3E3-OS20-3

Linked Dataを利用した対象文章の情報拡張への取り組み

Approach to Information Enhancement on Object Sentences using Linked Data

大西 可奈子[*1] 小林 一郎[*1] 岩爪 道昭[*2]
Kanako ONISHI Ichiro KOBAYASHI Michiaki IWAZUME

[*1]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
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

[*2]独立行政法人 情報通信研究機構 けいはんな研究所 知識処理グループ
National Institute of Information and Communications Technology, Knowledge Creating Communication Research Center

Recently, Linked Data has been main concern in the Semantic Web technologies and huge amount of the data has been constructed. However, there are not so many studies which develop a method to use the data. In this paper, we propose a technique to presume the meaning of links between data by taking the ideas of HITS and PageRank algorithms into the analysis of the links of the data. In concrete, we introduce three scores, i.e. Authority Score, Resource Score, and Hub Score, to analyse a target resource, and presume the meaning of links based on the values of scores. And then, we extract the information useful to users.

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年ごろにTim Berners-Lee氏によって提唱された技術[*1]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[*2] [*3]である．主要なLinked Dataとして，Wikipediaを構造化したDbpedia[Auer 07]，地理情報をLinked Dataで記述したGeonames[*4]，音楽のメタデータデータベースであるMusicBrainz[*5]，概念辞書であるWordNet[*6]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．

2. 関連研究

Linked Dataの利用法として，コンテンツをLinked Dataと結び付け，検索精度を従来よりも高める研究が数多く報告されている．例えばBBCは，BBCのコンテンツをLinked Dataで記述し，DBpediaやMusicBrainzとリンクさせるシステムを開発している[Kobilarov 09]．対象コンテンツをビデオコンテンツに特化したものとして，Waitelonisらはビデオデータのための意味検索を容易にするための手法を提案した[Waitelonis 09]．また，DBpedia Mobile[Becker 08]は，GPS情報を用いて携帯にユーザの位置情報に加えて，その位置情報に関連する情報をDBpediaから取得し，ラベルやアイコンで表示する．このような提案がなされる一方で，その有効な利用法はまだ多く報告されてない．そこで我々はLinked Dataの新たな利用法として，ユーザが興味ある事柄に対して，新たな気づきを与える様な情報の提供を目指した手法を提案する．

連絡先: 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，〒112–8610 東京都文京区大塚2–1–1，TEL:03-5978-5708，FAX:03-5978-5708，E-mail:onishi.kanako@is.ocha.ac.jp

∗1 http://www.w3.org/DesignIssues/Semantic.html
∗2 http://www.w3.org/DesignIssues/LinkedData.html
∗3 http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
∗4 http://www.geonames.org/
∗5 http://musicbrainz.org/
∗6 http://wordnet.princeton.edu/

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，それ以外の多くの物事と関係している．その知識を繋ぎ，関係を記述したものがLinked Dataである．このような知識からは様々な情報が得られる．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．Linked Dataは対象とする知識を様々な特徴において，他の知識とリンクすることにより記述し表現される．上記の様な様々な情報はLinked Dataの表現形式の中に直接記述されている訳ではない．そこで本研究では，Linked Dataのリンク構造を解析することにより，ユーザにとって興味がありそうな情報や意外と思われそうな情報を推定し，それに基づき対象知識に対して，“気づきを与える”情報を提供する情報拡張手法を提案する．

代表的なハイパーリンク解析として，まずHITSアルゴリズム[Kleinberg 99]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アルゴリズム[Page 99]が挙げられる．PageRankアルゴリズムの考え方はHITSアルゴリズムのそれに近いが，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の考え方をLinked Dataに適応し，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のリンクは HTMLのハイパーリンクとは異なり，その大きな違いとして，Linked Dataの場合，一方が関係性を示せばもう一方からも関係があると言える無向グラフで表現されるということが挙げられる．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” という関係も成り立つということを示す．すなわち Linked Data において，物事は常に相互リンク状態にある．これらの違いを考慮して，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 Linked Data の特性に合わせて以下のように定義し直した．

- Authority Score:
  対象とするリソースがどの程度，記述されるべき情報を持っているかを示す指標．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．他のリソースが対象となるリソースのみと関係を持つ場合が最も関係が強く，対象となるリソースの Hub Score の上昇幅は大きく，他のリソースがその他多くのリソースとも関係を持つ場合は，対象となるリソースの Hub Score の上昇幅は小さくなる．Hub Score が高いほど，情報が豊富なリソースと強い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.
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．

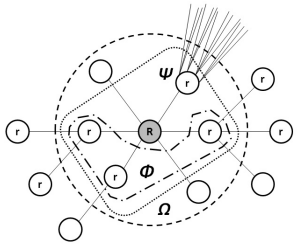

図 1: 注目リソース R を中心とした他のリソースとリンクの関係の概要

Step2.
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.
Ψ の各要素ごとに Step1〜2 を行い Authority Score と Resource Score を求める．

Step4.
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．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.
注目リソース R の Hub Score $z^{<R>}$ を以下のように定義する．

$$z^{<R>} = \sum_{r \in \Phi} \frac{x^{<r>}}{y^{<r>}} \tag{1}$$

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を，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される．

Step6.
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 $x^{<R>} = 6$. また，リンクしている 6 つのインスタンスのうち，統計情報など数値データ等を除くリソース（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 $y^{<R>} = 4$. 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．

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

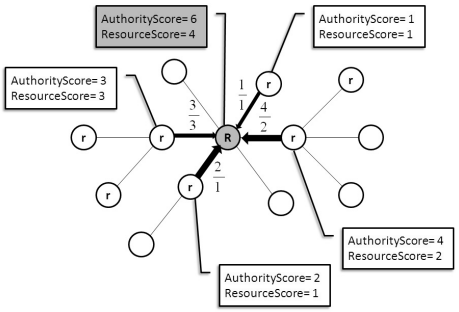

図 2: 注目リソース R および隣接するリソースの各スコア

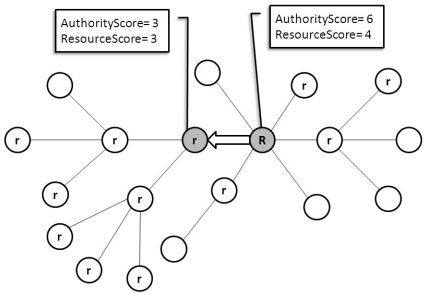

図 3: 注目リソースの変更

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注目対象を変更し（図 3 参照），同様にそのリソースについても同様に Hub Score を求める．

例えば，このリソースのリンク関係が図 4 のようになっていた場合，新たな注目リソース S の Hub Score は，

$$z^{<S>} = \frac{4}{3} + \frac{4}{4} + \frac{6}{4} = 3.8333$$

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 リソース抽出条件定義

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.

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

式 (2) を満たす値が大きいもの上位 t 件.

この値が大きい場合，“意外な情報” である傾向がある．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ない” という情報を示す傾向にある．

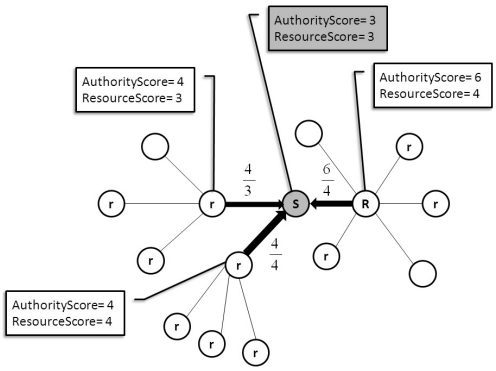

図 4: 注目リソース S および隣接するリソースの各スコア

条件 2.

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

式 (3) を 満たさない リソースで，$z^{<R>}(HubScore)$ が大きいもの上位 t 件.

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．

3.4 リソース間知識抽出

条件 1 または 3 を満たすリソースの集合（Φ の部分集合）の各要素 κ について以下のように SPARQL[*7] クエリを作成し，エンドポイント[*8]を通じて知識を取得する．

```
SELECT * WHERE {
{<注目リソース R> ?property <κ>}
UNION
{<κ> ?property <注目リソース R>}}
```

4. 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．なお，Linked Data には DBpedia を用いる．

∗7 http://www.w3.org/TR/rdf-sparql-query/
∗8 http://dbpedia.org/sparql

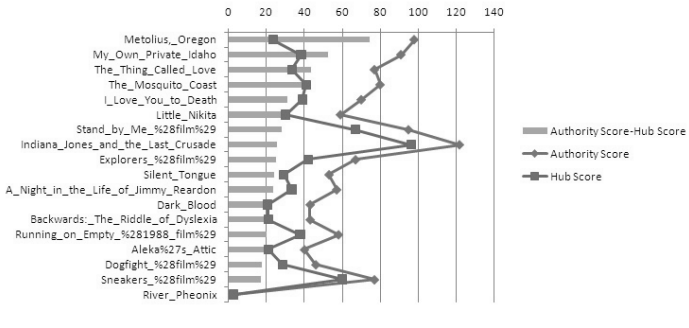

図 5: 条件 1 を満たすリソース

結果として，$x^{<RiverPhoenix>} = 121$, $y^{<RiverPhoenix>} = 45$, $z^{<RiverPhoenix>} = 84.041$ となり，中央値は 95，標準偏差は 764.097 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 5 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意外な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各スコアを図 6 に示す．カテゴリの他，“Phoenix” という “Phoenix という名前が入っている人や物のリスト” であるリソースが該当していることがわかる．

次に，式 (3) を満たすものを除き，Hub Score 順に並べたものを図 7 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

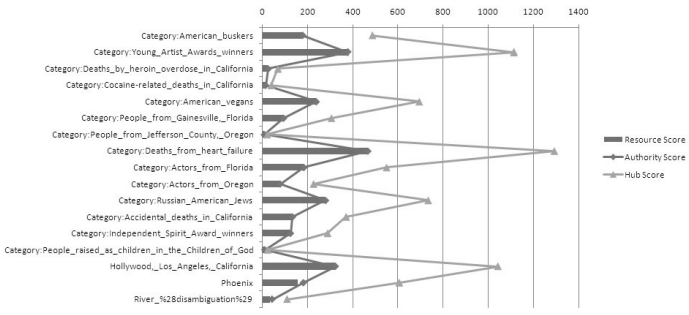

図 6: 式 (3) を満たすリソース

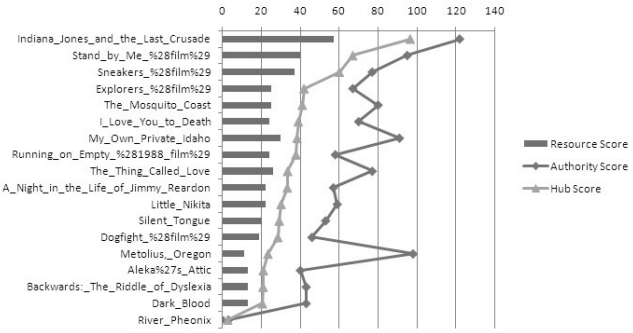

図 7: 条件 2 を満たすリソース

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，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_%28film%29 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かどうかは不明だが，少なくとも一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．

5. 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 Linked Data へ適応し，リソースからリソースへのリンクにどのような意味があるのかを推定する手法を提案した．そして，その手法によってどのような情報が抽出されるかの検証を行った．

今後の課題として，リソース抽出後の SPARQL による知識抽出方法の検討を行うと共に，被験者実験を行い，提案手法がどの程度有用であるかの調査を行いたいと考えている．

参考文献

[Auer 07] 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).

[Kobilarov 09] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, R. Lee: Media Meets Semantic Web — How the BBC Uses DBpedia and Linked Data to Make Connections, Heraklion Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications, (2009)

[Waitelonis 09] J. Waitelonis, H. Sack: Towards Exploratory Video Search Using Linked Data, 11th IEEE International Symposium on Multimedia, San Diego, CA, pp.540 - 545 (2009)

[Becker 08] C. Becker, C. Bizer: DBpedia Mobile: A Location-Enabled Linked Data Browser, 1st Workshop about Linked Data on the Web, (2008)

[Kleinberg 99] J.O.N.M. Kleinberg: Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, Vol. 46, No. 5, pp. 604-632 (1999)

[Page 99] L. Page, S. Brin, R. Motwani, and T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford InfoLab, (1999)