

論文 / 著書情報
Article / Book Information

論題(和文)	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
Title(English)	Extraction of Related Articles using Links and Sections in Wikipedia (O)
著者(和文)	近藤直樹, 渡辺陽介, 横田治夫
Authors(English)	Naoki KONDOH, Yousuke WATANABE, Haruo YOKOTA
出典(和文)	, , , A5-2
Citation(English)	, , , A5-2
発行日 / Pub. date	2010, 3

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 (O)

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 〒152-8552 東京都目黒区大岡山 2-12-1

^{††} 東京工業大学学術国際情報センター 〒152-8550 東京都目黒区大岡山 2-12-1

E-mail: [†]{naoki,watanabe}@de.cs.titech.ac.jp, ^{††}yokota@cs.titech.ac.jp

あらまし Wikipediaなどの不特定参加型の百科事典サイトの登録項目数は爆発的に増えている。その結果、内容的に関連する項目を探し出す場合に、キーワード検索や項目内のリンクをたどるだけでは探し出すことが困難になっている。本研究では、Wikipediaにおいてリンクの結びつきを関連とし、項目内のリンクだけでは得られない関連項目も検索できる手法を検討する。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と考えられる。提案手法では、項目内のセクションでリンク解析をすることによってよりの確かな関連項目を検索する手法を提案する。また、本稿では評価実験により提案手法の有効性を評価する。

キーワード Wikipedia、リンク解析、セクション構造

Extraction of Related Articles using Link Structure and Content Structure in Wikipedia (O)

Naoki KONDOH[†], Yousuke WATANABE^{††}, and Haruo YOKOTA^{††}

[†] Department of Computer Science, School of Engineering, Tokyo Institute of Technology 2-12-1

Oookayama, Meguro-ku Tokyo, 152-8552 Japan

^{††} Global Scientific Information and Computing Center, Tokyo Institute of Technology 2-12-1 Oookayama,
Meguro-ku Tokyo, 152-8550

E-mail: [†]{naoki,watanabe}@de.cs.titech.ac.jp, ^{††}yokota@cs.titech.ac.jp

Abstract The number of articles in web-based encycropedia like wikipedia is increasing. Since links between articles are also increasing, link structure of articles becomes more complicated. We often need important articles related to the specific article we focus on, but it is difficult to find them by only traversing links and using keyword search. In this paper, we propose a method to extract related articles using both link structure and content structure. An article consists of multiple sections based on subtopics, and each section has links to articles related to the subtopic. Our method performs section-level link analysis, and it extracts more focused results related to the section of the article. We also evaluate effectiveness of the method by experiments.

Key words Wikipedia, Link Analysis, Section

1. はじめに

Wikipedia [1] のような不特定参加者によるオンライン百科事典サイトが注目されている。これらは無料で使用でき、誰でも簡単に執筆、編集できる特徴がある。この特徴により、不特定参加者によるオンライン百科事典サイトは世間に浸透し、項目数は爆発的に増えている。

オンライン百科事典サイトの代表例として、Wikipedia があ

る。Wikipedia は項目の自由な利用を認めるコピーレフトなライセンス形態をとり、誰でも自由に文章の編集を行えるウィキシステムを採用している。Wikipedia は全世界で利用されており、約 15 万人が編集に関わっている。項目の分野は幅広く、一般の百科事典にない項目も多い。例えば、Wikipedia 日本語版 [1] の項目数は 64 万以上 (2009 年 7 月) となっている。

オンライン百科事典サイトの他の特徴として、本の百科事典とは違い、項目中に記述されている語がハイパーリンクされて

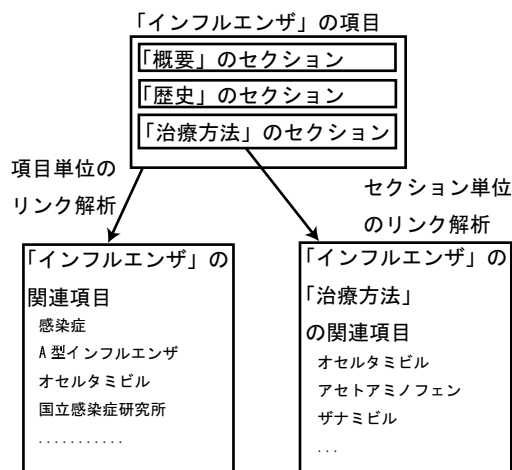


図1 項目単位での関連項目とセクション単位での関連項目

いるので、その語の項目に簡単に移動できることである。項目が違ふ項目をリンクしていることによって、オンライン百科事典サイトはグラフ構造と見ることができ、リンク解析が可能である。しかし、項目数の増加に伴って、関連項目が多すぎてどれが重要な関連項目かわかりにくくなったり、関連しているにもかかわらずリンクが張られてなかったり、といった問題が起きている。そのため、百科事典サイトにおいて関連項目を検索する仕組みを提供することが重要となってきた。

Wikipediaなどのオンライン百科事典サイトにおいては、項目間のリンク構造を解析することで、項目同士の結びつきの強さや、直接リンクされていない項目との関係も発見できると考えられる。これまでもWikipediaの関連項目をリンク解析等で探すことは試みられている[2]。しかし、詳細に解説された項目の場合、そこからリンクされた項目の数自体が多く、またそれらがカバーする分野も広いため、利用者が本当に必要な関連項目を探し出すことは難しい。例えば、Wikipediaの「インフルエンザ」の項目には、病原体の性質に関する記述や、病気の症状に関する記述、治療方法に関する記述などがあり、それぞれ扱う内容が異なっている。扱う内容の違いから、それぞれについての関連項目は異なっており、またそれらを検索したいシチュエーションも違う可能性が高い。既存研究で行われたような項目単位でのリンク解析では、このような内容ごとの関連項目の違いを考慮していなかった。

そこで本研究ではリンク構造だけでなく項目内の文書構造にも着目する。Wikipedia等では、多くの情報を含んだ項目は複数のセクションに構造化されており、各セクションは特定の内容・分野のみについて記述されていることが多い。そのため、項目単位でリンク解析を行う場合よりも、セクション単位でリンク解析をした方が内容を絞った関連項目の探索が可能である。例えば、図1のように項目インフルエンザのセクション治療方法を中心にリンク解析を行って関連項目を探索した場合、項目全体を対象にした場合よりも、治療方法に特化した関連項目が提示できると考えられる。

本研究では、リンク関係に基づく重要度を用いて、オンライン百科事典サイトにおいて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。まず、百科事典データからセクションとリンク情報を抽出しておく。入力、百科事典の項目中の利用者の着目するセクションとする。入力セクションを中心にセクションとリンク情報からリンク解析に必要なグラフデータを取得する。その際、年月のようにどんな内容からでもリンクされる一般的な項目をフィルターにより除去する。次に、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。最後に、重要なリンクに重みを付けてリンクを解析する。出力は、リンク解析の結果の上位 n 件を項目とセクションのペアで出力する。本稿では評価実験を行って、提案手法の有効性を評価する。

以降、2. 節に関連研究を述べる。3. 節では主な百科事典サイトをあげる。4. 節では提案手法について述べる。5. 節では提案手法の実験結果を述べる。6. 節ではまとめと今後の課題を述べる。

2. 関連研究

2.1 リンク解析

リンク解析は、グラフ構造のリンクを解析することによって、どのノードが重要であるかを調べる手法である。特に有名なアルゴリズムは、PageRank [3] と HITS [4] である。

HITS は Web 検索によく用いられ、Web ページを Authority と Hub で評価する。Authority とはページ自体の価値を表し、Hub とは価値の高いページをどのくらいリンクしているかを表す。HITS のアイディアは、Hub の高いページからリンクされると Authority が高くなり、Authority の高いページをリンクすると Hub が高くなるということである。

今回、HITS を使う理由に、百科事典のリンクを多く集めている項目が Hub の高い項目、一般的な項目で重要なものが Authority の高い項目になると考えたからである。HITS では本来リンクに重みは付いておらず隣接行列の値は 0 か 1 であるが、本研究では、相互リンク等の重要度を考慮しているので隣接行列に重みを付けて HITS の計算をしている。

2.2 Wikipedia におけるリンク解析

Wikipedia 英語版の全項目に PageRank や HITS を適用したらどのようなになるかが報告されている [2]。

HITS の Authority では、場所や日時の項目が上位を占めているのに対し、PageRank では、文化的な項目も上位にランキングされている。これは HITS によってカテゴリー、一覧やリストの項目はそれらを指すリンクが多いことによってランキングが高くなることと、PageRank では歴史や文化のような一般的な語が高くなることを示している。

この報告では項目単位にリンク解析を適用しているのに対し、本研究ではセクション単位でリンク解析を行う。その際、場所や日時の項目など一般的な内容の項目は in-リンクが多いことから、リンク解析の前に in-リンクの数によってこれらを除き、ランキングの上位に表れないようにする。

2.3 シソーラス辞書構築

Wikipedia や Web から類語辞書を構築する研究も行われている。類語辞書があることによって、キーワード検索で直接

キーワードが含まれないものを検索することや自然言語処理の形態素解析において未知語の解析精度向上が可能になる。

手法には、tfidf のキーワード抽出やリンク解析をする手法 [5] や HTML の木構造を利用する手法 [6] やリンク構造とセクション構造を利用して語の一般性・局所性・専門性を求める手法 [7] がある。

本研究では、関連項目を探す目的でリンク解析やセクション構造を利用しており、これらの研究とは目的が異なる。

2.4 コミュニティ発見

HITS はコミュニティ発見の手法でもあり、Authority と Hub を用いてグラフを 2 部グラフ構造と仮定している。他の 2 部グラフを使った手法 [8] では、2 部グラフによって fan と community center を表現する。他にも、グラフの最大フロー問題を使った手法 [9] では、最小カットでコミュニティを分割する。

本研究では、オンライン百科事典サイトに HITS を適用して、関連項目を発見する。

3. オンライン百科事典サイト

オンライン百科事典サイトの例として、Wikipedia [1]、はてなキーワード [10]、Yahoo!百科事典 [11]、コトバンク [12] などがある。

オンライン百科事典サイトは、様々な分野の語を網羅し、項目の追加が容易であるので、項目数は爆発的に増えている。これらの共通点は、項目中の語がハイパーリンクされて、サイト内の他の項目へ簡単に移動できることである。

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。

3.1 Wikipedia

Wikipedia 日本語版の統計によると、項目は一般的な項目以外にもカテゴリー、テンプレート、議論のページなどを含み約 168 万以上ある。また、一般的な項目いわゆる記事は約 64 万以上ある。

次に、Wikipedia 日本語版の 2009 年 11 月 17 日のダンプファイル (jawiki-20091117-pages-article.xml.bz2) による統計を述べる。

- トップレベルのセクション 約 243 万
- リンク先が存在する内部リンク 約 3 千万
 - このうち項目の特定のセクションを指すリンク 約 28 万
 - 他は項目へのリンク 約 3 千 92 万

トップレベルのセクションとは一段目のセクションのことであり、図 2 で言うとセクション 1,2,3 などのことである。統計を見ると、Wikipedia の内部リンクはほとんど項目へのリンクであることが分かる。セクション単位でのリンク解析を行う本研究において、セクションを直接指すリンクの数が少ないことは重要な問題となる。これについての対処は 4. 節で述べる。

本研究では、被リンクも合わせてリンクと呼ぶ。リンクを呼び分けるときは、図 3 のように普通のリンクを out-リンク、被リンクを in-リンクと呼ぶ。項目毎の in-リンク数で言うと、in-リンクが 100 以下の項目が 96% 以上、80 以下の項目が 95% 占

目次 [非表示]	
1	概要
2	語源
3	歴史
4	病原体
4.1	インフルエンザウイルス
5	症状
6	疫学
7	感染
8	予防手段

図 2 "インフルエンザ"のセクション

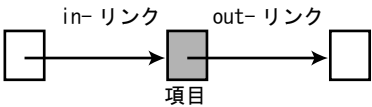


図 3 中心の項目から見たリンクの名称

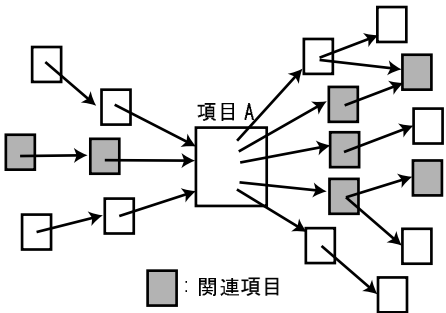


図 4 項目単位でのリンク解析

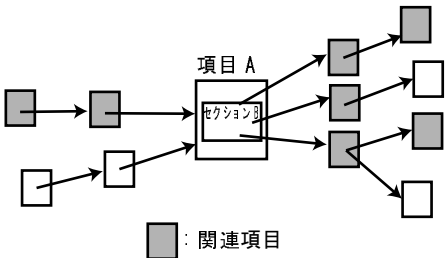


図 5 セクション単位でのリンク解析

める。

4. 提案手法

4.1 目的

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 に対して、項目 A 内の特定のセクション B との関連が強い項目 (およびセクション) のリストを提示することを目的とする。ただし、百科事典の関連項目を探したい時、図 4 のようにその項目内のリンク先を探すだけでは関連項目が得られない。その項目をリンクしている項目や、複数ステップ先のリンクの項目も関連が高い可能性がある。本研究では、これらの項目も関連項目として扱う。

4.2 アプローチ

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必要がある。図 4 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

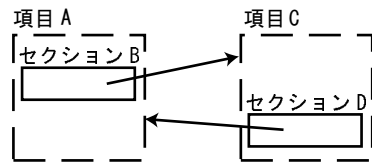


図6 相互リンクでセクションを特定する

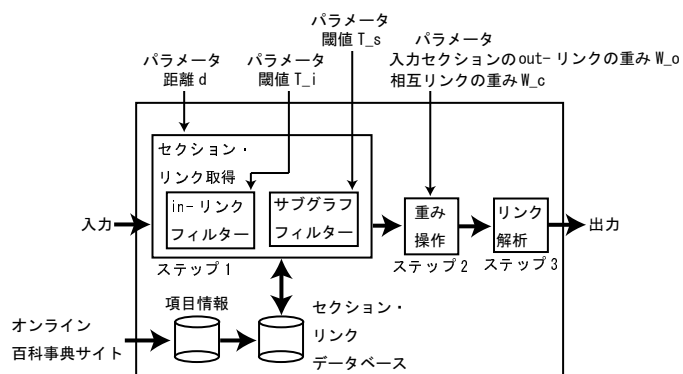


図7 システム概略図

で、図5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。セクションの関連項目を探すので、リンク解析の際に入力セクションからのリンクは重要である。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、本研究では入力セクションを起点として in-リンクおよび out-リンクから距離 d でたどれるセクション集合をリンク解析の対象とする。

しかし、3.1 節で述べたように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、リンク先が項目の場合は、まず項目内のすべてのセクションを指す弱いリンクの集合であるとみなし、さらにそれらのセクション内にリンク元への相互リンクがあれば、そのセクションを本来のリンク先であるとみなしてそのリンクの重みをより高くする。例えば図6のような、項目AのセクションBから項目Cにリンクがあり、項目CのセクションDから項目Aにリンクがあれば、リンクはそれぞれセクションDとセクションBを本来のリンク先であるとみなす。

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、百科事典の一般的な項目を Authority、カテゴリーや一覧の項目を Hub とみなせるので、HITS を用いるのがよいと考える。

4.3 実 現

システム概略を図7に示し、以下詳しく説明する。

本システムへの入力項目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。内部の処理は、関連項目抽出前の準備段階と、関連項目抽出のための3ステップからなっており、以降では各処理について述べる。

4.3.1 準 備

百科事典の情報を格納するデータベースを準備する。このデータベースは、事前に百科事典データを解析して構築する。

その際、百科事典データから項目毎にデータを抽出し、その後項目の情報からセクションやリンクの情報を抽出する。データベースには、百科事典の項目、セクション、リンクの情報、項目毎の in-リンク数情報がそれぞれテーブルとして格納されている。項目のテーブルには、タイトル、項目固有 ID、項目のテキストが含まれる。セクションのテーブルには、セクション名、項目固有 ID、セクション位置番号が含まれる。リンクのテーブルには、リンク元とリンク先のそれぞれの項目固有 ID とセクション位置番号が含まれる。項目毎の in-リンク数のテーブルには、項目固有 ID とその項目の in-リンク数が含まれる。

4.3.2 ステップ1

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。リンク解析のために取得する範囲は、入力セクションから in-リンク、out-リンクを使って距離 d でたどれるセクションとリンクである。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。また、項目へのリンクはその項目内のすべてのセクションへのリンクの集合に変換する。

上記の方法で取得したセクション集合には、年月や国名の項目といった多数の項目からリンクされている広い概念についての項目が含まれていることがある。このような項目は入力セクションとの関連が薄い上にリンク解析に影響を与えやすいのでフィルターにより取り除く。本研究では、2つのフィルターを用いる。

a) in-リンクフィルター

年月や国名の項目のようにどんな内容の項目からもリンクされる可能性のある項目は、in-リンク数は非常に多いが入力セクションとの関係は非常に弱いと考えられるので、本研究では in-リンクが閾値 T_i より多い項目の in-リンク側はたどらない。つまり、in-リンクが多い項目のセクションは out-リンク側だけをたどる。in-リンクフィルターは、セクションとリンクの取得途中で使用する。

b) サブグラフフィルター

サブグラフフィルターは、セクションとリンクの取得終了後に使用する。入力セクションが扱う話題に特化していないセクションを除くことを目的とする。その際、図8のように、取得したセクションとリンクのグラフを百科事典全体のサブグラフと考え、項目のサブグラフ内からの in-リンク数が全体の in-リンク数に比べて割合が少なければ、サブグラフ外の項目にも関連する広い概念の項目であると考えられる。つまり、全体からの in-リンク数と比べサブグラフ内からの in-リンク数の割合が閾値 T_s より小さい項目を除く。

これらのフィルターの効果については5.3 節で評価する。

4.3.3 ステップ2

入力セクションを起点に取得したサブグラフ $G(V, E)$ に対して、リンク解析のための隣接行列を作成する。HITS 等の通常のリンク解析では、ノード i とノード j の間にリンクがあれば、隣接行列の要素 $L_{i,j}$ の値は1にするが、ここでは重要なリンクを考慮するために重みを付ける。

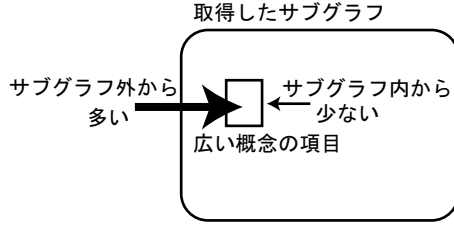


図 8 取得したセクション外にも関連する広い概念の項目

- 重み W_o : 入力セクションの out-リンク
- 重み W_c : すべての相互リンク

入力セクションの out-リンクであり、相互リンクするようなリンクの場合は、大きいほうの重みにする。まとめると以下の式のようになる。

$$L_{i,j} = \max \begin{cases} W_o \text{ if } (i,j) \in E \text{ and } i = \text{入力セクション} \\ W_c \text{ if } (i,j) \in E \text{ and } (j,i) \in E \\ 1 \text{ if } (i,j) \in E \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

これらの重みはパラメータであり、5. 節の実験では重みを変えた時の効果について評価する。

4.3.4 ステップ 3

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 を適用する。HITS のアルゴリズムを以下に説明する。グラフ $G = (V, E)$ を Authority と Hub で評価する。 L を隣接行列とする。ページ $i \in V$ の Authority を $a(i)$ 、Hub を $h(i)$ で表す。

$$a(i) = \sum_{(j,i) \in E} h(j) \quad (2)$$

$$h(i) = \sum_{(i,j) \in E} a(j) \quad (3)$$

k 番目の Authority をベクトル $A_k = (a(1), a(2), \dots, a(n))^T$ 、Hub をベクトル $H_k = (h(1), h(2), \dots, h(n))^T$ で表し、上の式を変形させる。

$$A_k = L^T H_{k-1} \quad (4)$$

$$H_k = L A_{k-1} \quad (5)$$

本来の HITS では初期値を $A_0 = H_0 = (1, 1, \dots, 1)^T$ とするが、本研究では、入力セクションを中心とした関連項目を求めるために、HITS の初期値には入力セクションに 1、それ以外は 0 とする。

最終的な出力は、この HITS で計算される Authority 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。

5. 実験

5.1 目的

本研究では提案手法により関連する項目・セクションを発見できるかを評価するため評価実験を行った。より具体的には以下の内容について評価を行った。

表 1 項目”デフレーション” セクション”デフレスパイラル” $W_o = 3, W_c = 10$

in-リンク フィルター	サブグラフ フィルター	結果	計算時間	HITS 適用時の セクション数
なし	なし	表 2	5 時間 5 分 28 秒	40325
あり	なし	表 3	7 分 28 秒	4774
なし	あり	表 4	56 分 58 秒	7651
あり	あり	表 5	5 分 15 秒	464

- in-リンクフィルターおよびサブグラフフィルターの効果の確認

- リンク重みのパラメータ W_o, W_c の影響の調査

前者については 5.3 節で、後者については 5.4 節で述べる。

5.2 実験環境

提案手法を適用する実験環境を以下に示す。

- CPU Intel Xeon x5570 2.93GHz (クアッドコア)
- Memory 16GB (4GBx4)
- HDD 450GBx4 (RAID5)
- 開発言語 Java (JDK1.6.0.17)
- 使用 RDBMS PostgreSQL 8.4.2
- Wikipedia データ Wikipedia 日本語版 2009 年 11 月 17 日時点の内容 (jawiki-20091117-pages-articles.xml.bz2)

提案手法では入力セクションから到達可能な距離 d のセクションを解析対象とする。今回の実験では計算量およびデータサイズの制限から $d = 2$ を用いた。

5.3 フィルター効果の比較

ここでは、4.3.2 節で提案した in-リンクフィルターおよびサブグラフフィルターにより、出力結果および処理時間がどのように変化するかを調べる。

各フィルターのパラメータを以下のように設定する。in-リンクフィルターが項目の除去を判定する際のパラメータは $T_i = 80$ にする。これは、in-リンクが 80 以上の項目は全体の 5% に過ぎず、その大部分が一般的な項目であると考えられるからである。サブグラフフィルターのパラメータは $T_s = 0.3$ にする。これは、全体のバランスを考慮して決定した。

2 種類のフィルターを適用した場合としない場合の計 4 通りの組合せに対して実験を行う。入力データとして項目”デフレーション”、セクション”デフレスパイラル”を与えた。入力データとして項目”インフルエンザ”、セクション”治療方法”では、取得セクション数が莫大でリンク解析に多くの時間がかかるため、今回は実験対象としなかった。

項目”デフレーション”、セクション”デフレスパイラル”、リンクの重み $W_o = 3, W_c = 10$ で実験した結果を表 1 に示す。各組合せの出力結果のランキングは表 2、表 3、表 4、表 5 である。出力のそれぞれの列は、順位、項目名、セクション名、入力セクションの out-リンク先の項目かどうかを表している。また、セクション名の null はその項目の見出しを表す。

2 つのフィルターを使わない表 2 の結果は、関係のない年号の項目が出力されよくないことが分かる。表 3 の結果では、年の項目は除去できたが、まだ国名等の入力セクションの分野と

表 2 項目”デフレーション” セクション”デフレスパイラル” $d = 2, T_i = \text{なし}, T_s = \text{なし}, W_o = 3, W_c = 10$

	項目名	セクション名	入力の out-リンク
1	1985 年	誕生	なし
2	1980 年	誕生	なし
3	1978 年	誕生	なし
4	1973 年	誕生	なし
5	1983 年	誕生	なし
6	1971 年	誕生	なし
7	1984 年	誕生	なし
8	2004 年	死去	なし
9	1968 年	誕生	なし
10	1979 年	誕生	なし

表 3 項目”デフレーション” セクション”デフレスパイラル” $d = 2, T_i = 80, T_s = \text{なし}, W_o = 3, W_c = 10$

	項目名	セクション名	入力の out-リンク
1	アメリカ合衆国	歴史	なし
2	アメリカ合衆国	政治	なし
3	第二次世界大戦	影響	なし
4	日本	安全保障	なし
5	日本	外交	なし
6	第二次世界大戦	概要	なし
7	日本	経済・産業	なし
8	日本	概要	なし
9	第二次世界大戦	経過（欧州・北アフリカ）	なし
10	フランス	経済	なし

表 4 項目”デフレーション” セクション”デフレスパイラル” $d = 2, T_i = \text{なし}, T_s = 0.30, W_o = 3, W_c = 10$

	項目名	セクション名	入力の out-リンク
1	経済学	専門分野	なし
2	経済学	概要	なし
3	経済学	分類	なし
4	経済学	歴史	なし
5	経済学	特徴	なし
6	経済学	経済学における主な用語・概念	なし
7	経済学	外部リンク	なし
8	経済学	論争	なし
9	経済学	脚注欄	なし
10	経済学	関連項目	なし

表 5 項目”デフレーション” セクション”デフレスパイラル” $d = 2, T_i = 80, T_s = 0.30, W_o = 3, W_c = 10$

	項目名	セクション名	入力の out-リンク
1	新自由主義	関連項目	あり
2	新自由主義	概要	あり
3	新自由主義	議論	あり
4	市場原理主義	関連項目	あり
5	グローバル資本主義	概要	なし
6	新保守主義	特徴	なし
7	新自由主義	null	あり
8	小さな政府	関連項目	あり
9	小さな政府	歴史	あり
10	グローバル資本主義	関連項目	なし

は無関係の項目が上位に来ている。これは、HITS が一般的な項目を上位にランキングするからである。サブグラフフィルターののみを使った表 4 の結果は、表 3 と比べ入力セクション

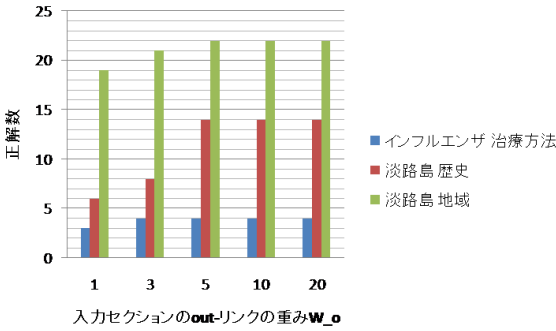


図 9 $W_c = 1$ に固定し、入力セクションの out-リンクの重み W_o を変化させる

の分野の一般的な項目が上位に来ている。2 つのフィルターを使った表 5 の結果は、他の 3 つの結果と比べ入力セクションの out-リンク先の項目のような入力と関係のある項目が出力された。また、計算時間がフィルターを使うほど短くなる点でフィルターの効果が確認できた。

5.4 リンクの重みパラメータの効果

ここでは、4.3.3 節で述べた”入力セクションからの out-リンクの重み W_o ”と”相互リンクの重み W_c ”が出力結果に与える影響について評価する。5.3 節の実験で 2 つのフィルターを用いることの有効性を確認したので、ここでは常に両フィルターを用いるものとする。フィルターの閾値をそれぞれ $T_i = 80, T_s = 0.30$ とする。

5.4.1 評価方法

出力の上位 40 件中どのくらい関連セクションが含まれているか人間が評価する。正解とする関連セクションは、セクションの記述に入力項目名またはその同義語（例:淡路島と淡路）を含み、また入力セクションの内容に適合した記述を含むセクション（例:歴史に関する記述であるか等）とする。

5.4.2 比較対象

リンク間の重み W_o, W_c をともに 1 にした場合は、すべてのリンクを等価値とみなして純粋に HITS を適用した場合と同等である。よって、これをベースラインとして比較を行う。

5.4.3 実験結果

実験する入力セクションを以下に示す。

- 項目”インフルエンザ” セクション”治療方法”
- 項目”淡路島” セクション”歴史”
- 項目”淡路島” セクション”地域”

相互リンクの重み $W_c = 1$ に固定して、入力セクションの out-リンクの重み W_o を変化させた結果を図 9 に示す。横軸は重み W_o の値、縦軸は上位 40 件中の正解数である。図 9 より、 W_o を大きくすることで関連するセクションがより多く発見できていることが分かる。また、ベースラインである $W_o = 1$ かつ $W_c = 1$ の場合よりも結果が改善されている。

次に、 W_c を変化させた場合の実験を行う。 W_o の値については、すでに上記実験から $W_o \geq 5$ の場合が良いことが分かっているので、結果の安定していると思われる $W_o = 10$ に固定する。 W_c を変化させた結果を図 10 に示す。結果より、「項目”イ

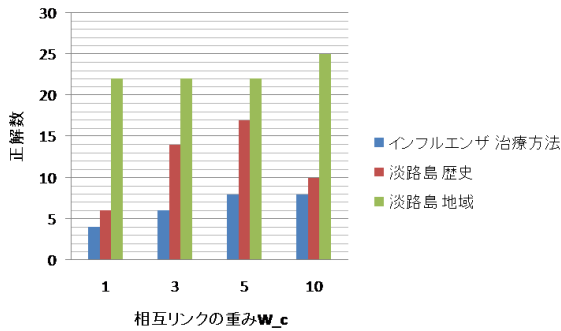


図 10 $W_o = 10$ に固定し、相互リンクの重み W_c を変化させる

「インフルエンザ」、セクション「治療方法」と「項目「淡路島」、セクション「地域」では W_c を上げることで正解数が伸びている。一方、「項目「淡路島」、セクション「歴史」では $W_c = 5$ が最高で、 $W_c = 10$ では逆に正解数が落ちている。しかし、いずれの場合でも $W_c = 1$ のケースよりも正解数が上がっており、相互リンクを重視する提案手法の効果があることが確認できる。

「項目「インフルエンザ」、セクション「治療方法」は $W_o = 10$ かつ $W_c = 10$ の場合でも正解が上位 40 件中に 8 件しかないが、実際のランキング (表 6) を見ると上位 10 位までに 6 件正解が提示できており、提案手法が関連するセクションに高い順位を与えられていることが確認できる。また、10 位の項目「エフ・ホフマン・ラ・ロシュ」、セクション「歴史」は正解であり、しかも入力セクションから直接リンクがない関連セクションである。このように提案手法では入力から直接リンクさせていない関連項目であっても発見可能である。

5.5 実験のまとめ

5.3 節の実験により、提案手法のフィルターによって一般的な項目が除去されて、関連項目が発見できることがわかった。また、フィルターによって計算時間の短縮ができることがわかった。

5.4 節の実験により、入力セクションの out-リンクの重み W_o と相互リンクの重み W_c を高くすると、 $W_o = 1$, $W_c = 1$ よりも良い結果となった。

これらの実験結果から、提案手法が多くの関連セクションをより効率的に探せることが確認された。

6. 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あるために、相互リンクに着目し重み付けをしたグラフに対して HITS を適用する方法を提案した。また、HITS を適用するセクションを取得する際の、項目の in-リンクの多さによって取得するセクションを制限するフィルターと一般項目を除去するフィルターを提案した。

実験システムを用いた実験の結果、提案手法の 2 つフィルターによる一般的な項目の除去が可能であることが確認された。一般的な項目が除去されたことによって、関連項目の発見が可能になった。また、2 つのフィルターは計算時間の短縮にも効果を示した。入力セクションの out-リンクの重みと相互

表 6 項目「インフルエンザ」セクション「治療方法」 $d = 2$, $T_i = 80$, $T_s = 0.30$, $W_o = 10$, $W_c = 10$

	項目名	セクション名	正解
1	オセルタミビル	null	○
2	オセルタミビル	概要	○
3	非ステロイド性抗炎症薬	概要	×
4	非ステロイド性抗炎症薬	NSAID の分類	×
5	非ステロイド性抗炎症薬	代表的な薬物	×
6	アセトアミノフェン	null	○
7	オセルタミビル	耐性	○
8	ザナミビル	概要	○
9	ノイラミニダーゼ阻害薬	null	×
10	エフ・ホフマン・ラ・ロシュ	歴史	○
11	非ステロイド性抗炎症薬	COX-3	×
12	後発医薬品	主な取扱会社	×
13	ジャーナル・オブ・ジ・アメリカン・メディカル・アソシエーション	他の主要な医学雑誌など	×
14	ザナミビル	null	○
15	非ステロイド性抗炎症薬	関連項目	×
16	非ステロイド性抗炎症薬	脚注	×
17	非ステロイド性抗炎症薬	副作用の例	×
18	非ステロイド性抗炎症薬	禁忌	×
19	非ステロイド性抗炎症薬	COX-2	×
20	非ステロイド性抗炎症薬	NSAID 使用の注意点	×
21	非ステロイド性抗炎症薬	歴史	×
22	非ステロイド性抗炎症薬	作用機序	×
23	非ステロイド性抗炎症薬	null	×
24	後発医薬品	オレンジブック	×
25	ジャーナル・オブ・ジ・アメリカン・メディカル・アソシエーション	編集の中立性	×
26	エフ・ホフマン・ラ・ロシュ	null	×
27	総合感冒薬	null	×
28	オセルタミビル	外部リンク	×
29	オセルタミビル	参考文献	×
30	オセルタミビル	合成法	×
31	オセルタミビル	副作用	○
32	オセルタミビル	一般的臨床成績	×
33	後発医薬品	外部リンク	×
34	後発医薬品	関連項目	×
35	後発医薬品	脚注	×
36	後発医薬品	その他	×
37	後発医薬品	処方箋様式	×
38	後発医薬品	日本での経緯	×
39	後発医薬品	品質再評価	×
40	後発医薬品	生物学的同等性試験	×

リンクの重みの効果も確認された。

今後の課題として、入力に対する最適なパラメーターの発見やより多くの多様な項目・セクションに適用させた場合の実験をする必要がある。また、計算量の削減によるさらなる高速化も重要な課題である。

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。

文 献

- [1] Wikipedia 日本語版. <http://ja.wikipedia.org/>.
- [2] F.Bellomi and R.Bonato. Network analysis for wikipedia. *Proceedings of Wikimania*, 2005.
- [3] S.Brin and L.Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, pp. 107–117, 1998.
- [4] J.Kleinberg. Authoritative sources in a hyperlinked environment. *the 9th ACM SIAM Symposium on Discrete Algorithms*, pp. 668–677, 1998.
- [5] 伊藤雅弘, 中山浩太郎, 原隆浩, 西尾章治郎. センテンスを考慮したリンク共起性解析による wikipedia からの連想シソーラス辞書構築に関する一考察. *DEWS2008*, 2008.
- [6] 渡邊陽太郎, 浅原正幸, 松本裕治. HTML の木構造を利用した条件付確率場による固有表現分類:wikipedia からのシソーラス半自動構築. 情報処理学会研究報告, 2007.
- [7] 中谷誠, 手塚太郎, Adam Jatowt, 田中克己. Wikipedia と web 検索を用いたトピック構造の抽出. *DEWS2008*, 2008.
- [8] R.Kumar, P.Raghavan, S.Rajagopalan, and A.Tomkins. Trawling the web for emergine cyber-communities. *In Proc. Of the 8th Intl. World Wide Web Conference*, 1999.
- [9] L.R.Ford Jr and D.R.Fullerson. Maximal flow through a network. *Canadian Journal Mathematics*, 1956.
- [10] はてなキーワード. <http://d.hatena.ne.jp/keyword/>.
- [11] Yahoo!百科事典. <http://100.yahoo.co.jp/>.
- [12] コトバンク. <http://kotobank.jp/>.