3E3-OS20-6

セマンティックWebにおける知識発見プロセス

Data Mining Process for Semantic Web

市瀬 龍太郎[*1] Ryutaro Ichise
Kappara Venkata Narasimha Pavan[*2] Venkata Narasimha Pavan Kappara
Vyas O.P.[*2] O.P. Vyas

[*1]国立情報学研究所 National Institute of Informatics
インド情報技術大学アラハバード校[*2] Indian Institute of Information Technology Allahabad

We discuss about data mining process for semantic web based on the traditional KDD process, and describe a novel data mining process which is adjusted for semantic web.

1. はじめに

ネットワークの高度化，コンピュータの高性能化に伴い，多様なデータがさまざまな場所に蓄えられるようになってきている．このようなデータは，一般的に，特定の用途で用いられることが多いため，それぞれの用途に応じて，個別に分散して管理されていることが多い．しかし，このような多様なデータを統合して扱うことにより，さまざまなことが可能となる．例えば，地理情報とのレストラン情報，レストランの評判情報を統合して扱うことにより，どのようなレストランがどの地域で人気があるのかなどの知識を発見することが可能となる．このようなことを容易に実現するために，これまでに，セマンティック Web [Berners-Lee 01] の研究が精力的に行われてきた．セマンティック Web では，データに意味的な情報を付与することによって，データ間の自動的な連携を可能とする．そのようなデータとして，近年，Linked Data [Christian 11] に注目が集まっている．Linked Data は，RDF で記述されたデータに対して，それぞれのデータを意味記述されたリンクで結びつけることによって，データの連携を試みる．すでに Linked Data として公開されているデータセットには，論文の書誌情報などの出版関連のデータ，薬や病気などに関する生命科学のデータ，音楽の情報などのメディアのデータ，公的機関が持つ統計情報などの政府のデータなどがある．それらのデータを統合して扱うことができれば，さまざまな知識を新たに見つけ出すことが期待できる．しかし，これらのデータから，新たな知識を発見しようとする際には，データの選別などの処理が必要となり，現状においては，容易に知識発見に利用できる状況ではない．そこで，本研究では，セマンティック Web における知識発見に必要なプロセスを，これまで研究されてきた知識発見プロセス [Fayyad 96] に基づき議論する．また，その考察に基づき，セマンティック Web で利用可能な知識発見プロセスを提案する．

2. セマンティック Web における知識発見

巨大なデータから知識を発見するために，データマイニングの研究が長年なされてきた．データマイニングにおける知識発見のプロセスは，Fayyad らにより，定式化されている [Fayyad 96]．そのプロセスは，図 1 のようになっている．まず，最初に，知識発見のための領域を理解し，目標の設定を行う．次に，その目的に必要なデータの収集，選択を行う．次に，データに対して前処理を施し，データの離散化などを実施する．次に，パターン発見手法に合わせたデータ変換を行う．そして，特徴的なパターンを発見し，それを評価，解釈することで，有用な知識を抽出する．

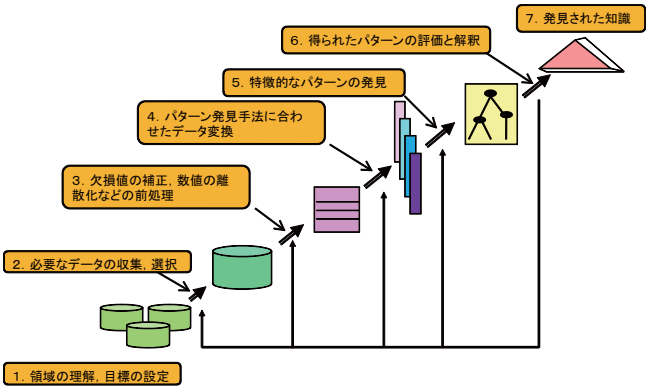

図 1: データベースからの知識発見プロセス

本研究では，Fayyad らの提案している知識発見プロセスをセマンティック Web で使われる Linked Data に対応させて，必要な技術を考察する．まず，図 1 の 1 ステップ目の，領域の理解，目標の設定は，Linked Data を用いるか否かに関わらず，データと独立に決定されるものであると考えられる．ステップ 2 の必要なデータの収集，選択では，分散して存在する Linked Data から，必要な部分を集約する必要がある．Linked Data では，必要なデータを SPARQL を使うことで，動的に取得することが可能である．しかし，現状においては，そもそも必要なデータがどこにあるのか，どのような語彙を使って探すことができるかなどが，分かっていなければならないため，それらを効率的に見つけるための技術が必要となる．また，このステップにおいては，必要なデータの統合を行わなければならない．たとえば，それぞれの国の人口のデータと穀物生産高のデータがあった時には，国名をもとにデータの統合を行わなければならない．それぞれのデータ間に，国名が一緒だとの RDF リンクがあれば，データの統合は容易に実行可能であるが，そのような場合でなければ，オントロジー写像 [市瀬 07] のような，データ統合技術が必要となる．このようにして生成された統合データには，不要なデータも含まれるため，データのフィルタリング機能なども必要となる．ステップ 3 では，欠

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，E-mail:ichise@nii.ac.jp

損値の補正や数値の離散化などの前処理が行われる．データマイニングにおいては，傾向や特徴をつかむために，詳細なデータの一般化をおこなわれることが多い．たとえば，人間を対象にしたデータでは，年齢の数値データを直接用いるのではなく，20 歳未満，20 歳代，30 歳代などに分け，データの一般化がおこなわれる．しかし，Linked Data では，詳細データが提供されていることは多いが，このようなカテゴリ化などに関するデータは，あまり提供されておらず，さらに，データマイングでは，発見する知識にカテゴリ化が依存する部分も大きいので，現状では自動的な対応は難しいと考えられる．ステップ４では，データマイニング手法に合わせてデータの変換を行う．これまでに様々なデータマイニング手法が開発されている．述語記述を用いた帰納論理プログラミングなどのデータマイニング手法も開発されているが，一般的には，表形式のデータに変換してデータマイニングをすることが多い．ステップ５では，特徴的なパターンの発見を行う．この段階で使うパターン発見手法としては，決定木や SVM などの一般的な手法が主に考えられる．しかし，Linked Data では，グラフ的な構造をデータが持っているため，そういった特性を利用したマイニング手法なども考えられるであろう．ステップ６の評価，解釈，および，ステップ７の発見された知識に関しては，Linked Data における特徴よりも，むしろ，対象領域における特徴の方が影響が大きいと考えられる．

3. Linked Data データマイニング

前章では，セマンティック Web における知識発見に必要な技術，課題に関して考察を行った．その考察に基づき，セマンティック Web の Linked Data を対象とする新たな知識発見プロセスの構築を行い，それに基づくツールの実装を行った [Kappara 11]．そのモデルは，図 2 のようになる．前章で考察したように，図 1 のステップ２は，SPARQL によるデータの収集と統合に分けられる．データの統合は，意味的な要素が絡むため，頻繁にユーザのインタラクションが必要となる．そこで，提案するモデルでは，図 2 のように，データ収集の部分と統合の部分を分離した．また，データのフィルタリングの機能も取り入れた．さらに，図 1 のステップ３にあるデータ離散化の機能も，前章で述べたように，ユーザとのインタラクションが多く必要であると思われるため，ユーザのインタラクションが多く必要な部分を Data Preprocessing として，まとめて，これらの作業を繰り返すことで，知識発見に必要なデータを作成することとした．そして，図 2 の提案モデルでは，図 1 のステップ４で行っていた，データマイニングツール用にファイルを変換するステップを経て，ステップ５のデータマイングを実施し，その結果を表示することで，ステップ６とステップ７を実施できるようにした．

図 2 のモデルに基づき，実際のデータマイニングツールの実装を行った．図 2 のデータマイニングには，Weka [Witten 05] を利用し，そのために，SPARQL でシステムに取り込んだデータを，ユーザインタラクションを通して，arff フォーマットに変換して利用するようにした．実装に関する詳細は，[Kappara 11] を参照してもらいたい．

4. おわりに

本論文では，セマンティック Web における知識発見プロセスを，従来から用いられてきたデータベースからの知識発見プロセスと対比させることで考察を行った．その結果，セマンティック Web で用いられるデータには，いくつかの特徴があるため，ユーザインタラクションが重点的に必要になる部分のあることが明確となった．そして，その特徴に基づいたセマンティック Web における知識発見プロセスを構築し議論を行った．今後は，このモデルに基づいて実装したシステムを利用して，その知見からフィードバックを得ることで，より実用的なモデルへ改良していく予定である．

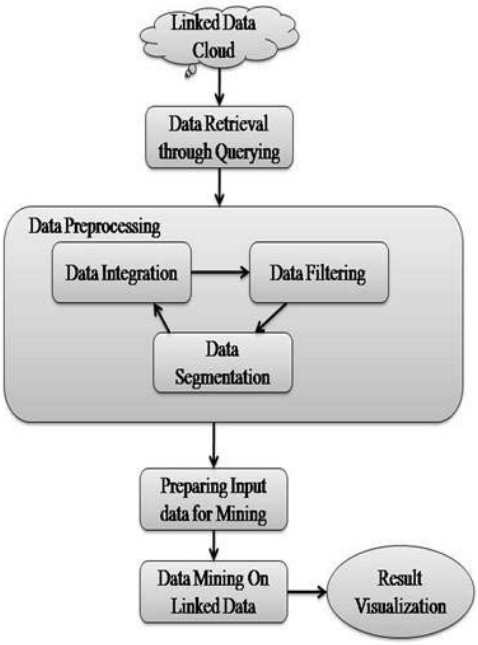

図 2: セマンティック Web における知識発見プロセス

参考文献

[Berners-Lee 01] Berners-Lee, T., Hendler, J., and Lassila, O.: The semantic web, *Scientific American*, Vol. 284, No. 5, pp. 34–43 (2001)

[Christian 11] Bizer, C., Heath, T., Berners-Lee, T., 翻訳：萩野達也：Linked Data の仕組み, 情報処理, Vol. 52, No. 3, pp. 284–292 (2011)

[Fayyad 96] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, Vol. 39, No. 11, pp. 27–34 (1996)

[Kappara 11] Kappara, V. N. P., Ichise, R., and Vyas, O.: LiDDM: A Data Mining System for Linked Data, in *Proceedings of the WWW 2011 Workshop on Linked Data on the Web* (2011)

[Witten 05] Witten, I. H. and Frank, E.: *Data Mining - Practical Machine Learning Tools and Techniques*, Elsevier, San Francisco, CA, USA, 2nd edition (2005)

[市瀬 07] 市瀬 龍太郎：情報の意味的な統合とオントロジー写像, 人工知能学会誌, Vol. 22, No. 6, pp. 818–825 (2007)