Wikipedia リンク構造を用いた歴史エンティティの重要度計算

高橋　侑久† 　大島　裕明† 　山本　光穂†† 　岩崎　弘利†† 　小山　聡†††
田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻　〒 604–8501 京都府京都市左京区吉田本町
†† 株式会社デンソーアイティーラボラトリ　〒 150–0002 東京都渋谷区渋谷 2-15-1 渋谷クロスタワー 28F
††† 北海道大学大学院情報科学研究科複合情報学専攻　〒 060–0814 北海道札幌市北区北 14 条西 9 丁目
E-mail: †{ytakahas,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp, ††{miyamamoto,hiwasaki}@d-itlab.co.jp,
†††oyama@ist.hokudai.ac.jp

あらまし　本稿では，人類史の要素を表す歴史エンティティの重要度評価手法を提案する．本研究では，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・空間的なインパクトを計算する．歴史エンティティのインパクトは時間・場所によって異なる．歴史エンティティは Wikipedia 記事を用い，Wikipedia 記事間のリンク関係が影響の伝播を表すものと考える．つまり，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．また，一部の Wikipedia 記事は時間や場所に関する情報を含んでおり，これらの情報を PageRank と似た反復計算アルゴリズムを用いて，Wikipedia リンク構造に沿って伝播させることで歴史エンティティのインパクトを計算する．そして，インパクトが及ぶ範囲が広く大きいほど，歴史エンティティは重要であると考える．このような仮説に対し，Wikipedia 記事を用いた評価実験及びインパクトの可視化を行った．

キーワード　PageRank，リンク構造分析，Wikipedia，歴史エンティティ

1. ま え が き

“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極めて興味深いが，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(注1)

本研究では，歴史エンティティが与えた時間的・地理的な影響という観点から評価することで，歴史エンティティの**重要度**を歴史エンティティ間の相対値ではなく，個々の歴史エンティティの絶対値として評価することを目的とする．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．そして，重要な歴史エンティティが与えた影響は，時間的にも空間的にも広い．これはつまり，異なる時代や地域の多くの歴史エンティティが影響を受けてているということである．このような仮説に基づき，我々は最初に歴史エンティティの時間的・地理的な**インパクト**を計算する手法を提案する．

歴史エンティティの与えた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場所から遠ざかるほど弱くなっていくと考えられる．例えば，*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 記事を歴史エンティティとみなし，記事の間の参照関係に着目する．Wikipedia を用いる理由は，記事の対象が広範であるため，様々な歴史エンティティを扱うことができるためである．一部の Wikipedia 記事は時間や場所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．このようにすることで，全ての歴史エンティティに対して，時間的・空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・空間的インパクトは “*Newton* in 1900” や “*Napoleon* on England” 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や場所に沿った変遷はグラフや地図を用いることでアプリケーションとして可視化を行うことが可能である．このようなサービスは，過去の歴史エンティティがどのようなインパクトを与えているかを知るだけでなく，これからの未来を考えることに用いることができる．例えば，訪れた先の土地が，自身が普段いる土地にどのような影響を与えたかを知る助けになる．

得られたインパクトを用いて各歴史エンティティの重要度を計算する．冒頭で述べたように，本研究では，歴史エンティティ

(注1)：例えば，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．

が，それ自身から見て時間や空間的に遠くにまで影響を与えているものが，重要であると考える．このような仮定に基づいた，歴史エンティティの重要度評価手法を提案する．

本稿では，複数の歴史教科書に共通して出現する単語を，歴史上の重要な単語とみなすことで正解セットを作成し，Wikipediaの実際のデータを用いて提案手法の評価実験を行った．

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ページの支持とみなし，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，という再起的な考えによって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に関する先攻研究は多数なされている．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)(注2) 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 biased PageRankを用いて地理的領域での重要度や時間区間での重要度を計算する場合のように，各要素の間に連続性がある重要度ベクトルを計算する．4. 章でこれらの具体的なアルゴリズムの計算法について触れる．また，我々は先行研究において，ハイパーリンクを用いて Webページが参照されている文脈を表すアスペクト特徴ベクトルを biased PageRankを用いて計算する手法を提案してきた [15].

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の記事は幅広い分野を網羅しており，その記事数は既に 350万 (2011年 1月英語記事) 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．そのような特徴の一つに，記事間の密なリンク構造が挙げられる．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].

Google Maps (注3) は地理的な Webサービスを簡単に構築できる APIを多数提供する，Web上の地図サービスである．Stoevらによって，Google Mapsを用いた歴史上の出来事を可視化する研究がなされている [12]．本研究では，空間的なインパクトの可視化を，本サービスを用いて行っている．

本研究では，地球表面上を Geohashアルゴリズムを用いて分割する．Martinsらはこのアルゴリズムを用いて XSLT/XQueryエンジンに地理情報を処理する機能を作成している [2].

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].

3. 問題定義

3.1 歴史エンティティ

歴史は人物や出来事だけでなく，様々なものから構成されている．歴史上の人物や出来事だけでなく，多くの物事にも歴史的な側面が存在する．例えば，“数字” のような一般概念にも，それ自体が発見された歴史が存在する．このように，一見すると歴史とは関わりのなさそうな物事にも，見方によって歴史的な側面を発見することが可能である．本研究では，歴史エンティティを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り，全てのエンティティ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．

3.2 時間的・空間的インパクト

歴史エンティティのインパクトは考慮する側面によって異なってくる．例えば，時間や場所を側面として考慮すると，ナポレオン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が，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．

3.3 重　要　度

歴史エンティティの重要度とは，他の歴史エンティティとの比較に用いられる．インパクトが，歴史エンティティを評価する観点によって異なる値を取るのに対し，歴史エンティティの重要度はただ一つの値を取るものとする．

3.4 データセット

Wikipediaはデータベースの情報を公開している(注4)．これらのスナップショットはほぼ毎月提供されており，本研究では 2010年 10月 11日に取得されたダンプファイルを用いた．

3.4.1 Wikipedia記事

1つの Wikipedia記事を用いて 1つの歴史エンティティを表現する．その際に，どのような Wikipedia記事を採用するかが問題となる．本研究では，3.1節で述べたように，様々な歴史エンティティが考えられる．そこで，我々は全ての Wikipedia記事が歴史エンティティに対応するものとした．

提案手法では，歴史エンティティ間のリンク関係を主に利用して歴史的な重要度の評価を行うが，同時にリンク元のページとの時間的な距離も評価尺度の一つとして用いる．そのため，データセットに含まれる全ての Wikipedia記事に対し，何年に関する記事か，という情報が必要となる．そこで，Wikipedia内に設定されている “Years” カテゴリを用いて，各記事に対

(注2)：Open Directory Project: http://www.dmoz.org/
(注3)：Google Maps: http://maps.google.com/
(注4)：http://download.wikipedia.org/enwiki/

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．例えば，“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．取得した Wikipedia 記事集合に対し，“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．そこで，今回はこれの記事集合をデータセットとして用いる．

3.4.2 正解セット

重要度を用いたランキング結果の評価を行うために正解セットを作成した．歴史において重要な語として，日本の世界史検定教科書 11 冊を調べ，その内の 6 冊以上で共通して出現している語を抽出した．これらの語に対応する Wikipedia 記事を取得した結果，1043 本の記事が得られた．本研究では，この Wikipedia 記事集合を歴史的に重要な歴史エンティティに対応する正解セットとした．

4. リンク構造分析

4.1 リンク構造分析

提案手法では，歴史エンティティの時間や空間的な情報を初期値として，これらの情報を PageRank と似た反復計算アルゴリズムを用いて，Wikipedia リンク構造に沿って伝播させることで歴史エンティティのインパクトを計算する．

PageRank は重要な Web ページを，Web のリンク構造を用いて発見するための手法で，背景には, 多くの重要な Web ページにリンクされている Web ページは重要である, という考えがある．これはページ u がページ v へのリンクを持つとき, このリンクは暗黙的に v にある種の重要さを与えるとみなすことを意味する．このとき v にどの程度の重要さが与えられるのかを考える．$i(u)$ をページ u の PageRank 値, F_u を u がリンクしているページの集合とすると, このとき, u は $|F_u|$ 本のリンクを持つ．簡単のためにこれらのリンクが全て等しく重要度を伝播させるとすると, リンク (u, v) は $i(u)/|F_u|$ だけの重要度を u から v に与える．このような考えによって, 次の式が導かれる．ここで，B_v をページ v へリンクしているページの集合とする．

$$i(v) = \sum_{u \in B_v} \frac{i(u)}{|F_u|} \tag{1}$$

しかしながら, 式 (1) の解の一意性はグラフ構造に依存する [11]. そのため, PageRank では通常，次のようなダンピングファクター α を導入する．ここで, N はページの総数を表す.

$$i(v) = \alpha \sum_{u \in B_v} \frac{i(u)}{|F_u|} + \frac{1-\alpha}{N} \tag{2}$$

式 (2) において導入された補正項は，全てのページに対して等しい．これをページによって偏らせたものを biased PageRank と言う．偏らせ方は複数考えられるので，ここでは複数のバイアスに対する biased PageRank 値をまとめてベクトルで表すと，ページ v の biased PageRank ベクトル $\mathbf{i}_v$ は次式のように表される：

$$\mathbf{i}_v = \alpha \sum_{u \in B_v} \frac{\mathbf{i}_u}{|F_u|} + (1-\alpha)\mathbf{b}_v \tag{3}$$

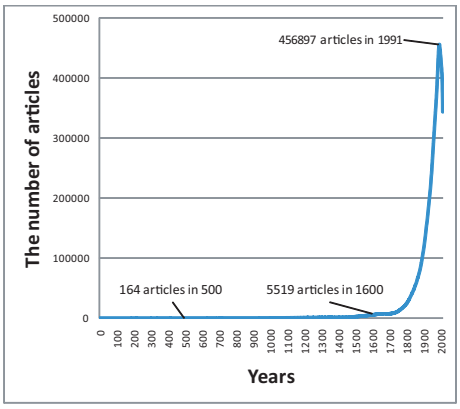

図 1　年に対する歴史エンティティの数.

$$\mathbf{b}_v = \{b_v(j)\} = \begin{cases} 0 & v \notin C_j \\ \frac{1}{|C_j|} & v \in C_j \end{cases} \tag{4}$$

ここで $\mathbf{b}$ はバイアスを表し，C_j はアスペクト A_j を持つページの集合とする．例えば，後述する時間的インパクトの場合，A_j は $j \times 10$ 年から $j \times 10 + 9$ 年の 10 年間を表す.

4.2 Wikipedia リンク構造を用いたインパクト計算

本研究では，Wikipedia 記事間のハイパーリンクによる参照関係に着目する．記事 v が，他の記事 u から参照リンクによって参照されているとする．Wikipedia のリンク構造では，記事間のアンカーテキストは常に記事と関連を持っている．リンクは全て Wikipedia のユーザグループによって作成 k されており，Wikipedia のポリシーにより，たとえ本文に他の記事の名前が出てきたとしても，その記事が文章の内容と関連がない場合はリンクは作成されない．したがって，u から v へのリンクがあった場合，u から v への強い関連が予想される．そのため，このようなリンクがあった場合，我々は v が u に影響を与えたものと見なす．一方で，このリンク関係から u から v に対する何事も判断することはできない．例えば，*Einstein* から *Newton* へのリンクがあった場合，*Einstein* は *Newton* から影響を受けたと考えられるが，その逆については判断することはできない.

4.2.1 時間的インパクト

仮に，*Newton* は，1900 年において重要な人物であったとすると，死後 200 年近くが経っているのにも関わらず，彼は当時の人々や出来事に影響を与えていたものと考えらえる．これは，ある時代における重要さというのは，その歴史エンティティがその時代の他のエンティティにどの程度影響を与えたかによって測ることができることを意味する．つまり，歴史エンティティのインパクトは次のように定まると考えられる：

- あるカテゴリの歴史エンティティに影響を与えた歴史エンティティは，そのカテゴリにインパクトを持っている.
- そのカテゴリでインパクトが持つ歴史エンティティに影響を与えた歴史エンティティはインパクトを持っている.

例えば，*Einstein* が “20 世紀” というカテゴリに属する他の歴史カテゴリに大きな影響を与えたとすると，*Einstein* は 20 世紀において重要である．また，*Einstein* が *Newton* から大きな

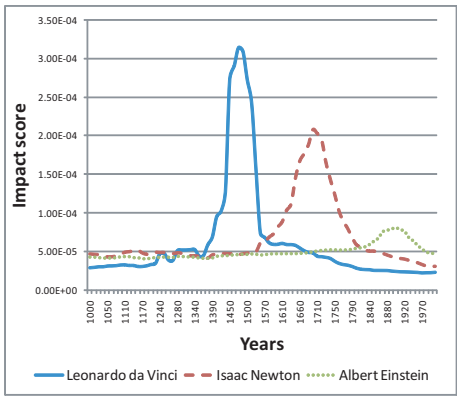

図 2　*Isaac Newton* と *Albert Einstein* と *Leonardo da Vinci* の西暦 1000 年以降におけるインパクトの遷移.

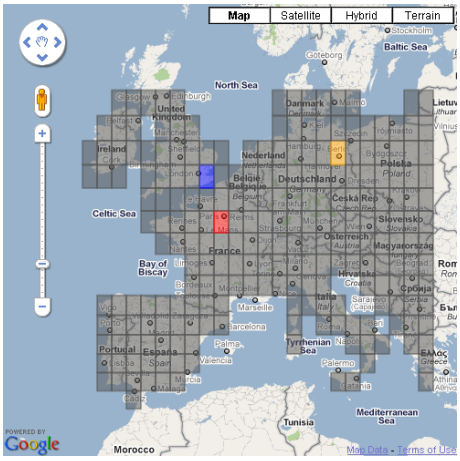

図 3　ジオハッシュを用いた，ヨーロッパ地理空間の分割

影響を受けていたとすると，20 世紀において重要な *Einstein* に影響を与えたという意味で，*Newton* もまた 20 世紀において重要である.

このような仮説に基づき，提案手法では，biased PageRank を用いて全ての歴史エンティティの時間的インパクトを求める．式 (4) におけるバイアス要素 C_j として，Wikipedia の “Years” カテゴリを用いる．これは，時間に関わるカテゴリのトップとなるもので，多くの具体的なカテゴリをサブカテゴリに持っている．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．例えば，*Newton* は “1643 births” と “1727 deaths” に属する．そこで，各歴史エンティティは “begin year” と “end year” の 2 つの属性を持たせる．カテゴリの内，最も古いものが “begin year”，最も新しいものが “end year” となる．*Newton* の場合，“begin year” は 1643 年，“end year” は 1727 年となる．また，“begin year” と “end year” が一致する歴史エンティティも存在する．例えば，*2008* 年夏季オリンピックは “begin year” と “end year” が 2008 年で一致する.

本研究では 201 個のアスペクト $A_0 \cdots A_{200}$ を Wikipedia カテゴリを用いて作成する．カテゴリ A_j は $j \times 10$ 年から $j \times 10 + 9$ 年の 10 年間に対応する．例えば，A_{200} は 2000 年から 2009 年を表す．各歴史エンティティは “begin year” と “end year” の間で活動的であったと表現すると，C_{200} は 2000 年から 2009 年の間で活動的であった歴史エンティティの集合となる.

このような C_j を用いて式 (3) によって計算されるインパクトベクトル $\mathbf{i}$ が各歴史エンティティの時間的インパクトを表す.

図 2 は *Leonardo da Vinci*，*Isaac Newton*，*Albert Einstein* の西暦 1000 年以降におけるインパクトの遷移を表したものである.

4.2.2　空間的インパクト

ここでは，地理空間をアスペクトとして用いた場合について説明する．地理空間を分割して用いるためにジオハッシュアルゴリズムを用いる．ジオハッシュは，Gustavo Niemeyer が geohash.org(注5) という Web サービスを作成中に発明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．ジオハッシュアルゴリズム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.

本研究では，ジオハッシュアルゴリズムを用いて，ヨーロッパを図 3 のように分割した．1 つの格子が，1 つのアスペクト A_j に対応し，その領域に含まれる Wikipedia 記事の集合が C_j となる．1 つ 1 つの格子は 3 文字の英数字で表される．例えば，パリを含む赤線で囲まれた格子はジオハッシュを用いて “u09”，ロンドンを含む青枠が “u12”，ベルリンを含む橙枠が “u33” である．ジオハッシュは末尾に文字列を追加していくことで，任意の精度を表現することが可能となる．そのため，“u09” で始まるジオハッシュは全て，赤線で囲まれた格子の内部に存在することになる.

次に，Wikipedia 記事の中から緯度経度情報を抽出する．Wikipedia 記事の中には，記事の対象の緯度経度情報を含むものが存在する．例えば，エトワール凱旋門は北緯 48 度 87 分 38 秒東経 2 度 29 分 50 秒に位置している．これは，Wikipedia の内部テキストでは，`{coord|48.87|N|2.29|E}` の様に記述される．次に，抽出された緯度経度情報を Geohash に変換する．エトワール凱旋門の緯度経度情報を Geohash に変換した値は，“u09wh1pnt2” となる．Geohash が u09 から始まっていることから，エトワール凱旋門は図 3 中の赤線で囲まれた格子の中に位置することが分かる．すなわち，エトワール凱旋門は C_{u09} に含まれる．同様に，エッフェル塔の Geohash は “u09tunqu1x” となるため，C_{u09} に含まれている．このようにして求められる C_{u09} を用いて biased PageRank を計算することで，歴史エンティティのパリ周辺でのインパクトを求めることが出来る.

以下にナポレオン，*Einstein*，ベルリンの壁の場合の結果を示す.

図 4 はナポレオン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繋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，ネマン川を超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 はベルリンの壁の結果を示している．ドイツを中心に強

(注5)：http://geohash.org

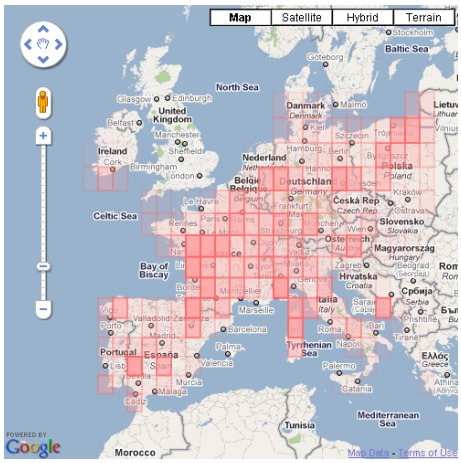

図 4　*Napoleon I*

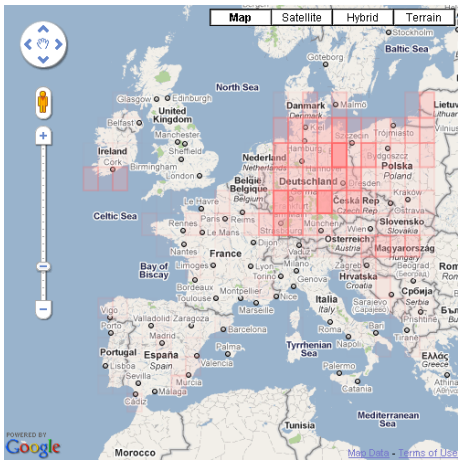

図 5　*Berlin Wall*

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．

5.　Wikipedia を用いた歴史エンティティの評価

Newton が *Hooke* と *Einstein* に影響を与えたとして，どちらの方がより *Newton* の重要性を表している言えるだろうか．一つの考え方は，*Newton* と *Hooke* は同時代の人物だが，*Einstein* は異なる時代の人物である，という事実に着目し，*Einstein* 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，次のような仮説を提案する：

- 時間や空間的に遠くでインパクトを持っていた歴史エンティティは重要

本節ではいくつかの歴史エンティティの重要度評価手法に言及する．まず，5.1 節で，上記の仮説に基づいた提案手法を説明する．ただし，本論文では，時間的な遠さに着目した方法について説明する．また，5.2 節で，その他のベースライン手法について説明する．

5.1　時間的インパクトを用いた重要度計算

上で述べたような仮説に基づいて，時間的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，その内の 1 つの方法は，時間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時間的インパクトの**ピーク**を考え，そこからの距離に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程，そのインパクトの重みを強くすることで，時間的に遠くにあ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．例えば，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時であると考えられるかも知れないが，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，もっとも大きなインパクトを示している時を，その歴史エンティティのピークとする．*Leonardo da Vinci*, *Newton*, *Einstein* のピークは図 2 よりそれぞれ，A_{149}，A_{170}，A_{190} となる．この方法を用いることで，全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．

e_{peak} を歴史エンティティ e のピーク，$\mathbf{i}_e$ を e のインパクトベクトルとする．その時，歴史エンティティ e の重要度 s_e は以下のように求められる：

$$s_e = \mathbf{f}_{e_{peak}}^T \mathbf{i}_e \tag{5}$$

ここでベクトル $\mathbf{f}_c$ は中心 c との距離の重み付けを行う．本研究では次の 3 種類について評価実験を行った：

$$\mathbf{f}_c^{linear} = \{f_c(i)\} = |c - i| \tag{6}$$

$$\mathbf{f}_c^{log} = \{f_c(i)\} = log(|c - i| + 1) \tag{7}$$

$$\mathbf{f}_c^{pow} = \{f_c(i)\} = (c - i)^2 \tag{8}$$

5.2　ベースライン手法

Wikipedia 記事を歴史エンティティとして用いる場合，歴史エンティティの重要度として記事の長さを用いる方法が考えられる．Wikipedia において，一般によく知られている対象に関する記事ほど，内容が充実している傾向があることは明らかである．しかし，Wikipedia はプロジェクトの性質上，誰でも記事を編集することが可能であり，記事が中立性と検証可能性を満たす限り，対象の重要さに関係なく，記事の分量を増やすことが可能である．そのため，このような尺度を重要さとして用いることは，手法の頑健性を損なうことになると考えられる．そのため，本研究においては，このような情報を対応する歴史エンティティの重要さを表す尺度として用いない．

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いくつか説明する．

5.2.1　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間の時間的な距離は考慮していない．

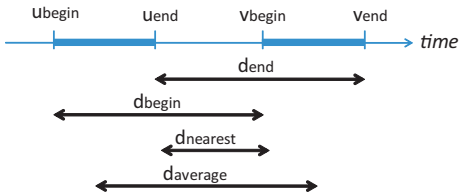

図 6　2 つの歴史エンティティ u と v の間の距離.

5.2.2　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．4.2.1 節で定めたように各歴史エンティティに “begin year” と “end year” が設定されているとき，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する．ただし，n_{begin} は歴史エンティティ n の “begin year”，n_{end} は “end year” を表す．任意の n に対し $n_{begin} \leqq n_{end}$ が成り立っている．

$$d_{begin}(u,v) = |v_{begin} - u_{begin}| \tag{9}$$

$$d_{end}(u,v) = |v_{end} - u_{end}| \tag{10}$$

$$d_{average}(u,v) = \left|\frac{v_{end}+v_{begin}}{2} - \frac{u_{end}+u_{begin}}{2}\right| \tag{11}$$

$$d_{nearest}(u,v) = \begin{cases} 0 & v_{begin} \leqq u_{begin} \leqq v_{end} \\ 0 & u_{begin} \leqq v_{begin} \leqq u_{end} \\ v_{end} - u_{begin} & v_{end} \leqq u_{begin} \\ u_{end} - v_{begin} & u_{end} \leqq v_{begin} \end{cases} \tag{12}$$

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．

$$d_{peak}(u,v) = |u_{peak} - v_{peak}| \tag{13}$$

データセットに含まれる Wikipedia 記事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 記事は時間的に見て近い相手にリンクをはる傾向があると言える．

その一方で，正解セット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．

5.2.3　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，次の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．

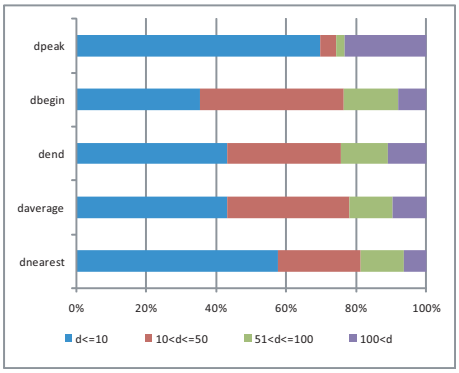

図 7　データセットの歴史エンティティ間の時間距離の分布.

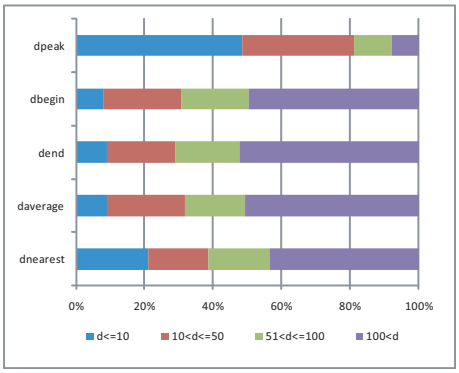

図 8　正解セットを指すリンクの時間距離の分布.

- 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.

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 記事とそれらの間のリンク構造を用いて，個々の Wikipedia 記事の 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．

5.2.4　Distance PageRank

5. 節の導入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：

$$s(v) = \alpha \sum_{u \in B_v} \frac{d(u,v)s(u)}{\sum_{w \in F_u} d(u,w)} + \frac{1-\alpha}{N} \tag{14}$$

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

3.4.2 節で述べた正解セットを用いて，5. 節の提案手法及び 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 $\mathbf{f}$ の取り方が 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．

その結果の適合率・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，もっとも結果のよかった d の場合だけを表示している．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，より効率よく正解セットが上位に来ており，提案手法の中でも $\mathbf{f}^{log}$ 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よくなっている．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．また，それぞれの結果の上位に現れる歴史エンティティを見たところ，国や地名が上位に現れる傾向が強かった．

本研究では歴史エンティティの対象が広い．そこで，対象を人物のページだけに絞った場合について調べた．データセットの中から人物だけを抽出した結果，474,680 個の歴史エンティティが該当した．また，正解セットの中には 392 個の人物が含まれていた．このような人物のみを対象として同様の評価を行った結果，提案手法では，適合率再現率共に同程度であったが，ベースラインでは若干の改善が見られた．このことから，提案手法はベースラインと比較して，人物以外の重要な歴史エンティティを効率的に発見することができると考えられる．

表 2 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．例えば，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常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 3 に示す．この結果から，*Napoleon I* 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．

表 3　*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．

	$\mathbf{f}^{log}$
Napoleon I	24
Leonardo da Vinci	511
Lobert Hooke	2640
Isaac Newton	150
Albert Einstein	192

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，歴史エンティティのインパクト計算の手法を提案した．歴史エンティティの時間情報を取得するために，Wikipedia の “Years” カテゴリを使用し，この情報をリンクを用いて再帰的に伝播させることで，歴史エンティティの時間的インパクトを計算した．しかしながら，この手法には 2 つの問題がある．まず初めに，Wikipedia 記事の網羅性の問題がある．“Years” カテゴリには 140 万記事程度含まれているが，これは Wikipedia に作成されている 350 万記事のほんの一部に過ぎない．これは，Wikipedia の量的優位性を十分に生かしきれていない可能性があるということを意味する．もう 1 つは，不正確な推定である．例えば *Japan* は “States and territories established in 660 BC” というカテゴリに属するため，“begin year” と “end year” が紀元前 660 年になる．これらの問題に対し，記事本文への自然言語処理をする．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．例えば，“*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重要度に対する仮説及び評価手法を提案した．提案手法では，時間的インパクトの分散具合と，次元の連続性に注目して重要度を計算する．様々なベースライン手法との比較実験を行った結果は我々の仮説を支持するものであった．前節で論じたように，提案手法を用いると，ある特定のカテゴリに属する歴史エンティティが偏って強い重要度を持つ傾向がある．並べ替え結果の多様性は，人類史をまとめる場合に有用であると考えられる．この問題に取り組むには，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ようなページを発見する必要があると考えられる．Kleinberg による HITS アルゴリズム [7] の特徴を利用することが可能なのではないかと考えられる．

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），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

表 1　適合率 – 再現率

Name	@10	@100	@1000	@10000	@100000
proposed $\mathbf{f}^{log}$ 0.15	0.4 — 0.00401606	0.28 — 0.0281124	0.092 — 0.0923695	0.028 — 0.281124	0.00379 — 0.380522
proposed $\mathbf{f}^{log}$ 0.5	0.4 — 0.00401606	0.28 — 0.0281124	0.136 — 0.136546	0.0449 — 0.450803	0.00713 — 0.715863
proposed $\mathbf{f}^{linear}$ 0.85	0.4 — 0.004016	0.29 — 0.029116	0.182 — 0.182731	0.0681 — 0.683735	0.00956 — **0.959839**
proposed $\mathbf{f}^{log}$ 0.85	0.5 — 0.005020	0.33 — 0.033132	0.17 — 0.170683	0.0663 — 0.665663	0.0096 — **0.963855**
proposed $\mathbf{f}^{pow}$ 0.85	0.4 — 0.004016	0.25 — 0.025100	0.19 — 0.190763	0.065 — 0.65261	0.00948 — **0.951807**
被リンク数	0.4 — 0.004016	0.28 — 0.028112	0.077 — 0.077309	0.0204 — 0.204819	0.00329 — 0.330321
$d_{average}$ 被リンク数	0.5 — 0.005020	0.19 — 0.019076	0.11 — 0.110442	0.0274 — 0.2751	0.0038 — 0.381526
PageRank	0 — 0	0.09 — 0.009036	0.08 — 0.080321	0.0205 — 0.205823	0.00282 — 0.283133
d_{peak} PageRank	0.1 — 0.002551	0.05 — 0.012755	0.078 — 0.19898	0.019 — 0.484694	0.00269 — 0.686224

表 2　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．

順位	proposed $\mathbf{f}^{linear}$ 0.85	proposed $\mathbf{f}^{log}$ 0.85	proposed $\mathbf{f}^{pow}$ 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**

術」（研究代表者：田中克己）によるものです．ここに記して謝意を表します．

文　献

[1] Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proc. of the 7th International Conference on World Wide Web (WWW 1998). pp. 107–117 (April 1998)

[2] Bruno Martins, N.F., Borbinha, J.: Complex data transformations in digital libraries with spatio-tmporal information. In: Proc. of the 11th International Conference on Asia-Pacific Digital Libraries (ICADL 2008). pp. 174–183 (December 2008)

[3] Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007). pp. 708–716 (June 2007)

[4] Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007). pp. 1606–1611 (January 2007)

[5] Haveliwala, T.H.: Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. IEEE Transactions on Knowledge and Data Engineering (TKDE 2003) 15(4), 784–796 (July/Aug 2003)

[6] Jing, Y., Baluja, S.: Visualrank: Applying pagerank to large-scale image search. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 30, 1877–1890 (2008)

[7] Kleinberg, J.: Authoritative sources in a hyperlinked environment. Jounal of the ACM 46(5), 604–632 (1999)

[8] Larson, R.R.: Geographic information retrieval and spatial browsing. GIS and Libraries: Patrons, Maps and Spatial Information pp. 81–124 (April 1996)

[9] Larson, R.R., Frontiera, P.: Geographic information retrieval (gir): Searching where and what. In: Proc. of the the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004). p. 600 (July 2004)

[10] Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Proc. of the 2004 Conference of the Empirical Methods in Natural Language Processing. pp. 404–411 (July 2004)

[11] Motwani, R., Raghavan, P.: Randomized algorithms. ACM Computing Surveys (CSUR) 28(1), 33–37 (March 1996)

[12] Stoev, S.L., Feurer, M., Ruckaberle, M.: Exploring the past: a toolset for visualization of historical events in virtual environments. In: Proc. of the the ACM Symposium on Virtual Reality Software and Technology (VRST 2001). pp. 63–70 (November 2001)

[13] Strötgen, J., Gertz, M., Popov, P.: Extraction and exploration of spatio-temporal information in documents. In: Proc. of the 6th ACM Workshop on Geographic Information Retrieval (GIR 2010). pp. 1–8 (February 2010)

[14] Zoltan Gyöngyi, H.G.M., Pedersen, J.: Combating web spam with trustrank. In: Proc. of the 30th International Conference on Very Large Data Bases (VLDB 2004). pp. 576–587 (August 2004)

[15] 高橋侑久, 大島裕明, 小山聡, 田中克己: アスペクト特徴ベクトルとリンク元ページ情報に基づく web ページの特徴量計算. In: Proc. of the 2nd Data Engineering and Information Management (March 2010)