで，英国での初の BSE 発見からの歴史などの内容がある．2001/10 に現れた新規ページ群の中心は厚生労働省のページと農林水産省のページ（中心が一つではなく，両方のページにリンクが集まっている）で，リンク元のページは全国の農協などが多い．

以上の結果は，全時間についてのページのスコア S(p)を用いてランキングを行い，その上位を切り取ったものである．図 23 から分かるように，この結果では 2002/02 以降に大きな変化は見られない．この例のキーワード"BSE"の場合，各ページは BSE に関する研究報告・調査報告などが多く，互いにリンクでつながっているものが多いので，日本初の BSE が報告されてからは，あまり大きな変化は見られないということになる．スコア S(p)は，各時間のスコア s(p,t)の時間積分であるから，継続的にリンクを集めているページが上位にランクされる．よって，長期にわたる持続的な時系列変化の推移を見るのに適していると言える．

これに対して，1 時系列前後の短期的な変化を見るには，各時間 s(p,t)を用いてランキングした方が都合が良いと考えられる．本研究では 1 時系列分のスコアは第 6 章で述べたようにリンク元アンカーテキストにキーワードを含むページからのインリンク数 a(p,t)を用いる．

以下，2003/07～2004/01 の期間，キーワード"BSE"について，全時間におけるスコア S(p)を用いてランキングしたものを，図 25 に，1 時系列分のスコア a(p,t)を用いてランキングした結果を図 26 に，それぞれ示す．

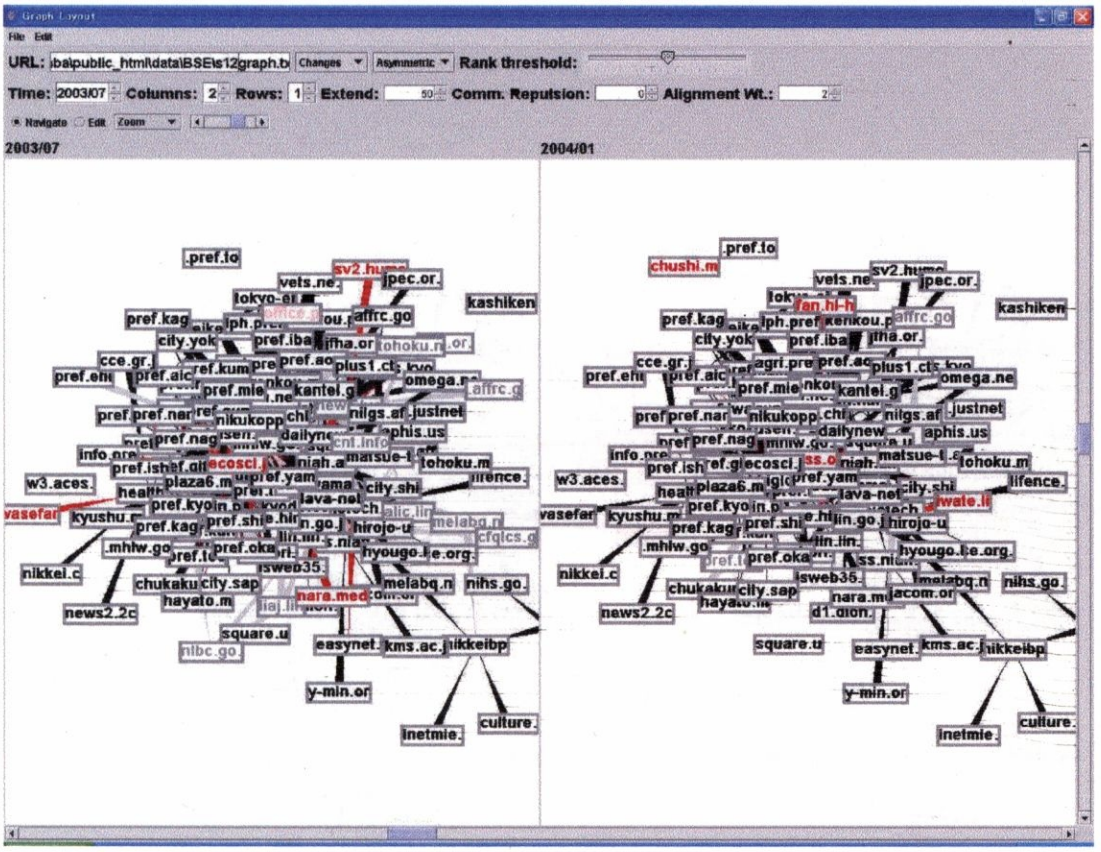

図 25：部分時系列グラフ（2003/07～2004/01，キーワード：BSE，スコア S）

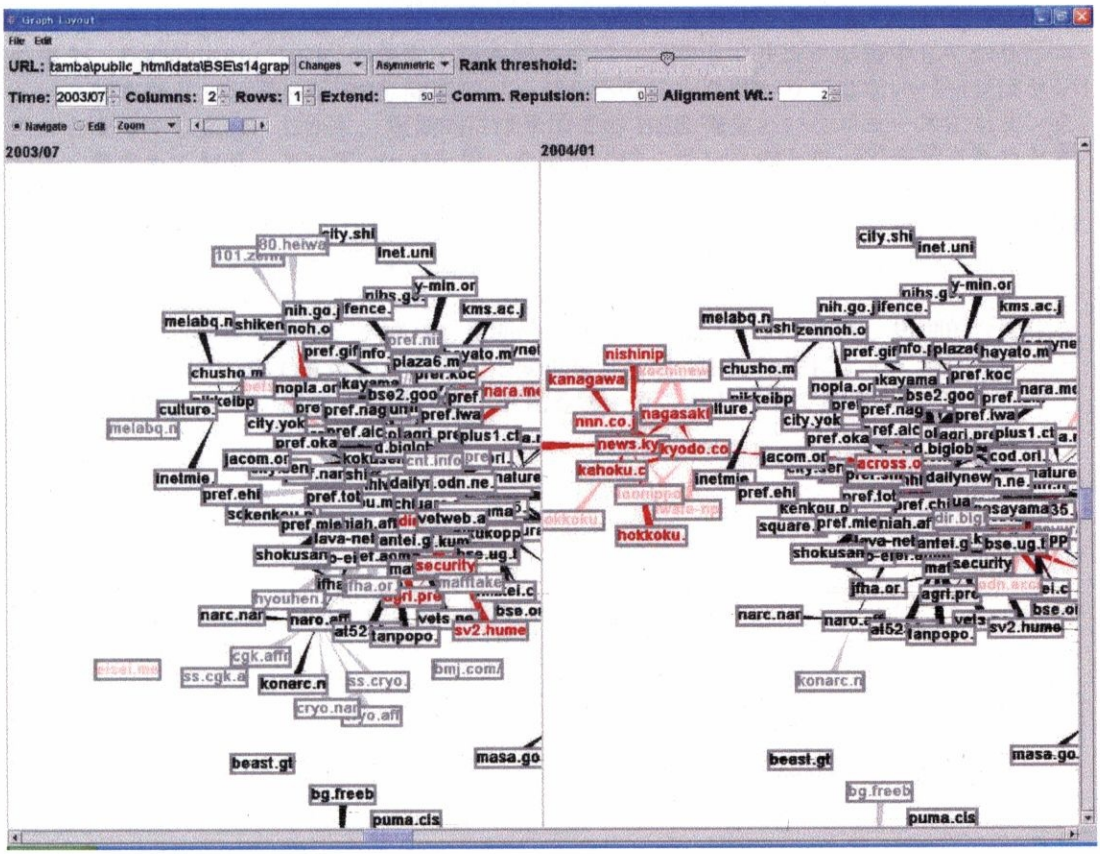

図 26：部分時系列グラフ（2003/07～2004/01，キーワード：BSE，スコア a）

図 25 と図 26 を比較すると，スコア S(p)による結果では表示されていなかったページ群がスコア a(p,t)による結果では現れていることが分かる．
まず，この新規ページ群について，実社会の事象との関連付けの観点から分析すると，2003/12 に米国で BSE が発生しており，その事象を反映しているものである．実際にページの内容を見てみると，中心は米国での BSE 発生を伝えるニュース記事である．
次に，これがなぜスコア S(p)と a(p,t)によって違った現れるかという考察であるが，既に述べたように S(p)を用いることは，長期的な時系列変化を見るときに適している．この例の場合，長期的には，BSE の研究・調査報告のページ同士でリンクが集中しており，ニュース記事など，必ずしも BSE 関連のページからのリンクばかりではないページはランク外となってしまう．しかし，短期的には米国での BSE 発生というニュースにもリンクが多く集まっており，スコア a(p,t)を用いた場合には，ランク内となって表示されると考えられる．

参考として，この米国の BSE に関するニュース記事を URL 指定してその周辺を展開した場合のグラフを図 27 に示す．実際には，多くの関連ページ（主にニュースサイト）がリンクでつながって発生していることがわかる．

このように，継続的なリンク構造を長期的に見るときはスコア S(p)を，短期的な変化を見る場合はスコア s(p,t)を用い，発見した変化をさらに URL 指定によって展開することで，より詳細にリンク構造の時系列変化を見るということが可能である．

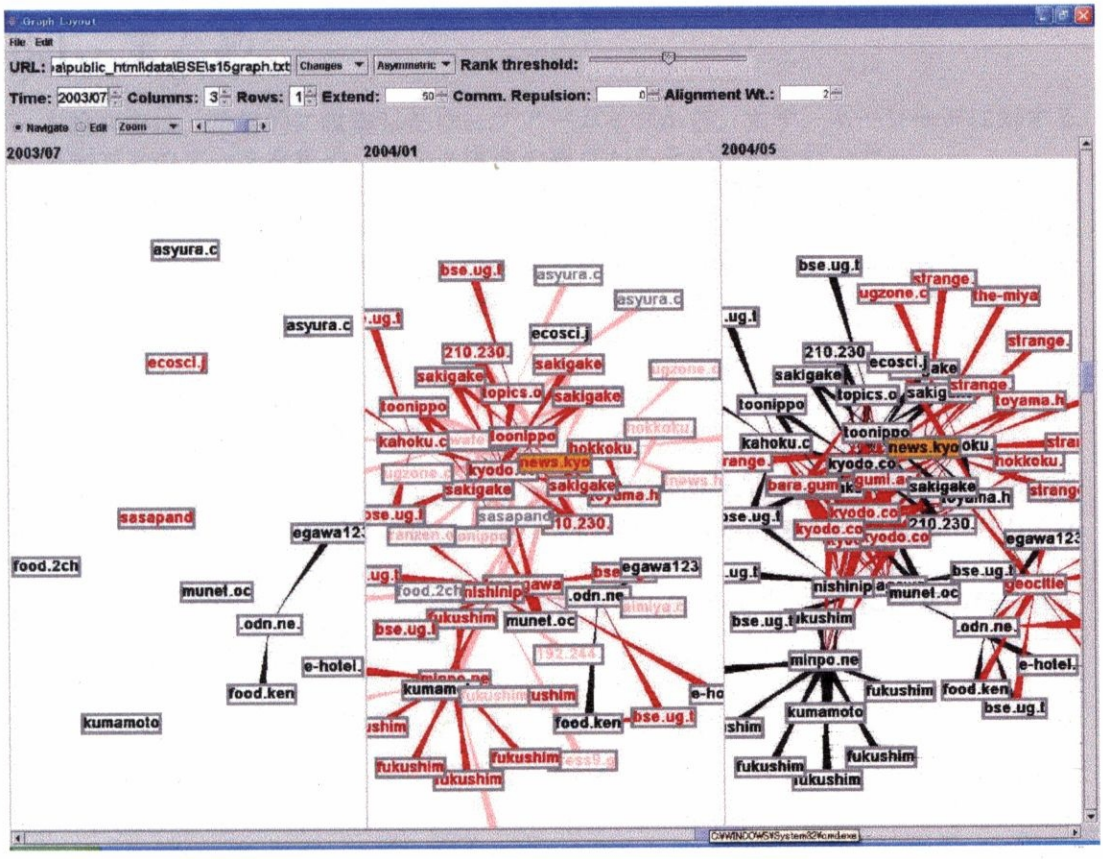

図 27：部分時系列グラフ（2003/07～2004/05，キーワード：BSE，　url 指定：http://news.kyodo.co.jp/kyodonews/2003/bse/ ）

第8章　結論

8.1. まとめ

本論文では，定期的に収集したウェブアーカイブから，まず，キーワードに対する時系列ごとのデータを集め，リンク構造の観点からその分析を行った．

次に，それらのデータから抽出された時系列グラフを可視化し，過去の実社会の事象の反映という観点から分析を行った．その際，部分グラフを切り取り，またそのためのランキングスコアを導入した．ランキングスコアを様々に変えた場合やあるページだけに注目した場合についても可視化と分析を行った．

8.2. 今後の課題

本研究では部分グラフを切り取ることで，グラフの時系列変化の可視化が一歩進んだが，グラフの変化と実社会の過去の事象とについてさらに多くの例を調べてみる必要がある．

また，膨大な数のウェブページのグラフを扱うにあたって，さらに高速にデータを抽出する工夫が今後の課題である．

謝辞

本研究を行うにあたり，研究テーマから研究生活に至るまで，終始ご指導とご助言をいただきました喜連川優教授に厚くお礼申し上げます．

豊田正史先生には研究内容について本当によく指導していただきました．助手の中野美由紀さんにも助けていただきました．心からお礼申し上げます．

また，同研究室の皆さまにも様々なアドバイスをしていただきました．お世話になった方々に深く感謝いたします．

参考文献

[1] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web" , Stanford Digital Libraries Working Paper , 1998.

[2] N. Eiron, K. S. McCurley, J. A. Tomlin, IBM Almaden Research Center, "Ranking the Web Frontier" ,World Wide Web Conference, 2004.

[3] Sepander D.Kamver, Taher H. Haveliwala, Christopher D. Manning, Gene H.Golub, "Exploiting the block structure of the web for computing pagerank" , Stanford University, 2003.

[4] Wayback Machine, The Internet Archive. http://www.archive.org/

[5] 豊田正史, 喜連川優 "WebRelievo: ウェブにおけるリンク構造の発展過程解析システム" Workshop on Interactive Systems and Software: WISS2004, pp.89-94, 2004.12.1-3

[6] J. M. Kleinberg. "Authoritative Sources in a Hyperlinked Environment." In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms,pp. 668-677, 1998.

[7] J. Dean and M. R. Henzinger. "Finding related pages in the WorldWideWeb." In Proceedings of the 8th World-Wide Web Conference, pp. 389-401, 1999.

[8] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.

[9] Jeffrey Dean and Monika R. Henzinger, "Finding related pages in the World Wide Web", In Proceedings of the 8th International World Wide Web Conference, 1999.

[10] M. Toyoda and M. Kitsuregawa. "Creating a Web Community Chart for Navigating Related Communities." In Conference Proceedings of Hypertext 2001, pp. 103-112, 2001.

[11] C. Chen and L. Carr. "Visualizing the evolution of a subject domain: A case study." In D. Ebert,M. Gross, and B. Hamann eds., IEEE Visualization '99, pp. 449-452, San Francisco, 1999.

[12] C. Chen and S. Morris. "Visualizing Evolving Networks: Minimum Spanning Trees versus Pathfinder Networks." In IEEE Visualization 2003, pp. 67-74, 2003.

[13] E. H. Chi, J. Pitkow, J. D. Mackinlay, P. Pirolli,R. Gossweiler, and S. K. Card. "Visualizing the Evolution of Web Ecologies." In Proceedings of ACM SIGCHI '98, pp. 400-407, 1998.

[14] C. Erten, S. G. Kobourov, V. Le, and A. Navabi. "Simultaneous Graph Drawing: Layout Algorithms and Visualization Schemes. In The 11th Symposium on Graph Drawing, pp. 437-449, 2003.

[15] P.Eades. "A Heuristic for Graph Drawing", Congressus Numerantium, 42, 149–160, 1984.

[16] T. M. J. Fruchterman and E. M. Reingold. "Graph drawing by force-directed placement." Software - Practice and Experience,21(11):1129-1164, 1991.

[17] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web page. In Proceedings of the 12th International World Wide Web Conference, pages 669--678, May 2003.
[18] A. Ntoulas, J. Cho, and C. Olston. "What's new on the web? the evolution of the web from a search engine perspective." In WWW Conference, pages 1--12, New York, New York, May 2004.