web リソースからのメタモデル導出法に関する提案

代表者　鈴木裕利　　研究協力者　伊藤誠　石井成郎

1. はじめに

現在，ICT 技術の急速な発展により，web に含まれる情報量は爆発的に増加しつつある．とりわけその増加量は 21 世紀に入り著しく，IDC は 2020 年に 40 ゼッタバイトにも達すると予想している[1]．さらに，ネットショッピング，ブログ，SNS，スマートフォンの普及等により，創出される情報の内容も一層多様化が進み，その情報の利用者であるインターネット利用者数，及び，人口普及率も増加している[2]．この増加した利用者をターゲットとして，企業が情報宣伝等の形態として，web サイトを利用する機会が増えていることも情報量の爆発的増加に拍車をかけている[3]．この情報爆発の解決への一歩は必要な情報を容易に取得できる検索技術の高度化であり，実際に，web ナビゲーション，マイニング等，関連技術の研究も多く行われている．しかしながら，現在，web 上の膨大な情報量が要因となり，web マイニングの精度が高いとはいえない状況である．そこで本研究では，web マイニングの精度を高めるために，対象となるリソースが一定の規則で記述されている Wikipedia に着目して，その活用を検討する[4]．

2. 目的

現状における web 検索は，検索者が入力する検索キーワードに対して，web 上のどのリソースが有効かについての判断に重点をおいている．本研究でのリソースとは，検索目的を達成するための要素，または，必要となる要素の意味である．しかし，この方法では，キーワードに対する関連性のみが評価されるために，web 上のリソースの内容に対する多面的な評価が不可能である．そこで，本研究は，web 上，および，現実世界で取り上げられている事象の関連を整理し，知識階層での事象の関連を導き出し，検索行為に対する多面的な結果の提供，効率的な web マイニングシステムの提供を目的とする(図 1 参照)．そして，Wikipedia を対象のリソースとして，その特徴を活用する．本研究は，第 1 段階として web マイニングの基盤構築，第 2 段階として構築された基盤を利用した web ナビゲーションの提供の実現を目指して進める．

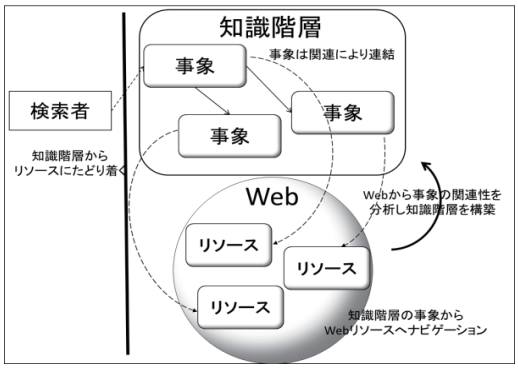

図 1　本研究の提案システム

本稿では，第 1 段階の web マイニングの基盤構築において，Wikipedia の構文を分析，解析したデータをデータベースに永続化する知識階層の構築システムについて，web リソースからのメタモデル導出法に関する提案として報告する．

3. 調査・解析

本章では，Wikipedia の構造に関する調査，解析について説明する．

Wikipedia の構造を明確にする際に，メタデータを有益な情報として利用できることが重要である．よって，その意味が共通の認識となっている語彙が必要である．そのため Dublin Core における Wikipedia の共通の認識，web や文書の作者，タイトル，作成日等の書誌的な情報をメタデータとして記述するためのボキャブラリとして定める必要がある．

本研究ではメタデータを，Wikipedia にある膨大な量のデータから目的のデータを取得するために作成する．個々の情報にメタデータを付加することにより，データの性質を的確に反映した検索が可能となる．本研究では，基本的なメタデータを規格に合わせて整形し，どのようなメタデータ記述方式にも対応可能に設計する．

Wikipedia のデータを解析するには，主として 3 つの手法がある．HTML を収集して解析する手法，最近更新された内容として配信される RSS を解析する手法，さらに，無償で公開されているダンプデータを解析する手法である．本研究では，ダンプデータの解析を行う手法を導入する．無償で公開されているファイルは言語別に jawiki-latest-pages-meta-current.xml という形で用意されている．このファイルは XML 形式であり，ユーザが独自のタグを指定可能である．また XML 言語は，タグを用いた規則的な記述であるために，構造把握が容易であり，膨大な量のデータでも，構造の理解を行えば，解析分析が容易である．

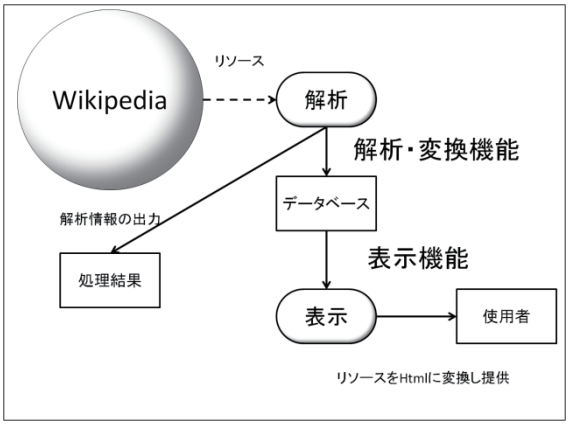

図 2　構築システム

4.　構築システム

本章では，前述した調査結果を踏まえて構築した提案システムについて言及する（図 2 参照).

第 1 の機能は，解析・変換機能である．これは，Wikipedia の構文解析を行い Wikipedia に存在するリソースを取得した後，それらのリソースに含まれる情報に基づいて，知識階層を構築するリソース選択してデータベースに変換する機能である．第 2 の機能は，データベースに変換されたリソースを表示する機能である(図 3 参照).

5.　実験・結果

web マイニングの対象となる Wikipedia のデータが膨大なため，データベースに構築された知識階層のデータが有用かどうかについての確認が十分とはいえない．そこで本研究では，知識階層の有用性の確認のために，データベースに変換されたリソースの件数とそのデータの内容を取得するプログラムを実装して，有用性の評価を行った．テストデータとして 13752 件のリソースに対して評価を行った結果，Base Resource については，文字数の制限による 4 件のデータを除いた 13748 件のデータ内容の取得が確認された.

6.　おわりに

今後は，取得したデータの有用性について，より詳細な評価を実施して，その結果に基づいてシステムの改善を行う．その後, web ナビゲーション機能の実装，実装システムの評価，改善，そして，web ナビゲーションの公開を目的として研究を進める予定である.

参考文献

[1] International Data Corporation, http://www.idc.com/
[2]総務省，情報通信白書 http://www.soumu.go.jp/johotsusintokei/whitepaper/
[3] 喜連川 優，情報爆発のこれまでとこれから，電子情報通信学会誌 Vol.94，No.8，pp.662-666(2011)
[4] 中山浩太郎，伊藤雅弘，Erdmann Maike，白川真澄，道下智之，原隆浩，西尾章治朗，Wikipedia 研究のサーベイ，情報処理学会論文誌データベース Vol.2No.4，pp.49-60(2009)

URI	Subject	Description	TemplateID
http://ja.wikipedia.org/wiki/九州帝国大学	九州帝国大学	#REDIRECT 九州大学 Category日本の旧制大学きゅうしゅうていこく	2024
http://ja.wikipedia.org/wiki/東京帝国大学	東京帝国大学	#REDIRECT 東京大学 Category日本の旧制大学とうきょうていこく	2025
http://ja.wikipedia.org/wiki/北海道帝国大学	北海道帝国大学	#REDIRECT 北海道大学 Category日本の旧制大学ほっかいどうていこく	2026
http://ja.wikipedia.org/wiki/京都帝国大学	京都帝国大学	#REDIRECT 京都大学 Category日本の旧制大学きょうとていこく	2027
http://ja.wikipedia.org/wiki/東北帝国大学	東北帝国大学	#REDIRECT 東北大学 Category日本の旧制大学とうほくていこく	2029
http://ja.wikipedia.org/wiki/名古屋帝国大学	名古屋帝国大学	#REDIRECT 名古屋大学 Category日本の旧制大学なごやていこく	2031
http://ja.wikipedia.org/wiki/大阪帝国大学	大阪帝国大学	#REDIRECT 大阪大学 Category日本の旧制大学おおさかていこく	2032
http://ja.wikipedia.org/wiki/帝國大学	帝國大学	#REDIRECT 帝国大学	2082
http://ja.wikipedia.org/wiki/	北海道	#REDIRECT 北海道大学	2083

図 3　変換リソースの表示機能