Wikipediaのセクションを考慮したリンク解析による関連項目検索手法の提案

近藤 直樹† 渡辺 陽介†† 横田 治夫††

†東京工業大学工学部情報工学科 ††東京工業大学学術国際情報センター

1 はじめに

近年、Wikipedia[1] のような不特定参加者による百科事典サイトが普及した。これらの主な特徴は、誰にでも簡単に執筆、編集できることである。この特徴により、百科事典の項目数は爆発的に増えている。また、他の特徴として他の項目とリンクすることによって項目中の語を簡単に調べることができる。

これまでに、関連項目をキーワードの共起やリンク解析などで探す手法は提案されている。本研究では、キーワード検索や共起性では探すことの難しい項目、つまり項目内に語やリンクとして出てこない項目を探すことを目標とする。また、項目を対象としてリンク解析をすると、出てくる候補が多すぎるため、欲しい情報が探せない場合が増えている。Wikipediaの各項目はセクション毎に内容が整理されており、セクションの情報を用いることで利用者の検索意図に合った関連項目の絞り込みが可能である。

本研究では、リンク関係に基づく重要度を用いて、百科事典サイトにおける項目内のセクションの関連項目を検索する手法を提案する。その際、項目間のリンクをセクション間のリンクに変換する処理を行う。次に、重要なリンクに重みを付けてリンクを解析する。

2 百科事典サイト

インターネット百科事典例として、Wikipedia[1]、はてなキーワード [2]、Yahoo!百科事典 [3]、コトバンク [4] などがある。これらの共通点は、項目中にサイト内リンクが含まれていることである。項目中のキーワードにリンクがあることによって、そのキーワードの項目へ容易に移動できる。本研究では、百科事典サイトにリンクだけでなくセクション構造を必要とする。実験対象としてWikipediaを用いるが、提案手法は条件に合う百科事典サイトなら適用可能である。

3 提案手法

3.1 目的

百科事典のセクションで関連項目を探すために、項目をセクション単位に分けて考える。図1左では、項目A全体の関連項目として探すと欲しい項目が含まれているが、利用者にとってあまり興味のない方面の項目も多く含まれる。そこで、図1右のように利用者の興味のあるセクションBに限ると、欲しい項目を探しつつ欲しくない項目を減らすことができる。

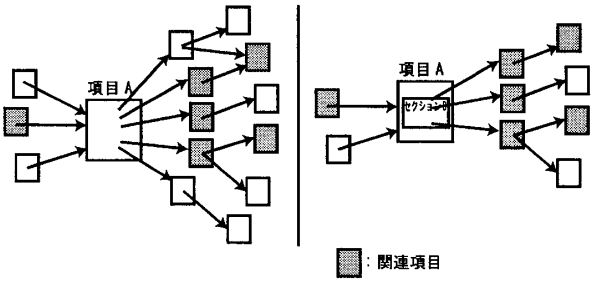

図 1: 項目単位 (左) とセクション単位 (右) のリンク解析

本研究では、利用者が閲覧している百科事典サイトの項目Aのセクション Bに対してセクション内から直接リンクのないものも含めて関連が強いセクションの関連度に従ったリストを提示することを目的とする。このため、対象セクション内のリンク先を探すだけではなく、その項目をリンクしているセクションや、複数ステップ先のリンクのセクションも関連項目として扱い、関連度を算出し、順位付けを行う。

3.2 アプローチ

セクションの関連項目を探すため、元のセクションからのリンクは重要であり、重みを付ける。また、セクション中のリンク先の項目以外の関連項目を探し出すため、元のセクションからout-リンクで距離 d のセクションと元のセクションのin-リンク先のセクションをリンク解析する。しかし、リンクの多くは項目全体を指していて、本来どこのセクションを指しているのかを考慮する必要がある。そこで、項目へのリンクはリンク先のすべてのセクションへのリンクとみなし、あるセクションと相互リンクする場合、そのセクションが本来のリンク先と考え、他のセクションへのリンクより重みを高くする。

このようにして得られる重み付きグラフに対して、リンク解析をする。本研究では、Authorityの高い項目を探したいのでHITS[5] を用いる。

Proposal of an Extraction Method for Related Articles using Link Structure and Content Structure in Wikipedia

Naoki KONDOH†, Yousuke WATANABE††and Haruo YOKOTA††

†Faculy of Engineering, Department of Computer Science, Tokyo Institute of Technology ††Global Scientific Information and Computing Center, Tokyo Institute of Technology

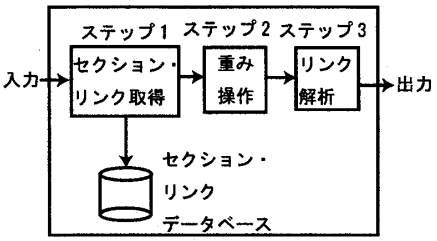

図 2: システム概要

4 システム

システムの概要を図 2 に示し、以下簡単に説明する。

百科事典の項目内のセクションの関連項目を探したいので、本システムへの入力は項目名とその項目内のセクション名である。出力はリンク解析の上位 n 件の項目名とセクション名のペアをリストとして返す。

ステップ 1 では、リンク解析をするために、データベースからセクションとリンク情報を取得する。このデータベースは、事前に辞書データを解析して構築する。データベースには、辞書の項目、セクション、リンクの情報が格納されている。

取得する範囲は、入力セクションから out-リンクを距離 d でたどれるセクションとそのセクションの in-リンク先のセクションである。その際、セクションはトップレベルのセクションを考え、下位のセクションはそのトップレベルのセクションの中に属しているとみなす。

ステップ 2 では、リンク解析のための隣接行列を作成する。その際、入力セクションの out-リンクに重み ow、セクションを特定する相互リンクに重み cw を付ける。

ステップ 3 では、ステップ 2 で求めた隣接行列を用いてリンク解析を行う。ここでは、HITS[5] を適用し、Authority の高い n 件を出力する。

システムの詳細については、文献 [6] を参照されたい。

4.1 出力例

辞書に Wikipedia 日本語版 (2009 年 11 月 17 日時点) を使った場合の出力例を示す。入力を項目「デフレーション」のセクション「デフレスパイラル」とし、重みを $ow = 10, cw = 100$ とする。出力の表示はセクション単位でも項目単位でもできる。ここでは見やすさのため項目単位で表 1 に示す。

上位は入力セクション内にリンクがある項目である。その中でランキングされていて、項目の関連度が分かる。また、4,8 位に入力セクションに直接リンクのない項目がある。これらも入力セクションに十分関係している。15 位以降にも直接リンクがない項目が出力されたが、これらも入力セクションと関連している。このように、直接リンクがなくても関連項目を探すことができた。

なお、14 位以内に入力セクション内にリンクがある項目がすべて入っている。これは、入力セクションの out-リンクを重視したことによる影響と考える。今後は、パラメータを変えた実験をする必要がある。

表 1: 出力例

	項目名	入力からリンク
1	市場原理主義	あり
2	小さな政府	あり
3	新自由主義	あり
4	クラウディングアウト	なし
5	世界金融危機 (2007 年-)	あり
6	サッチャリズム	あり
7	レーガノミックス	あり
8	グローバル資本主義	なし
9	ワシントン・コンセンサス	あり
10	物価	あり
11	金利	あり
12	累進課税	あり
13	設備投資	あり
14	ビルト・イン・スタビライザー	あり
15	民営化	なし
16	リバタリアニズム	なし
17	インフレーション	なし
18	新保守主義	なし
19	規制緩和	なし
20	経済政策	なし

5 おわり

本論文では、百科事典サイトの関連項目検索をセクション単位でするために、相互リンクに着目し重み付けしたグラフに対して HITS を適用する方法を提案した。この結果、セクションからリンクされている項目の重要度の分析とセクションからリンクされていない関連項目の発見が可能となる。

今後の課題として、パラメータを変更させた実験をする必要がある。また、さまざまな項目に対する相対的な評価も行う予定である。

謝辞

本研究の一部は文部科学省科学研究費補助金特定領域研究 (#21013017) の助成により行われた。

参考文献

[1] Wikipedia 日本語版, http://ja.wikipedia.org/

[2] はてなキーワード, http://d.hatena.ne.jp/keyword/

[3] Yahoo!百科事典, http://100.yahoo.co.jp/

[4] コトバンク, http://kotobank.jp/

[5] J.Kleinberg. Authoritative Sources in a Hyperlinked Environment. In Proc. Of the 9th ACM SIAM Symposium on Discrete Algorithms, 1998

[6] 近藤直樹、渡辺陽介、横田治夫:Wikipedia のセクションを考慮したリンク解析による関連項目検索の実現, DEIM2010, 2010