E-06

リンク構造分析と時空間詳細度制御に基づく イベント情報の一般性・専門性発見と提示

高橋 侑久† 大島 裕明† 小山 聡‡ 田中 克己†
Yuku Takahashi Hiroaki Ohshima Satoshi Oyama Katsumi Tanaka

1. はじめに

近年，インターネット上の情報の量はますます増加し，何かの情報を探すためにインターネットを用いる人も多く，検索エンジンは情報を探すために必要不可欠なツールとなっている．このような検索エンジンはユーザが入力したクエリに適合するページ群を取得した後，ユーザにとって“良い”と思われる順番に並べ替えて結果を提示する．Web ページの“良さ”はこれまで様々な観点から評価されてきたが，昨今最もよく用いられているアプローチはページ間のリンク関係に着目したものである．代表的なアルゴリズムである PageRank[3]や HITS[4]はリンクをリンク先ページへの支持と見なし，「多くのページからリンクされているページほど良い」という考えに基づいている．これらの手法は，ページの“良さ”を社会的な支持度という観点から評価したものであり，商用的にも一定の成功を収めている．しかし，全 Web ページを対象としてよく機能するこれらの手法が，ある特定のカテゴリに属するページのみを対象にしてページの“良さ”を測る場合に，必ずしも適しているとは限らない．

本研究では，イベント情報について考える．イベント情報とは属性として時間と場所を持つ情報のことで，例えば，「関ヶ原の戦い」は“1600 年”という時間的情報と“関ヶ原で起きた”という地理的情報を持つイベント情報である．本研究では，イベント情報の“良さ”を表す観点として“一般性”と“専門性”の 2 つについて考える．そして，イベント情報として特に日本史上の出来事を扱い，それに対応する Wikipedia 記事の他の記事とのリンク関係と時空間情報，特に時間的な情報から，一般性及び専門性の高いイベント情報の発見手法を提案する．

2. 日本史上のイベント情報

日本史上のイベント情報を表すために，Wikipedia の日本史に関連する記事を用いる．Wikipedia の日本史記事の中には，1 つの記事で複数の歴史イベントを含む場合があるが，あくまでも Wikipedia の 1 つの記事で 1 つの日本史上のイベントを表すものとして扱う．以後，簡単のために Wikipedia の記事とそれが表す日本史イベントのことをまとめて**イベント情報**と呼び，イベント情報のリンクとは対応する Wikipedia 記事に対するリンクのことを表すとする．

ここで，日本史イベントとして採用した Wikipedia 記事には，日本の歴史上の事件や出来事などの，いわゆるイベントを表す物だけでなく，歴史上の登場人物や，歴史上の物，概念なども含まれている．

†京都大学大学院 情報学研究科 社会情報学専攻
‡北海道大学大学院 情報科学研究科 複合情報学専攻

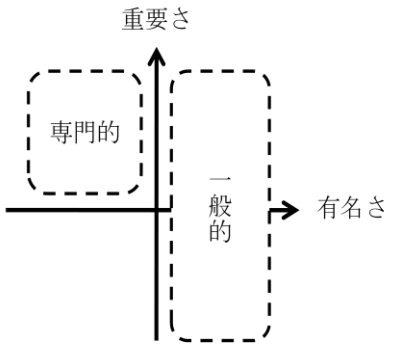

図1. イベント情報の有名さ及び重要さと一般性及び専門性の関係

本研究では，イベント情報の時間的情報を用いて一般性，専門性の評価を行う．そこで，日本史を 16 個の主要な時代に区分し，それぞれのイベント情報をこれらの内の 1 つに割り当てている．

これらイベント情報及びリンク関係と，16 個の時代区分についての詳しい説明は 4 節で述べる．

3. イベント情報の一般性と専門性

3.1 概要

本研究では，日本史上のイベント情報に対し，一般性と専門性という 2 つの観点に基づく評価を行う．本節ではどのようなイベント情報の一般性が高い・低い，また専門性が高い・低いと言えるのかについての本研究における態度を説明する．

歴史的な出来事について，“有名さ”や“重要さ”を考えることができる．ここで言う有名さとは，その出来事が広く一般的に知られている度合いのことを指し，重要さとはその出来事が日本の歴史に与えた影響の度合いを指すとする．例えば「関ヶ原の戦い」は広く知られているだけでなく，後に 300 年続く江戸時代の幕開けとなった出来事であるということから，有名かつ重要な出来事であると言える．本稿では，イベント情報に対して有名さと重要さは，それぞれ独立に考えることができるものとする．本研究では，このような有名さ及び重要さの観点を用いて，イベント情報の一般性及び専門性を次のように考える．（図 1）

- 重要であるかに依らず，有名なイベント情報は一般性が高い．
- 重要であるが，あまり知られていないようなイベント情報は専門性が高い．

3.2 イベント情報の一般性

本研究では，イベント情報の一般性を，その出来事の有名さ，すなわち広く知られている度合いとして考える．有名なイベント情報の特徴として，Web 上にその出来事についての記事が多い，Wikipedia 上で多くの記事からリンクされている，などが考えられる．

本研究では一般性を測るための指標としてイベント情報の Wikipedia 中での被リンク数，PageRank 及び HITS を候補として用いる．これらの手法は，リンク関係をリンク先への支持であると見なすものであり，有名な出来事であるほど，多くの被リンクを持っているという予測に基づいている．

3.3 イベント情報の専門性

3.3.1 概要

本研究では，重要だが有名でないようなイベント情報を専門性が高いとみなす．本節ではまず，どのようなイベント情報が重要と考えられるか説明し，次にそのような性質を満たすようなイベント情報があったときに，どのようにして重要なイベント情報を発見するかを説明する．そして，得られた重要なイベント情報の中から，専門的なイベント情報を選別する方法について説明する．

3.3.2 イベント情報の重要さ

歴史的な出来事の重要さを測る観点はさまざまなものが考えられるが，3 節で述べた様に，本研究ではその出来事が歴史に与えた影響の大きさであると考える．つまり，多くの出来事が生じる原因となった出来事や，多くの出来事の結果として生じた出来事が，歴史的に重要であると考えられる．

3.3.3 重要なページの発見手法

このような出来事間の因果関係を発見するために Wikipedia ページ間のリンクに着目する．あるページからあるページへのリンクが存在した場合，それら 2 つのページが表す歴史上の出来事の間に何かしらの関係性があると見なすことができる．従って，多くのページにリンクされているようなページは，歴史的に重要である傾向が強いと考えられるが，一般にイベント情報間のリンクには，因果関係だけでなく，様々な意味合いがあると考えられ，特に同時代のイベント情報間ではリンクが貼られることが多い傾向があるため，単純なリンク数でそのイベント情報の重要さを測ることはできない．

そこで，本研究では，時代を超えたリンク関係，特に時代を超えた被リンクに着目する．これは，同じ時代の出来事に対してリンクがあるのは珍しいことではないが，異なる時代区分の イベント情報にわざわざリンクがあるということは，それらの間には何かしらの重要な関係があるのではないか，という考えに基づいている．

ここで，あるイベント情報 e とは異なる時代区分に属するイベント情報集合をW_e^+，同じ時代区分に属するイベント情報集合をW_e^-とする．また，イベント情報集合 W の要素のうち，e へのリンクを持っているものの数を $LF(e, W)$ と表現すると，次の条件式を満たす場合にイベント情報 e へのリンクを持つ記事が他の時代区分に偏って存在していると言うことができる．

$$\frac{LF(e,W_e^+)}{LF(e,W_e^-)} > \frac{\sum_i LF(i,W_i^+)}{\sum_i LF(i,W_i^-)} \quad (1)$$

イベント情報 e がこの条件式を満たすとき，さらにその偏りが統計的に有意なものであるかを検証するため，「あるイベント情報がある時代区分に含まれるかどうかと，そのイベント情報に e へのリンクが出現するかどうかは互いに独立である」という帰無仮説に対してχ^2独立性検定を行う．この場合χ^2検定量は自由度 1 のχ^2分布に従う．全てのイベント情報についてχ^2検定量を計算し，有意水準αで棄却されたものを重要なイベント情報として検出する．本論文では$\alpha = 0.01$として実験を行った．

3.3.4 専門的なイベント情報の発見手法

上記の結果，得られた重要なイベント情報が，専門的なイベント情報の候補となる． 3.1 節で述べたように，本研究では専門的なイベント情報とは，重要であるが，有名でないようなイベント情報のことである．そこで，3.3 節で用いたイベント情報の一般性の尺度を用いて，候補語の中から一般性の低いイベント情報を抽出し，これを持って専門的なイベント情報とする．

4． 実験

4.1 Wikipedia 上の日本史記事について

ここでは，実験において用いた Wikipedia 記事とそれぞれの時代区分，そしてそれらの間のリンク構造について説明する．

Wikipedia の中から日本史の記事を取得するために，Wikipedia の「日本の各時代のカテゴリ」とそのサブカテゴリを取得し，それらのカテゴリに属するページを日本史上の記事として採用した．

日本史の時代区分としては「日本の歴史」の記事を参考に 16 個の時代に分けた．上記の Wikipedia 記事をこの 16 個の時代区分に割り当て，これを持ってイベント情報の時間詳細度情報として扱っている．記事の中には，複数の時代に該当する物もある．そのような場合，より適当であると考えられるものを 1 つ選択し対処している．どれか 1 つの時代に区分することが不適当と考えられるような場合，本研究においては，イベント情報ではないとした．例えば，「京都」という記事には様々な時代に関する歴史的記述があるが，その性質上どれか 1 つの時代に割り当てることが不適当であると判断し，イベント情報の中から取り除かれている．各時代区分と，Wikipedia 記事の数の関係は表 1 のようになった．

本研究では，リンク関係を用いてイベント情報の一般性及び専門性を考えるが，特に専門的なイベント情報を考える際には，リンク関係の中でも異なる時代区分に属する記事間のリンク関係に着目している．抽出された記事間のリンク数及びそれらの内で同時代の記事を結ぶリンクと，異なる時代を結ぶリンクの数を表 2 に示す．

4.2 一般性発見

被リンク数，PageRank，HITS を用いて一般性の評価を行う．ここで，被リンク数にはイベント情報の記事からのリンクのみを数え，Wikipedia 上のその他の記事からのリ

表1. 時代区分とイベント情報数

時代区分	イベント数
旧石器時代	16
縄文時代	125
弥生時代	73
古墳時代	237
飛鳥時代	328
奈良時代	160
平安時代	548
鎌倉時代	263
南北朝時代	66
室町時代	1082
安土桃山時代	210
江戸時代	5352
明治時代	255
大正時代	262
昭和時代	1322
平成時代	108
合計	10607

表2. イベント情報間のリンク

同じ時代同士のリンク	159284
違う時代へのリンク (未来→過去) (過去→未来)	41727 (22751) (18976)
合計	201011

ンクは数えていない．また，PageRank はイベント情報記事全体に対して行った値であり，HITS は各時代毎に記事を絞った上で，それらの記事間のリンク関係のみを用いて計算を行った．この時，明治時代に区分されるイベント情報について，各々の評価尺度で並べ替えた結果，表 3 のようになった．

他の 15 の各時代区分についても同様のランキングを作成することができる．それらを考慮した結果，主観的に PageRank による評価が最もイベント情報の一般性を評価する上で適していると判断した．まず，被リンク数に関しては下位に行くほど多くのページが同数の被リンクを持っており，それらのページ間に差がほとんどなくなってしまうことが問題点として挙げられる．3.4.3 節で述べたように，提案手法において，専門的なページを抽出する際に重要なイベント情報の中から一般性評価の低いものを選ぶ．その

表3. 一般性評価指標による明治時代のイベント情報の並べ替え結果上位10件

被リンク数	PageRank	Auth(HITS)	Hub(HITS)
明治	坂本龍馬	明治	桂太郎
廃藩置県	西郷隆盛	日露戦争	伊藤博文
戊辰戦争	廃藩置県	戊辰戦争	大隈重信
明治維新	西南戦争	伊藤博文	山縣有朋
版籍奉還	明治	日清戦争	田中義一
日露戦争	日清戦争	西南戦争	西郷従道
西郷隆盛	賞典禄	大隈重信	明治
大政奉還	戊辰戦争	山縣有朋	児玉源太郎
奥羽越列藩同盟	日露戦争	大久保利通	西園寺公望
坂本龍馬	江戸開城	井上馨	井上馨

ため，下位のイベント情報間の差が専門的なイベント情報を抽出する際に重要な要素となる．故に，被リンク数に基づく一般性評価は適していないと考えられる．また，HITS アルゴリズムから求まる Auth 値と Hub 値は時代区分毎に上位に来るイベント情報の性質が大きく異なるような印象を受けた．また，これらの値はカテゴリ依存で求まるものであるため，候補となるイベント情報の属する時代区分が異なると，一般性を用いて互いに比較することができなくなるという問題がある．以上のことから PageRank を用いて一般性を評価することが，今回考えた 4 つの指標の中では，最も優れていると考えられる．

4.3 専門性発見

時代区分を持つイベント情報 10607 個に対して式（1）と表 2 のリンク数から

$$\frac{LF(e,W_e^+)}{LF(e,W_e^-)} > \frac{41727}{159284} \qquad (1)'$$

を満たすイベント情報 e に対して，有意水準$\alpha = 0.01$でχ^2検定を行った．その結果「原爆ドーム」「大化の改新」「関ヶ原の戦い」などを含む 78 個のイベント情報が他の時代からのリンク数が有意に多いと検定された．これらが本手法によって発見された重要イベント情報であり，また専門的なイベント情報の候補である．

次に，求められた候補を PageRank 値の低い順に並べ替える．その結果を得られた専門的イベント情報の内，上位 10 件を表 4 に示す．

表4. 提案手法によって抽出された
専門性の高いイベント情報

専門的イベント情報
幕臣
土豪
埴輪
石山本願寺
大宝律令
一向一揆
幕藩体制
検非違使
彰義隊
執権

5. 関連研究

中谷ら[5]は Wikipedia 上の記事と，それにリンクしている記事集合のカテゴリ情報から，ある単語の専門性を評価する手法を提案している．これは，ある 1 つの分野の記事から有意に多くリンクされているような記事を，その分野の専門語や業界語として抽出する．本研究におけるイベント情報の重要性発見手法は，中谷らの手法の拡張となっている．Haveliwala[2]の提案した Topic-Sensitive PageRank は特定のトピックに偏った PageRank ベクトルを作り，それらを個人のプロファイルに応じて組み合わせることで，パーソナライズされた PageRank ベクトルを作るものである．歴史コンテンツというトピックに応じたランキングを考えるという点において，本研究と似ている点があり，今後参考にしたいと考えている．Gyongyi[1]らの TrustRank はスパムページのリンク特性を用いてスパムページを検出する手法である．本研究においても，同様にイベント情報間のリンク特性を上手く利用することで，専門的なイベント情報などを効率良く検出することが可能になるのではないかと考えている．

6. まとめ

本論文では Wikipedia の日本史コンテンツのリンク構造と，時間情報を用いた，イベント情報の一般性・専門性評価手法の提案を行った．提案手法では，まず，PageRank アルゴリズムを用いて，イベント情報の一般性の評価を行う．次に，各イベント情報が他の時代からリンクされている度合いを考えることで，他の時代に大きな影響を与えた重要なイベント情報を抽出する．抽出された重要イベントを一般性によって並べ替え，一般性は低いが重要なイベント情報を抽出することで，専門的なイベント情報を発見する．

実験では，一般性の評価手法として PageRank アルゴリズムが妥当であろうという結論を導いたが，定量的な評価法を考える必要がある．また，イベント情報の重要性，専門性の評価も行っていきたい．

今回は，日本史上のイベント情報の時間的情報を大きく 16 段階で捉え，時代を超えたリンク関係に着目することで歴史的に重要なイベント情報を発見し，そこから専門的なイベント情報を抽出するというアプローチを取った．時代を超えたリンクには過去から未来と未来から過去への 2 方向を考えることができるが，本稿ではそれらを区別せずに用いている．今後は，このようなリンクの時間的方向を用いていきたいと考えている．また，イベント情報は地理的な情報も含まれており，本稿では考慮されていない．今後はイベント情報の地理的情報も交えた手法を考えていきたい．

謝辞 本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，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），および，文部科学省科学研究費補助金若手研究（B）「オンデマンド利用を目的とする Web からの知識発見に関する研究」（研究代表者：大島裕明，課題番号：21700105），および，独立行政法人情報通信研究機構の高度通信・放送研究開発委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発 課題ア Web コンテンツ分析技術」（研究代表者：田中克己），および，文部科学省科学研究費補助金若手研究（B）「時間変化するオブジェクト情報のＷｅｂからの収集と管理方式の研究」（研究代表者：小山聡，課題番号：21700106）によるものです．ここに記して謝意を表します．

文献

[1] Gyongyi, Z., Garcia-Moline, H., Pedersen, J.: “ Combatting web spam with TrustRank ”, Proc. 30th InternationalConference on Very Large Data Bases (VLDB) Toronto,Canada, Augast 2004.

[2] Taher H. Haveliwala.:“Topic-sensitive PageRank”, In TheEleventh International World Wide Web Conference.

[3] S.Brin and L.Page.:“ The anatomy of large-scale hypertextual web search engine. ”, Proc. of Seventh InternationalWorld Wide Web Conference.

[4] Kleinberg, J.M.: “ Authoritative sources in a hyperlinkedenvironment. ”, Journal of the ACM 46 pp.604-632(1999).

[5] 中谷誠，アダムヤトフト，田中克己：“理解容易性を考慮した用語説明のランキング手法”, WebDB Forum 2009