情報処理学会論文誌 Vol. 52 No. 12 3542–3557 (Dec. 2011)

インパクトを考慮した歴史エンティティの重要度計算手法

高橋侑久[†1] 大島裕明[†1] 山本光穂[†2]
岩崎弘利[†2] 小山聡[†3] 田中克己[†1]

本稿では，Web 百科事典 Wikipedia の中から歴史的観点から重要な項目を発見する手法を提案する．本研究では，歴史上の出来事や人物などを歴史エンティティと呼び，歴史エンティティの重要度を他の歴史エンティティに与えた影響の大きさと考える．我々の提案手法では，まず歴史エンティティの時間的なインパクトを計算する．Wikipedia 項目間のリンク関係が，歴史エンティティ間の影響関係を表すものと見なし，反復計算アルゴリズムを用いて，歴史エンティティのインパクトを様々な時間に対し計算する．そして，各エンティティが持つインパクトが広く大きいほど重要であると考え，歴史エンティティの重要度計算を行う．空間インパクトに着目した同様の手法に対し，実現するうえでの課題を考察する．提案手法といくつかのベースライン手法に対し，Wikipedia データを用いた評価実験を行う．

Evaluating Significance of Historical Entities Based on Impacts Calculation

Yuku Takahashi,[†1] Hiroaki Ohshima,[†1]
Mitsuo Yamamoto,[†2] Hirotoshi Iwasaki,[†2]
Satoshi Oyama[†3] and Katsumi Tanaka[†1]

We propose a method to find a historically significant article from Wikipedia. We treat an article as a historical entity and evaluate the significance of historical entities (people, events, and so on.). Here, the significance of a historical entity means how it affected other historical entities. Our proposed method first calculates the temporal impact of historical entities. The impact of a historical entity varies according to time. We assume that a Wikipedia link between historical entities represents an impact propagation. That is, when an entity has a link to another entity, we regard the former is influenced by the latter. Historical entities in Wikipedia usually have the date and location of their occurrence. Our proposed iteration algorithm propagates such initial tempo-spacial information through links in the similar manner as PageRank, so the tempo-spacial impact scores of all the historical entities can be calculated. We assume that a historical entity is significant if it influences many other entities that are far from it temporally or geographically. We demonstrate a prototype system and show the results of experiments that prove the effectiveness of our method.

1. はじめに

“人類史上最も重要な出来事は何か”，“*Albert Einstein* と *Isaac Newton* では，どちらのほうが偉大か” このような歴史にまつわる疑問はきわめて興味深いが，問題の内容が相対的であり，回答者によって歴史の見方が異なるため，答えを出すことは難しい[⋆1]．

Wikipedia は，世界中の参加者によって構築された，多言語の大規模 Web 百科事典である．Wikipedia の項目は幅広い分野を網羅しており，その項目数はすでに 350 万（2011 年 1 月英語版）を超えている．Wikipedia は幅広い分野の網羅性以外にも注目すべき特徴をいくつか持つ．そのような特徴の 1 つに，項目間の密なリンク構造があげられる．このような特徴を用いることで，Wikipedia は知識抽出のコーパスとして利用できることが示されている[3),4)]．

本研究では，Wikipedia を歴史的な出来事や人物などに関する知識抽出のコーパスとして利用し，歴史的な観点から重要な出来事や人物に関する Wikipedia 項目を発見することを目的とする．このような歴史上の出来事や人物などを表す概念を，本研究は**歴史エンティティ**と呼ぶ．ここに，歴史エンティティとは，人物，組織，出来事，場所，建造物などの歴史を形作る要素を表す概念である．この問題に取り組むにあたり，我々はまず歴史エンティティを特定の観点から評価したときの重要性を評価する手法を提案する．このような観点を**アスペクト**，重要度を**インパクト**と呼ぶ．本研究では，アスペクトとして特に時間と空間に基づくものを取り上げる．

†1 京都大学大学院情報学研究科社会情報学専攻
Graduate School of Infomatics, Kyoto University
†2 株式会社デンソーアイティーラボラトリ
Denso IT Laboratory
†3 北海道大学大学院情報科学研究科複合情報学専攻
Graduate School of Information Science and Technology, Hokkaido University

⋆1 たとえば，2005 年 12 月に英国王立協会が数千人の科学者を対象に行ったアンケートでは，*Newton* の方が *Einstein* よりも人類の発展に貢献しているという結果が出ている．

歴史エンティティのインパクトは時と場所によって異なる．基本的に，歴史エンティティのインパクトは，それ自身が存在した時間や空間から遠ざかるほど弱くなっていくと考えられる．たとえば，*Napoleon* は彼が生きていた時代において，ヨーロッパで非常に強いインパクトを持っていた．しかしながら，その当時の日本ではインパクトはほとんどなく，現代のヨーロッパでも影響力は小さくなっている．提案手法では Wikipedia 項目の時間情報および，項目の間の参照関係に着目する．一部の Wikipedia 項目は時間に関わる情報を含んでおり，このような情報を Wikipedia のリンク構造を用いて伝播させることでインパクト計算を行う．ある歴史エンティティが他の歴史エンティティへのリンクを持っている場合，前者は後者から影響を受けたと見なす．提案手法では，PageRank に似た再帰的なアルゴリズムを用いることで，すべての歴史エンティティに対して，時間的インパクトを計算することが可能となる．各歴史エンティティの時間的インパクトは “*Newton* in 1900” のように表すことができる．このようなインパクトの時間に沿った変遷はグラフを用いることで可視化を行うことが可能である．また，同様に空間インパクトを求め，地図を用いた可視化を行うことができる．

得られたインパクトを用いて各歴史エンティティの歴史的観点からの重要度を計算する．本研究では，このような歴史エンティティの重要度を**シグニフィカンス**と呼ぶ．歴史エンティティが与えた影響の範囲が時間的に広いほど，その歴史エンティティのシグニフィカンスは大きくなる，という仮説に基づくシグニフィカンス評価手法を提案する．同様のアプローチを影響の空間的広がりに対して適用することが考えられるが，直線で表される時間軸に対し，平面または球面で表される空間を用いる場合，いくつかの克服すべき課題があることを見る．

本稿では，複数の歴史教科書に共通して出現する単語を，歴史的に重要な用語であると見なすことで正解セットを作成し，Wikipedia の実データを用いて提案手法といくつかのベースライン手法との比較評価実験を行う．

2. 関 連 研 究

本研究では，Wikipedia のリンク構造を用いて，歴史エンティティの重要度の算出を行う．Web のリンク構造を用いて Web ページの重要度を計算する従来研究に，Brin らの PageRank [1] がある．PageRank ではハイパーリンクによる参照を Web ページの支持と見なし，多く参照されている Web ページほど有用であり，そのような有用なページに参照されている Web ページはさらに有用である，という再帰的な考えによって Web ページの重要度を計算する．

リンク構造に対して一意に定まる PageRank に，偏りを持たせた biased PageRank に関する先攻研究が多数なされている．Gyöngyi らの TrustRank [14] は biased PageRank を Web ページのスパム発見に用いている．これは，人手で判定された非スパムのページからハイパーリンクによって，トラスト値と呼ばれる非スパムらしさを伝播させることで，スパムページを Web のリンク構造を用いて発見する手法である．Haveliwala の topic-sensitive PageRank [5] は複数のトピックに偏った biased PageRank を用いて Web ページの重要度を計算する．topic-sensitive PageRank では Open Directory Project（ODP）[⋆1] に現れる 16 のトピックを用いているが，個々のトピックの間に連続性がないことが本研究との違いである．本研究では時間や空間のように，互いに連続性や距離を与えることができる属性に対し，biased PageRank 計算を行う．

著者らの研究グループでは，“Virtual History Tour” と呼ばれるアプリケーションの開発を行ってきた [11]．本アプリケーションは，歴史上の出来事を時間軸に沿って，地図インタフェース上で再現することで，ユーザに仮想的な時間旅行を経験してもらうことを目的とする．たとえば，坂本龍馬の一生で起こった出来事を順番に表示することで，ユーザは坂本龍馬の一生を仮想的に体験し学習することができる．対象への興味の有無が学習効率に大きな影響を与えることから，このような学習型アプリケーションでは，興味を持たないユーザに対しいかに興味を喚起し学習意欲を持たせるかが重要となる．そのうえで，単に重要な出来事の羅列を提示するのではなく，時代背景や出来事の間の関連性に着目し，歴史の流れを提示することが興味喚起につながるのではないかと考えられる．提案手法は影響を与えた時代の広がりに着目し，歴史エンティティの重要度評価を行う．そのため，一過性の出来事ではなく，前後の時代とのつながりの中で重要な出来事を発見することができると期待されるため，このようなアプリケーションにおいて提案手法は有用であると考えられる．

歴史を考えるうえで，いつ頃起こったか，という時間に関する情報は重要であるが，どこで起こったか，という空間に関する情報も重要である．本稿では，時間に着目した手法を拡張し，歴史エンティティの影響の空間的広がりを扱うことを試みる．時間に着目した手法では，数年単位で時間範囲を区切り，各区分ごとに重要度計算を行う．一方，空間を用いる手法では，ジオハッシュアルゴリズムを用いて空間を分割し，それぞれ重要度計算を行う．ジオハッシュを用いた研究としては，Martins らによるジオハッシュを用いた XSLT/XQuery

⋆1 Open Directory Project: http://www.dmoz.org/

エンジンに地理情報を処理する機能を実現する研究がある[2]．ジオハッシュによる区分けおよび，計算結果の可視化を Web 上の地図サービスである Google Maps[⋆1]を用いて行った．その他の，地理空間を用いた可視化環境としては，Google Earth[⋆2]や Stoev らが提案した物が存在する[12]．これらは 3 次元空間上でのオブジェクトの表現などを行うことができるが，ジオハッシュが地球表面を平面としてとらえ分割するアルゴリズムであるため，Google Maps を用いた平面上での可視化を行った．

Larson によって定義された[8),9)] 地理情報検索分野にも多数の関連研究が存在する．Strötgen らは単一の文書から時空間情報を抽出する手法を提案している[13]．また，岡本らは特に Wikipedia 文書を対象とした時空間情報を抽出する手法を提案している[15]．これらの研究は，各文書に出現する単語から，その文書に特に関係性の高い時空間情報の発見を試みるが，本研究では参照関係に基づいて時間インパクトおよび空間インパクトと呼ばれる数値の計算を行う．

3. 歴史エンティティ

3.1 定　　義

歴史は，歴史上の人物や建造物，出来事など様々な要素が集まって構成されていると見なすことができる．本研究では，このような歴史を構成する 1 つ 1 つの要素を “歴史エンティティ” と呼び，次のように定義する．

- 存在していて，他と区別できるものはエンティティである．
- エンティティの中で，時間的な情報を持つものは歴史エンティティである．
- 今は消滅して存在しないエンティティも，歴史エンティティに含まれる．

このような定義から，すべての Wikipedia 項目が歴史エンティティになるのではなく，時間情報を持つものが歴史エンティティとなる．たとえば，*Einstein* は “1879 年生” や “1955 年没” のような時間情報を持つため，歴史エンティティの 1 つである．

3.2 歴史エンティティの重要度

歴史エンティティの重要度と一言でいっても，着目する観点によって様々な種類が考えられる．たとえば，時間や空間に着目して歴史エンティティの重要度を評価することができる．“1900 年における *Einstein* の重要度” や “1940 年における *Einstein* の重要度” は時間に着目した評価であり，“ヨーロッパにおける *Einstein* の重要度” は空間に注目した歴史エンティティの重要度評価である．また，歴史的観点から見た重要度も考えられる．これは，その歴史エンティティの総合的な評価と見なすことができる．

本研究では，このように着目している観点によって定まる歴史エンティティの重要度を，歴史エンティティの**インパクト**と呼び，このとき注目している観点をインパクトの**アスペクト**と呼ぶ．また，特に歴史的観点から見たインパクトは，前述のように，対象となる歴史エンティティの総合的な評価を表すものであり，区別のために，歴史エンティティの**シグニフィカンス**と呼ぶ．

本研究の目標は，歴史エンティティのシグニフィカンスを客観的に求めることである．

3.3 Wikipedia 項目

Wikipedia はデータベースの情報を公開している[⋆3]．これらのデータには記事本文だけでなく，項目間のリンク関係，カテゴリ情報などが含まれている．

本研究では，Wikipedia 項目を用いて歴史エンティティの重要度計算を行う．3.1 節で定めたように，歴史エンティティは時間情報を持つ．そのため，Wikipedia 項目の中から，時間情報を持つものを抽出し，それをもって歴史エンティティを表す．

時間情報を持つ Wikipedia 項目を抽出する方法はいくつか考えられる．1 つは，各項目の本文中に，時間と関わる表記のあるものをとってくる方法である．たとえば，*Einstein* の項目には次のような時間表現が含まれるため，歴史エンティティとして抽出することが可能である．

> **アルベルト・アインシュタイン**（*Albert Einstein*，1879 年 3 月 14 日–1955 年 4 月 18 日）は，ドイツ生まれのユダヤ人理論物理学者．

その他の手法としては，Wikipedia カテゴリを用いて抽出する方法が考えられる．たとえば，*Einstein* の項目は “1879 年生” と “1955 年没” というカテゴリに属し，これらのカテゴリは時間に関わるものであるため，*Einstein* を歴史エンティティとして抽出することができる．

4 章や 5 章で述べるように，提案手法では，各歴史エンティティに対して有効時間や，歴史エンティティ間の時間軸上での距離を用いる．Wikipedia 項目には，項目の内容とは直接かかわらない背景説明などが含まれる場合がある．そのため，一般的に，本文中の記述を用いて時間情報を抽出する手法は，カテゴリを用いる手法と比べてノイズが含まれやすいと

⋆1 Google Maps: http://maps.google.com/
⋆2 Google Earth: http://earth.google.com/
⋆3 http://download.wikipedia.org/enwiki/

考えられる．そのため，本研究では，カテゴリを用いた手法を用いて歴史エンティティとなる Wikipedia 項目を抽出し，この時間情報を用いて，歴史エンティティの有効時間や距離を求める．

4. インパクト

3.2 節で述べたように，あるアスペクトから見たときの歴史エンティティの重要度を，インパクトと呼ぶ．本章ではまず，4.1 節で歴史エンティティのインパクトをどのように定めるか，また，Wikipedia を用いてどのように計算するか述べる．本研究では，アスペクトとして，特に時間と空間について考える．4.2 節で時間，4.3 節で空間をアスペクトとしたときの，インパクト計算について説明する．また，具体的な計算結果示し，考察を行う．

4.1 インパクト計算

あるエンティティ e のアスペクト a におけるインパクトを $i_e(a)$ と表すとする．たとえば，“1900 年における *Einstein* のインパクト” は $i_{Einstein}(1900)$ と表される．

ここで，$i_{Einstein}(1900)$ について考える．仮に *Einstein* が “1900 年” において大きなインパクトを持っていたとしたら，*Einstein* は 1900 年当時に生きていた多くの人物や出来事に大きな影響を与えたといえるのではないだろうか．また，“1900 年” において大きなインパクトを持つ *Einstein* に影響を与えたエンティティもまた，“1900 年” において大きなインパクトを持っていたと考えられるのではないだろうか．このような考えから，$i_e(a)$ は，a において**有効**であるエンティティ集合に対して e が与えた影響の大きさと，e が影響を与えた相手の e' のインパクト $i_{e'}(a)$ を用いて表すことができるのではないかと考えた．

歴史エンティティ間の影響関係を対応する Wikipedia 項目間の参照関係を用いて考える．ここで，仮に項目 v が，他の項目 u から参照されているとする．Wikipedia のリンク構造では，項目間のアンカテキストはつねに項目と関連を持っている．リンクは Wikipedia のユーザグループによって管理されており，Wikipedia のポリシにより，たとえ本文に他の記事の名前が出てきたとしても，その項目が文章の内容と関連がない場合，リンクは作成されない．したがって，u から v へのリンクは，u から v への強い関連が存在するととらえることができる．本研究では，このようなリンクがあった場合，v が u に影響を与えたものと見なす．たとえば，*Einstein* から *Newton* へのリンクがあった場合，*Newton* は *Einstein* に影響を与えたと見なす．

以上から，$i_{Einstein}(1900)$ は，*Einstein* が Wikipedia リンク構造上で，“1900 年” において有効であるエンティティ集合からリンクされている度合いを用いて表す．提案手法では $i_e(a)$ を次式で求める．ここで，C_a は a において有効なエンティティの集合であり，F_e は e が参照するエンティティ集合，B_e は e を参照するエンティティ集合をそれぞれ表す．

$$i_e(a) = \alpha \sum_{e' \in B_e} \frac{i_{e'}(a)}{|F_{e'}|} + (1-\alpha) b_e(a) \tag{1}$$

$$b_e(a) = \begin{cases} 0 & e \notin C_a \\ \frac{1}{|C_a|} & e \in C_a \end{cases} \tag{2}$$

これは biased PageRank アルゴリズムと等しい．biased PageRank は，与えられた初期ノード集合に対して個々のノードがグラフ上で離れている度合いが計算されると解釈することができる．提案手法では，C_a が初期集合に対応している．たとえば，C_{1900} は，日本語版 Wikipedia であれば “1900 年生” や “1900 年の音楽”，英語版 Wikipedia であれば “1900 births” や “1900 in economics” などのカテゴリに属する項目の集合となる．この結果，式 (1)，(2) は，a において有効であるエンティティ集合から個々のエンティティがリンクされる度合いを計算することを意味する．リンクを影響関係ととらえる立場から解釈すると，これはすなわち，個々のエンティティが初期集合に含まれるエンティティに与えた影響の大きさを表す．また，biased PageRank は，大きな値を持つノードにリンクされるほど，大きな値をとる，という再帰的な性質を持つため，大きなインパクトを持つエンティティに影響を与えたエンティティも大きなインパクトを持つ，という考えを反映したものとなっている．

4.1.1 議　　論

Wikipedia 項目間の参照関係を影響関係の表れととらえるには，いくつかの議論すべき問題がある．

まず，過去の項目から未来の項目へのリンク，たとえば *Newton* から *Einstein* へのリンクについて考える．前述の規則に従うと，このリンクは *Einstein* は *Newton* に影響を与えたことを意味するが，明らかに *Newton* は *Einstein* のことを知らないため，このような影響関係は実際には存在しえない．そもそも，このようなリンクが存在するのは，Wikipedia が歴史を俯瞰する視点から記述されるためである．そのため，*Newton* の項目中で，たとえば物理学というコンテキストで，*Einstein* について言及される場合がある．過去から未来へのリンクを排除することも 1 つの手段であるが，本研究では，Wikipedia は歴史を俯瞰する視点から記述されるという特性から，このようなリンクも他と同様に扱うこととした．

次に，リンクの内容に関する問題がある．上で述べたように，Wikipedia には内容と関係ない場合はリンクにしない，というポリシがあるが，必ずしも守られているわけではな

い．また，内容と関係があったとしても，影響関係と見なせない場合も存在する．たとえば，2011 年 7 月 7 日時点において，日本語版 Wikipedia の *Einstein* の項目には，次のような記述がある．

> 音楽学者 で モーツァルト 研究者の アルフレート・アインシュタイン は従弟とされる場合があるが、異説もある。

下線はリンクを表す．明らかに，モーツァルトが *Einstein* に影響を与えたと，この文書から読み取るのは誤りである．しかしながら，リンクが項目間の影響関係を意味するかどうか判定することは容易ではない．そのため，本稿の範囲では，簡単のためにエンティティ間の参照関係はすべて影響関係を表すものとして扱う．リンクの意味による項目間の関係性の分類は，今後の課題である．

4.2　時間インパクト

各歴史エンティティが，そのエンティティが生きていた期間や，起こった時間を表す，有効時間という概念を持っているとする．4.1 節で例示したように，アスペクトとして時間に関するものを用いた場合，式 (2) の C_a は，時間 a を有効時間に含むエンティティの集合を用いて表すことができる．そこで，各歴史エンティティに対し，属する時間カテゴリの中で最も古いものから，最も新しいものまでの間を，そのエンティティの有効時間として用いる．たとえば，日本語版 Wikipedia において “1879 年生” と “1955 年没” というカテゴリに属するため，*Einstein* の有効時間は 1879 年から 1955 年となる．同様に英語版 Wikipedia では “1879 births” と “1955 deaths” に属している．別のいい方をすれば，*Einstein* は “1879 年” から “1955 年” において C_a に含まれることを意味する．このように，有効時間を定めることで任意の時間 a に対して C_a を作ることが可能となり，これを用いて biased PageRank を計算することで，すべての歴史エンティティの $i(a)$ を計算することができる．

ここで，Wikipedia のデータを使って実際に時間インパクトを計算した結果をいくつか示す．

図 1 は英語版 Wikipedia を用いて計算した場合の *Leonardo da Vinci*，*Newton*，*Einstein* の西暦 1000 年以降におけるインパクトの遷移を，グラフ上にマッピングして，可視化したものである．*Leonardo da Vinci* は 15 世紀，*Newton* は 17 世紀，*Einstein* は 20 世紀の人物であることから，それぞれの有効時間近辺で大きなインパクトをとっていることが分かる．日本語版 Wikipedia でも同様に 3 者のインパクトを求めたが，ほぼ同様の結果が得られた．このように，日米 Wikipedia ではおおむね同じような結果が得られる場合が多かった．

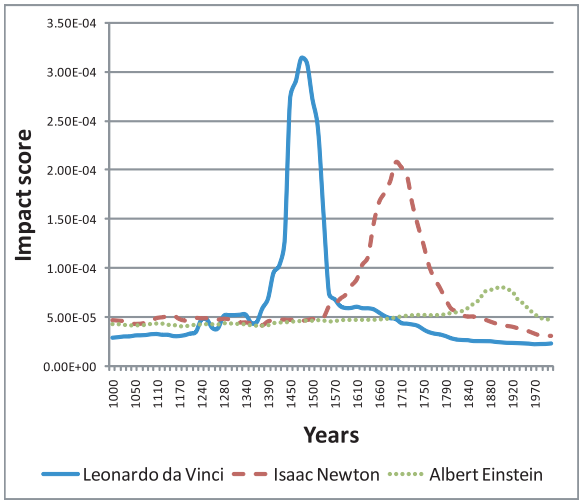


図 1　西暦 1000 年以降における *Leonardo da Vinci*，*Newton*，*Einstein* のインパクトの遷移
Fig. 1　The evolution of the impact of *Leonardo da Vinci*, *Isaac Newton* and *Albert Einstein* after year 1000.

日米で結果が異なった事例として，*Oda Nobunaga*，*Toyotomi Hideyoshi*，*Tokugawa Ieyasu* の場合を取り上げる（**図 2**）．日本語版を使った場合，江戸時代（1603 年–1868 年）において *Tokugawa Ieyasu* の方が明らかに大きい値をとっている．これは，*Tokugawa Ieyasu* が江戸幕府を開いた人物であるという事実から，妥当な結果であると考えられる．一方，英語版の結果では，日本語ほどの明確な違いは得られなかった．このような違いが生じたのは，江戸時代の歴史エンティティの中で，日本語 Wikipedia にのみ存在するものが多く存在し，それと同時にこれらの歴史エンティティの多くが *Tokugawa Ieyasu* を直接的・間接的に参照しているためである．

英語版 Wikipedia は他の言語に比べて最も記事の数が多く，内容も充実しているが，特定の国の歴史にまつわる話題を扱う場合は，その国の言語の Wikipedia を用いた方が，より細かなインパクトの違いを発見することが可能になるのではないかと考えられる．

4.3　空間インパクト

ある時代における個々のエンティティのインパクトの大きさを，その時代で有効であったエンティティからのリンク度合いを用いて計算したように，ある空間におけるインパクトの

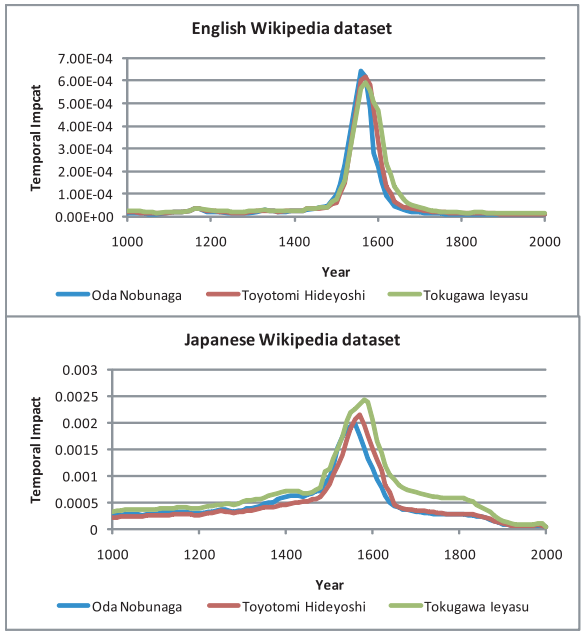


図 2　西暦 1000 年以降における *Oda Nobunaga*，*Toyotomi Hideyoshi*，*Tokugawa Ieyasu* のインパクトの遷移．上は英語版 Wikipedia を用いた場合，下は日本語版 Wikipedia を用いた場合の結果を表す
Fig. 2　The evolution of the impact of *Oda Nobunaga*, *Toyotomi Hideyoshi* and *Tokugawa Ieyasu* after year 1000.

大きさを，その空間に存在したエンティティからのリンク度合いを用いて計算する．エンティティが空間内に位置するかどうかを判定するために，本研究ではジオハッシュアルゴリズムを用いた．

ジオハッシュは，Gustavo Niemeyer が geohash.org[⋆1] という Web サービスを作成中に考案した緯度経度に基づくジオコーディング手法の 1 つで，パブリックドメインとして公開されている．ジオハッシュは階層的な空間データ構造であり，空間を格子状に分割する特徴を持つ．階層ごとに空間を 32 分割にするため，ジオハッシュは 32 進数文字列として表され，文字列の長さが階層の深さを表す．

⋆1 http://geohash.org

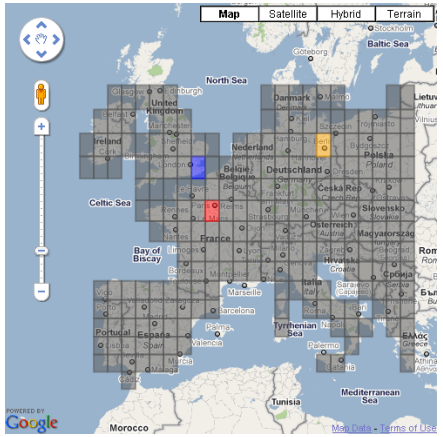


図 3　ジオハッシュを用いた，ヨーロッパ地理空間の分割
Fig. 3　Geo-spatial divide of Europe using geo-hash algorithm.

まず，世界地図を 32 分割することで第 1 階層の格子を得る．これらはそれぞれ 1 文字の 32 進数文字で表される．次に，第 1 階層の格子を分割することで 32^2 個の第 2 階層の格子を得る．これらは 2 文字の 32 進数文字で表され，先頭の文字が親となる第 1 階層の格子を表す．すなわち，先頭文字が共通の格子どうしは，共通の第 1 階層の格子を分割することで得られることを意味する．同様の分割操作を繰り返すことで，ジオハッシュは非常に細かく空間を分割することができ，また，個々のジオハッシュに対し先頭からの文字列を見ることで，与えられたジオハッシュを包含する上位階層の格子を得ることができる．このようなジオハッシュの特徴を用いることで，特定の空間に対し，与えられたジオハッシュが内部に存在するかどうかを容易に判定することができる．たとえば，“u09” という格子が表す空間の内部に個々のジオハッシュが位置するかどうかは，それが “u09” で始まるかどうかで判断することができる．

本研究では，まずヨーロッパを**図 3** のように分割した．これらはすべて第 3 階層の格子であり，たとえば，パリを含む赤線で囲まれた格子は “u09”，ロンドンを含む青枠が “u12”，ベルリンを含む橙枠が “u33” とそれぞれジオハッシュで表される．これら 1 つ 1 つの格子が，1 つのアスペクト a であるとする．そして，格子内に位置する項目の集合 C_a からのリンク度合いによって，a における空間インパクトを求める．

個々のエンティティに対し有効空間を定め，それをジオハッシュを用いて表すことで，C_a

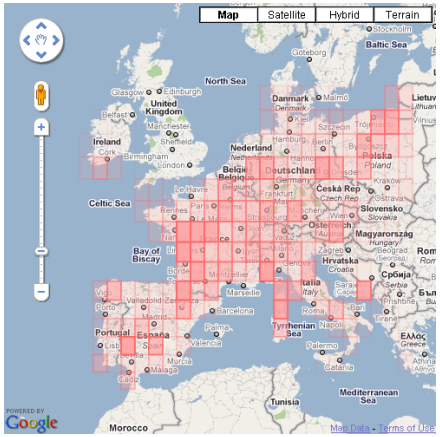


図 4　*Napoleon I*
Fig. 4　*Napoleon I.*

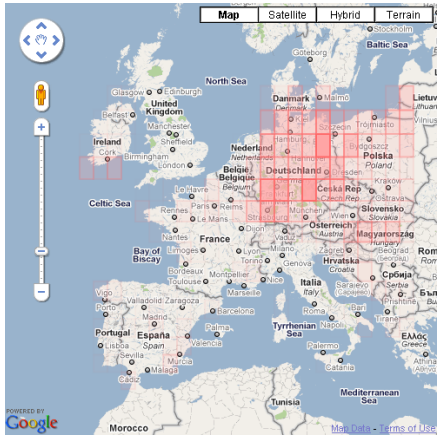


図 5　*Berlin Wall*
Fig. 5　*Berlin Wall.*

を求める．歴史エンティティの有効空間を定める 1 つの方法は，対応する Wikipedia 項目の本文中に出現する地名を抽出して用いることである．この方法にはいくつかの問題がある．まず，項目とあまり関係のない地名が本文に含まれている可能性がある．その土地で長年活動したのか，それとも旅先で一時的に訪れただけなのか，Wikipedia の文脈から自動的に判定することは容易ではない．また，得られた地名から緯度経度情報を求めるには，外部サービスが提供する API を利用することが一般的であるが，6 章で説明するように，本研究で扱う Wikipedia 項目数は数百万単位である．そのため，これらすべてを処理するには膨大な時間を必要とする．

本研究では，Wikipedia 項目に直接付与されている緯度経度情報を用いることとした．たとえば，*Eiffel Tower* は Wikipedia のテキストに {`coord|48.8583|N|2.2945|E`} という文字列を含んでいる．これは *Eiffel Tower* が北緯 48 度 86 分東経 2 度 29 分に位置することを表しており，この緯度経度をジオハッシュに変換すると “u09tunqu1” となる．このことから，*Eiffel Tower* は “u09” 内に位置することが分かる．これはすなわち，アスペクト “u09” に対し，*Eiffel Tower* が C_{u09} に含まれることを意味する．

このようにして求められた C_a を用いて空間インパクトを計算した．以下に，*Napoleon I* と *Berlin Wall* に対する英語版 Wikipedia における計算結果を示す．

図 4 は *Napoleon I* に対する地理的影響力の分布を可視化したものである．皇帝に即位したフランスを中心に分布していることが分かる．またロシアとフランスをつなぐ方向に強い値を示しているのは，ナポレオンがロシア遠征の際，ネマン川を越えてモスクワに向かったことの影響だと考えられる．

図 5 は *Berlin Wall* の結果を示している．ドイツを中心に強い影響が現れており，西ヨーロッパと比較して，東ヨーロッパの方が大きな値を示している．

4.3.1　議　　論

歴史エンティティに対し，空間インパクトを求めるうえで問題になる点として，計算量が膨大になる点があげられる．今回のヨーロッパ地方を対象とした計算では，第 3 階層の格子を用いたが，この階層の格子は全部で 32^3 個存在する．このうち海洋上に位置するため，事実上無視できる格子も多いが，格子を小さくしていくと，格子数が指数爆発を起こしてしまうため，計算を行うことが困難となる．一方，時間軸では分割数の増加は大きな問題とはならない．

5.　シグニフィカンス

本章では，歴史エンティティの歴史的な重要度を意味する，歴史エンティティのシグニフィカンスを計算する手法について説明する．

まず，5.1 節で，歴史的に重要な歴史エンティティが持つ性質について仮説を提案し，こ

の仮説に基づき，歴史エンティティのシグニフィカンスを時間インパクトを用いて計算する手法を 5.2 節で提案する．提案手法では，歴史エンティティの時間インパクトの広がりを用いてシグニフィカンスを評価することを試みる．同様に空間インパクトを用いるアプローチについても 5.3 節で議論を行う．

5.4 節では，歴史エンティティのシグニフィカンスを求めるための，いくつかのベースライン手法を提案する．ベースライン手法は大きく分けて，被リンク数に基づく手法と PageRank アルゴリズムに基づく手法に分けられる．被リンク数は，各歴史エンティティが影響を与えた相手の数をシグニフィカンスとして用いる手法である．4 章で，アスペクトに対して有効なエンティティ集合からの参照の度合いを用いて，インパクトを求める手法を提案したが，歴史エンティティはすべて “歴史的” というアスペクトに対して有効であるため，これは初期集合 C_a にすべてのエンティティを指定することに等しい．このように，すべてのノードを初期集合として biased PageRank に与えた場合，PageRank と呼ばれる．PageRank は本来，Web ページの重要度評価を行うアルゴリズムである．それぞれの単純な手法に，5.1 節の仮説に従って変更したものを加え，計 4 種類のベースライン手法を提案する．

5.1　シグニフィカンス計算の仮説

ニュートンがホックとアインシュタインに影響を与えたとする．このとき，どちらの影響関係の方がよりアインシュタインの重要性を表しているいえるだろうか．1 つの考え方は，ニュートンとホックは同時代の人物だが，アインシュタインは異なる時代の人物である，という事実に着目し，アインシュタインへの影響は時代をまたいでいるためより重要だ，とすることである．このような考えに基づき次のような仮説を立てた：

- 重要な歴史エンティティは多くの歴史エンティティに影響を与えている．
- 特に，様々な時代や地域の相手に影響を与えたものは重要である．

5.2　時間インパクトを用いたシグニフィカンス計算

仮説に基づいて，時間インパクトを用いて歴史エンティティの重要度を評価する方法は複数考えられ，そのうちの 1 つの方法は，単純にすべての時刻 t におけるインパクトの合計を用いることである．この方法では，様々な時代において高いインパクトを持つ歴史エンティティが高い重要度を持つ傾向があると考えられるが，その一方で，ある一瞬でのみ巨大なインパクトを持つものも高い値を持つ可能性がある．

その他の方法の 1 つは，各歴史エンティティのインパクトの**ピーク**を考え，そこからの距離で重み付けを行うことである．インパクトがピークから時間的に遠いほど，そのインパクトの重みを強くすることで，離れた時刻におけるインパクトを重要視するという仮説に沿った計算を行うことができる．しかしながら，この手法の問題点は，歴史エンティティのピークという概念が曖昧だということである．たとえば，歴史上の人物を対象とした場合，ピークはこの人物が生前最も輝いていたときであると考えられるかもしれないが，そのような瞬間を定めることは容易ではない．

本研究では，最も大きなインパクトを示しているときを，その歴史エンティティ e のピーク p_e とする．レオナルド・ダ・ヴィンチ，ニュートン，アインシュタインのピークは図 1 よりそれぞれ，$p_{daVinci} = 1490$，$p_{Newton} = 1700$，$p_{Einstein} = 1900$ となる．この方法を用いることで，すべての歴史エンティティに対してピークを定めることが可能となる．

以上のように定まるピーク p_e を用いて，歴史エンティティ e のインパクト $i_e(a)$ を重み付けし，重み付けされたインパクト $i'_e(a)$ の和をとることで，シグニフィカンス s_e を求める：

$$s_e = \sum_a i'_e(a) = \sum_a i_e(a) f(a, p_e) \tag{3}$$

ここで，関数 f は引数を 2 つとり，それらの差が小さいほど小さな値を返す重み関数である．本研究では次の 3 種類について評価実験を行った：

$$f_{linear}(x, y) = |x - y| \tag{4}$$

$$f_{log}(x, y) = \log(|x - y| + 1) \tag{5}$$

$$f_{square}(x, y) = (x - y)^2 \tag{6}$$

5.2.1　議　　論

提案手法では，時間インパクトが単峰形状をしていることを暗黙的に仮定し，歴史エンティティの時間的な中心点を，時間インパクトのピークを用いて表した．このような方法を用いることができるのは，6.3 節で示すように，一般的に歴史エンティティは時間的に近い相手から参照される度合いが高いため，ピークが有効時間の近くに位置するためである．実際，今回確認した範囲では，すべての歴史エンティティに対して時間インパクトは単峰形状をとり，ピークは有効空間内に収まっていた．一方，時間インパクトが複数の山を持つ場合，ピークから遠い位置で大きなインパクトを持つことから，このようなエンティティのシグニフィカンスは大きくなる．上記の事実から，複数のピークを持つエンティティは稀であると考えられるが，仮に存在したとしても，様々な時代に大きなインパクトを持つエンティティが歴史的に重要である，という仮説に基づいて考えた場合，このようなエンティティは大きなシグニフィカンスを持つべきであるといえるので，大きな問題にはならない．

5.3　空間的広がりを用いたシグニフィカンス計算

時間インパクトを用いたアプローチでは，より遠い時代からの参照を重視するが，同様に

空間インパクトに対し，より遠い場所からの参照を重視することを考える．このとき，いくつかの解決されるべき問題がある．

まず，エンティティの空間的中心の与え方の問題である．時間インパクトからシグニフィカンスを求める際に，各エンティティに対して時間インパクトのピークを求め，この時点を時間的中心とし，そこからの距離によって重み付けを行った．このような手法を用いたのは，先述したように，多くのエンティティで時間インパクトが単峰型をとるためである．一方，4.3 節で示した *Napoleon I* の例のように，地理インパクトは綺麗な単峰型にならない場合が多い．このような違いは，人間は連続した時間でしか生きられないが，活動した地域は飛び地になりうる，という事実から生じるものと考えられる．そのため，空間インパクトが最大となる地点を単純に歴史エンティティの中心ととらえることはできない．そこで，最も高い値を示す 1 か所ではなく，ある閾値より大きな値をとっている領域や，上位 n 個の領域の集合として，歴史エンティティを表すことが考えられるが，この場合，閾値や n をどのように設定するのかが問題となる．また，エンティティどうしの距離を測る方法として，領域集合どうしの最短距離，最遠距離，重心間の距離などが考えられ，またそれぞれに対して物理距離や，時間距離を用いて測る方法がある．このようなパラメータに対して考えられるあらゆる組合せに対し，評価実験を行うことで優劣を評価することが可能であるが，4.3.1 項で述べたように，空間インパクトを求めること自体に計算量的問題が存在するため，実際の数値演算や有効性の評価には至っておらず，これらは今後の課題である．

5.4　ベースライン手法

以降で歴史エンティティの重要度評価に対する比較手法を説明する．

5.4.1　被リンク数

本研究での仮定に従えば，歴史エンティティの被リンク数は，その歴史エンティティが影響を与えた相手の数を意味する．この手法は，歴史エンティティのシグニフィカンスとして，影響を与えた数を用いることと等しい．本手法では，エンティティ間の時間的な距離は考慮していない．

5.4.2　重みつき被リンク数

歴史エンティティの年情報を用いることで，歴史エンティティ間に時間的距離を導入することが可能となる．すなわち，各リンクに対しリンク相手との時間的距離に比例した重みを与える．こうすることで，単なる被リンク数を用いる方法と比較して，5.1 節で立てた仮説により即した手法となる．

各歴史エンティティが “begin year” から “end year” の間で有効であるとする．このとき，2 つの歴史エンティティ u，v の間の距離関数 $d(u,v)$ を**図 6** に表されるように導入した．ただし，e_{begin} は歴史エンティティ e の “begin year”，e_{end} は “end year” を表す．任意の e に対し $e_{begin} \leq e_{end}$ が成り立っている．

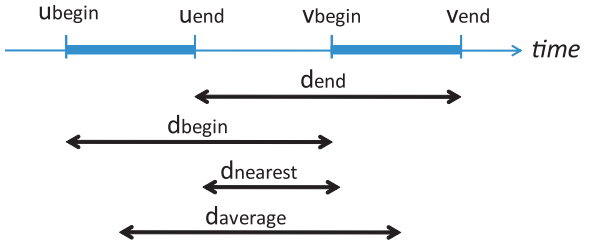


図 6　2 つの歴史エンティティ u と v の間の距離
Fig. 6　Temporal distances between two historical entities u and v.

$$d_{begin}(u,v) = |v_{begin} - u_{begin}| \tag{7}$$

$$d_{end}(u,v) = |v_{end} - u_{end}| \tag{8}$$

$$d_{average}(u,v) = |\frac{v_{end} + v_{begin}}{2} - \frac{u_{end} + u_{begin}}{2}| \tag{9}$$

$$d_{nearest}(u,v) = \begin{cases} 0 & v_{begin} \leq u_{begin} \leq v_{end} \\ 0 & u_{begin} \leq v_{begin} \leq u_{end} \\ v_{end} - u_{begin} & v_{end} \leq u_{begin} \\ u_{end} - v_{begin} & u_{end} \leq v_{begin} \end{cases} \tag{10}$$

また，歴史エンティティのピーク間の距離を用いることも可能である．

$$d_{peak}(u,v) = |u_{peak} - v_{peak}| \tag{11}$$

5.4.3　PageRank

本研究では，歴史エンティティの重要度は，その歴史エンティティが影響を与えた歴史エンティティによって規定されると考える．すなわち，より多くの歴史エンティティに影響を与えた歴史エンティティほど重要である．しかしながら，重要でない歴史エンティティに多く影響を与えたとしても，その歴史エンティティが与えた影響度は小さいと考えられる．そこで，影響を与えた歴史エンティティの重要度を考慮することで，“重要な歴史エンティティに影響を与えた歴史エンティティは重要である” というように歴史エンティティの重要度を考えることができる．そこで，本研究では，このような再帰的な重要度の伝播を計算す

るために PageRank アルゴリズムを用いる．すなわち，Wikipedia 項目とそれらの間のリンク構造を用いて，個々の Wikipedia 項目の PageRank 値を計算することで，対応する歴史エンティティの重要度が計算される．

5.4.4 Distance PageRank

5.1 節で述べた仮説を用いることで，PageRank を改良することができる．リンクの重みを，それによって接続された 2 つの歴史エンティティの時間的距離とする．$d(u,v)$ を u から v へのリンクの重みとし，重要度の伝播する量がリンクの重みに比例する場合，PageRank 計算は次のように表される：

$$s(v) = \alpha \sum_{u \in B_v} \frac{d(u,v)s(u)}{\sum_{w \in F_u} d(u,w)} + \frac{1-\alpha}{N} \tag{12}$$

これにより，時間的に遠くからリンクされている歴史エンティティを発見することができる．このような PageRank の拡張は TextRank [10] や VisualRank [6] でも応用されている．

6. 実　　験

6.1 データセット

本研究では歴史エンティティのシグニフィカンス計算手法の評価実験を行うために，2010 年 10 月 11 日に取得された Wikipedia 英語版のデータセットを用いた．各歴史エンティティの有効時間を定めるために，本研究では，Wikipedia カテゴリを用いる．Wikipedia 内に設定されている “Years” カテゴリを用いて，各項目に対する年情報を抽出する．“Years” カテゴリの下には具体的な年に関わるサブカテゴリが設定されている．たとえば，“Years” にはサブカテゴリ “2010” が含まれ，さらに “2010” には “2010 births” などの 2010 年に関するカテゴリが登録されている．1 つの記事が複数のこのような時間カテゴリに属する場合がある．たとえば，*Newton* は “1643 births” と “1727 deaths” に属する．そこで，各歴史エンティティが属する “Years” カテゴリのうち，最も古いものを “begin year”，最も新しいものが “end year” とする．取得した Wikipedia 項目集合に対し，“Years” カテゴリとそのサブカテゴリに属するものは 1,424,616 本存在した．これをデータセット 1 とする．次に，より絞られたデータ集合に対して評価を行うために，データセット 1 の中から，人物に関する Wikipedia 項目だけを抽出した．この 474,680 本の項目からなる項目集合をデータセット 2 とする．

6.2 正 解 集 合

シグニフィカンス評価手法の評価を行うために，歴史的に重要な歴史エンティティを集めた正解集合を作る．本研究では，日本の世界史検定教科書 11 冊を調べ，そのうちの 6 冊以上で共通して出現している語を，歴史的に重要な用語として抽出した．そして，それぞれに対応する Wikipedia 項目をデータセット 1 から取得した結果，1043 本の項目が得られた．これをデータセット 1 に対する正解セット 1 とする．また，正解セット 1 の中から人物に関する項目のみを抽出し，これを正解セット 2 とした．正解セット 2 は 392 本の項目からなる．

6.2.1 議　　論

歴史的に重要な歴史エンティティを，教科書に基づいて求めることについて，以下のような批判が考えられる．

- 教科書は，歴史の一面からの評価を与えるにすぎず，教科書で言及されない歴史エンティティの中にも重要なものは存在しうる．
- 日本の検定教科書は，日本の学習指導要領に基づいて内容が定められている．そのため，内容にバイアスがかかっている可能性がある．

まず，歴史教科書の内容は歴史学者の意見も取り入れられて作られている．そのため，歴史教科書に多く出現するものは，歴史学の観点から重要であると判断されたものであると考えることができる．たしかに，教科書に現れない中にも重要な歴史エンティティは存在しうるが，教科書の観点から抽出された用語を効率良く集められる手法が高い重要性を与えるエンティティの中で，教科書に出てこないが上位にきた用語を調べていくことで，このような問題に対応できるのではないかと考えられる．

日本の学習指導要領に基づくことによるバイアスの可能性は，複数の教科書を調べているとはいえ，否定しきれない．今後は，他の国の教科書から正解集合を作り，同様の評価を行った場合の相違点について，調べていきたいと考えている．

6.3 歴史エンティティ間の時間距離の分布

データセット 1 に含まれる Wikipedia 項目間のリンクに対して 5 つの距離関数の集計した結果，**図 7** のようになった．この図からも分かるように，Wikipedia 項目は時間的に見て近い相手にリンクをはる傾向があるといえる．

正解セット 1 に含まれる歴史エンティティへのリンクだけを対象とした場合，距離関数は**図 8** のようになる．図 7 と比較して，明らかに $100 < d$ の割合が高くなっており，正解セットに含まれる歴史エンティティは，通常の歴史エンティティと比べて様々な時代からリ

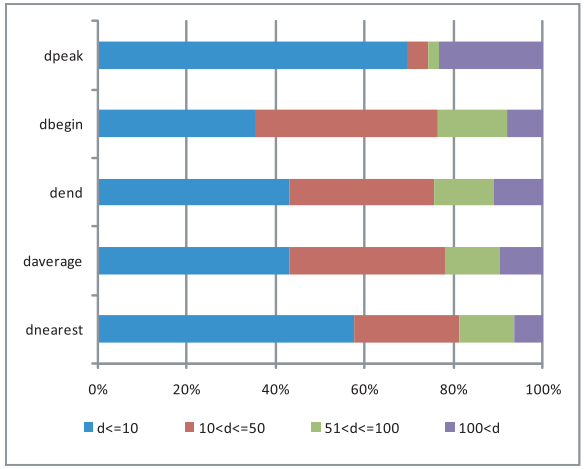


図 7　データセット 1 の歴史エンティティ間の時間距離の分布
Fig. 7　Ratio of result of the four distance function.

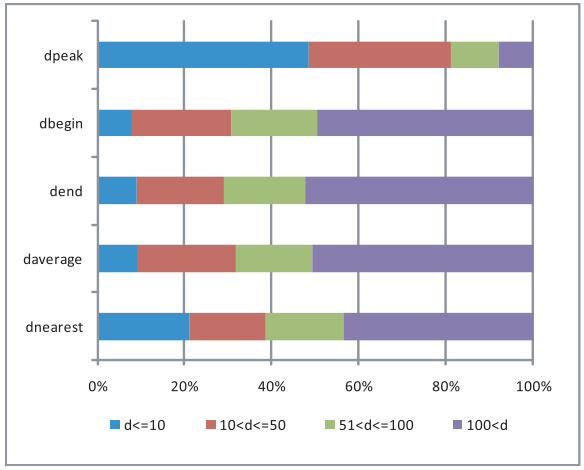


図 8　正解セット 1 を指すリンクの時間距離の分布
Fig. 8　Ratio of value of four distance of links point to item of the correct answer set.

ンクされる傾向があることが分かる．これは，異なる時代の歴史エンティティへ与えた影響の方が，同時代のものへの影響より歴史的に重要である，という考えを支持するものであると考えられる．

6.4　ランキング評価

5 章の提案手法および 4 種類のベースライン手法の評価実験を行った．提案手法では重み関数 f のとり方が 3 種類，そして，インパクト計算の際のダンピングファクタ α として 0.15，0.5，0.85 の 3 種類で評価を行った．また，ベースライン手法では 5 種類の距離関数 d で評価を行った．これらを合わせると，全部で 21 通りとなる．

まず，データセット 1 および正解セット 1 を用いた場合の，各手法における適合率・再現率を**表 1** に示す．重みつき被リンク数と Distance PageRank については，最も結果の良かった d の場合だけを表示している．また，提案手法とベースライン手法の f 値を**図 9** に示す．この表から，$\alpha = 0.85$ のときの提案手法は結果が良かった．提案手法 $\alpha = 0.85$ では *Recall*@100000 で 0.95 近く値を示している．提案手法の方が，より効率良く正解セットが上位にきており，提案手法の中でも f_{log} が最も精度が高かった．また，ベースライン手法でも，inbound link と重みつき被リンク数を比較した場合と PageRank と distance PageRank を比較した場合に，それぞれ歴史エンティティ間の時間的距離を考慮した手法の方が結果が良くなっている．これらの結果は，重要な歴史エンティティは様々な時代の歴史エンティティに影響を与えている，という仮説を支持するものであると考えられる．

次に，データセット 2 および正解セット 2 を対象とした場合の，結果が**表 2** および**図 10** である．この場合も，ベースライン手法よりも提案手法の方が結果が良かった．提案手法では，表 1 と比較して大きな変化がないが，ベースライン手法は大きく結果が改善された．このことから，全歴史エンティティを対象とした場合，ベースライン手法では人物以外の歴史エンティティを抽出する能力が低いと考えられる．

表 3 は，人物だけを対象とした場合の主要な手法の上位 10 件の結果を表したものである．太字のものは正解セットに含まれることを示している．この表から提案手法では，アメリカ大統領やローマ法皇，古代ローマ帝国の皇帝などのように，特定のグループに属する人物が高い順位をとりやすいことが分かる．これは，これらのグループを表す歴史エンティティがハブとなり，そのグループに属するページに高い重要度を与えているものと考えられる．たとえば，Wikipedia には *President of the United States* というページが存在しており，このページはアメリカの歴史を通してつねに参照され続けているページであると考えられる．

最後に，本稿で取り上げた 5 名の人物の順位を**表 4** に示す．この結果から，*Napoleon I*

表 1　適合率–再現率データセット 1
Table 1　Precision and Recall of sorting all historical entities.

Name	@10	@100	@1000	@10000	@100000
proposed f_{linear} 0.15	0.300 — 0.003	0.310 — 0.031	0.113 — 0.113	0.027 — 0.275	0.004 — 0.381
proposed f_{log} 0.15	0.400 — 0.004	0.280 — 0.028	0.092 — 0.092	0.028 — 0.281	0.004 — 0.380
proposed f_{square} 0.15	0.200 — 0.002	0.270 — 0.027	0.108 — 0.108	0.027 — 0.270	0.004 — 0.379
proposed f_{linear} 0.5	0.300 — 0.003	0.310 — 0.031	0.138 — 0.138	0.045 — 0.455	0.007 — 0.689
proposed f_{log} 0.5	0.400 — 0.004	0.280 — 0.028	0.136 — 0.136	0.045 — 0.450	0.007 — 0.715
proposed f_{square} 0.5	0.200 — 0.002	0.270 — 0.027	0.128 — 0.128	0.042 — 0.425	0.006 — 0.635
proposed f_{linear} 0.85	0.400 — 0.004	0.290 — 0.029	0.182 — 0.182	0.068 — 0.683	0.010 — **0.959**
proposed f_{log} 0.85	0.500 — 0.005	0.330 — 0.033	0.170 — 0.170	0.066 — 0.665	0.010 — **0.963**
proposed f_{square} 0.85	0.400 — 0.004	0.250 — 0.025	0.190 — 0.190	0.065 — 0.652	0.010 — **0.951**
被リンク数	0.400 — 0.004	0.280 — 0.028	0.077 — 0.077	0.020 — 0.204	0.003 — 0.330
$d_{average}$ 被リンク数	0.500 — 0.005	0.190 — 0.019	0.110 — 0.110	0.027 — 0.275	0.004 — 0.381
PageRank	0.000 — 0.000	0.090 — 0.009	0.080 — 0.080	0.020 — 0.205	0.003 — 0.283
d_{peak} PageRank	0.100 — 0.003	0.050 — 0.013	0.078 — 0.199	0.019 — 0.485	0.003 — 0.686

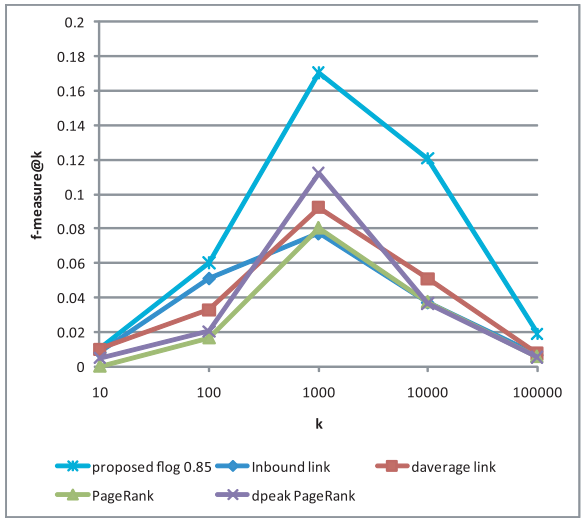


図 9　表 1 の f 値
Fig. 9　F-measure score of Table 1.

が 5 人の中で最も重要な人物であるといえる．また，冒頭で取り上げた “*Albert Einstein* と *Isaac Newton* では，どちらのほうが偉大か” という問題に対する，提案手法による解答は “*Isaac Newton* の方が偉大である” という結果となった．

7. まとめと今後の課題

本稿では，歴史エンティティを様々な複数の観点から評価する手法を提案した．

まず，時間情報を持つ Wikipedia 項目を用いて歴史エンティティを表す方法を提案した．本研究では，時間に関わるカテゴリを用いて，時間情報を持つ項目の抽出を行っている．しかしながら，この手法では，本来時間情報を持つが，上記のようなカテゴリに属さないために，抽出されない項目が生じる場合が考えられる．この問題に対し，今後は本文中の時間に関する記述を用いて歴史エンティティを抽出を行うことが考えられる．

次に，歴史エンティティのアスペクトに対するインパクト計算の手法を提案した．提案手法では，インパクトのアスペクトに対して有効な歴史エンティティをまず抽出し，それらのエンティティ集合から参照されている度合いを，反復的なアルゴリズムを用いて求める．本稿では，アスペクトとして，時間に関わるものと空間に関わるものについて言及した．歴史エンティティの有効時間は，対応する Wikipedia 項目が属するカテゴリの情報を用いて定め，また，本文中に記載されたジオタグを用いて有効空間を定めたが，いくつかの問題点が残されている．まず，Wikipedia カテゴリを用いて有効時間を定める手法は，たとえば

表 2　適合率–再現率データセット 2
Table 2　Precision and Recall of sorting all historical people.

Name	@10	@100	@1000	@10000	@100000
proposed f_{linear} 0.85	0.600 — 0.012	0.310 — 0.079	0.147 — 0.375	0.034 — 0.875	0.004 — 0.987
proposed f_{log} 0.85	0.600 — 0.015	0.260 — 0.066	0.133 — 0.339	0.035 — 0.898	0.004 — 0.995
proposed f_{square} 0.85	0.700 — 0.018	0.370 — 0.094	0.148 — 0.378	0.032 — 0.827	0.004 — 0.982
inboud link	0.400 — 0.010	0.280 — 0.071	0.077 — 0.196	0.020 — 0.520	0.003 — 0.839
$d_{average}$ link	0.500 — 0.013	0.190 — 0.048	0.110 — 0.281	0.027 — 0.699	0.004 — **0.969**
PageRank	0.000 — 0.000	0.090 — 0.023	0.080 — 0.204	0.020 — 0.523	0.003 — 0.719
d_{peak} PageRank	0.100 — 0.003	0.050 — 0.013	0.078 — 0.199	0.019 — 0.482	0.003 — 0.686

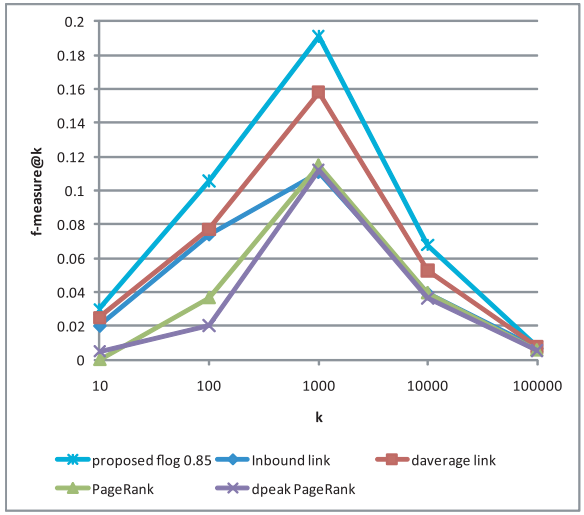


図 10　表 2 の f 値
Fig. 10　F-measure score of Table 2.

Japan は “States and territories established in 660 BC” というカテゴリに属し，他に時間に関するカテゴリを持たないため，“begin year” と “end year” が紀元前 660 年になってしまう．しかしながら，日本は現在でも明らかに有効であると考えられる．また，有効空間についても，Wikipedia 中の緯度経度情報は，建物や都市などのように，位置情報が明確に定まるような項目にしか設定されていない．そのため，たとえば人物の項目は緯度経度情報を含まないが，本文中には国や街などが含まれる場合がある．このように，本文中の情報を用いる発展が考えられる．また，我々の手法では，同一人物の異なる時間的・空間的インパクトや，異なる人物の同じ時間・場所でのインパクトは比較できるが，異なる人物の異なる時間的・空間的インパクトの比較を行うことはできない．たとえば，“*Newton* in 1700” や “*Einstein* in 1900” を比較することはできない．このような問題へも今後取り組んでいく予定である．

また，歴史エンティティの歴史的な重要度を意味するシグニフィカンスに対し，時間的・空間的に影響が広く及ぶほど大きくなる，という仮説を立て，また仮説に従い，インパクトの分散具合と次元の連続性に着目した計算手法を提案した．提案手法では，時間的インパクトの最も大きくなるアスペクトに注目し，その時点から連続的に離れるに従って大きな重み付けを行い，それらを合計することでシグニフィカンスを計算する．一方，空間インパクトを用いた手法は，議論されるべき問題が多くあり，今後取り組んでいく予定である．

提案手法を評価するために，様々なベースライン手法との比較実験を行った．その結果，提案手法が最も良い結果となっただけでなく，ベースライン手法の中でも，仮説に則った手法の方が，仮説を考慮しない単純な手法よりも良い結果となった．この結果は，重要な歴史エンティティが与えた影響は時間的に広い，という仮説を支持するものであると考えられる．提案手法によって高い値を得るエンティティを見てみると，ローマ法王のように，長い有効時間を持つ役職についていた人物のような歴史エンティティが得られた．これは，ローマ法王という役職が，これらの歴史エンティティに対して重要度を供給するハブとして機能しているのだと考えられる．このハブとして機能しているページを発見するには Kleinberg による HITS アルゴリズム[7] を利用することを考えている．このように，ある歴史エンティティの重要度を高めるのに大きく寄与している項目を発見できれば，その歴史エンティティの重要性の理由付けを行うことができるようになるため有用である．

表 3　人物を対象とした場合の順位．太字になっている人物は正解セットに含まれていることを表す
Table 3　Sorting result of historical people. Bold person is included in correct answer set.

順位	proposed f_{linear} 0.85	proposed f_{log} 0.85	proposed f_{square} 0.85	被リンク数	$d_{average}$ 被リンク数
1	**Augustus**	Charlemagne	Augustine of Hippo	**George W. Bush**	**Augustus**
2	Pope John Paul II	**Augustus**	Pope John Paul II	Bill Clinton	**William Shakespeare**
3	Charlemagne	**Muhammad**	**Plutarch**	Barack Obama	**Saint Peter**
4	Ptolemy	Pope John Paul II	Ptolemy	**Ronald Reagan**	John Chrysostom
5	**Plutarch**	**Augustine of Hippo**	**George W. Bush**	Bob Dylan	Jerome
6	**Saint Peter**	Diocletian	**Saint Peter**	Robert Christgau	**Muhammad**
7	**George W. Bush**	**Plutarch**	**Adolf Hitler**	**Adolf Hitler**	Athanasius of Alexandria
8	**Augustine of Hippo**	**Saint Peter**	**Napoleon I**	**John F. Kennedy**	Gregory of Nazianzus
9	**Muhammad**	**Justinian I**	Charlemagne	Michel Jackson	Hilary of Poitiers
10	Diocletian	Ptolemy	**Augustine of Hippo**	Elvis Presley	**Pope Gregory I**

表 4　*Napoleon I*，*Leonardo da Vinci*，*Lobert Hooke*，*Isaac Newton*，*Albert Einstein* の提案手法での順位．ここでは $\alpha = 0.85$ を用いている
Table 4　Significance of *Napoleon I*, *Leonardo da Vince*, *Lobert Hooke*, *Isaac Newton* and *Albert Einstein* calculated by our proposed method.

	proposed f_{log} 0.85
Napoleon I	24
Isaac Newton	150
Albert Einstein	192
Leonardo da Vinci	511
Lobert Hooke	2640

実験においては，教科書を利用して作成された正解セットを用いて各手法の評価を行った．教科書は “古代”，“中世” などの複数の章立てで構成されており，取り上げられる内容は，各章の中で決定されている．そのため，すべての歴史エンティティを並べ替えて評価するのではなく，各時代ごとに歴史エンティティを並べ替え，その時代で教科書に取り上げられている歴史エンティティを用いて評価を行う方が適切であると考えられる．今後は，このような評価実験を行っていきたい．

本研究は，計算機を用いて歴史を考える Computational History に関する研究であり，従来の歴史研究と異なる点として，計算機を用いることで人手では困難だった，複数の分野を横断した大規模なデータマイニングを，本稿で Wikipedia をコーパスとして用いたように，行うことができる点があげられる．しかしながら，これは従来の歴史学と相対するものではなく，その上に築かれていくべき学問である．今後はこのような歴史学者を交えたうえで議論を行っていきたいと考えている．

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，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

参　考　文　献

1) Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proc. 7th International Conference on World Wide Web* (*WWW 1998*), pp.107–117 (1998).
2) Martins, B., Freire, N. and Borbinha, J.: Complex Data Transformations in Digital Libraries with Spatio-Tmporal Information, *Proc. 11th International Conference on Asia-Pacific Digital Libraries* (*ICADL 2008*), pp.174–183 (2008).
3) Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data, *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL 2007*), pp.708–716 (2007).

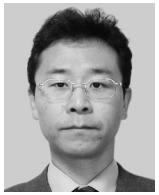
4) Gabrilovich, E. and Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, *Proc. 20th International Joint Conference on Artificial Intelligence* (*IJCAI 2007*), pp.1606–1611 (2007).
5) Haveliwala, T.H.: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, *IEEE Trans. Knowledge and Data Engineering* (*TKDE 2003*), Vol.15, No.4, pp.784–796 (2003).
6) Jing, Y. and Baluja, S.: VisualRank: Applying PageRank to Large-Scale Image Search, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*TPAMI*), Vol.30, pp.1877–1890 (2008).
7) Kleinberg, J.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).
8) Larson, R.R.: Geographic Information Retrieval and Spatial Browsing, *GIS and Libraries: Patrons, Maps and Spatial Information*, pp.81–124 (1996).
9) Larson, R.R. and Frontiera, P.: Geographic Information Retrieval (GIR): Searching Where and What, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 2004*), p.600 (2004).
10) Mihalcea, R. and Tarau, P.: TextRank: Bringing Order into Texts, *Proc. 2004 Conference of the Empirical Methods in Natural Language Processing*, pp.404–411 (2004).
11) Ymamoto, M., Takahashi, Y., Iwasaki, H., Oyama, S., Ohshima, H. and Tanaka, K.: Extraction and Geographical Navigation of Important Historical Events in the Web, *Proc. 10th International Symposium on Web and Wireless Geographical Information Systems* (*W2GIS 2011*), pp.21–35 (2011).
12) Stoev, S.L., Feurer, M. and Ruckaberle, M.: Exploring the Past: A Toolset for Visualization of Historical Events in Virtual Environments, *Proc. ACM Symposium on Virtual Reality Software and Technology* (*VRST 2001*), pp.63–70 (2001).
13) Strötgen, J., Gertz, M. and Popov, P.: Extraction and Exploration of Spatio-Temporal Information in Documents, *Proc. 6th ACM Workshop on Geographic Information Retrieval* (*GIR 2010*), pp.1–8 (2010).
14) Gyöngyi, Z., Garcia-Molina, H. and Pedersen, J.: Combating Web Spam with TrustRank, *Proc. 30th International Conference on Very Large Data Bases* (*VLDB 2004*), pp.576–587 (2004).
15) 岡本章裕，黒井星良，横山昌平，福田直樹，石川　博：Wikipedia を対象とした地理情報と時間情報の抽出手法の提案，*Proc. 1st Data Engineering and Information Management* (*DEIM 2009*) (2009).

(平成 23 年 4 月 11 日受付)
(平成 23 年 9 月 12 日採録)

高橋　侑久（学生会員）
2011 年京都大学大学院情報学研究科博士前期課程修了．主に Web マイニングの研究に従事．日本データベース学会学生会員．

大島　裕明（正会員）
京都大学大学院情報学研究科社会情報学専攻助教．2007 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主にウェブ，情報検索，データベースの研究に従事．電子情報通信学会，日本データベース学会，ACM 各会員．

山本　光穂
（株）デンソーアイティーラボラトリ研究企画部シニアエンジニア．2003 年長岡技術科学大学大学院電気電子システム工学専攻修了．主に車載機器向けサービスの開発および情報検索の研究に従事．



岩崎　弘利（正会員）
（株）デンソーアイティーラボラトリ研究開発部ジェネラルマネージャ．1990 年名古屋大学大学院工学研究科博士課程前期課程電気・電子工学専攻修了．博士（工学）．1990 年日本電装株式会社（現在の（株）デンソー）入社．2000 年より（株）デンソーアイティーラボラトリ出向．車の知能的ユーザインタフェースの研究開発に従事．人工知能学会，電子情報通信学会各会員．

小山　聡（正会員）

北海道大学大学院情報科学研究科複合情報学専攻准教授．1994 年京都大学工学部数理工学科卒業．1996 年京都大学大学院工学研究科数理工学専攻修士課程修了．2002 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．1996〜1998 年日本電信電話株式会社．2001〜2002 年日本学術振興会特別研究員（DC2）．2002〜2007 年京都大学大学院情報学研究科社会情報学専攻助手．2003〜2004 年スタンフォード大学 Visiting Assistant Professor．2007〜2009 年京都大学大学院情報学研究科社会情報学専攻助教．2009 年北海道大学大学院情報科学研究科複合情報学専攻准教授，現在に至る．機械学習，データマイニング，情報検索等に興味を持つ．2005 年度人工知能学会論文賞．2009 年度日本データベース学会上林奨励賞．電子情報通信学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田中　克己（正会員）

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．京都大学工学博士．主にデータベース，マルチメディアコンテンツの処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会等各会員．