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

2004－DBS－133　(9)
2004－ＦＩ－ 75　(9)
2004／5／14

Web 上の「活動の場」に着目した人物の特徴付け

佐藤 進也　原田 昌紀　風間 一洋
NTT 未来ねっと研究所
東京都武蔵野市緑町 3-9-11

Web は人工的に作られた情報空間であるが，そこに記述されている内容は実世界を反映したものである．実社会における Web の担う役割が今後ますます重要なものとなることを考えると，Web の “実世界の鏡” としての傾向は今後もさらに強まっていくと思われる．このことを鑑みると，実世界の知識を積極的に用いた Web の解析，いわば実世界指向 Web マイニングというアプローチが有望であると考えられる．本論文では，この考え方にもとづき，実世界における人とその活動の場との関係を Web に当てはめ，与えられた名前をもつ人物の Web 上での活動の場を抽出する方法を導く．そして，その妥当性を具体例を用いた実験により検証する．このようにして得られた Web 上の活動の場は，当該人物の特徴を与える情報としての利用が期待される．本論文では，その一つの試みとして，活動の場の相互関係を可視化するという手法を紹介する．本手法により，同姓同名の異なる人物をおおまかに弁別できる．

Figuring out People by their Workspaces on the Web

Shin-ya SATO, Masanori HARADA and Kazuhiro KAZAMA
NTT Network Innovation Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo

Although the Web is an artificial space for information utilization, many of the pages on the Web are mentioning things or concepts in the real world. As the Web is going to play more important role in our daily lives, it will more precisely reflect the real world. In consideration of these facts and the trend, an approach for analyzing the Web that makes good use of knowledge on the real world, so to say “real world oriented Web mining”, is considered to be promising. Based on this idea, this paper presents a method to extract workspaces for persons with a given name on the web. The method is derived by applying knowledge on the relationship between people and their workspaces in the real world to the Web. The validity of the method is verified through experiments. A set of workspaces for a person extracted by the method from the web is expected to be put to use for characterizing the person. As an example, the paper introduce a technique to visualize relationships between the workspaces, which can be used for roughly distinguishing between persons with the same family and personal name.

1 はじめに

Web マイニング [1] は，Web という多様性に富んだ広大な情報空間から必要な情報を効率よく取り出すための一つのアプローチである．これは，Web を解析してその特徴を抽出し情報の取捨選択などに利用することを狙ったもので，その例としては，ハイパーリンクリンクを介したページ間のつながりを解析して得られる，PageRank[2] という Web ページの評価尺度が挙げられる．Web マイニングでは，Web の特徴抽出の精度を向上させるためや，抽出された特徴の妥当性・有用性を評価するために実世界の知識を用いることはあるが，基本的には Web の中身 (文書，ハイパーリンク，ユーザの利用履歴など）に注目し，そこに潜在する特徴的なパタンの “採掘” を行う．Web は人工的に作られた空間であり，そこで情報を記述するための独自のルール[1]があるため，当然，解析にあたっては Web に特化した方法が必要となる．

しかし，Web は実世界から分離されているわけではない．むしろ，Web において記述されている内容は実世界を反映したものであると考えるべきだろう．実社会における Web の担う役割は，情報流通だけでなく，経済的にも，ますます重要なものとなることが予想され，“実世界の鏡” としての傾向は今後もさらに強まっていくと思われる．

このことを鑑み，我々は実世界の知識を積極的に用いた Web の解析，いわば実世界指向 Web マイニングというアプローチを提案してきた．人などの実世界のエンティティを Web から発見する実世界指向情報検索手法 NEXAS[3] はその一例である．

本論文では，実世界における人とその「活動の場」との関係の Web への当てはめを試みる．具体的には，まず，実世界における人の活動の場の要件を明らかにする．そして，その要件に合致する Web ページの集合を Web 上の活動の場と考える．本論文では，与えられた人（人名）からその Web 上の活動の場を抽出するアルゴリズムを述べるとともに，2 つのケース（人名）に適用した結果を示す．このアルゴリズムにより得られる Web 上の活動の場は，当該人物の特徴を与える情報としての利用が期待される．本論文では，その一つの試みとして，活動の場の相互関係を可視化するという手法を紹介する．前述の 2 つのケースに対しては，この手法の適用により，活動の場が（同姓同名の異なる）人物の単位でおおまかに分類されることを確認した．

以下，2 章では，実世界における活動の場の要件を明らかにする．それらが Web に対してどのように適用できるかを 3 章で議論し，与えられた人物の活動の場を Web から抽出するアルゴリズムを導く．4 章では，このアルゴリズムを具体的な人物（人名）に対して適用し，その妥当性を確認する．さらに，5 章では，活動の場の相互関係を解析し可視化する方法と，それを用いた同姓同名人物の弁別について述べる．

[1]例えば，文書は HTML で書く，など．

2 実世界における活動の場の要件

本章では，実世界において人が活動する場とはどのようなものであるか，その要件を明らかにする．

国語辞典によれば，活動とは『目的に応じた積極的・精力的な行動やその業績』を指す言葉とされている．ここに「目的」「積極的・精力的な行動」そして「業績」という 3 つの特徴を見い出すことができる．本論文では，この 3 つの特徴を使って活動の場を次のように定義する．すなわち，活動の場とは，(1) 組織やプロジェクトなどとしての一貫した目的を持っており，(2) 人がそこで継続的に行動し，(3) その結果として目的に沿う何らかのものが生産されるような場所とする．ここで「積極的・精力的な行動」は,「人と活動の場に継続的な関係がある状態」であると読み替えている．

3 Web における人と活動の場

本章では，前章で示した実世界における活動の場の要件を Web に当てはめることにより，ある人物（人名）が与えられたときに，その Web におけ

る活動の場を抽出する方法を導く．

3.1 Webにおける「人」

2章で示した活動の場の要件の1つに，人と継続的関係を持つことがある．この条件をWebに当てはめるためには，Webにおいて何をもって人物とみなすか（あるいは，どのようなものを人物を指し示すものとするか）を明確にしなければならない．

本手法では，NEXASのアプローチと同様に，Webページに現れる人物を表す固有表現，すなわち人名を実世界の人物に対応するものと考える．具体的な人名の抽出方法もNEXASに従い，Webページの内容を形態素解析し，品詞が人名として同定された形態素の並びを人名として採用する．

Webページに人名が出現する確率は比較的高く，我々の調査では，Webロボットで収集した4000万を越えるWebページのうち2割強は人名を含んでいた[3]．各人名について，出現ページの数と出現Webサーバ数を対比させてみるたところ，それぞれ固有のサーバに偏って現れる傾向，すなわち人名の偏在性が認められた．図1は，無作為に選んだ500の人名と同数の普通名詞の出現傾向を比較したものであり，横軸，縦軸はそれぞれ各語の出現するページ数，サーバ数である（対数スケール）．人名に対応する点は，普通名詞に比べてグラフ下方にばらついていることが，その偏在性を示している．

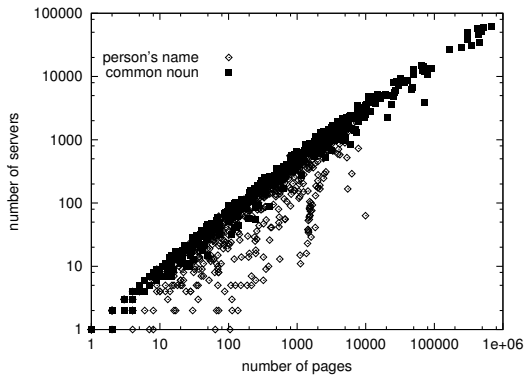

図1: 語の偏在性

Webサーバは組織やプロジェクトがその活動の一環として立ち上げており，人名の偏在は，実世界において各人物がそれぞれ固有の分野で活動していることに対応していると考えられる．これは，Webが実世界を反映しているというここでの議論の前提を支持する事実としてとらえることができる．ただし，図1を見て分かるように，偏在性の度合いは人名によりまちまちであり，普通名詞と同レベルのものもある．この，いわゆる“普通名詞化”現象は，有名人に多く見受けられる．このような人物に対して本手法の効果の有無，あるいはもっと一般的に，偏在性の高低と本手法の効果との関係について，4章のケーススタディにおいて調べる．

3.2 活動の場の抽出

Webにおける「人」の定義とその特徴が明らかになったところで，2章で示した要件をWebに当てはめてみる．

まず，(1) および (3) を満たすものとしてWebサーバが挙げられる．組織やプロジェクトがその活動の一環としてWebサーバを立ち上げていること，Webで提供されている情報はそこでの生産物とみなせることがその理由である．しかし，一つの組織で複数のプロジェクトを立ち上げていることも珍しくない[2]．注目している人物の活動が，組織全体には直接関係せず，あるプロジェクトに閉じているような場合には，そのプロジェクトに対応するWebサーバ（が提供するページ全体）の一部分を活動の場と考えるべきであろう．一方，Webサーバの一部を特定用途に割り当てるためには，多くの場合，ファイルシステムの階層構造を利用して，あるディレクトリ以下を割り当てるという方法がとられている．そこで，本論文では，Webサーバ内で階層構造上互いに隣接して存在するWebページの集合体を活動の場（の候補）と考

[2]たとえば，独立行政法人情報処理推進機構 http://www.ipa.go.jp/はセキュリティセンター(http://www.ipa.go.jp/security/)や未踏ソフトウェア創造事業(http://www.ipa.go.jp/jinzai/esp/)といった複数の事業活動を推進している．

える．

そして，人との継続的な関係があるという (2) の条件には，Web ページ集合体の中に当該人物を示す人名が複数回出現する，という条件を対応させる．

以上をまとめると，Web[3]から *name* という姓名を持つ人物の活動の場を抽出するアルゴリズムは以下のようになる．

1. *name* を（文字列として）含むすべての Web ページ（を指し示す URL）の集合を U とする．

2. URL の集合を要素とする集合 T を以下のように初期化する：

$$T = \{\{u\} | u \in U\}$$

3. T の 2 つの要素 C_1, C_2 に対して，ある $u_1 \in C_1$ と $u_2 \in C_2$ が存在して，u_1, u_2 が置かれているディレクトリ階層の隔たりが高々 1 である場合には，C_1 と C_2 をマージする．例えば，

$$C_1 = \{\texttt{http://a.jp/x/v.html}\}$$
$$C_2 = \{\texttt{http://a.jp/x/y/z.html}\}$$

であったとき，T から C_1 と C_2 が削除され，新しい要素 C_{new} が追加される：

$$C_{new} = \{\texttt{http://a.jp/x/v.html}, \texttt{http://a.jp/x/y/z.html}\}$$

4. マージされるものがなくなるまで 3. を繰り返す．

5. T の要素 C で，その要素数が 2 未満であるものを T から削除する．

ここで最終的に得られた T の要素が人物 *name* の活動の場に対応する．以降，本論文では，T の要素をクラスタと呼ぶことにする．

[3]正確には，ロボットにより収集した Web ページの集合．

4 ケーススタディ

上記アルゴリズムの妥当性を検証するため，具体的な 2 つの人名を適用して活動の場を抽出する実験を行った．本章ではその内容と結果を紹介する．

4.1 人名の選択

本実験に用いる人名としては，その人物の実績が広く知られていること，同姓同名の人物が存在する場合に発生する問題を調べる手がかりとなることを考慮して「竹内郁雄」と「江川卓」を選んだ[4]．以下，この 2 つのケースをそれぞれ TAK, EGW と略記することにする．

アルゴリズムの適用実験に先だって，偏在性の観点から TAK と EGW を比較しておく．図 2 は，図 1 の一部を拡大したもので，TAK と EGW に対応する点が示されている．TAK が普通名詞の領域の境界にあるのに対し，EGW は普通名詞の領域内に含まれていおり，EGW は TAK より偏在性が低いことがわかる．

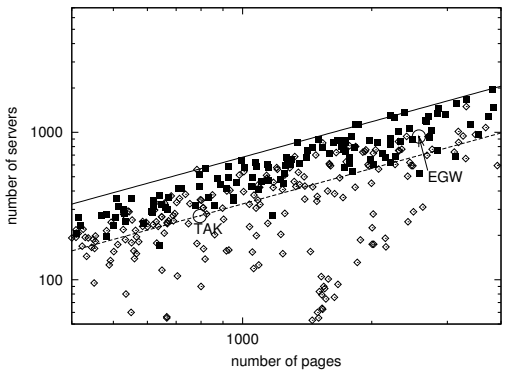

図 2: TAK と EGW の偏在性の違い

[4]Web 検索で調べた範囲では，「竹内郁雄」という名の人物には，少なくとも電気通信大学教授，独立行政法人森林総合研究所研究調整官の二氏が，「江川卓」には，元プロ野球選手，ロシア文学者，バスプロ（ブラックバスを釣る競技のプロ選手）の三氏が存在する．

4.2 実験結果

TAK と EGW に対してそれぞれ活動の場を抽出するアルゴリズムを適用した結果を以下に示す.

4.2.1 クラスタ

まず，実験により得られたクラスタの数を表 1 に示す．表には，それぞれの人名が出現する Web ページの総数（ページ数）とクラスタあたりの平均ページ数（平均ページ数）も併せて示した.

表 1: 抽出されたクラスタの数

人名	ページ数	クラスタ数	平均ページ数
TAK	795	58	6.93
EGW	2575	46	5.46

EGW は出現ページ数では TAK を大きく上回っているにもかかわらず，得られたクラスタ数は TAK よりも少ない．これは，TAK が EGW に比べて偏在性が高く（すなわちページが局所に集中している），クラスタが生成されやすいためと考えられる.

4.2.2 活動の場としての妥当性

ここで得られたクラスタの活動の場としての妥当性を評価するため，クラスタを構成しているそれぞれのページの内容を調べた．その結果，クラスタは以下の 3 タイプに分類できることが分かった.

I. 当該人物が直接関与しているもの．その人物が所属・活動している組織やプロジェクトなどに対応するもの.

II. 当該人物は直接関与していない（あるいは関与が明らかでない）組織等がその活動の一環として当該人物の活動や業績に言及しているもの．その人物の活動と目的を同じくする別組織や，その分野の情報を扱うメディアなど.

III. 当該人物の活動とは直接関係ない目的で作成されているページでのその人物に関する言及．個人の日記や時間軸で事実を列挙した記録など.

この他にも，第 4 のタイプとして，クラスタが当該人物と関係無い場合（例えばスパム的行為によるものなど）が考えられるが，今回の実験にはこのタイプに該当するものはなかった.

この 3 タイプのなかで実世界における人の活動の場に厳密に合致するのは I であるが，広義の活動の場として II を含めることができる．これは，たとえば，ある人物が著書を多く出している出版社があったとき，その出版社も広い意味でその人物の活動の場とみなせる，という考えにもとづくものである.

TAK，EGW のクラスタ群をそれぞれタイプ毎に分類すると表 2 のようになる．TAK では 7 割強クラスタ，EGW では 6 割強のクラスタが広い意味で活動の場として妥当なものであった.

表 2: クラスタの分類

人名	I	II	III
TAK	36 (62.1%)	8 (13.8%)	14 (24.1%)
EGW	5 (10.9%)	25 (54.3%)	16 (34.8%)

以下，TAK，EGW それぞれのクラスタの例を 10 づつ挙げる．リストの各エントリは，

(x) *description* (*type*)
representative_url

というフォーマットで記述されており，*type* はクラスタのタイプ（I〜III），*representative_url* はクラスタに属する URL の表記で最長一致する部分，*description* は *representative_url* で指し示されるページのタイトルである.

TAK

(a) NTT Basic Research Laboratories (I)
http://www.brl.ntt.co.jp/

(b) プログラミングシンポジウム ホームページ (I)
http://www.ipsj.or.jp/prosym/

(c) 平成１２年度「未踏ソフトウェア創造事業」の公募について－トップページ (I)
http://www.ipa.go.jp/NBP/12nendo/12mito/

(d) 平成１３年度「未踏ソフトウェア創造事業」の公募について－トップページ (I)
http://www.ipa.go.jp/NBP/13nendo/13mito/

(e) 電気通信大学トップページ The University of Electro-Communications Top Page (I)
http://www.uec.ac.jp/

(f) 森林総合研究所 四国支所 Forestry and Forest Products Research Institute (I)
http://www.ffpri-skk.affrc.go.jp/

(g) サイエンス社 & 新世社 Web サイトにようこそ (II)
http://www.saiensu.co.jp/

(h) 林業科学技術振興所ホームページ (II)
http://www.rinsin.or.jp/

(i) TAKANO Ryousei's Homepage (III)
http://www.os-omicron.org/ takano/

(j) 幻想・アングラ DADA 書房 (III)
http://www.m-net.ne.jp/ dada/

EGW

(a) 日テレ・ホームページ (I)
http://www.ntv.co.jp/

(b) 全国講演会講師派遣協会 (I)
http://www.eventk.co.jp/koushihaken/

(c) JB・NBC 公式サイト「NBCNEWS」 (I)
http://www.jbnbc.jp/

(d) スポニチアネックス 野球 トップ (II)
http://www.sponichi.co.jp/baseball/

(e) 激闘の記憶と栄光の記録 / 甲子園 / 高校野球 (II)
http://www.fanxfan.jp/bb/

(f) 花島書店 (II)
http://www.hanasima.gr.jp/

(g) ウェブの書斎 (III)
http://www.shosai.ne.jp/shincho/

(h) 今日は何の日？ (III)
http://www.fmfukuoka.co.jp/koyomi/

(i) 奈古中学校ホームページ (III)
http://www.nago-j.shinminato.ed.jp/

(j) ☆便利屋 寅さん トクトク情報☆ (III)
http://www.ucatv.ne.jp/ shuumei/

上記例のなかで，たとえばTAKの(c)と(d)などは，組織やプロジェクトを抽出するという意味では本来マージされるべきものである．逆に，過度にマージされるといったケースもあり，より適切なクラスタ生成のための工夫が今後の課題である．

5 活動の場による人物の特徴付け

このようにして得られたWeb上の活動の場は，当該人物の特徴を与える情報としての利用が期待される．本章では，その一つの試みとして，活動の場の相互関係を可視化するという手法を紹介する．後に示すように，本手法により，同姓同名の異なる人物をおおまかに弁別できる．

本手法は，以下に示す3つの手順からなる．

1. 活動の場の特徴づけ
2. 上記特徴にもとづく構造化
3. 上記構造の可視化

手順1.では，活動の場に対応するクラスタを一つの文書とみなし，そこから特徴的な語を抽出する．具体的には，クラスタを構成するWebページの内容を形態素解析して名詞と未定義語を取り出し，tf・idf法によって重みづけし，重みの大きい上位N語を特徴語として採用する．

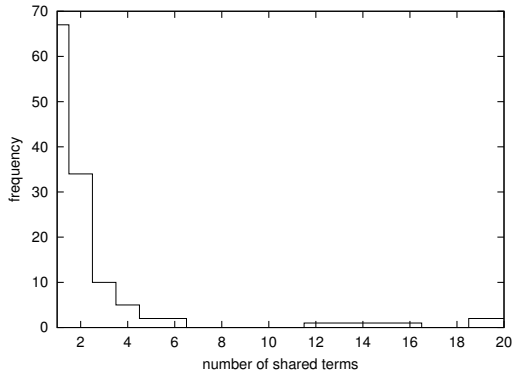

図3: クラスタ間で共有される特徴語の数（EGW）

図 3 は，EGW のクラスタのすべてのペアに対して二者間で共有されている特徴語の数を調べ集計した結果で，特徴語がクラスタ間でどのぐらい共有されているかを示したものである．ここで，$N = 20$ とした．図では省略されているが，共通の語が 1 つもない場合が 912 ケース（全体のおよそ 88%）あり，クラスタ間の関係は疎であることがわかる．

なお，図 1 で示したとおり，同じ文書頻度をもつ 2 語の間でサーバ頻度が異なることがある．偏在性を重視する場合には，文書頻度の代りにサーバ頻度を用いて重みを計算する方法も考えられる（これを tf・isf とする）．

手順 2. では，各クラスタと特徴語をノードとし，クラスタとそのそれぞれの特徴語に対して（無向）リンクを張ってグラフ構造を与え，手順 3. では，グラフ構造を 2 次元平面に配置するアルゴリズムを用いて可視化する．

図 4 は，語の重みの計算に tf・isf を用い，グラフ配置には Fruchterman-Reingold のアルゴリズム [4] を用いて，EGW の活動の場の相互関係を可視化したものである．ただし，図を 3 つの部分に分けている太い線は，後から人手で加えたものである．3 つに分けられたそれぞれの部分は，江川卓という同じ姓名をもつ元プロ野球選手，ロシア文学者，ブラックバス釣り競技のプロ選手の三氏に対応しており，Web から抽出した活動の場が，実世界の状況を正しく反映して適切に配置されていることがわかる．グラフ構造自体では，元プロ野球選手とロシア文学者の間に複数のつながりがあるなど，厳密な分離はできていない．しかし，特徴語を理解の助けとすることで，同姓同名人物の弁別に役立てることができると考えられる．

6 むすび

本論文では，実世界指向 Web マイニングという考え方にもとづき，実世界における人とその活動の場との関係を Web に当てはめ，与えられた名前をもつ人物の Web 上での活動の場を抽出する方法を示し，具体例を用いた実験によりその妥当性を検証した．

抽出精度の向上など，改善すべき点は少なからずあるものの，実世界の知識を Web に適用するという基本的なアプローチの妥当性，有効性は確認できた．

今後は，既存のマイニング手法に対する本アプローチの優位性の検証，実世界の複数知識の多面的な適用の検討をすすめていく予定である．

参考文献

[1] R. Kosala, H. Blockeel, “Web Mining Research: A Survey,” SIGKDD Explorations, Vol. 2, No. 1, pp. 1–15, 2000.

[2] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” Proc. of the 7th International World Wide Web Conference, 1998.

[3] 原田昌紀, 佐藤進也, 風間一洋, “Web 上のキーパーソンの発見と関係の可視化,” 情報処理学会研究報告, 2003-FI-71, pp. 17–24, 2003.

[4] T. M. J. Fruchterman, E. M. Reingold, “Graph Drawing by Force-directed Placement,” Software - Practice and Experience, Vol. 21, No. 11, pp. 1129–1164, 1991.

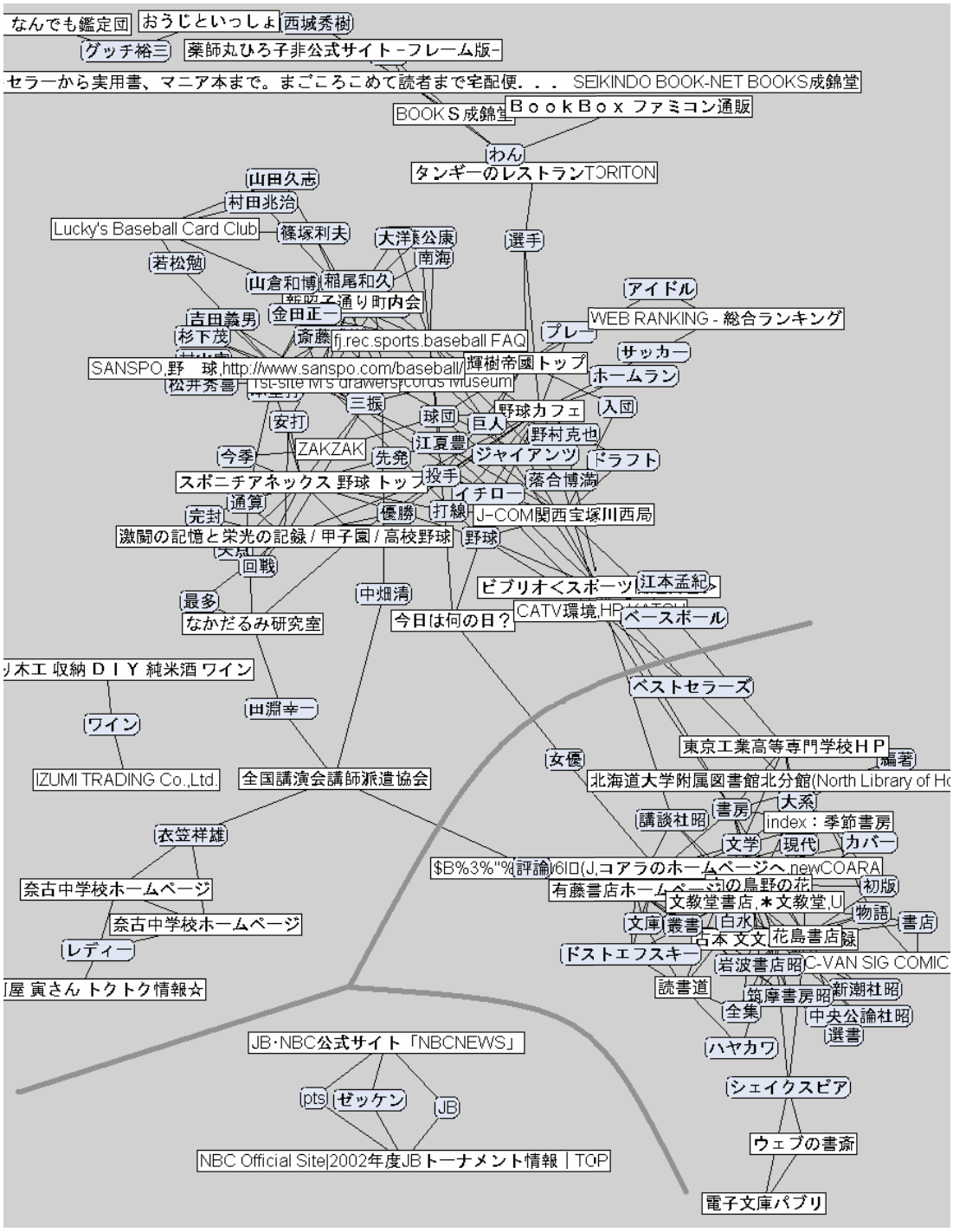

図 4: 語を介したクラスタ間の関係（EGW）