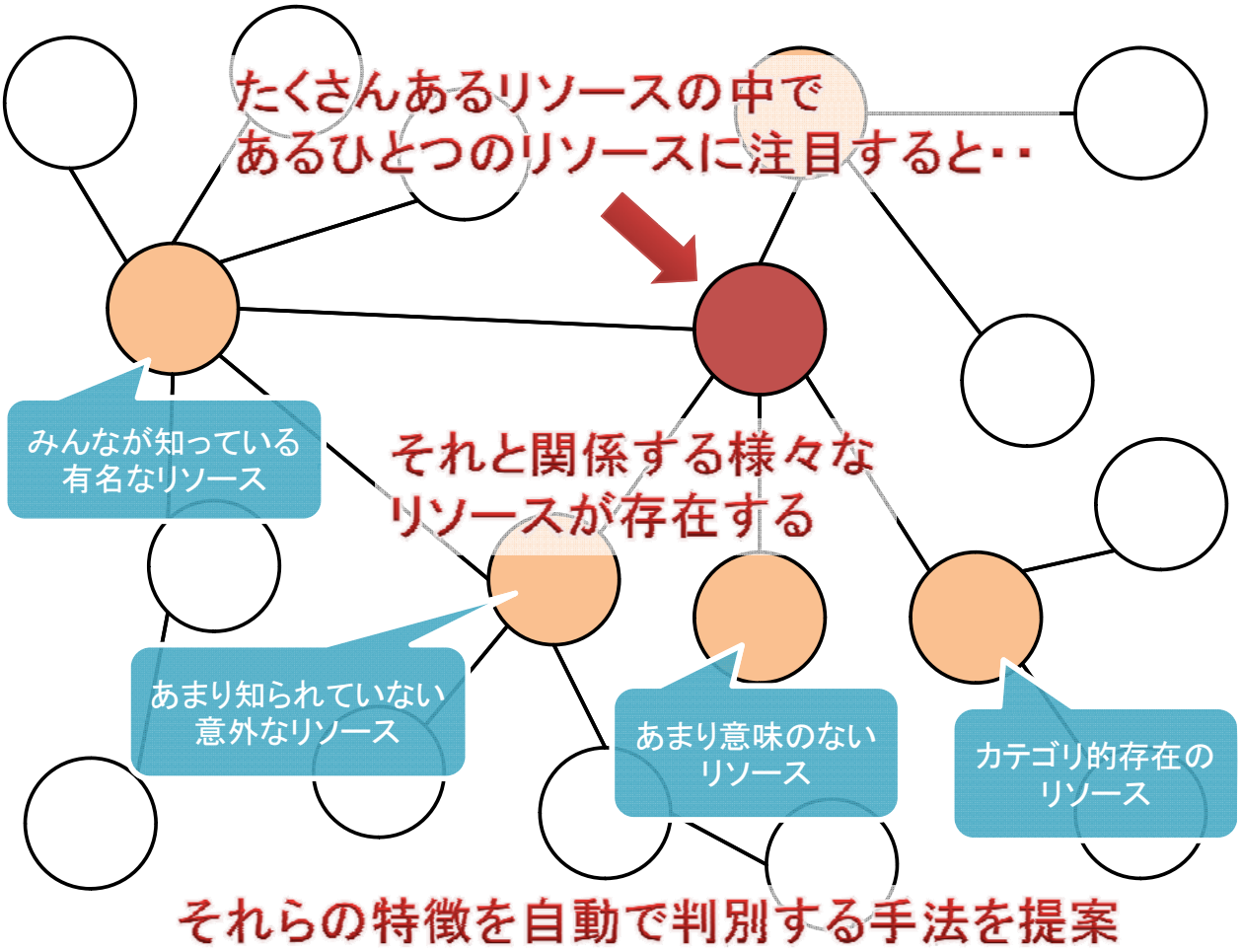
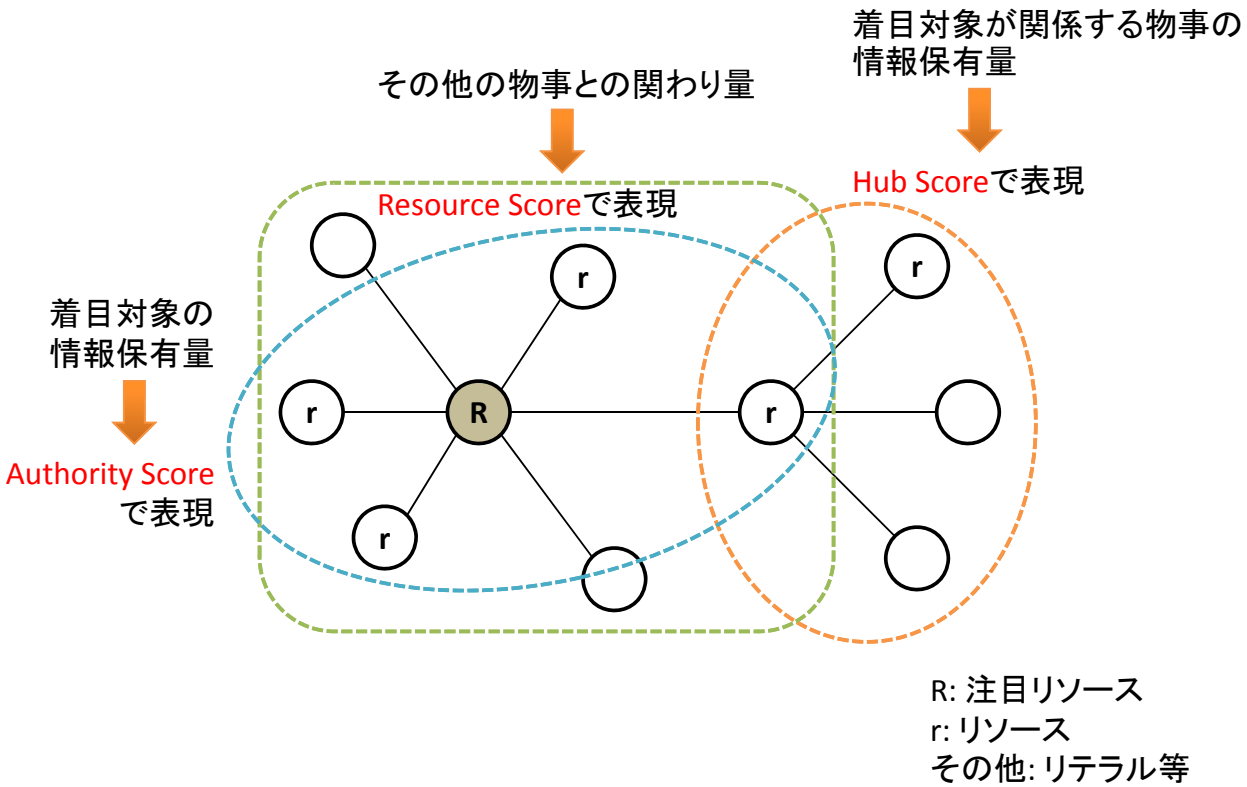
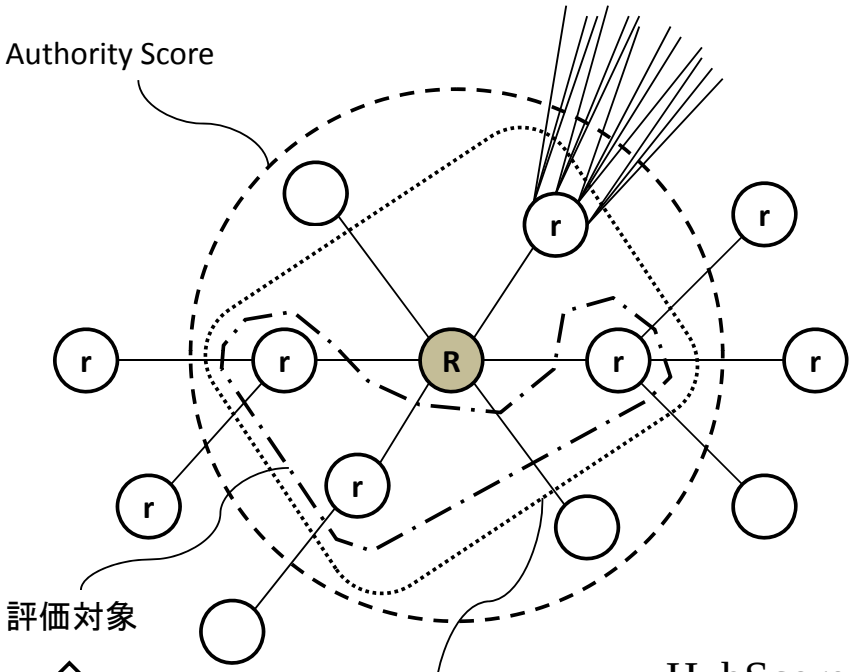

たくさんあるリソースの中で
あるひとつのリソースに注目すると・・
みんなが知っている
有名なリソース
それと関係する様々な
リソースが存在する
あまり知られていない
意外なリソース
あまり意味のない
リソース
カテゴリ的存在の
リソース
それらの特徴を自動で判別する手法を提案


着目対象が関係する物事の
情報保有量
その他の物事との関わり量
Hub Scoreで表現
Resource Scoreで表現
着目対象の
情報保有量
Authority Score
で表現
r
R
r
r
r
r
R: 注目リソース
r: リソース
その他: リテラル等

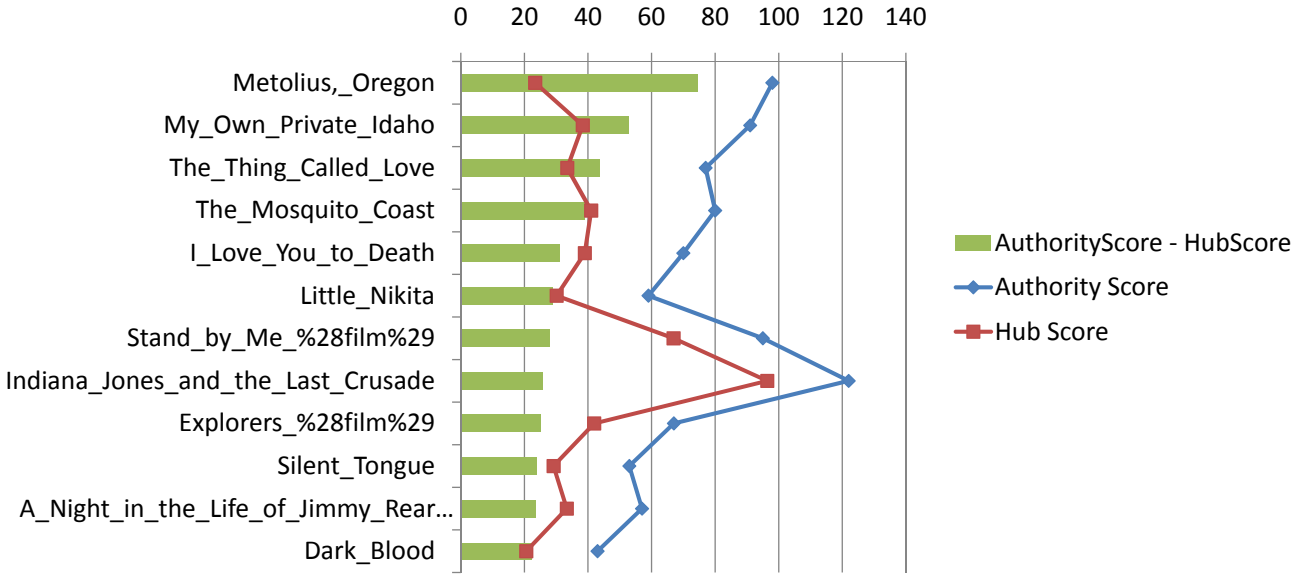


$$HubScore = \sum \frac{AuthorityScore}{ResourceScore}$$

Authority Scoreが
Authority Scoreの中央値±1*（Authority Scoreの標準偏差）
に収まっているものの集合

注目リソース: River Phoenix（俳優）
調査対象データ: DBpedia

River Phoenix固有の傾向があるリソースは…



RDFで記述されたリソースの傾向抽出手法の提案 –DBpediaを例に–

大西可奈子
お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻

1 はじめに

近年，大容量かつ多様化するWebドキュメントをどのようにして有効に扱うかが大きな課題となってきている．そこで，この問題の有効的な解決方法に成り得ると考えられるメタデータやセマンティック・ウェブの技術が，現在改めて注目されている．セマンティック・ウェブは1998年頃にTim Berners-Lee氏によって提唱された技術*であり，従来のHTMLでは伝えきれなかった，語彙の意味なども記述できる．セマンティック・ウェブが注目を浴びる中，セマンティック・ウェブ技術のひとつとしてTim Berners-Lee氏が新たに提唱したのがLinked Data† ‡である．

Linked Data[1][2]は，Web上に情報源のデータを意味の着いたリンクを使って結び付けることであり，技術的にいえば，機械的処理可能で，意味が明示的に定義されていて，外部のデータとリンクしリンクされる形で公開されたWeb上のデータのことである．また，Linked Dataの構成要素は，RDF（Resource Description Framework）形式のデータである．Linked Dataでは，これらのデータを意味の付いたRDFのリンクによって結び付ける．主要なLinked Dataとして，Wikipediaを構造化したDBpedia[3]，地理情報をLinked Dataで記述したGeonames§，音楽のメタデータデータベースであるMusicBrainz¶，概念辞書であるYago‖などがあり，これら以外にも多くのLinked Dataが作成されている．

我々は，このようなLinked Dataを集合知と考え，情報拡張にLinked Dataを利用する研究を行った[4]．この時問題となったのが，注目しているリソースと関係のあるリソースの内，ユーザの求めるものに合致するリソースがどれかわからない，ということである．例えば“東京”に着目し，そこで生まれた人にユーザが興味を持った時，多数存在する東京で生まれた人の内，ユーザの興味に合致した人を選ばなくてはならない．この時，例えばユーザが，その中でも特に有名な人物の情報が欲しいと思った場合，その傾向にある特定のリソースをいくつか提示することを目的とする．ここで注意したいのは，ユーザに提示するリソースは，“類似しているもの”や“関連が強いもの”ではない，ということである．

そこで本研究では，RDFにおけるインスタンスのリンク構造を解析をすることにより，そのリソースが，あるLinked Dataのネットワーク内でどのような存在にあるのかを推測するための手掛かりとなるスコアを定義し，かつ，そのスコアを使って特定の情報を抽出する方法を述べる．この手法は特に，地理情報等のように，どのリソースにおいても一定数のプロパティを持つLinked Dataではなく，リソース毎にプロパティ数に特徴のあるLinked Dataを対象とする．そこで本研究では，DBpediaを例に検証と被験者実験を行う．

*http://www.w3.org/DesignIssues/Semantic.html
†http://www.w3.org/DesignIssues/LinkedData.html
‡http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
§http://www.geonames.org/
¶http://musicbrainz.org/
‖http://www.mpi-inf.mpg.de/yago-naga/yago/

2 関連研究

RDFにおけるインスタンスネットワークに対するリンク解析に研究においては，まずDING(Dataset Ranking)[7]が挙げられる．これはWebデータをランキングするためにPageRank[6]を拡張した手法である．DINGはデータセットのランクを計算し，2レイヤーモデルに基づいた意味依存エンティティランキング戦略によって得た値を結び付けるために，データセット間のリンクを利用する．また，Passant[8][9]は，Linked Dataにおけるリソース間の関係の強さを表わすセマンティック距離を計測するための手法を提案した．セマンティック距離については，リソース間の直接リンクに関わる距離を測る直接距離，リソース間の間接的な距離に基づく関節距離，それらを合わせた距離など複数の計算手法を提案している．さらに，Mirizziら[10]は，DBpediaに着目し，リソースをランクするための新しいハイブリッド手法を提案した．この手法は，RDF構造に内在する性質に基づいたグラフ解析，グラフ上の意味関係から独立した背景，およびサーチエンジンの結果やソーシャルタギングシステムのような外部の情報源を活用し，二つのリソースの関連類似性を計算する．

一方，我々の手法は，HITSアルゴリズム[5]を拡張して定義した三つのスコアをリソース毎に計算する．このスコアの特徴は，計算が非常に単純であり，かつリソース固有のものであるため，どのリソースに着目した場合も再利用可能であるということが挙げられる．すなわちこれらのスコアは上記研究のようにリソース間の関係の強さや類似性を計算するためのものではなく，あくまで個々の独立した存在である．そして，これらの数値を組み合わせることにより，着目対象にとって色々な観点から評価ができ，意外なリソースや周知のリソースを抽出することができる．本研究では二つの条件定義を行ったが，それ以外にも組合せ次第で様々なリソースが抽出できるものと考える．

3 リンク解析に基づく情報提供

世の中に存在する物事は，それ以外の物事と関係している．関係の多さは物事によって異なる．その関係

を記述するのがリンクである．

リンクには複数ある．例えば，HTML で記述されるリンクは定義が緩いため，必ずしも “関係がある＝リンクする” とはならない．一方，RDF で記述された Linked Data において，“関係” はプロパティによって定義される．ここでいう “関係” は，例えば “日本の首都は東京である” のような，不変の事実である場合が多く，例えばその事実が変わった場合は適宜修正が行われる．そのため，HTML の雑多としたリンクよりも，より世の中の物事のつながりを純粋に反映しているものと考えられる．よって本研究では，RDF におけるインスタンのリンク解析を行うことにより，物事をランク付けするための再利用可能なスコアを定義する．ここで “再利用可能” とは，そのスコアを利用して，任意の傾向を持つ情報を推定するための条件を定義することができることを言う．

関係を記述した，すなわちプロパティによって記述されたリンク構造には，記述されているもの以上に様々な情報を含んでいるが分かる．例えば，“誰もが知っている有益な情報”，“知る人ぞ知る意外な情報”，“情報を知るための手がかりとなる情報” 等である．しかしどのような情報かは人が見て初めてわかるものであり，インスタンスやプロパティ自体にそのような記載はない．それを発見するために，RDF を対象にリンク解析を行う．

ここで，代表的なハイパーリンク解析として，HITS アルゴリズムが挙げられる．これは “被リンクの多いページは被リンク数の少ないページよりも優良ページである”，“優良ページは，優良ページへ多くリンクしている” という考え方に基づいたものである．もう一つのハイパーリンク解析として，PageRank が挙げられる．PageRank アルゴリズムの考え方は HITS アルゴリズムのものに近いが，HITS アルゴリズムと違い，“それ自身からリンクする” ことはそのページが優良かどうかに影響しないものと考えている．また，被リンクにおいては，リンク元のリンク数に応じて重みを決定する．

本研究では，これら二つのハイパーリンク解析を参考に，知識同士のリンクの仕方を反映した情報抽出を行う．Linked Data は RDF 形式で記述されるため，ラベル付きの有向グラフで表現される．しかし本研究では，関係の量を見るため，このようなプロパティの性質は捨象し，リソース間のリンクのみに着目することにより単純化することができる．これは例えば，“ある俳優がある映画を演じた” という関係が成り立つ場合，“ある映画はある俳優に演じられた” という関係が成り立つためである．以上を踏まえ，次節でスコアの定義を行う．

3.1 スコア定義

HITS アルゴリズムでは各ページに Authority Score および Hub Score が定義される．Authority Score は重要な情報を発信しているページであるかを示す指標となる．Authority Score が高いほど，優良なハブから多くリンクされていることを示す．Hub Score は重要な情報を発信しているページに，どの程度リンクしているかという指標となる．Hub Score が高いほど，優良なページへリンクしていることを示す．

本研究では，これらの数値を RDF の特性に合わせて以下のように再定義を行った．

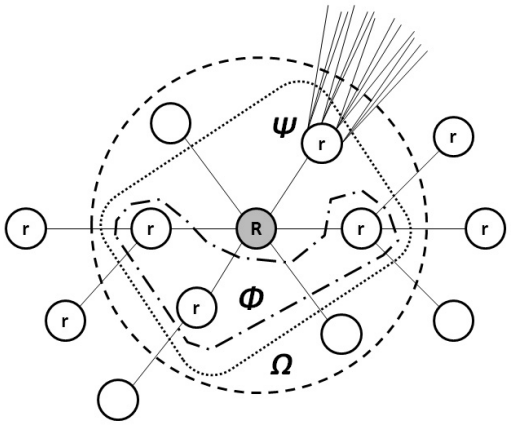


図 1: 注目リソース R を中心としたリソース概要

- Authority Score
  対象とするリソースについて，どの程度，情報が記述されているかを示す指標．Authority Score が高いほど，そのリソースは情報が豊富であることを示す．

- Hub Score
  対象とするリソースが関わる他のリソース群について，どの程度，情報が記述されているかを示す指標．他のリソースがその他のリソースと多くリンクしている場合，対象とするリソースはその他大勢の一つに過ぎないためハブとしての意義は小さくなると考えられるが，他のリソースが対象とするリソースのみと繋がっている場合は，対象とするリソースはそのリソース固有のリソースとなり，ハブとしての意義は大きくなると考えられる．従って，Hub Score が高いほど，情報が豊富なリソースと固有の関係を持っていることを示す．

また，リソース間関係の強さを測るため，新たに Resource Score を以下のように定義した．

- Resource Score
  注目リソースがどの程度，他のリソースと関わっているかを示す指標．Resource Score が高いほど，多くのリソースと関係を持っていることを示す．

3.2 アルゴリズム

前節で定義したスコアは以下のアルゴリズムに従って求められる．

Step1. Authority Score 計算
注目リソース R がリンクしている全てのリソースの集合を $\Omega = \{r_1, r_2, ..., r_\alpha\}$ とする．Ω は重複を許さない α 個のリソースの要素からなる集合とする．ここで，Authority Score $x^{<R>}$ を注目リソース R がリンクしている全てのリソースの数とする．この時，リソース R の Authority Score は，$x^{<R>} = \alpha$ と示される．

Step2. Resource Score 計算
Ω の要素のうち，それ自身から別のリンクが張られている要素の集合を $\Psi = \{r_1, r_2, ..., r_\beta\}(\Psi \subseteq \Omega)$ とする（図 1 参照）．ここで，Resource Score $y^{<R>}$ を注

目リソース R がリンクしている全てのリソースの中で，それ自身から別のリンクが張られているリソースの数とする．従って，リソース R の Resource Score は，$y^{<R>} = \beta(\beta \leq \alpha)$.

Step3. Step1〜2 反復

Ψ の各要素ごとに Step1〜2 を行い Authority Score と Resource Score を求める．

Step4. 特定リソース除去

Ψ の各要素の Authority Score の中央値を M，Ψ の各要素の Authority Score の標準偏差を SD とするとき，Authority Score が $M \pm 1SD$ の範囲内である要素の集合を $\Phi = \{r_1, r_2, ..., r_\gamma\}(\gamma \leq \beta, \Phi \subseteq \Psi)$ とする（図 1 参照）．

Φ の設定は，Ψ の要素のうち，Authority Score が極端に大きいリソースを除くためである．ここで予備実験により，通常，Authority Score の値はほぼ全て中央値付近に集まっていることが判明されたことから，その値を $\pm\mathbf{1}SD$ とした．Authority Score が極端に大きいリソースは注目リソース以外の多くのリソースと関係を持っているため，注目リソースにとっての重要度は低いと考えられる．これには例えば，“London” 等のような地名や，“1900 年代生まれの人物” 等のようなカテゴリを表すリソース等があてはまる．

Step5. Hub Score 計算

注目リソース R の Hub Score $z^{<R>}$ を以下のように定義する．

$$z^{<R>} = \begin{cases} \sum_{r \in \Phi} \frac{x^{<r>}}{y^{<r>}} & (y \neq 0) \\ 0 & (y = 0) \end{cases} \quad (1)$$

Hub Score は対象となるリソースが関わる他のリソース群がどの程度記述されるべき情報を持っているかを示す指標であり，リソース間関係の強さによって求める数値と定義した．今，記述されるべき情報の量は Authority Score $x^{<R>}$ で記述され，リソース間の関係の強さは，リンク数 Resource Score $y^{<R>}$ で記述されている．従って，Hub Score は注目リソース R がリンクしているリソースで Φ に属する要素がもつ Authority Score を Resource Score で割った値の総和で表わす．

Step6. Step1〜5 反復

集合 Φ の要素をそれぞれ注目リソースとして Step1〜5 の手順を繰り返し，各要素の Hub Score を求める．

例えば，図 2 において注目リソース R は 6 つのリンクを保持していることから，R の Authority Score は $x^{<R>} = 6$ となる．また，リンクしている 6 つのリソースのうち，統計情報など数値データ等リテラルを除くリソース（図 2 中，○の中に r が記されているものに相当）は 4 つであることから，R の Resource Score は $y^{<R>} = 4$ となる．同様に，リソース R とリンク関係にある各リソースの Authority Score も求める（図 2 参照）．従って，注目リソース R の Hub Score は以下の式で計算できる．

$$z^{<R>} = \frac{3}{3} + \frac{2}{1} + \frac{4}{2} + \frac{1}{1} = 6$$

注目リソース R の Hub Score が求まった段階で，注目リソースを R がリンクしている別のリソースへと変更し，同様にそのリソースについても Hub Score を求める．

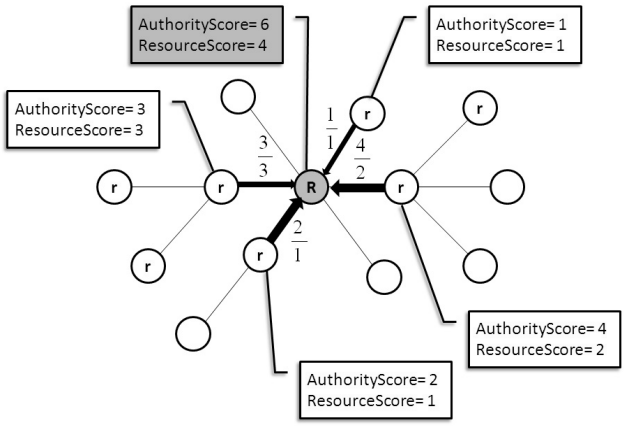


図 2: 注目リソース R および隣接するリソースの各スコア

この手順を繰り返すことにより，注目リソースとリンク関係にある全てのリソースの Hub Score を求める．

3.3 スコアを利用したリソース傾向抽出条件定義

前節で求めたスコアを利用し，以下の条件に基づいてリソースを抽出する．

条件 1.（固有な情報の発見）

$$AuthorityScore(x^{<R>}) - HubScore(z^{<R>}) \quad (2)$$

注目リソース R に隣接リソースについて式 (2) により求めた値を降順にソートした上位 t 件．

この値が大きい場合，“固有の情報” である傾向があると考えられる．なぜなら，“Authority Score が大きい” ということは “注目リソースに対する記述が多い” ということであり，“Hub Score が小さい” ということは “情報が豊富なリソースと関係を持っていない，もしくは持っている場合でも，その情報が豊富なリソースはその他大勢のリソースと関係を持っているため，注目リソースとの関係は薄い傾向にある” ということを示す．すなわち，条件 1 を満たすものは “注目リソースにとっては重要だが一般的ではない” という情報を示す傾向にあると考えられる．

条件 2.（周知の情報の発見）

$$\{HubScore(z^{<R>}) > AuthorityScore(x^{<R>})\} \wedge \{\frac{ResourceScore(y^{<R>})}{AuthorityScore(x^{<R>})} > \sigma\} \quad (3)$$

注目リソース R に隣接するリソースについて式 (3) を満たさないリソースにおける $HubScore(z^{<R>)}$ を降順にソートした上位 t 件．

ここで，$HubScore(z^{<R>}) > AuthorityScore(x^{<R>})$ は，自身からリンクを張るリソースが持つ情報量（Hub Score）が，自身が持つ情報量（Authority Score）よりも大きい．すなわち，“自身からリンクを張るリソースの情報は豊富である”

ということを示す．また，$\frac{ResourceScore(y^{<R>})}{AuthorityScore(x^{<R>})} > \sigma$ は，Authority Score に対する Resource Score の割合を閾値によって判定するためのものである．ここで，式 (3) を用いて “それ自身の情報は僅少である” もの，すなわち “Resource Score の Authority Score における割合が，通常よりも高いもの” ものを特定する．“通常よりも高い” を定義する閾値 σ は，予備実験より $\sigma = 0.6$ と定義した．従って式 (3) は，自身からリンクを張るリソースの情報は豊富であり，かつ，それ自身の情報は僅少であるリンク集のようなものを特定する．このため，式 (3) を満たすリソースはユーザに提示するためのリソースとはならない．

式 (3) は，カテゴリのような一定の特徴を持つリソースを集めるためのリソースを特定する．特定されるリソースは，リンク集のような特徴を持っている．すなわち，それ自身の情報は僅少であり，自身からリンクを張るリソースの情報は豊富である．ここで，“それ自身の情報は僅少である” とは，すなわち，Authority Score と Resource Score がほぼ同じであると言い換えることができる．従って，$\frac{ResourceScore(y^{<R>})}{AuthorityScore(x^{<R>})} > \sigma$ と表せる．ここで，“ほぼ同じ” を定義する閾値 σ は，予備実験より $\sigma = 0.6$ と定義した．また，“自身からリンクを張るリソースの情報は豊富である” は，$HubScore(z^{<R>}) > AuthorityScore(x^{<R>})$ と表せる．従って，式 (3) を満たすものは，一定の特徴を持つリソースを集めるためのリソースであると判断され，ユーザに提示するためのリソースとならない．

また，Hub Score は高ければ高い程それ自身の情報の豊富さに関わらず，情報の豊富なリソースと固有の関係を持っていることを示す．従って，式 (3) を満たさないリソースで Hub Score $z^{<R>}$ が大きいもの，すなわち条件 2 を満たすものは “誰もが知っている情報” を示す傾向にあると考えられる．

4 検証

本提案手法で抽出される知識が，実際に想定した傾向にあるかを調べるために 2 つの検証を行う．いずれの場合も Linked Data には DBpedia[††] を用いる．

4.1 リソース抽出条件に関する検証

例として，“River Phoenix（人物名）” を注目リソースとして検証を行う．

結果として，$x^{<RiverPhoenix>} = 121$，$y^{<RiverPhoenix>} = 45$，$z^{<RiverPhoenix>} = 84.04$ となり，中央値は 95，標準偏差は 764.09 となった．この時，条件 $M \pm 1SD$ により，Authority Score が極端に大きいものが除かれる．これは例えば，“River Phoenix” においては，Category:American film actors や Los Angeles, California 等が該当する．Category:American film actors はリソースが属するカテゴリであり，属するリソース数はカテゴリ毎に異なる．

条件 1 に基づく情報提示

Authority Score から Hub Score を引いた値が大きいものを順に図 3 に示す．注目リソース “River Phoenix” においてこの値が最も大きかった Metolius Oregon は River Phoenix の生まれた場所である．また，次に値が

[††] http://dbpedia.org/sparql

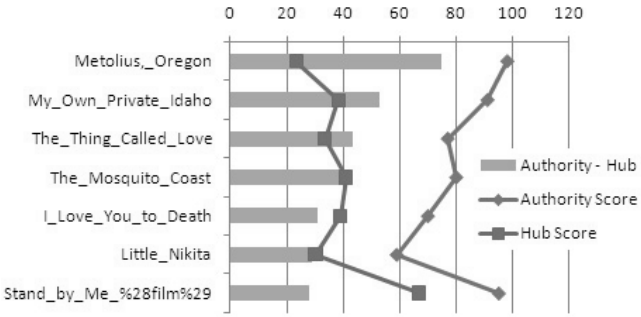


図 3: 条件 1 を満たすリソース

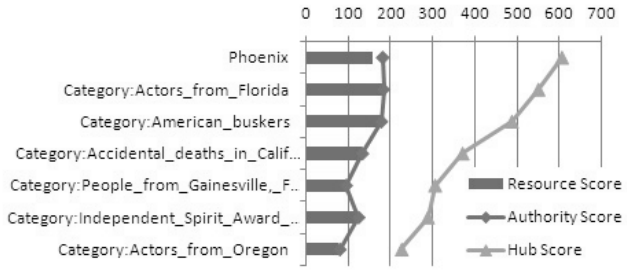


図 4: 式 (3) を満たすリソース

大きかった My Own Private Idaho は，River Phoenix が出演した映画の中では比較的知られていない異色作である．これらの情報は彼を語る上であまり頻繁に語られないものであるが，知る人のみ知っている固有の情報であることが被験者予備実験によって確認されている．

条件 2 に基づく情報提示

条件 2 の式 (3) では，リンク集のようにそれ自身には意味がない “情報を知るための手がかりとなる情報” を特定する．“River Phoenix” において，式 (3) を満たすリソースの内，ランダムに選んだ一部のリソースの各スコアを図 4 に示す．カテゴリの他，“Phoenix という名前が入っている人や物のリスト” であるリソース “Phoenix” 等も該当していることがわかる．

また，“River Phoenix” とリンク関係にあるリソースの，Authority Score に対する Resource Score の割合を図 5 に示す．

図 5 より，カテゴリ等と一般的なリソースを比べた場合，Authority Score に対する Resource Score の割合で顕著な違いが見られることがわかる．

次に，式 (3) を満たすものを除き，Hub Score 順に並べたものを図 6 に示す．“River Phoenix” において Hub Score が最も大きかった Indiana Jones and the Last Crusade は，スティーヴン・スピルバーグ監督，ハリソン・フォード主演の誰もが知っている映画と言ってよい．River Phoenix はこれに出演しているが主演ではなく，出演していること自体あまり知られておらず，River Phoenix にとって重要な映画であるとは考えにくい．次に Hub Score の大きかった Stand by Me(film) は，River Phoenix 主演の映画であり，River Phoenix が一躍有名になった作品でもある．また，社会的にも名の知れた映画であると考えられる．

このように，条件 2 を満たすものは，注目リソースにとって重要であるか判断は難しいが，少なくとも一

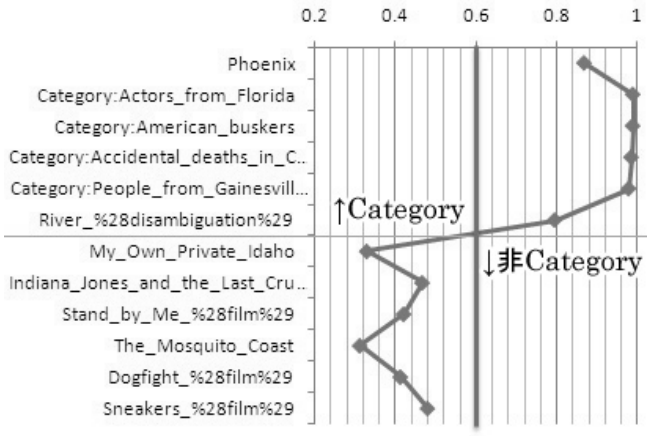


図 5: Authority Score に対する Resource Score の割合

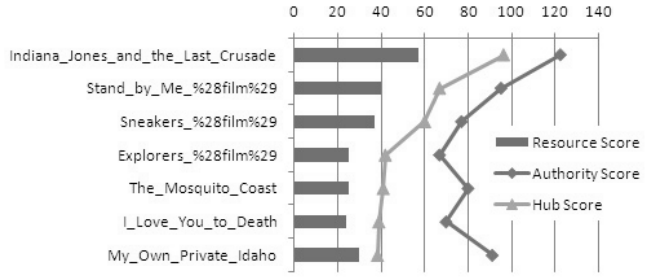


図 6: 条件 2 を満たすリソース

般的に有名かつ有益な情報である傾向を満たしていることがわかる．

4.2 注目対象像把握に関する検証

漫画を例に全体像把握の検証を行う．調査した漫画は，ドラゴンボール，One Piece，NARUTO など日本ではいずれも名の知れた 7 作品，および海外で絶大な人気を誇る X-men を加えた 8 作品である．DBpedia のリンク構造を解析することによって得た結果であるため，この結果は特に海外での各作品の全体像を写すものであると考えられる．各作品のスコアをグラフで表したものを図 7 に示す．

ドラゴンボールは，Authority Score の高さによって，それ自身に対する記述が突出して多いことが示されている．これはドラゴンボールの知名度が世界的に高いことを表している．実際にドラゴンボールのアニメーションはこれまでに 40 カ国以上で放映され，コミックスは 24 カ国以上で発売されている（2011 年 5 月現在; Wikipedia より）．

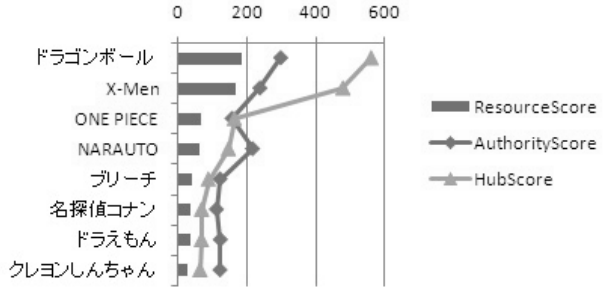


図 7: 漫画 8 作品における各スコア値

NARUTO と One Piece を比較すると，Authority Score と Hub Score の合計値が One Piece よりも NARUTO の方が大きいことがわかる．すなわち海外での人気において，NARUTO の方が One Piece よりも上であると考えられる．実際に，NARUTO は 80ヵ国以上でアニメーションの放送がなされているが，One Piece は 28 カ国以上にとどまる（2011 年 5 月現在; Wikipedia より）．

また名探偵コナンは，全体的にスコアが低い結果であった．これは日本での知名度に比べ海外での知名度が低いため，DBpedia を利用したこの調査では各スコアがそれぞれ低くなったものと思われ，海外での名探偵コナンの全体像を示すものと考えられる．

5 被験者実験

本論文で定義した二つの条件，すなわち “注目リソースにとって固有の情報を抽出する条件 1” と，“周知の情報を抽出する条件 2” が本当にその傾向の情報を抽出しているのかを検証するため，20 代から 60 代までの 17 人（男性 3 人，女性 14 人）を対象に被験者実験を行った．実験の手順は以下の通りである．

[実験手順]

Step1. 聞き取り調査

被験者から，好きな物事や詳しい物事の名前のみを聞き取り調査を行う．本調査では被験者 17 人から，オスカー・ワイルド，AKB48 等 13 個のリソースが挙げられた

Step2. リソース情報抽出

被験者によって挙げられた物事を注目リソースとして，前述した手順で情報抽出を行う．抽出は各条件を満たすもの上位 5 件とした．

Step3. アンケート調査

抽出した情報を元に，被験者毎にアンケート調査を行う．複数回行った被験者がいるため，最終的に被験者数 17 人，出力リソース 13 個でアンケートは 18 回行われた．

アンケートの作成については以下に示す．

[アンケート作成について]

以下，条件 1 によって抽出された知識を “条件 1 知識”，条件 2 によって抽出された知識を “条件 2 知識” と呼ぶ．この時，各知識に対して以下の 2 つの作業を行う．

作業 1. 注目リソースに対する記述が少なかった場合，すなわち注目リソースとリンクするリソースが少なかった場合，条件 1 知識と条件 2 知識で共通して現れるものがある．条件 1 知識と条件 2 知識で共通して現れるものは，条件 1 および条件 2 の傾向を両方持っているものと判断できるため，抽出傾向の違いを見るのに相応しくないためアンケート項目から削除するあくまで条件 1 知識と条件 2 知識で共通して現れるものを削除しているのであり，恣意的に削除しているわけではない．これは，注目リソースに対する記述が多かった場合はこの問題は起きないが，現在の Linked Data にはまだ記述が少ないリソースもあるため，一部のリソースに対して作業 1 を行う必要があった．

作業 2. “ABBA は Dancing Queen の musicalBand（property）である” と “ABBA は Dancing Queen の musicalArtist（property）である．” のように，知識として同一と思われるものは重複しないように一つの知識に纏める．これは，現在の Linked Data にはまだ情報が精査されていないものがあり，アンケートを作る際に同じ質問になることを防ぐために行った．この作業は人手で行っているが，注目リソースに対する知識のない人間が行っているため実験に有利なように進めることは不可能である．

システムで抽出されなかったリソース全ての内，誤記によるリソースのような意味のないリソース以外で，条件 1 から大きく外れたリソースを “条件 1 下位リソース”，条件 2 から大きく外れたリソースを “条件 2 下位リソース” と呼ぶ時，以下のような質問と二択の選択肢を設定する．各質問に対する想定される回答には選択肢に下線をつけた．

問 1，問 2. 世間一般ではどちらがよく知られている情報だと思いますか.
• 条件 1 知識または 条件 2 知識
問 3，問 4. 知らない情報はどちらでしたか.
※両方知っていた場合は，より世間的に知られていないと思われる方を選んでください.
• 条件 1 知識 または条件 2 知識
問 5，問 6. 注目リソースにとって，どちらがより大切で詳細な情報だと思いますか.
• 条件 1 知識 または条件 1 下位リソース
問 7，問 8. 知らない情報はどちらでしたか.
※両方知っていた場合は，より世間的に知られていないと思われる方を選んでください.
• 条件 2 知識または 条件 2 下位リソース

問 1 から問 4 は，条件 1 および条件 2 が，各々想定された抽出傾向を判断するものであるかを調査するものであり，被験者が想定される回答の選択肢を選んだ場合その傾向が強いと判断し，選ばなかった場合はその傾向が弱いと判断する．問 5 から問 8 は，条件 1 および条件 2 によって抽出された情報は，選ばれなかった情報よりも大切な情報であるか，あるいは認知度の高い情報であるかを問うものであり，被験者が想定される回答の選択肢を選んだ場合その傾向が強い，選ばなかった場合はその傾向が弱いと判断する．なお，統計的処理を行いやすくするため，質問は二つ一組とした．

アンケートの一例を以下に示す.

問 4 知らない情報はどちらでしたか.
（　）オスカー・ワイルドはジェームス・ジョイスに影響を与えた.
（　）オスカー・ワイルドはジョリス＝カルル・ユイスマンスの影響を受けた.

[実験結果]

結果を表 1 に示す．想定回答を 1 点，想定外回答を 0 点として，想定回答得点（想定回答得点＝想定回答数× 1）および想定外回答得点（想定外回答得点＝項目数× 1 −想定回答得点）を算出した.

問 1，問 2 の想定回答得点（平均 1.33，SD0.58），想定外回答得点（平均 0.67，SD1.42）を算出した．想定回答得点と想定外回答得点の差の検討を行うために，対応のある t 検定を行った．その結果，想定外回答より想定回答の方が有意に高い得点を示していた（両側検定：$t(17) = 2.38, p < .05$）.

表 1: アンケート結果

	想定回答数	想定外回答数
問 1	10	8
問 2	14	4
問 3	14	4
問 4	12	6
問 5	9	9
問 6	10	8
問 7	12	6
問 8	16	2

問 3，問 4 の想定回答得点（平均 1.44，SD0.68），想定外得点（平均 0.56，SD1.32）を算出した．想定回答得点と想定外回答得点の差の検定を行うために，対応のある t 検定を行った．その結果，想定外回答より想定回答の方が有意に高い得点を示していた（両側検定：$t(17) = 2.68, p < .05$）.

問 5，問 6 の想定回答得点（平均 1.06，SD0.78），想定外得点（平均 0.94，SD1.22）を算出した．想定回答得点と想定外回答得点の差の検定を行うために，対応のある t 検定を行った．その結果，想定回答得点と想定外得点に有意な差は見られなかった（両側検定：$t(17) = 0.29, n.s.$）.

問 7，問 8 の想定回答得点（平均 1.56，SD0.60），想定外得点（平均 0.44，SD1.40）を算出した．想定回答得点と想定外回答得点の差の検定を行うために，対応のある t 検定を行った．その結果，想定外回答より想定回答の方が有意に高い得点を示していた（両側検定：$t(17) = 3.83, p < .001$）.

[考察]

問 5，問 6 において，想定回答得点と想定外得点に有意な差は見られなかったことについて考察を行う．問 5，問 6 の設問において，“どちらがより大切か” という部分に主観的な意見を，“どちらがより詳細な情報か” という部分に客観的な意見を求めてしまった．実際は，注目リソース固有の知識だと思う方を選ばせることを目的とした質問であったが，二つの意見を一つの質問に纏めてしまったことが，この結果を招いた原因の一つであると考えられる．また選択肢にカテゴリ知識を入れたことも結果に影響を与えたと思われる．カテゴリ知識とは例えば，River Phoenix が Category:American film actors に属する場合，“River Phoenix はアメリカの映画俳優である” のような知識のことである．質問で，より詳細な情報を選ぶように指示していたものの，被験者はより大切な情報を選ぶ傾向にあった．これが，この結果を招いた二つ目の原因であると考えられる．しかしその他の質問では全て，想定回答の方が有意に高い得点を示していた．これは，本研究が提案する手法の有意性を示すものである.

6 おわりに

本研究では HITS アルゴリズムおよび PageRank アルゴリズムを RDF のリンク構造へ適用し，リンク解析することによりリソースに再利用可能な三つのスコアを与えた．これらのスコアを組み合わせることにより，注目リソースと隣接しているリソースの内，ある一定の傾向にあるリソースを特定する手法を提案した．また，その手法によって抽出したリソースは，想定した傾向にあるかの検証と被験者実験を行った．

今後は，今回定義したもの以外のスコア定義を検討し，よりリソースの全体像を把握できようなスコアを設定することを目指す．また，実験結果を踏まえ，リソース抽出条件の設定の再検討を行うと共に，抽出したリソースを利用した SPARQL による知識抽出部にも工夫を加えたい．

参考文献

[1] Christian Bizer, Tom Heath and Tim Berners-Lee: Linked Data - The Story So Far, International Journal on Semantic Web and InformationSystems, Vol. 5(3), pp. 1-22 (2009)

[2] Christian Bizer, Tom Heath, Tim Berners-Lee, 翻訳：萩野達也: Linked Data の仕組み Linked Data-The Story So Far，会誌「情報処理」, Vol.52 No.3 pp. 284-292 (2011)

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives: Dbpedia: a nucleus for a web of open data, In Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, pp. 722-735 (2007).

[4] 大西可奈子，小林一郎: Linked Data を利用したユーザの興味に基づく情報拡張手法の開発, 第 19 回 Web インテリジェンスとインタラクション研究会, WI2-2011-03, pp. 13-18 (2011)

[5] J.O.N.M. Kleinberg: Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, Vol. 46, No. 5, pp. 604-632 (1999)

[6] L. Page, S. Brin, R. Motwani, and T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford InfoLab, (1999)

[7] Delbru, R., Toupikov, N., Catasta, M., Tummarello, G., Decker, S.: Hierarchical Link Analysis for Ranking Web Data, The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC (2), pp. 225-239 (2010)

[8] Passant, A.: Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations, In Proceedings of the AAAI Spring Symposium “Linked Data Meets Artificial Intelligence”, (2010)

[9] Passant, A.: dbrec: music recommendations using DBpedia, Proceedings of the 9th international semantic web conference on The semantic web - Volume Part II, ISWC’10, pp. 209-224 (2010)

[10] Mirizzi, R., Ragone, A., Noia, T. D., Sciascio, E. D.: Ranking the Linked Data : The Case of DBpedia, Management, pp. 337-354 (2010)