意味を考慮したWeb文書検索に関する研究

M2015MM015 坪根祐治
指導教員：野呂昌満

1 はじめに

Web 文書検索には，カテゴリ検索やキーワード検索など複数の方法が提供されているが，現状では，キーワード検索が主流である．キーワード検索においては，入力されたキーワードを文字列として Web 文章内の文字列と比較し，同一文字列が存在する Web ページを検索結果とする．他方で，文字列ではなく Web ページの意味を考慮した検索を行なうことを目的として，セマンティック Web に関する研究が行われてきた [1, 4, 6, 7, 8]．セマンティック Web では，各 Web ページごとにオントロジーを用意し，そのオントロジーを用いることで意味検索を支援している [1]．

キーワード検索では，関係のない多くの Web ページが検索結果となる場合がある．一方，セマンティック Web では，すべての Web ページにオントロジーを用意することは事実上不可能である．これらの問題を解決するために他のアプローチによる意味検索（Semantic Search）の研究が行われてきた [3, 5]．Guha らは，入力されたキーワードとオントロジーを比較し，意味検索を行なっている [3]．しかし，検索者が該当のキーワードで何を意味するのか（以下，意図と呼ぶ）を平均化したオントロジーを構築して利用しているので，必ずしも検索者の意図が反映されるとは限らない．

本研究の目的は，検索者の意図を反映した Web 文章検索方法を提案することである．すなわち，検索者の意図をオントロジーとしてモデル化し，入力されたキーワードに関連するキーワードをこのオントロジーをもとに探し出す．Web ページ側にオントロジーを構築するのではなく，検索者ごとにオントロジーを構築する．検索者から入力されたキーワードをオントロジーを介して，キーワード置き換えを行なう．置き換えられたキーワードを用いて，キーワード検索を行なう．

提案する方法で，検索者の意図を適切に反映可能であることが確認できた．

2 背景技術

2.1 セマンティック Web

セマンティック Web は，Tim Berners-Lee が 1998 年に提唱した [2]．現在の Web 文書は HTML（Hyper Text Markup Language）を用いて記述されており，Web 文書の意味については記述されていない．セマンティック Web では，Web 文書の意味を示す意味データを付加し，意味検索構築の基盤を提供している [1]．

セマンティック Web のアーキテクチャを図 1 に示す [9]．図 1 のように，セマンティック Web に関する技術や仕様を階層構造で示している．最下層に位置する URI（Uniform Resource Identifier）/IRI（Internationalized Resource Identifier）は，リソースを特定するために使われる識別子である．RDF（Resource Description Framework）は，Web 文書の意味の表現形式であり，URI で場所が示されているリソースとリソース間の関係をプロパティで示している．これを図 2 に示す．この関係を本稿では基本関係と呼ぶ．RDFS（RDF Schema）にリソース間の関係やその関係の意味がプロパティの種類として定義されている．RDF は，XML（Extensible Markup Language）でシリアラズされることを念頭に置いている．

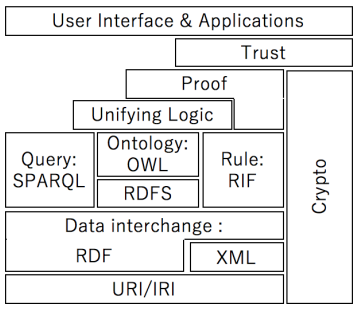

図 1 セマンティック Web のアーキテクチャ

RDF で記述した基本関係を組み合わせて Web ページの意味を記述するために，OWL（Web Ontology Language）が定義されている．RIF（Rule Interchange Format）は，複数のオントロジー間の交換ルールを定義している．

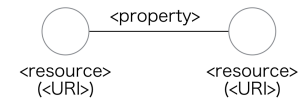

図 2 RDF

SPARQL（SPARQL Protocol and RDF Query Language）は，OWL で記述されたオントロジーに対して，検索を行なうための問い合わせ言語である．その上位層として，概念の統合を行なうための Unify Logic が定義されている．Proof 層は，Unify Logic における証明に関連する事項を定義している．最下層の URI/IRI 層から Proof 層まで横断的に，暗号化のポリシーを記述した Crypto 層が設けられている．アプリケーション層の直下には Trust 層があり，ここでは Unify Logic の記述方法が定義されている．

2.2 オントロジー

オントロジーとは，哲学用語で存在論を意味する．コンピュータ科学や人工知能の分野では，これが転じて，知識の客観的な存在を示すものとされ，諸概念の明示的な表現を指す．すなわち，概念を表現するエンティティとそれらの関係を体系化したものである．

一般にオントロジーとは，世の中の事実を概念体系として表現したものである．ここでは，これを大域オントロジーと呼ぶ．

本研究では，検索者の意図を表現したオントロジーを局所オントロジーと呼ぶ．局所オントロジーでは，概念を抽象概念と例示概念に分類する．抽象概念は“抽象概念”というエンティティでラベル付けする．概念の包含関係を記述するために subclassOf 関係を定義する．同義の関係を記述するために sameAs 関係を定義し，抽象概念と例示概念を instanceOf 関係で関連付ける．

図 3 に音楽の嗜好に関する局所オントロジーの例を挙げる．検索者にとって，“楽しい”と“好きな音楽”と“楽しい音楽”は抽象概念であり，“EDM”と“アヴィーチーの曲”と“クラブミュージック”は例示概念である．“好きな音楽”と“楽しい音楽”が同義の関係であることを sameAs 関係で，“好きな音楽”の例示概念が“EDM”であることを instanceOf 関係で表現している．さらに，“アヴィーチーの曲”も“好きな音楽”の例示概念であるので instanceOf 関係で表現している．また，“EDM”と“クラブミュージック”は subclassOf 関係にあり，“クラブミュージック”の方が上位概念なので，矢印を用いて表現している．

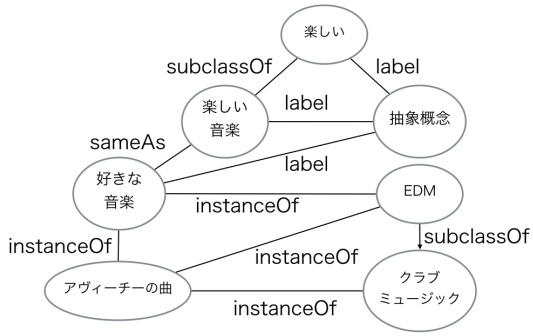

図 3　音楽の嗜好に関する局所オントロジーの例

本研究では，検索者の意図は，抽象概念が何を意味するかを示したものである．したがって，検索者の意図を考慮した検索においては，抽象概念を instanceOf 関係にある例示概念に置き換え，キーワードの組を特定することによって意味検索を行なえる．さらに，その先の例示概念を辿ることで関連しているキーワードを追加することが可能である．

3　関連研究

Web 文書の意味検索（Semantic Search）についての研究が行われている [3, 5]．Guha[3] らは，オントロジーを使ってキーワードに関連するエンティティを探し出し，キーワードを置き換える検索方法を提案している．検索結果を検索者に評価してもらい，評価結果の可否に基づいてオントロジーを洗練している．このオントロジーは，一定相当量の評価結果を基に洗練されているので，キーワードの意味を平均的に表現するオントロジーである．

Lei[5] らは，オントロジーに対して検索を行なうさいに用いる SPARQL の問い合わせの生成を試みている．検索者がキーワードとそのキーワードが属するカテゴリを入力することで，SPARQL の問い合わせを生成している．

4　提案する検索方法

4.1　検索方法

検索者の意図をオントロジーとしてモデル化した局所オントロジーを用いることで意味検索を行なう．

図 4 に検索プロセスを示す．

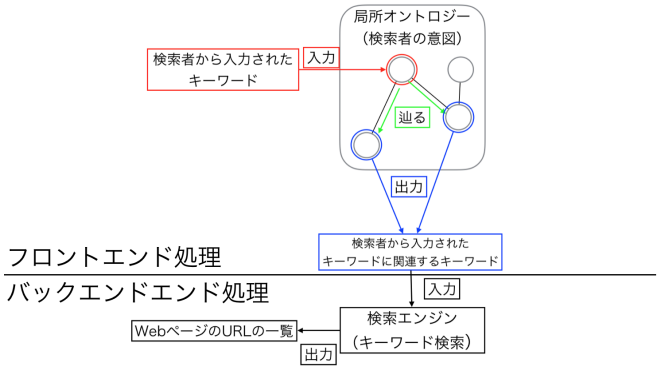

図 4　検索プロセス

フロントエンド処理では，局所オントロジーを辿ることで関連するキーワードを見つけ，キーワード置き換えを行なう．局所オントロジーの抽象概念から例示概念へ辿り，キーワードの置き換えを行なう．さらに，例示概念と instanceOf 関係または subclassOf 関係で関連付けられている 1 つ隣の例示概念へ辿り，関連するキーワードの追加が可能である．現在，instanceOf 関係または subclassOf 関係を 1 回だけ辿ることを想定しているが，これを何段階辿るかによって，検索結果が最良になるかは，実験をしなければならない．

バックエンド処理では，置き換えられたキーワードを入力としてキーワード検索を行なう．

4.2　実現

4.1 の検索プロセスの実現方法を記述する．我々は，オントロジーを記述する標準的な技術である，セマンティック Web 技術を用いて実現する．

事前に検索者の意図を局所オントロジーとして定義しているものとする．局所オントロジーを RDF と OWL を用いて記述する．図 3 の局所オントロジーを，セマンティック Web に習い，図 5 のように記述した．ここでは，instanceOf 関係を“rdf:type”で記述している．局所オントロジーの問い合わせには，SPARQL を利用する．“好きな音楽”に対する例示概念の問い合わせは SPARQL で図 6 のように記述した．

フロントエンド処理では，SPARQL を用いて局所オントロジーのエンティティを辿り，入力されたキーワードに関連するキーワードを探し出し，キーワード置き換えを行なう．フロントエンド処理での出力を図 7 に示す．

フロントエンド処理で置き換えられたキーワードを入力としてキーワード検索を行ない，Web ページの URL の一覧を検索結果とする．

この試作により，本研究の検索方法が実現可能であることを確認できた．

```
<rdf:RDF
  xmlns:owl ="http://www.w3.org/2002/07/owl#"
  xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <rdf:Description rdf:about="楽しい音楽">
    <owl:sameAs rdf:resource="好きな音楽"/>
  </rdf:Description>
  <rdf:Description rdf:about="好きな音楽">
    <rdf:type rdf:resource="EDM"/>
  </rdf:Description>
  <rdf:Description rdf:about="好きな音楽">
    <rdf:type rdf:resource="アヴィーチーの曲"/>
  </rdf:Description>
</rdf:RDF>
```

図 5　図 3 のオントロジーの RDF，OWL による記述

```
select distinct * where
{ <http://lodcu.cs.chubu.ac.jp/好きな音楽>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
?o .  }
```

図 6　SPARQL による問い合わせの記述

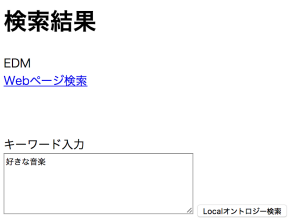

図 7　フロントエンド処理での出力

5　考察

我々は，意味を考慮した Web 文書検索を実現するための仕組みや原理を上述で示した．これらを今後，運用していく上で必要な事項について，以下に考察する．

5.1　関連研究との比較

[3] で用いられているオントロジーは，キーワードの意味を平均的に表現するオントロジーである．[5] で用いられているオントロジーは，Web ページごとのオントロジーである．特に，[3] では，一般的に含意する Web 文書を検索することができる．一方，我々は，検索者の意図に対応できるようオントロジーをモデル化している．したがって，検索者の意図は反映できるが，一般的に含意する Web 文書を検索することはできるとは限らない．一見すれば，局所オントロジーを洗練していけば，両者は，両立させることができると考えられる．しかし，本来は，平均的な意図を反映した大域オントロジーと検索者の意図を反映した局所オントロジーは矛盾するものである．例えば，先の例で，“EDM”は“好きな音楽”という意味が，一般的な嗜好でなかった場合，大域オントロジーと局所オントロジーで示す意味は互いに矛盾したものとなる．

本研究で提案する検索方法は，定義された局所オントロジーを用いてキーワードを置き換え，検索を行なっている．すなわち，局所オントロジーは変化するが，大域オントロジーは変化しないことを前提とした検索方法である．一方，[3] は，検索者の評価結果の可否に基づいてオントロジーを洗練している．これは，局所オントロジーは存在せず，大域オントロジーだけ変化することを前提としている．これらを両立させる方法を 5.4 節で議論する．

5.2　局所オントロジーの洗練

現在の検索方法では，事前に局所オントロジーを定義している．対話的に局所オントロジーを洗練する方法として，フロントエンド処理で，局所オントロジーを洗練する．以下の 2 種類の洗練が考えられる．

局所オントロジーを無から構築する場合

抽象概念のキーワードに対して，例示概念のキーワードが定義されていない場合，例示概念となるキーワードの入力を促す．入力されたキーワードで抽象概念に対する例示概念を作成し，局所オントロジーを構築する（図 8）．

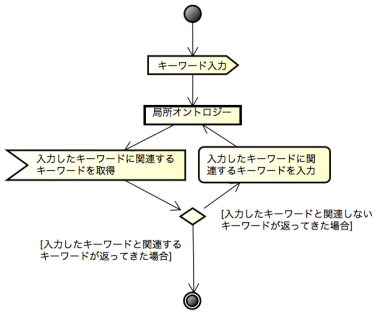

図 8　局所オントロジーを無から構築する場合

既存の局所オントロジーを洗練する場合

抽象概念のキーワードに対する例示概念のキーワードを検索者に提示する．提示されたキーワードを検索者に取捨選択させる，すなわち，検索者の嗜好に合うキーワードは残して，それ以外は削除する．取捨選択されたキーワードで抽象概念に対する例示概念を修正し，局所オントロジーを洗練する（図 9）．

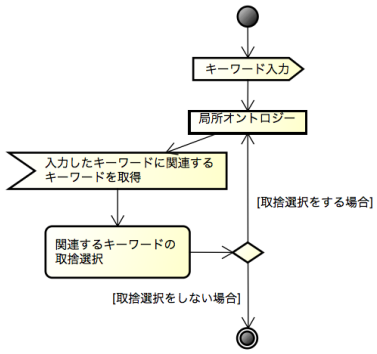

図 9　既存の局所オントロジーを洗練する場合

5.3 Web ページにオントロジーが用意されている場合の検索方法

Web ページにオントロジーが用意されているさいに，本研究の検索方法を適応することを考える．フロントエンド処理で，局所オントロジーを介して，関連する局所オントロジーの部分木をバックエンド処理の入力とする．局所オントロジーの部分木と Web ページのオントロジーの部分木の不完全一致を行なうことで，検索できると考える．

不完全一致の方法は，局所オントロジーの例示概念の部分木と Web ページのオントロジーの部分木の，2 つのエンティティとそのエンティティを結ぶ関係を比較する．一般に木の不完全一致は NP 困難問題であるが，2 つのエンティティと 1 つの関係に一致する部分を選び出すことは，P 問題となるので計算可能であると考えている．

5.4 提案する検索方法の拡張

局所オントロジーと大域オントロジーを両立させる方法として，[3] と組み合わせることを考える．フロントエンド処理として本研究の検索方法を用いてキーワードの置き換えを行ない，バックエンド処理で [3] の検索方法を用いて意味検索を行なう．

組み合わせた検索方法を用いて，両者を洗練することができると考えている．フロントエンド処理では，5.2 節で述べた洗練方法で局所オントロジーだけを洗練する．バックエンド処理では，検索者に検索結果を評価させ，その評価結果を用いて，局所オントロジーを洗練すると同時に，バックエンド処理で用いた大域オントロジーも洗練する（図 10）．

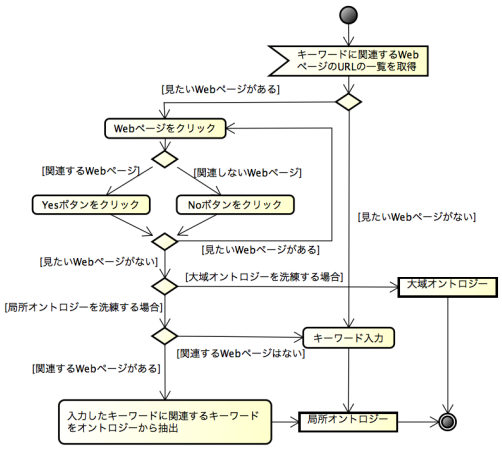

図 10 検索結果の検証によるオントロジーの洗練

局所オントロジーの洗練は，検索結果の中に関連する Web ページがある場合，バックエンド処理で用いた大域オントロジーから入力したキーワードに関連するキーワードを抜き出して，検索者に提示する．検索者は提示されたキーワードに対して，検索者の嗜好に合うキーワードは残して，それ以外は削除することで，キーワードの抽象概念に対する例示概念を修正し，局所オントロジーを洗練することができる．大域オントロジーは，[3] のオントロジー洗練の仕組みを用いて，検索者の評価結果を集計し，それらを解析することで洗練する．

[3] と組み合わせて，フロントエンド処理の結果を局所オントロジーの洗練に用い，バックエンド処理の結果を両者の洗練に用いることで両立ができると考えている．

両者の収束については，相当数の検索を通じて，洗練することが必要と考える．

6 おわりに

本研究では，検索者の意図を反映した Web 文書検索方法を提案した．検索者の意図を局所オントロジーとしてモデル化し，局所オントロジーを介することでキーワード置き換えを行なった．置き換えられたキーワードでキーワード検索を行なうことで，適切に検索者の意図を反映可能であることが確認できた．

参考文献

[1] Amato, Flora, et al, ”An RDF-based framework for Semantic Indexing of web pages,” Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on. IEEE, 2013.

[2] Berners-Lee, Tim, James Hendler, and Ora Lassila, ”The semantic web, ” Scientific american, pp.28-37, 2001.

[3] Guha, Ramanathan, Rob McCool, and Eric Miller, ”Semantic search,” Proceedings of the 12th international conference on World Wide Web, ACM, 2003.

[4] Jain, Vishal, and Mayank Singh, ”Ontology based information retrieval in semantic web: A survey,” International Journal of Information Technology and Computer Science (IJITCS), vol.5, no.10, pp.62, 2013.

[5] Lei Yuangui, Victoria Uren, and Enrico Motta, ”Semsearch: A search engine for the semantic web,” International Conference on Knowledge Engineering and Knowledge Management, Springer Berlin Heidelberg, 2006.

[6] Menemencioglu, Oguzhan, and Ilhami M. Orak, ”A Review on Semantic Web and Recent Trends in Its Applications,” Semantic Computing (ICSC), 2014 IEEE International Conference on. IEEE, 2014.

[7] Singh, Gagandeep, and Vishal Jain, ”Information Retrieval (IR) through Semantic Web (SW): An Overview,” arXiv preprint arXiv:1403.7162, 2014.

[8] Sudeepthi, G. Anuradha, and M. Surendra Prasad Babu, ”A survey on semantic web search engine,” IJCSI International Journal of Computer Science Issues vol.9, no.2, 2012.

[9] World Wide Web Consortium, ”sw-stack-2009,” http://www.w3c.it/talks/2009/athena/images/layerCake.png, 2009.