2N-6

構造/テキストWebデータ用ハイブリッド問合せ言語を用いた問合せ構築支援

安永 ゆい † 袖山 広輝 † 森嶋 厚行 ‡* 只石 正輝 †

筑波大学大学院 図書館情報メディア研究科 † 筑波大学 図書館情報メディア系/知的コミュニティ基盤研究センター ‡
JST さきがけ *

1. はじめに

近年，テキストデータとそれと並存する構造データの2つのデータの組で構成されたWebデータが広く普及しつつある．このようなデータの具体例として，Wikipediaとそれと並存するDBpediaデータ[3]の組が挙げられる．Wikipediaの各記事がテキストデータとして存在する一方，DBpediaはWikipediaの各記事に関する様々な情報をRDF形式の構造データとして提供している．SemanticWebビジョンにおけるLinked Dataの取り組み等と相まって，テキストデータとそれと並存する構造データの組からなるデータは今後ますます一般的に利用されると考えられる．

本稿では，**テキストページ**（本稿では，テキストを含むWebページなどをテキストページと呼ぶ）と**グラフデータ**(本稿では，主にRDFデータを指す)の組からなるWebデータに対する問合せ言語であるGradation問合せ言語[1](Gradation Query Language. 以下，**Gradation**)の概要を説明し，それを利用した問合せの支援について議論する．

これまで，先述のようなWebデータに問合せを行う方法には，**キーワード問合せ**(本稿では，一つ以上のキーワードと，それらをつなぐ論理演算子，及び括弧で記述された問合せをキーワード問合せと呼ぶ)を用いるか，**構造化問合せ**(SPARQLなどの構造化問合せ言語に従った問合せ)を用いるか，という2つの選択肢しか存在しなかった．キーワード問合せは，ライトユーザに普及しているが表現力が低く，Webデータのセマンティクスを利用した問合せを記述することができない．一方，構造化問合せは，高度な問合せを行えるが言語の学習コストが大きく，また，問合せ対象のデータに関する知識も必要となるため，ライトユーザが習得・利用することは難しい．したがって，ユーザにとっては，構造化問合せを使うか使わないか，というオール・オア・ナッシングの状況であった．

我々が提案しているGradationは，キーワード問合せと構造化問合せの溝を埋めることによって，高度な問合せ能力をライトユーザにとって身近なものにすることを目標に設計された問合せ言語である．Gradationではキーワード問合せをベースとし，簡単な追加記述によって構造化データに対する問合せ条件を表現することで，キーワード問合せと構造化問合せとのシームレスな融合を実現する．これにより，単純なキーワード問合せから高度な構造化問合せまでを広くカバーするとともに，“pay-as-you-go” スタイルの問合せを実現する．すなわち，問合せ記述にかけるコストや問合せ条件の精確さを，ユーザが自身の要求やスキルに応じて選択し，利用することを可能にする．

先述した通り，一般に，構造化問合せを行うためには，ユーザはクラス名や属性名などのスキーマに関する情報を知っている必要がある．Gradationはキーワード問合せと構造化問合せをシームレスに融合しているため，それらの情報をユーザに提示することで，キーワード問合せを行ったユーザが徐々に精確な構造化問合せを作成していくという過程を支援できると考えられる．本稿では，この支援手法について議論する．

支援手法はGradationを用いることが前提となる．まず，Gradationの概要を説明し，次に支援手法の説明を行う．

2. Gradation問合せ言語

Gradationは，キーワード問合せと構造化問合せとのシームレスな融合を実現している言語である．これは，キーワード問合せをベースとし，簡単な追加記述によってグラフデータに対する問合せ条件を表現することで実現している．

Gradationは，キーワード問合せのみを利用するようなユーザであっても，キーワードに加えて簡単な検索オプションをしばしば利用することに着目して設計されている．例えば，Web検索エンジンのユーザは，指定したサイトのみを検索対象とするオプション (`site:ac.jp` など) を記述して検索することがある．Gradationでも，キーワードに加えて検索オプションのような簡単な追加記述を利用することができる．この簡単な追加記述によって構造化問合せの問合せ条件を表現する．例えば，グラフデータに各人物の年齢 (age) を表すデータが含まれる場合，40歳以上のTomという人物のテキストページを取得するという問合せは，追加記述 `age>=40` を用いて `Tom age>=40` と記述する．

2.1 問合せ対象データ

Gradationの問合せ対象データは，テキストページとそれと並存するグラフデータ (RDFデータ) の組である．図1は問合せ対象データの一例である．図1中の点線は，テキストページとノードの対応関係を表す．

2.2 問合せ処理モデル

Gradationの問合せ処理モデルを図2に示す．入力はGradationで記述した問合せ q (図2左上)，問合せ対象データは2.1節で説明した問合せ対象データ S (図2右)，出力は，特別な指定がなければ，q にマッチするテキストページのURIの集合を表すリレーション R (図2左下) である．

2.3 Gradationの問合せ例

図1を問合せ対象データとし，問合せ例Q1-5を説明する．

(Q1) TomとActorという文字列を含むテキストページを取得する問合せ．これはキーワード問合せである．

```
Tom Actor
```

(Q2) 俳優 (Actor) である人物のテキストページで，かつTomという文字列を含むテキストページを取得する問合せ．

Support of Query Construction with a Hybrid Query Language for Structured and Text Web Data
Yui Yasunaga† Hiroki Sodeyama† Atsuyuki Morishima‡* Masateru Tadaishi†
Grad. Sch. of Library, Information and Media Studies, Univ. of Tsukuba.† Faculty of Library, Information and Media Science/Research Center for Knowledge Communities, Univ. of Tsukuba. ‡ PRESTO, JST*

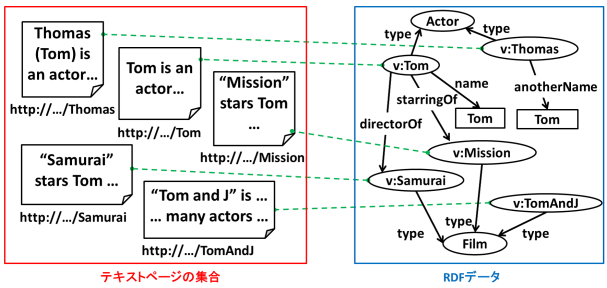

図 1 問合せ対象データの例

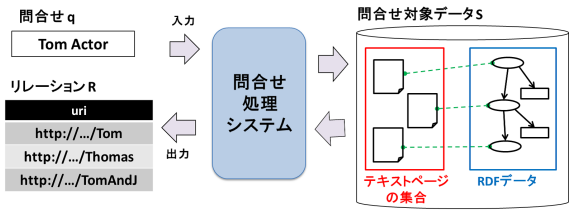

図 2 問合せ処理モデル

この問合せは，キーワード問合せの条件と構造化問合せの条件が混在した問合せである．

```
Tom  type=Actor
```

(Q3) 名前 (name) が Tom である俳優のテキストページを取得する問合せ．この問合せは構造化問合せである．

```
name=Tom type=Actor
```

(Q4) 名前が Tom である俳優と，彼と何らかの関係 (*) がある映画のテキストページの組を取得する問合せ．Gradation では，2 つのノード n_1, n_2 間の関係 rel を記述する場合は，$n_1.rel.n_2$ と記述する．この場合，検索結果はテキストページの組の集合となる．

```
(name=Tom type=Actor).*.type=Film
```

(Q5) 名前が Tom である俳優と，彼が主演を務めた (starringOf) 映画のテキストページの組を取得する問合せ．

```
(name=Tom type=Actor).starringOf.type=Film
```

以上の問合せ例のように，Gradation を用いると，キーワード問合せの条件と構造化問合せの条件を混在させ，簡易なレベルの問合せから高度な問合せまでが表現可能である．

2.4 問合せの構成要素

問合せの構成要素には，キーワード (例えば Q1 の `Tom`, `Actor`)，属性名 (例えば Q2 の `type`)，属性値 (例えば Q2 の `Actor`)，エッジ名 (例えば Q5 の `starringOf`)，クラス名，エイリアス，URI，予約語 (例えば Q4 の `(`, `)`, `=`, `.`, `*`) がある．予約語には，クラス指定やパス指定の記号，ワイルドカード，集合・大小比較演算子などが存在する．以降，構成要素のうち，予約語以外の要素をまとめて**値**と呼ぶ．

3. Gradation を用いた構造問合せ作成支援

Gradation のユーザは，単純なキーワード問合せを行い，次第により精確な構造化問合せへと変更していくことが可能である．例えば，Q1 を Q2, ···, Q5 と順に変更していくことが考えられる．本節では，この変更の過程を支援する手法について述べる．

本手法のアイデアは，問合せが与えられた際に，問合せの意味とは直接関係なくても，問合せに含まれる値がグラフデータ中にどのような形で存在するかを調べた結果 (以下，

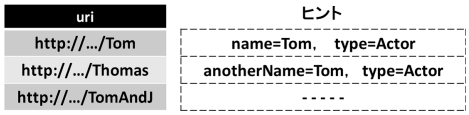

図 3 検索結果とヒントの例

ヒント) を表示することである．本手法により，ユーザは，自分が入力した値がグラフデータ中にどのように存在するのかを把握し，より精確な問合せを作成することが可能になると考えられる．図 3 は，問合せ対象データを図 1 として問合せ Q1 を実行した場合の，検索結果とヒントの例である．Q1 を書いたユーザは図 3 を見ることで，Q2 や Q3 を容易に作成可能となると考えられる．

ヒントは次のように取得する．まず，問合せ q の検索結果を得る．次に，検索結果に含まれる各 URI に対応するノード v を得る．最後に，ノード v に隣接する各ノード v_i，もしくは v と v_i 間のエッジのエッジラベル l_{v_i} が，q に含まれる値と一致するかを調べる．どちらか一方でも一致する場合，v_i および l_{v_i} をヒントとして出力する．具体例として図 3 の http://.../Tom (図 3 左 1 行目) のヒント (図 3 右 1 行目) を取得する場合を説明する．この URI に対応するノードは v:Tom (図 1) である．v:Tom と隣接するノードおよび v:Tom を端点とするエッジのエッジラベルのうち，問合せに含まれる値 (Tom, Actor) と一致するものは，ノード Tom とノード Actor (それぞれエッジラベルが name, type のエッジで繋がっている) である．したがって，図 3 右の 1 行目のヒントが出力される．

4. 関 連 研 究

ユーザとのインタラクティブなやりとりにより，キーワード問合せを元に構造化問合せを構築する研究として IQP[2] がある．IQP では完全な構造化問合せが作られるまでインタラクションが必要であり，Gradation で記述できる様な一部が構造化された問合せを構築することはできない．これに対して，本研究は，ユーザが自身の要求やスキルに応じた程度の構造化問合せの条件を記述することを支援する研究である．

5. まとめと今後の課題

本稿では，構造/テキスト Web データ用ハイブリッド問合せ言語を用いた問合せ構築支援手法を議論した．今後の課題としては，本提案手法の有効性の実験的評価，ヒント発見プロセスにおける類義語への対応などが挙げられる．

謝辞 ゼミなどにおいてコメントいただきました，筑波大学杉本重雄教授，阪口哲男准教授，永森光晴講師に感謝いたします．本研究の一部は JST さきがけ「情報環境と人」および科学研究費補助金 (#21240005) による．

参 考 文 献

1) 袖山広輝, 只石正輝, 安永ゆい, 品川徳秀, 森嶋厚行. “構造/テキスト Web データのためのハイブリッド問合せ言語”. WebDB Forum2011, 9 pages, 東京, 2011.
2) Elena Demidova, Xuan Zhou, Wolfgang Nejdl. IQP: Incremental Query Construction, a Probabilistic Approach. ICDE2010, pp.349-352, 2010.
3) “wiki.dbpedia.org : About”. DBpedia. http://www. dbpedia.org/.