4B-4

Linked Dataを用いた関係抽出に基づく情報拡張

大西可奈子† 小林一郎†
†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．この問題の有効的な解決方法に成り得ると考えられるセマンティック・ウェブが注目を浴びる中，その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[1]である．Linked Dataの例として，DBpedia[2]は，Wikipediaから構造化された情報を抽出し，その情報をWebで利用可能なRDFの形にして提供している．抽出した語彙には、それぞれURIが与えられており，そのURIに語彙の概念や，固有名詞が持つ情報などが記述されている．

本研究では，このようなLinked Dataの有効な活用法として，ユーザが興味を持った文章を対象に，Linked Dataを用いてユーザに提供する情報を拡張する手法を提案する．

2 情報拡張手法

2.1 情報拡張手法プロセス

提案手法のプロセスを図1に示す．ユーザがWebページに記述されたある文章内のある一部分に興味を持った時，システムはまず，その文章の選択された範囲の内容を最もよく表わしていると考えられる名詞（“内容語”と呼ぶ）をひとつ抽出する（①）．次に，内容語に対してその他の名詞の関連の強さを求め（②），関連度が最大となる名詞（“関連語”と呼ぶ）を取得する．ここで，内容語のURIが指すRDFデータの中に関連語を含む知識を探す（③）．その知識にしたがって，システムはRDFクエリ言語SPARQLクエリを自動作成し，エンドポイントを通してLinked Dataにアクセスし（④），知識の抽出を行う（⑤）．ここで抽出される知識はRDF言語で記述されているため，必ず「関係（property）」を持つ（⑥）．本研究では，その関係を持つもので，かつ内容語を含まないものを拡張情報とする．ここでもシステムは，SPARQLクエリを自動作成し（⑦），再度エンドポイントを通してLinked Dataにアクセス，新たな知識の抽出を行う．

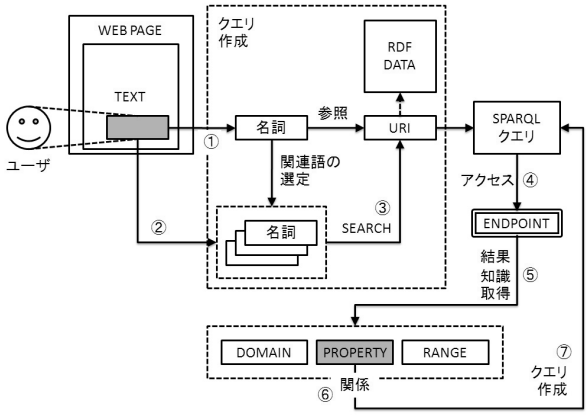

図 1: 情報拡張手法プロセス

2.2 内容語抽出

本研究では,「重要な単語は一箇所に偏らず文章全体に繰り返し現れる」という仮説の下，そのような名詞こそ，ユーザが興味を持った文章を最もよく表わしている名詞と考える．そこで，そのような名詞を抽出するために，名詞nの散らばりの程度$W(n)$を不偏分散を用いて以下の式で求める．

$$W(n) = \frac{1}{(f_D(n)-1)^{\alpha}} \sum_{i=1}^{n} (|p_{i+1} - p_i| - \bar{p})^2$$

ここで，$\frac{1}{f_D(n)-1}$をα乗しているのは，上記の仮説に基づき出現回数を重要視するためであり，αは経験的に3とする．また，$\bar{p} = \frac{X}{f_D(n)}$．$X$は文章Dに含まれる全単語数を表わす．$W(n)$は$n$が広範囲かつ均等に出現している時，最小となる．したがって，$W(n)$を最小とするnを対象文章の内容語とする．

2.3 関連語抽出

関連語とは，対象文章中に出現し，内容語と強い関連を持ち，かつ重要だと思われる名詞のことである．本研究では，内容語との関連の度合いは相互情報量を用

An Information Enhancement Technique based on the Relationship obtained from Linked Data

†Kanako ONISHI(onishi.kanako@is.ocha.ac.jp),
†Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
†Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

いて表し，重要度は単語の出現回数の平方根を用いて表す．すなわち，選択された文章中に出現するある名詞 n_l の，内容語 n_k に対しての関連語らしさ $K(n_k, n_l)$ は以下のように表す．

$$K(n_k, n_l) = I(n_k, n_l) \times \sqrt{f_D(n_l)}$$

ここで，$I(n_k, n_l)$ は内容語 n_k と関連語候補 n_l の相互情報量，すなわち $I(n_k, n_l) = \log \frac{p(n_k, n_l)}{p(n_k)p(n_l)}$ $(n_k \neq n_l)$．ここで，$p(n_k, n_l) = \frac{f_D(n_k, n_l)}{X}$，$p(n_k) = \frac{f_D(n_k)}{X}$，$p(n_l) = \frac{f_D(n_l)}{X}$．$X$ は文章 D の名詞総数を表し，$f_D(n_k, n_l)$ は n_k と n_l が同時に一文に出現する頻度を表す．$I(n_k, n_l)$ が大きいほど n_k と n_l は強い関係で結びついているとみなせる．したがって，$K(n_k, n_l)$ を最大とする n_l を内容語 n_k の関連語とする．

2.4 関係抽出

内容語と関連語の関係抽出として，内容語の RDF データに記述されている三つ組において，関連語を含む domain，property，range の数をカウントし，最多となったものを解析対象とする．すなわち，関連語が property に多く含まれていた場合は，内容語をリソース，関連語を含むプロパティを関係として，知識の抽出を行う．関連語が domain または range に多く含まれていた場合は，内容語をリソース，関連語をもう一つのリソースとし，その二つのリソースの間にある関係を抽出する．

2.5 クエリの作成

前節において関連語を含む property が解析対象であると判断された場合，ある文章から抽出される知識は「内容語 R に対して，関連語を含むプロパティP をもつリソース」になる．したがって，文章 D から抽出できる知識は，SPARQL のコマンドで表現すると，`SELECT ?hasValue WHERE {<R> <P> ?hasValue}` または，`SELECT ?isValueOf WHERE {?isValueOf <P> <R>}` で求められる．前節において関連語を含む domain または range が解析対象であると判断された場合，文章 D から抽出される知識は「内容語 R が、その他の関連語を含むリソース R' との間に持つプロパティ」である．したがって，文章 D から抽出できる知識は，`SELECT ?property WHERE { <R> ?property <R'>}` または，`SELECT ?property WHERE {<R'> ?property <R>}` で求められる．

また，上記二つのクエリから抽出された知識が持つ関係を P とする時，新たに抽出される情報は「プロパティP を持つ domain と range の組」である．したがって，文章 D に対して拡張できる情報は，`SELECT ?isValue ?hasValue WHERE {?isValue <P> ?hasValue}` で求められる．

3 ケーススタディ

ユーザが，Wikipedia の Supermarine_Spitfire の項目の一部分に興味を持ったとする．この時，その文章に対する内容語 “Spitfire”，関連語 “Fighter” に対して作成される行列 T が，条件 $(\sum T_{1j} \geq \sum T_{0j}) \wedge (\sum T_{1j} \geq \sum T_{2j})$ を満たす．よって，この文章から抽出される知識は「Supermarine_Spitfire に対して関係 `aircraftFighter` をもつ domain または range」である．結果，以下の知識（抽出されたものの一部を掲載）がユーザに提示される．

```
aircraft fighter of
No._303_Polish_Fighter_Squadron is
Supermarine_Spitfire.
aircraft fighter of No._453_Squadron_RAAF is
Supermarine_Spitfire.
```

次に拡張情報として,「aircraftFighter という関係を持つリソースの組」を得る．結果，以下の知識（抽出されたものの一部を掲載）を得，ユーザに提示される．

```
aircraft fighter of United_States_Air_Force
is F-16_Fighting_Falcon.
aircraft fighter of Turkish_Air_Force is
F-16_Fighting_Falcon.
```

4 おわりに

本研究では，ユーザが興味をもった文章から内容語と関連語を抽出し，それらの間にある関係を Linked Data を用いて抽出・提示する．また，そのようにして抽出した関係を持つ別のリソースを，Linked Data を用いて取得することにより，ユーザの興味に沿った情報拡張を行う．

今後の課題として，内容語や関連語が RDF ファイルに含まれないような単語であった場合でも，有益な情報を取得することが可能になるような手法に改良していくことが必要と考える．さらに，本研究で提案した手法の有用性の評価方法の検討と評価実験を行うことも重要であり，今後対処すべき課題として進めていくつもりである．

参考文献

[1] Berners-Lee, T.: Design Issues: Linked Data (2006), http://www.w3.org/DesignIssues/LinkedData.html

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives, ”DBpedia: a nucleus for a web of open data”, ISWC'07/ASWC'07, November 2007.