Wikipediaを用いた
概念間の関連度測定に関する研究

2011年1月

伊藤　雅弘

関連発表論文

1. 学会論文誌発表論文

1. 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎: Wikipedia のリンク共起性解析によるシソーラス辞書構築, 情報処理学会論文誌:データベース (TOD), Vol. 48, No. SIG 20 (TOD 36), pp. 39–49 (Dec. 2007).

2. 中山浩太郎, 伊藤雅弘, Maike ERDMANN, 白川真澄, 道下智之, 原 隆浩, 西尾章治郎：Wikipedia マイニング -近未来チャレンジキックオフ編-, 人工知能学会論文誌, Vol. 24, No. 6, pp. 549-557 (2009).

3. 中山浩太郎, 伊藤雅弘, Maike ERDMANN, 白川真澄, 道下智之, 原 隆浩, 西尾章治郎：Wikipedia マイニング: Wikipedia 研究のサーベイ, 情報処理学会論文誌:データベース (TOD), Vol. 2, No. 4, pp. 49–60 (Dec. 2009).

4. Ito, M., Nakayama, K., Hara, T., and Nishio, S.: Semantic Relatedness Measurement based on Wikipedia Link Co-occurrence Analysis, *International Journal of Web Information Systems (IJWIS)*, (2011, 採録決定).

2. 研究会等発表論文（査読付）

1. Nakayama, K., Ito, M., Hara, T., and Nishio, S.: Wikipedia Mining for Huge Scale Japanese Association Thesaurus Construction, *in Proceedings of IEEE International Symposium on Mining And Web (MAW 2008)*, pp. 1150–1155 (Mar. 2008).

2. Nakayama, K., Pei, M., Erdmann, M., Ito, M., Shirakawa, M., Hara, T., and Nishio, S.: Wikipedia Mining –Wikipedia as a Corpus for Knowledge Extraction–, *in Proceedings of Annual Wikipedia Conference (Wikimania 2008)* (July 2008).

3. Ito, M., Nakayama, K., Hara, T., and Nishio, S.: Association Thesaurus Construction Methods based on Link Co-occurrence Analysis For Wikipedia, *in Proceedings of ACM International Conference on Information and Knowledge Management (CIKM 2008)*, pp. 817–826 (Oct. 2008).

4. Nakayama, K., Ito, M., Hara, T., and Nishio, S.: Wikipedia Relatedness Measurement Methods and Influential Features, *in Proceedings of IEEE International Symposium on Mining And Web (MAW 2008)*, pp. 738–743 (May 2009).

3. その他の研究会等発表論文

1. 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎: Wikipedia のリンク共起性解析によるシソーラス辞書構築のスケーラビリティ向上, 情報処理学会研究報告 (データベースシステム研究会 2007-DBS-143), Vol. 107, No. 131, pp. 539–544 (July 2007).

2. 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, 電子情報通信学会データ工学ワークショップ (DEWS 2008) (Mar. 2008).

3. 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎: Wikipedia からの連想シソーラス構築プロジェクト, 第 20 回セマンティックウェブとオントロジー研究会 Wikipedia ワークショップ (Jan. 2009) .

4. 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎: Wikipedia の概念に基づく連想関係テストコレクション「WikiSimi3000」, 第 23 回人工知能学会全国大会 (JSAI 2009), CD-ROM (June 2009).

5. 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎: Web 上の情報を用いた Wikipedia 記事の信頼性評価に関する検討, 第 24 回人工知能学会全国大会 (JSAI 2010), CD-ROM (June 2010).

6. 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎: Wikipedia の多様な特徴を利用した概念間関連度と有用な特徴の調査, 電子情報通信学会技術研究報告 (データ工学研究会 DE2010-26), Vol. 110, No. 328, pp. 7–12 (Dec. 2010).

以上

内容梗概

近年，情報爆発と言われるように，人類によって生成される情報が爆発的に増大している．この情報爆発時代に向けて，莫大なデータを処理する技術やデータを活用したサービスなど，国内外で多様な研究が行われている．このような研究の中で重要な技術の 1 つに，自然言語処理がある．自然言語処理を行うためには，構文解析などの文法知識などと共に，「ことば」に関する知識が非常に重要になってくる．特に日々，新しい語が生み出される人名や固有名詞などの，いわゆる固有表現（Named Entity）に関する知識は，自然言語を処理する上で重要な役割を果たす．これら固有表現を含む概念（語の意味）に関する知識を抽出する重要な技術の 1 つに，概念間の意味的関連度（Semantic Relatedness）の測定がある．たとえば，複数のテキスト文書を整理するために分類処理を行う際，その方法の 1 つに各テキスト中に出現する単語の出現頻度を用いる手法があるが，それぞれの単語の意味的関連度を考慮することにより，より精度の高い分類を行うことが可能となる．

一方，WWW の爆発的な普及に伴い，Wikipedia に代表される Web 事典が公開されてきた．Wikipedia は，Wiki を利用して構築された百科事典であり，文化，歴史，数学，科学，社会，テクノロジなどの幅広い分野の多くの概念（記事）をカバーしている．また，Wikipedia などの Web 事典と通常の電子事典の最大の違いは，記事（概念）どうしがハイパーリンクで互いに参照されていることである．このような特徴を持つ Wikipedia は，近年，知識抽出のためのコーパス（解析用データ）として研究者から注目を集めている．概念間の関係解析に関する研究においても，概念の網羅性や自然言語処理における未知語の対応，同義語，多義語の判別など，自然言語解析における諸問題を意識することなく概念に関する解析が行えるため，該当分野の研究者から注目を集めている．このような背景の下，これまで Wikipedia をコーパスとした概念間関連度の測定に関する研究がいくつか行われてきた．しかし，従来手法は解析データ量に対するスケーラビリティや精度の観点から問題があった．たとえば，各概念（記事）ペアの概念間関連度を測定する際，従来手法である記事間のリンク構造を n ホップ先まで再帰的に解析する手法では，膨大な計算が必要となる．また，特定の記事内の情報だけを用いて概念

（記事）の特徴情報を生成する手法では，記事内の記述が極端に少ない，記事が荒らされているなど，特定の記事の質に大きく精度が影響される．

そこで本論文では，計算量が少なく高精度な概念間関連度の測定を行うことを目的として，Wikipedia をコーパスとした概念間関連度の測定手法を議論する．具体的には，Wikipedia に存在する他の記事へのハイパーリンク（記事間リンク；以下，特に明示しない限り Wikipedia における「リンク」は「記事間リンク」を意味する）の共起情報，つまり記事中の近い位置に存在するリンクペアの情報を用いた概念関関連度の測定手法を提案する．次に，Wikipedia 全体のリンクの共起情報に加えて，記事内に存在するリンク情報という 2 つの情報を融合した概念間関連度の測定手法を提案する．最後に，既存手法による概念間関連度や，記事内に存在するリンク数など，Wikipedia から取得可能な記事間の関係性を特徴付けうる複数の情報の中で，どの情報が概念間関連度を測定する上で有効に働くかを網羅的に調査すると共に，機械学習器を用いて，それらの情報を融合することによる高精度な概念間関連度の測定手法を提案する．

本論文は，5 章から構成され，その内容は次の通りである．まず，第 1 章において，序論として研究の背景と動機について述べる．

第 2 章では，リンクの共起性解析に基づいた概念間関連度の測定方法を提案する．提案手法であるリンク共起性解析は，Wikipedia 全体の記事から抽出したリンクデータをシーケンシャルに解析するアルゴリズムの特性上，従来手法のリンク構造を解析する手法より計算量が大幅に少なくスケーラビリティが高い．また，Wikipedia 全体の情報を概念の特徴情報として活用するため，特定の記事内の情報だけを用いて概念（記事）の特徴情報を生成する手法に比べて，高精度な関連度測定が行える．具体的には，記事間リンクに対して近傍リンクとの共起をカウントし，その処理を Wikipedia のすべての記事で行うことによって，2 つの記事（概念）の共起性を求める．そして，その共起性の値を用いて記事の共起ベクトルを生成し，ベクトル間の距離をコサイン類似度によって求めることで，2 つの記事間，つまりそれらの記事が表す概念間の関連度を測定する．また，提案方法の性能評価のために行った評価実験の結果を示し，その有効性について検証する．

第 3 章では，異なる 2 つの情報を組み合わせることによって，より高精度な概念

間関連度の測定方法を提案する．これは，第 2 章の Wikipedia 全体のリンクの共起情報のみを扱う手法における，共起情報が不足している概念間の関連度測定精度の低下を防ぐためである．この方法では，概念の特徴情報を得る際に，リンク共起性解析による Wikipedia 全体のリンク情報を用いる大域的情報と，記事内に存在するリンク情報のみを用いる局所的情報の両方を活用する．本手法によって，それぞれの情報の性質が持つ情報量不足を補い合うことによって，より高精度な概念間関連度の測定を目指す．具体的には，概念（記事）の特徴を表すために，Wikipedia 全体でのリンクの共起情報を用いた特徴ベクトルと，記事内に存在するリンクを用いて tfidf の考え方に基づく特徴ベクトルを生成し，それら 2 つのベクトルを合成する，この処理によって，双方の手法によって得た特徴情報を補い合い，概念間関連度の測定精度向上を図る．また，提案方法の性能評価のために行った評価実験の結果を示し，その有効性について検証する．

第 4 章では，既存手法による概念間関連度や，記事内のリンク数，記事へのリンク数，Wikipedia に存在するカテゴリへの所属情報など，Wikipedia から取得可能な記事や記事間の関係性を特徴付けうる複数の情報の中で，どの情報が概念間関連度の測定にとって有用であるかを網羅的に検証する．これは，従来研究では種々の情報が個別に提案・検証されており，それぞれの情報の概念間関連度測定への貢献度やパフォーマンス，統合方法を総合的に検証する研究が行われていなかったためである．さらに，それら複数の情報を機械学習によって組み合わせた高精度な概念間関連度の測定方法を提案する．これは，概念間関連度は，単独の情報ではなく，種々の情報による複合的条件が影響しているという仮説を実証するためである．具体的には，機械学習における各学習情報（素性）の貢献度を測定する F-score によって，概念間関連度に重要な影響を与える情報の検証を行う．また，回帰問題を解くことのできる機械学習手法である SVR を用いて，それら複数の情報を素性として，その素性セットのパターンと概念間関連度の関係を学習させることにより，未知の関連度を素性セットから予測させる．さらに，提案方法の性能評価のために行った評価実験の結果を示し，その有効性について検証する．

第 5 章では，本論文の成果を要約したのち，今後の研究課題について述べ，本論文のまとめとする．

目 次

第1章　序章

1.1　研究背景

近年，情報爆発と言われるように，人類によって生成される情報が爆発的に増大している．この情報爆発時代に向けて，莫大なデータを処理する技術や膨大なデータを活用したサービスなど，国内外で多様な研究が行われている．このような研究の中で重要な技術の 1 つに，テキスト解析がある．Web に代表されるコンテンツは，多くのテキスト情報を含んでおり，日々増大を続けている．これらのテキストをいかに効率よく解析し，情報を整理し，有用な情報を抽出するかが非常に重要になっている．これまで，XML [7] に代表される機械が処理可能な形式で情報を生成することによって，解析処理の効率を高める研究が行われてきたが，Web 上に存在する情報の多くは依然として自然言語を含むため，古くから行われてきた自然言語処理（NLP: Natural Language Processing）に関する研究が重要な役割を果たしている．

自然言語処理技術において，構文解析などの文法知識などと共に,「ことば」に関する知識が非常に重要になってくる．特に日々，新しい語が生み出される，いわゆる固有表現（Named Entity）に関する知識は，自然言語を処理する上で重要な役割を果たす．本論文では，この固有表現の中で，ある固有の意味を持つ対象を「概念（Concept）」として扱う．たとえば，フルーツの「Apple」と企業の「Apple」は個別の概念を持つが，逆に「Apple」「Apple Inc.」「Apple Computer」と記述されていても，それらが同じ企業としての「Apple」を対象としていれば，同一概念であると見なす．これら概念に関する知識を抽出する重要な技術の 1 つに，概念間の意味的関連度（Semantic relatedness）の測定がある．この概念間の意味的関連度は，自然言語処理，語義曖昧性解消（WSD: Word Sense Disambiguation），文書分

類，情報検索（IR: Information Retrieval），複合名詞解析など，様々なアプリケーションで利用することができる [24, 41, 64]．たとえば，複数のテキスト文書を整理するために分類処理を行う際，各テキスト中に出現する単語の出現頻度を用いる手法があるが，それぞれの単語の意味的関連度を考慮することにより，より精度の高い分類を行うことが可能となる．この意味的関連度の測定に関する研究は，これまで主にニュース記事や Web などのテキストコーパスを解析する方法 [5, 21] や，WordNet [23] や Roget's Thesaurus [68] などの構造化された既存データセットのグラフ構造を解析する方法 [10, 24, 27, 29, 39, 40, 47, 67] が提案されてきた．テキストコーパスを解析する方法では，膨大な自然言語から多くの概念（語彙）を取り扱うことができる反面，解析コストが高いことや，また未知語や多義語，同義語の識別などの自然言語処理に関する技術的課題があり，語義曖昧性解消などの多くの研究が行われている [31, 83]．一方，WordNet や Roget's Thesaurus などのグラフ構造を解析する方法では，主に人手によって定義されている概念やその関係（上位語，下位語など）をノードとして扱うことができるため，自然言語処理における諸問題は考慮する必要がない．しかし，定義されている概念が一般語に偏っており，人名や専門用語などの固有表現に関する概念が少ないことが指摘されている [75].

一方，WWW の爆発的な普及に伴い，Wikipedia[1] に代表される Web 事典が公開されてきた．Wikipedia は，Wiki を利用して構築された百科事典であり，文化，歴史，数学，科学，社会，テクノロジなどの幅広い分野の概念（記事）をカバーしている．Wikipedia では，Web ブラウザを通じて，他のユーザと議論しながら自由に記事を投稿できることが大きな特徴である．Wikipedia には，2008 年 8 月の段階で約 250 万もの膨大な数の記事（英語のみ）が公開されており，市販の百科事典の記事数が数万から 10 万であることと比較してもその規模が膨大であることが分かる．Nature 誌の調査によると，Wikipedia の記事数および精度は，多くの専門家が集まって作成した百科事典「Britannica」と同等であると報告している [26]．また，Wikipedia などの Web 事典と通常の電子事典の最大の違いは，記事どうしがハイパーリンク[2] で

[1] http://www.wikipedia.org

[2] インターネット上のリソースを一意に表す URL によって，そのリソースへの参照を提供する．

互いに参照されていることである．記事から他の記事へのハイパーリンクを，本論文では「記事間リンク[3]」とする．このような Wikipedia の持つ特徴は，近年知識抽出のためのコーパスとして研究者から注目を集めており，Wikipedia のデータを用いた様々な研究が行われている [2,4,19,22,25,49,52,53,54,55,66,70,75,76]．概念間の関係解析に関する研究においても，概念の網羅性や自然言語処理における未知語の対応，同義語，多義語の判別など，自然言語解析における諸問題を意識することなく概念に関する解析が行えるため，該当分野の研究者から注目を集めている．筆者の属する研究グループでは，Wikipedia のこれらの特性にいち早く着目し，Wikipedia に対して Web マイニングを行い，有益な情報を抽出する Wikipedia マイニングに関する研究を行ってきた．これまでの研究において，Wikipedia のリンク構造を解析することで，概念どうしの関連度を定義した連想シソーラス辞書を高精度で構築できることを示してきた [56,57,58].

以下では，第 1.2 節で知識抽出のためのコーパスとして有効な Wikipedia が持つ特徴について詳説し，第 1.3 節では，概念間関連度の測定手法の関連研究を紹介する．第 1.4 節では，本研究の目的について述べ，最後に第 1.5 節で本論文の構成について説明する．

1.2 知識抽出のためのコーパスとしての Wikipedia

Wikipedia は，閲覧して知識を得るという活用の他，研究者にとっては機械処理によって知識抽出可能なコーパスとして，非常に注目されている．Wikipedia が持つ知識抽出に有効な特徴の主なものを以下に挙げる．

- 密なリンク構造
- コンテンツの網羅性
- 質の高いアンカーテキスト
- URL による概念の一意性

[3]以下，特に明示しない限り Wikipedia における「リンク」は「記事間リンク」を意味する

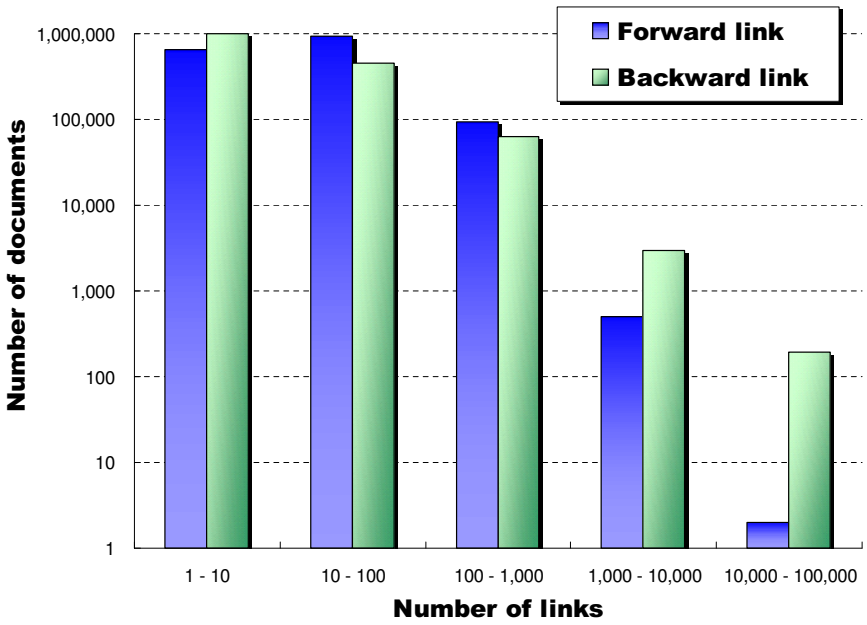

図 1.1: リンク数の分布

- 多様なリンク構造

以下では，それぞれについて解説し，最後に技術的課題について考察する．

1.2.1　密なリンク構造

まず，2008 年 8 月の段階での Wikipedia 内における約 250 万記事（英語のみ）の記事間リンクの数は，約 8,573 万であることが分かっている．これは，1 記事あたり平均 35.6 のリンクを持つ計算となる．また，他の記事からのリンク数の分散は，1 万以上を持つ記事が 426 件，1,000 以上を持つ記事が 6,874 件，100 以上を持つ記事にいたっては 98,650 件も存在することが分かっている．さらに，15,184 記事が 500 以上の他の記事へのリンクを持っており，185,814 記事が 100 以上の他の記事へのリンクを持っている．しかも，これらのリンクはサイト内に対するリンクのみをカウントしたものであり，サイト外へのリンクは含まれていない．これは，Wikipedia では閉じられた語彙（記事）空間の中で密なリンク構造を持っており，リンク構造を解析することで有用な情報を抽出できる可能性が高いことを示している．図 1.1 と図 1.2 にリンク数の分布を示す．

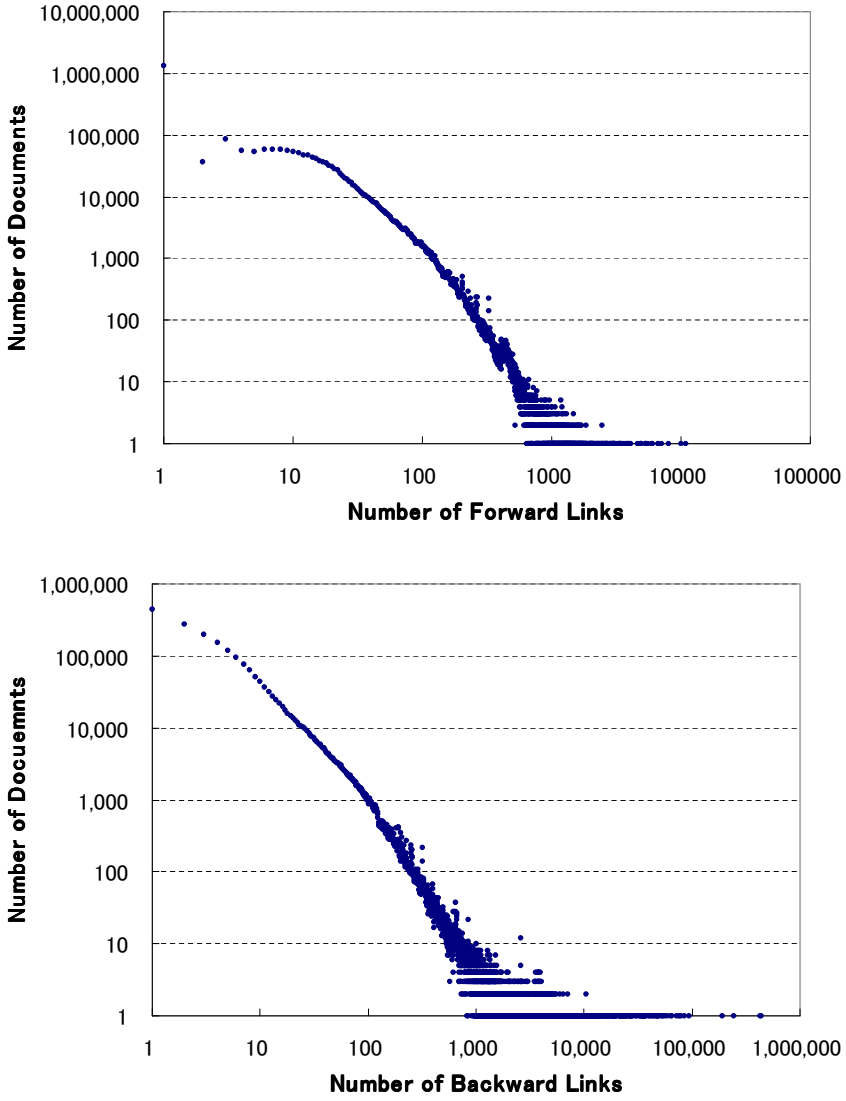

図 1.2: フォワードリンクとバックワードリンク数が示すべき乗法則

非常に興味深いのは，図 1.2 が示すとおり Wikipedia のリンク構造は，一部の記事に極端に多くのリンクが集中する Zipf 分布 [84] に従う点である．このような分布は，人気の Web サイトに対するリンク関係やアクセス頻度 [8]，図書館における図書の貸し出し数，有名論文の参照関係などに見られる分布であり，一様にリンクが分布しているわけではないことを示している．この傾向は特にバックワードリンクに顕著であり，全体数から見るとごく少量の記事が非常に多くの記事から参照されている．

1.2.2　コンテンツの網羅性

従来の辞書では，一般的な語からトップダウン的に追加されていくのが通常であり，一般的でない語や専門的な語は辞書に追加されるのが遅れる，もしくはいつまでも登録されないのが一般的である．しかし，Wikipedia では，インターネットを通じてリアルタイムに記事が公開・アップロードされ，リンクが構築されていくため，極めて即時性と網羅性が高い．たとえば，ある企業から最新の技術の発表があった数時間後には，記事が生成され，その説明や詳細なスペック，画像などが他の語へのリンク付きで公開されたというケースもある．このような新しい概念に対する網羅性の高さは Web コーパスとして見たときの重要な特徴の 1 つである．

1.2.3　質の高いアンカーテキスト

アンカーテキストは，HTML 文書におけるリンクが設定されたテキストで，HTML の<A>タグで囲まれたテキスト部分である．通常の Web ページにおけるリンクのアンカーテキストは，リンク先のページ内容を表す語が用いられている場合があると同時に,「最新情報はこちらをクリック」といったようにリンク先の概念とは無関係な情報も多く含まれている．一方，Wikipedia においては，他の記事へのリンクのアンカーテキストは，リンク先の概念を端的に表す語が利用される [59]．図 1.3 にアンカーテキストの例を示す.「Microsoft」「Paul Allen」「Bill & Melinda Gates Foundation」といったリンク先の概念を端的に示す語がアンカーテキストに利用されており，ノイズが少ないことは統計を取るまでも無く明らかである．このアンカーテキストの統計を取ることによって，リンク先記事（概念）の同義語や多義語を判別することができる [58]．たとえば，企業である「アップル社」に関する記事へのリンクのアンカーテキストは「Apple」「Apple Inc.」「Apple Computer」などが多く用いられており，これらは同義語であると判断できる．多義語に関しても，アンカーテキストが「Apple」であるリンクでも，リンク先が「フルーツである Apple の記事」か「企業である Apple の記事」かで判別することができる．

Bill Gates

From Wikipedia, the free encyclopedia
(Redirected from Bill gates)

For other people named Bill Gates, see Bill Gates (disambiguation).

William Henry "Bill" Gates III (born October 28, 1955)[2] is an American business magnate, philanthropist, author and chairman[3] of Microsoft, the software company he founded with Paul Allen. He is consistently ranked among the world's wealthiest people[4] and was the wealthiest overall from 1995 to 2009, excluding 2008, when he was ranked third.[5] During his career at Microsoft, Gates held the positions of CEO and chief software architect, and remains the largest individual shareholder with more than 8 percent of the common stock.[6] He has also authored or co-authored several books.

Gates is one of the best-known entrepreneurs of the personal computer revolution. Although he is admired by many, a number of industry insiders criticize his business tactics, which they consider anti-competitive, an opinion which has in some cases been upheld by the courts.[7][8] In the later stages of his career, Gates has pursued a number of philanthropic endeavors, donating large amounts of money to various charitable organizations and scientific research programs through the Bill & Melinda Gates Foundation, established in 2000.

Bill Gates stepped down as chief executive officer of Microsoft in January 2000. He remained as chairman and created the position of chief software architect. In June 2006, Gates announced that he would be transitioning from full-time work at Microsoft to part-time work and full-time work at the Bill & Melinda Gates Foundation. He gradually transferred his duties to Ray Ozzie, chief software architect and Craig Mundie, chief research and strategy officer. Gates' last full-time day at Microsoft was June 27, 2008. He remains at Microsoft as non-executive chairman.

参照 URL: http://en.wikipedia.org/wiki/Bill_Gates

図 1.3: アンカーテキストの例

1.2.4 URL による概念の一意性

URL により概念の一意性が確立されている点は，Wikipedia の大きな特徴の 1 つである．電子辞書では，通常 1 つの見出し語が 1 つの記事に割り当てられており，その中で複数の意味について詳述される．一方，Wikipedia では 1 つの URL に 1 つの概念（記事）が割り当てられており，多義性が URL によって解決されている点が大きな特徴である．たとえば，「Football」は強いコンテキスト依存性を持つ単語であり，アメリカンフットボールを示す場合もサッカーを示す場合もある．Wikipedia では，これら 2 つの概念は別の記事で管理されており，それぞれ

「http://en.wikipedia.org/wiki/American_Football」（図 1.4 上）「http://en.wikipedia.org/wiki/Football_(soccer)」（図 1.4 下）という別々の URL が割り当てられている．

このように，概念と URL が一対一で対応していることは，概念の関連性を解析する際に多義性やコンテキスト依存性の影響を受けずに解析できることを示している．

1.2.5　多様なリンク構造

Wikipedia には，記事から記事へ移動するための通常のリンク以外に，カテゴリリンクやリダイレクトリンクなどいくつかの特殊なリンクが存在する．本項では，これらの特殊なリンクについて解説する．

カテゴリリンク

カテゴリリンクは，ある記事（概念）がどのようなカテゴリに属するかを指定するためのリンクである．すべてのカテゴリには専用のページ（カテゴリページ）が用意されており，カテゴリページはさらに別のカテゴリページに属することが可能である．このカテゴリ構造は，一種のタクソノミー（分類辞書）としての役割を有しており，カテゴリを絞り込みながら記事を検索するような機能を実現するために利用されている．Wikipedia が提供しているカテゴリ検索システム「CategoryTree」[4]では，カテゴリを検索することや，カテゴリの階層構造をブラウジングすることが可能である．しかし，一見階層構造に見える Wikipedia のカテゴリ構造は，実際にはネットワーク構造となっている．これは，1 つのカテゴリページが複数のカテゴリページに属することが可能であり，一部にはループも存在するためである．Wikipedia の英語版（2008 年 5 月）には，約 997 万のカテゴリリンクが存在していることが分かっている．

[4] http://en.wikipedia.org/wiki/Special:CategoryTree

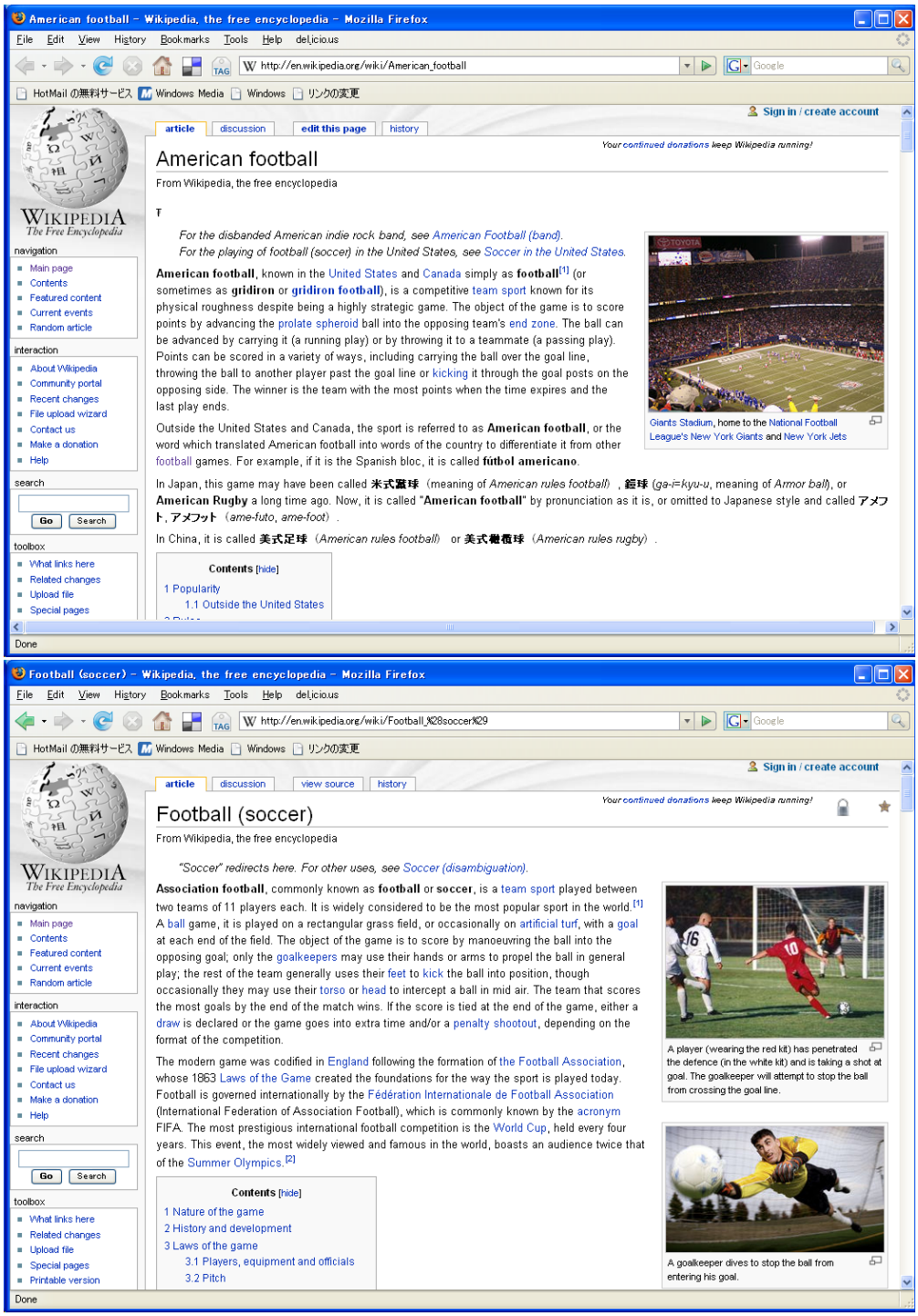

図 1.4: Wikipedia における URL による概念の一意性

リダイレクトリンク

リダイレクトリンクは，ある記事が参照されたときに別の記事へとリダイレクトする機能を提供するリンクである．たとえば，記事「Movie」を参照した場合，記事「Film」へと自動的にリダイレクトされる．リダイレクトリンクは，主に同義語や表記のゆれを表現し，同一内容の記事が散在することを防ぐために利用される．

Wikipedia の中でリダイレクトリンクは重要な役割を果たし，通常のリンク構造解析とは別途考慮して解析する必要がある．たとえば，リダイレクトリンクが同義語や表記のゆれを表現するために利用されているのであれば，リダイレクトされているページは同じ意味を持つ概念として解析するなどの工夫が必要となる．Wikipedia の英語版（2008 年 8 月）には リダイレクトページを含めると約 500 万のページ（カテゴリページなどの特殊ページを除く）が存在するが，その中の実に約 250 万ページがリダイレクト専用のページであることが分かっている．

1.3　関連研究

概念間関連度は，自然言語処理だけでなく幅広い研究領域で利用されてきた．特に，情報検索（IR）の分野では，検索クエリに含まれない語であっても意味的に関連する語を含む文書を検索することや，関連文書を発見する際に，それぞれの文書に含まれる語の関連度を考慮することによって検索性能を向上させることに利用されてきた [73, 82]．概念間関連度を測定する最も単純な方法は，人間の手によるものが考えられる．しかし，このような概念間関連度を蓄積したデータベースを人手で構築するには，概念の追加・更新に手作業による膨大な手間がかかるため，最新の概念や一般的でない語彙などへの対応が難しいのが現状である．そのため，自動的に概念間関連度を高精度に測定する手法が必要とされている．

本節では，概念間関連度の測定に関する従来研究について述べる．

1.3.1 自然言語処理による概念間関連度

自然言語処理による概念間関連度の測定に関する研究の歴史は古く，コーパス解析により自動的に行う手法は数多く提案されてきた．たとえば，語の共起関係に基づく手法 [73] や，語のフィルタリングやクラスタリング手法を用いる研究 [13,18,77] などがある．しかし，自然言語処理における語義や係り受けなどの曖昧性および多義性の解消，同義語の同定などの課題が存在し，語義曖昧性解消などの多くの研究が行われている [31,83].

また，形態素解析の問題もある．自然言語処理を行う場合，前処理として，入力文を意味のある最小の言語単位である形態素に分け，品詞タグを付与する必要がある．形態素解析および品詞タグを付与するツールとしては，Brill の Tagger [9]，Stanford NLP tools[5] [43]，OpenNLP[6]が有名であるが，未知語への対応や曖昧性の取り扱いなどが課題となっている．

1.3.2 Web マイニングによる概念間関連度

Web コーパスと通常の文書コーパスの性質の最も大きな違いは，ハイパーリンクである．ハイパーリンクは，単に他ドキュメントへ移動するための機能を提供するだけでなく，トピックの局所性やアンカーテキストなど重要な情報を豊富に有している [17].

ハイパーリンクは，ある 1 つのページに着目した場合，フォワードリンク（Forward Link）とバックワードリンク（Backward Link）に分類できる．フォワードリンクは対象のページから別のページに移動するためのリンクであり，バックワードリンクは別のページから対象のページへと移動するためのリンクである．HITS アルゴリズム [44] や PageRank アルゴリズム [63] など，最近の Web 構造解析のアルゴリズムでも，客観的な情報を得るためにはバックワードリンク解析が有用であることが示されている．これは，バックワードリンクは対象ページに対する「投票」と見なすことができるためである．

[5]http://nlp.stanford.edu/software/
[6]http://opennlp.sourceforge.net/

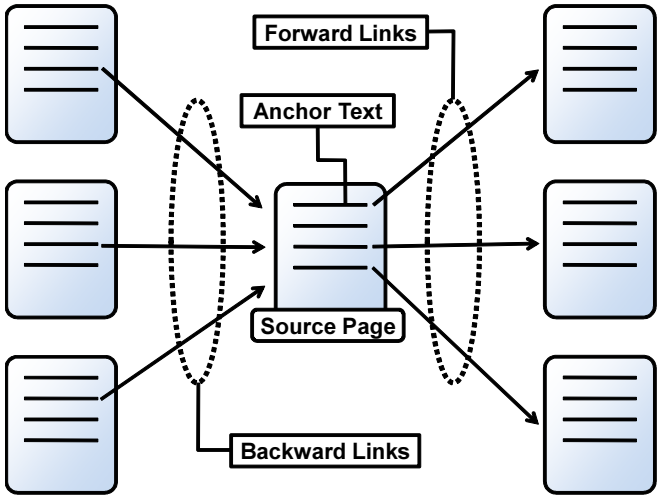

図 1.5: リンクの概念

トピックの局所性とは，ハイパーリンクで繋がっているページどうしは，繋がっていないページどうしに比べて同じトピックに関する記述である場合が多いという性質を表す．Davison らの研究 [20] は，このトピックの局所性が多くの場合に正しいことを示している．また，アンカーテキストも Web マイニングにおいて重要な役割を果たす．アンカーテキストは一般的に被リンクページの内容（要約）を表現していることが多い．図 1.5 にフォワードリンク（Forward Link），バックワードリンク（Backward Link），アンカーテキスト（Anchor Text）の概念を示す．

上記のような Web コーパスの特徴を生かし，リンク構造を解析することで，関連性の高い語を抽出する研究が行われている [15]．Web サイトの情報を用いた概念間関連度の測定では，Web コンテンツの増加・更新に従い，新しい語や他の語との関係などの情報を更新することができるというのが大きな特徴である．しかし，これらの手法は，解析対象とするコーパスに関する考察がなく，依然として自然言語処理を利用した解析による精度の問題などが残されている．また，膨大な Web 空間をコーパスとして用いた場合，探索空間が広すぎることから，リンク構造解析を行った場合などに，解析内容が収束せず特徴的な意味情報を取得できない場合がある．一方，ドメインを限定した場合には内容が偏るといった問題がある．

1.3.3 Wikipediaマイニングによる概念間関連度

Wikipediaを概念間関連度を測定するためのコーパスとすることは，1.2節であげたような様々な特徴による多くのメリットがある．Wikipediaを解析して概念間関連度を測定する先行研究として，大別するとカテゴリリンクに基づく手法，記事テキストに基づく手法，記事間リンクに基づく手法がある．以下にその手法を解説する．

カテゴリリンクに基づく手法

Wikipediaマイニングにおいて，最も一般的な関連度計算のアプローチの1つがWikipediaのカテゴリ構造を利用する方法である．前述のとおり，Wikipediaのカテゴリ構造は記事（概念）を分類するための階層的構造であるが，本アプローチはカテゴリ構造において記事間のパスの長さが短いほど関連度が高くなる，という考えに基づいている．代表的な研究として，Strubeらの WikiRelate! [75] が挙げられる．Strubeらは，WordNetに用いられてきた関連度算出の手法 [45,46] がWikipediaのカテゴリ構造に適用できることを証明し，複数の指標を統合することで精度が向上することを示した．WikiRelate!では，カテゴリ構造の解析手法をさらに2つの手法に分類している．

1. カテゴリ構造におけるパスの長さに基づく手法
2. カテゴリ構造における情報の共有度（直近の共通の祖先が持つ子概念が少ないほど関連度が高い）に基づく手法

以下に，評価実験において最も精度が良かった，カテゴリ構造におけるパスの長さに基づいて関連度を測定する手法lchを示す．

$$lch(c_1, c_2) = -\log \frac{length(c_1, c_2)}{2D} \tag{1.1}$$

ここで，c_1, c_2 を2つの記事（概念），$length(c_1, c_2)$ は，ノード c_1, c_2 間のカテゴリ構造を介した最も短いパスの長さ，D はカテゴリ構造の深さである．WordSimilarity-353 Test Collection [24] などを利用した実験結果では，Wikipediaを利用した手法

が WordNet に匹敵する評価を残し，Wikipedia が知識抽出のためのコーパスとして有用であることを示した．しかし，この手法で用いるカテゴリ情報は，記事数に対して少なく，各記事に記述されたテキスト情報や他の記事へのリンク情報に比べて，各記事を特徴付ける情報として十分とは言えないため，他の手法に比べて高い精度を実現できていない．

記事テキストに基づく手法

2 つ目の関連度計算手法は，テキスト内容を比較し，その類似度を利用する手法である．テキストを用いた手法は，概念に関する説明文（記事内容）が充実している場合に有効な手法であり，一般にテキストに出現する単語が重複していればいるほど関連度が高くなるという戦略に基づく手法である．Gabrilovich らの研究 [25] では，単語やテキストの意味を表現するための手法として，Explicit Semantic Analysis（ESA，明示的意味解析）を提案している．ESA では，特定の単語（文字列）またはテキストの意味を，Wikipedia の概念を基底とする高次元ベクトルで表す．Explicit（明示的）という由縁は，背景知識を用いない純粋な統計的手法としての Latent Semantic Analysis（LSA，潜在的意味解析）[21] と比較して，人間の認識に基づく明らかな概念を用いているからである．単語の意味を表すベクトルは，各概念を tfidf で重み付けられた単語のベクトルで表した後，これらのベクトルの逆索引を作成することで得られる．単語の関連度は，ベクトル間のコサイン類似度 [85] によって求められる．テキストの意味を表すベクトルも，出現する全ての単語のベクトルを合成することで求められ，文脈を考慮した多義語の解消が可能となる．また，前節で述べた Strube らの研究 [75] では，カテゴリ構造を用いた手法だけでなく，テキストの重複度に基づく指標も関連度計算に有効であると報告しているが，これもテキスト内容の比較による関連度抽出の手法の一種である．しかし，これらの記事テキストを利用した手法では，言語によっては高度な自然言語処理が必要であり，特に日本語では形態素解析や構文解析，語義解析などが精度に大きく影響するといった側面も持つ．また，リンクデータに比べて記事テキストは膨大であり，解析コストが非常に大きいという問題点もある．

記事間リンクに基づく手法

Wikipedia は Wiki をベースにしており，記事の中に他の概念（を意味する単語）が出現するとその概念に対して対応する記事にリンクが張られるため，全体としてみると，概念をノード，ハイパーリンクをエッジとした一種のネットワークと見なすことができる．通常の Web サイトと異なり，ノード（記事）は概念を表し，リンクは意味的な関係を表す上に，Wikipedia 内部で概念どうしが密なリンク構造を形成している（内部リンクが多い）ため，リンク構造を解析することで概念間の関係性を抽出することが可能である．この特徴を生かし，リンクの構造を解析して関連度計算を行うのが記事間リンクの解析手法である．この分野において，主な手法としては，記事内のリンクの重み付けを行う tfidf に基づく手法や，ネットワーク構造における 2 つの記事のリンク数やホップ数に基づく手法である pfibf がある．

tfidf [71] は，Salton らによって提案された文書中の特徴的なキーワードを抽出するための手法である．tfidf は tf（Term Frequency）と idf（Inverse Document Frequency）の 2 つの指標を利用し，それらの積によって文書中の各語の重要度を計算する．tf は文書中における特定の語の出現頻度であり，文書中に多く含まれる語が特徴語とされる．idf は全文書中に，特定の語が出現する文書数の逆数であり，出現する文書数が多い語は idf の値が小さくなる．つまり，広く使われている一般的な語ほど特徴語としての重要度が低くなる．この tfidf の考え方を Wikipedia のリンクに対して適用した手法に，Milne らによる研究 [51] がある．Wikipedia においては，一記事が一概念に対応し，リンクは他の概念に対する意味的かつ明示的な関係を示す．そのため，tfidf における文書内の語の代わりに，Wikipedia における記事内のリンクを用いることによって，記事内の重要なリンクを抽出し，概念どうしの関係性を求めることができる．tfidf によって記事中の各リンクの重要度を以下の式によって与える．

$$tfidf(l,d) = tf(l,d) \cdot idf(l) \tag{1.2}$$

$$idf(l) = \log \frac{N}{df(l)} \tag{1.3}$$

ここで，$tf(l,d)$ は記事 d におけるリンク l の出現回数であり，$df(l)$ はリンク l を含

む記事数，N は全記事数である．そして各概念をベクトル空間モデル [72] によって，リンクを次元，その各リンクの重要度（重み）を要素としたベクトルを生成する．各概念の関連度の算出は，それらのベクトル間のコサイン類似度 [85] によって求められる．この手法では，1 つの概念の特徴ベクトルを抽出するには 1 つの記事に存在するリンク情報だけを解析すれば良いため，解析データ量に対してスケーラビリティが高い．しかし，それ故に記事の内容に信頼性がない場合やリンク数が少ない場合に，精度が低下する．

pfibf [57, 58] は，ある記事 v_i から v_j の関連度を算出する手法である．pfibf は pf（Path Frequency）と ibf（Inverse Backward link Frequency）の 2 つの指標を利用し，それらの積によって関連度を算出する．pf は記事 v_i から v_j へのパスの多さと，各パスの短さによって決定され，全経路 $T = \{t_1, t_2, \cdots, t_n\}$ によって以下の式で表わされる．

$$pf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)} \tag{1.4}$$

ここで，d は経路 t_k の経路長に応じて増加する関数であり，単調増加関数や指数関数を利用することができる．ibf は全記事中の記事 v_j が参照された数，つまり記事 v_j が持つ Backward リンク数の逆数である．この指標は，記事 v_j に対するリンクが多いほど小さい値になる．したがって，記事 v_i から記事 v_j への関連度は pfibf によって以下の式で与えられる．

$$pfibf(v_i, v_j) = pf(v_i, v_j) \cdot ibf(v_j) \tag{1.5}$$

$$ibf(v_j) = \log \frac{N}{df(v_j)} \tag{1.6}$$

N は全記事数，$df(v_j)$ は記事 v_j が持つ他の記事からのリンク数とする．つまり，pfibf に基づく記事 v_i から v_j への関連度は，v_i から v_j へ多くの短いパスを持ち，v_j の Backward リンク数が少ない場合に高い値を示す．pfibf では，ある記事から n ホップ先の記事までのリンク構造を再帰的に解析し，語彙どうしの関連度を算出している．そのため，1 つの概念に対する計算量が多く，全体として多量の計算が必要になる．

1.4 研究内容

1.2 節で述べたように，Wikipedia はこれまでのコーパスには無い，多くの知識抽出のために有用な特徴がある．このような特徴は，概念間関連度の測定においても，概念の網羅性を向上させることができ，また自然言語処理における未知語の対応，同義語，多義語の判別など，これまでの手法における課題を意識せずに解析可能とすることができる．1.3 節で述べたように，これまで Wikipedia をコーパスとした概念間関連度の測定に関する研究が行われてきた [25,51,55,58,75]．しかし，これまでの手法は解析データ量に対するスケーラビリティや精度の観点から問題があった．たとえば，各概念（記事）ペアの概念間関連度を測定する際，従来手法である記事間のリンク構造を n ホップ先まで再帰的に解析する手法 [58] では，膨大な計算が必要となる．また，特定の記事内の情報だけを用いて概念（記事）の特徴情報を生成する手法 [51] では，記事内の記述が極端に少ない，記事が荒らされているなど，特定の記事の質に大きく精度が影響される．

そこで本論文では，計算量が少なく高精度な概念間関連度の測定を行うことを目的として，Wikipedia をコーパスとした概念間関連度の測定手法を議論する．具体的には，以下の 3 つの研究課題に取り組む．

- リンク共起性解析による概念間関連度

  これまでの Wikipedia を用いた概念間関連度の測定手法においては，tfidf [71] に基づく手法 [25,51] などのように，各記事内に存在するリンク情報（記事内の局所的情報）を用いてその記事（概念）の特徴情報を生成し，概念間関連度を測定する研究が行われてきた．しかし，このような手法の特定の記事の質に大きく精度が影響される問題や，リンク構造を解析する手法 [58] での，計算量に関する問題があった．

  そこで，記事（概念）の特徴情報を生成する際に，Wikipedia 全体に存在するその記事へのリンクに関する統計情報（Wikipedia 全体における大域的情報）を活用することで，特徴情報が特定の記事の質に影響を受けにくく，また計算量の少ない概念間関連度の測定手法を提案する．具体的には，Wikipdia における記事間リンクの共起関係を解析することによって，記事間つまり概念

間の関連度を測定する．本手法で用いる共起性解析は，Wikipedia 全体の記事から抽出したリンクデータをシーケンシャルに解析するアルゴリズムの特性上，リンク構造を n ホップ先まで再帰的に解析する手法に比べて，大幅に計算量を削減することができる．

- 大域的情報と局所的情報を活用した概念間関連度

  リンク共起性解析のような Wikipedia 全体の大域的情報を用いた場合でも，特徴情報を生成する記事（概念）へのリンクが少ない場合，その記事に関する特徴情報が十分取れない場合がある．一方で，そのような記事中にリンクがある程度存在した場合，記事内の局所的情報としてのリンク情報は，その記事の特徴を端的に表す情報として依然として有用である．

  そこで，大域的情報である Wikipedia 全体のリンクの統計情報と，記事内のリンクの統計情報の双方を融合することによって，概念間関連度の測定精度向上を図る．この手法では，リンク共起性解析による記事の特徴ベクトルと，記事内リンクを tfidf で重み付けすることによって求めた特徴ベクトルを合成し，双方の手法によって得た特徴情報を補い合うことによって，概念間関連度の測定精度向上を図る．

- SVR による多様な情報を活用した概念間関連度

  これまで Wikipedia を活用して概念間関連度を測定する手法はいくつか提案されているが，いずれも単独の情報を特徴情報として用いているものが多かった．しかし，Wikipedia から関係性を特徴付けることができる情報は，それぞれの手法で用いられている共起性情報，リンク構造情報，カテゴリ構造情報，その他にもバックワードリンク数やフォワードリンク数など多く存在するため，それらの情報の取捨選択，融合手法が必要とされてきた．

  そこで，Wikipedia から取得可能なそれらの情報を網羅的に提案し，どの情報が概念間関連度の測定に対して有用な特徴情報であるかを，機械学習における各学習情報（素性）の貢献度を計算する方法を用いて調査する．さらに，機械学習手法によってそれら複数の情報を組み合わせることによって，計算量と精度の両面を考慮した概念間関連度の測定手法を提案する．

1.5 本論文の構成

本論文は，5 章から構成され，本章以降の内容は次の通りである．

まず，第 2 章では，リンクの共起性解析に基づき，高精度で計算量の少ない概念間関連度の測定方法を提案する．この方法では，記事内における記事間リンクに対しての共起回数をカウントし統計量を取ることで，2 つの記事の共起性を求める．そして，共起性の値を用いて記事を特徴付ける「共起ベクトル」を生成し，そのベクトルの類似度を求めることで，2 つの記事間（概念間）の関連度を測定する．最後に，提案方法の性能評価のために行った評価実験の結果を示し，その有効性について検証する．

第 3 章では，Wikipedia 全体の大域的情報と記事内の局所的情報の両方を活用することによって，それぞれの手法が持つ弱点を補完し，より高精度な概念間関連度の測定を可能とする方法を提案する．具体的には，記事の特徴を表すために，Wikipedia 全体でのリンクの共起性情報と，tfidf の考え方に基づいた記事内のリンクの重要度をそれぞれベクトル化し，ベクトルの合成を行う．この際，ベクトルの合成方法を多角的に検討し，実験により最適な合成方法を発見することを目指す．さらに，提案方法の性能評価のために行った評価実験の結果を示し，その有効性について検証する．

第 4 章では，Wikipedia から取得可能な記事間の関連性を特徴付ける様々な情報の中で，どの情報が概念間関連度に有用であるかを検証し，さらに複数の情報を組み合わせた高精度な概念間関連度の測定方法を提案する．この方法では，機械学習における素性の重要度を測定する F-score によって重要な情報の検証を行う．また，回帰問題を解くことのできる機械学習手法である SVR を用いて，素性セットのパターンと概念間関連度の関係を学習させる．これにより，以下に挙げる 3 つの効果が得られる

- 未知の概念ペアに対する関連度が高精度に予測できる．
- 1 つの素性で十分に精度が得られない時に他の素性で補完ができる．
- 精度を高めるのに重要な素性の組み合わせや，計算量の低減に効果のある素性の組み合わせなどが検討できる．

さらに，提案方法の性能評価のために行った評価実験の結果を示し，その有効性について検証する．

第 5 章では，本論文の成果を要約したのち，今後の研究課題について述べ，本論文のまとめとする．

なお，第 2 章は，文献 [32,33,34,35,37] で公表した結果に基づき論述する．第 3 章は，文献 [33,34,37] で公表した結果に基づき論述する．第 4 章は，文献 [38,61] で公表した結果に基づき論述する．

第2章　リンク共起性解析による概念間関連度

2.1　まえがき

これまでのWikipediaを用いた概念間関連度の測定手法において，各記事内の情報によってその記事が表す概念を特徴付ける手法の場合，精度が記事の質に影響される場合がある．また，記事間のリンク構造を解析して記事間つまり概念間の関連度を求める手法の場合，膨大な計算量が発生するという問題があった．

そこで本章では，Wikipedia全体の統計情報を活用することで，特徴情報が特定の記事の質に影響を受けにくく，また計算量の少ない概念間関連度の測定手法を提案する．具体的には，Wikipdiaにおける記事間リンクの共起関係を解析することによって，記事間つまり概念間の関連度を測定する．Wikipediaにおける記事間（概念間）の関係を表すリンクの共起が概念間関連度の測定に有効であるかは，これまでの研究において調査されたことがなく，本研究においてそれを確認する．提案手法は，様々な人に編集されたWikipedia全体の記事におけるリンクペアの共起性を用いるため，リンクの特徴情報が特定の記事の情報に依存しないという特徴を持つ．また，Wikipedia全体の記事から抽出したリンクデータをシーケンシャルに解析するアルゴリズムの特性上，バックワードリンク，フォワードリンクのリンク構造を各記事からnホップ先の記事まで再帰的に解析する手法に比べて，大幅に計算量を削減することができる．さらに本章の最後では，提案方法の性能評価のために行った評価実験の結果を示し，その有効性について検証する．

以下では，第2.2節で提案手法について述べ，第2.3節で評価実験の結果を示す．最後に第2.4節で本章のまとめを行う．

2.2　リンク共起性解析

2.2.1　概要

1.3.3 節に挙げたように，Wikipedia から概念間関連度を測定する際，従来手法ではいくつかの問題があった．まず，カテゴリリンクを活用する手法では，記事に対してカテゴリ情報が不足しているため，精度が良くない．また，記事テキストを用いる方法では，膨大な記事テキストを解析する計算コストの問題や，高度な自然言語解析が必要であるという問題がある．そこで，Wikipedia の記事間のリンクを活用する手法が提案されてきたが，この手法にも 2 つの問題がある．1 つ目の問題は，tfidf に基づく手法のように記事（概念）の特徴情報を生成する際に，記事内のリンク情報（局所的情報）のみを用いるため，その記事の質が悪い場合に精度が低下する点である．2 つ目の問題は，pfibf では n ホップ先までのリンク構造を再帰的に解析しているため，膨大な計算が必要であるという点である．そこで著者らは，リンクの共起性に着目した．Wikipedia 全体を通したリンクの共起性による特徴情報は，tfidf に基づく手法のような記事内の局所的情報ではなく，Wikipedia 全体における大域的統計情報を用いており，ある特定の記事の質に大きく左右されることはない．その理由は，tfidf に基づく手法の場合は特定の記事に関する特徴情報を求めるとき，その記事内のリンク情報を用いるため，特定の記事の質に大きく左右されるが，Wikipedia 全体における特定の記事へのリンクの共起性を解析した場合，その特定の記事の特徴情報は Wikipedia 全体の統計的情報に基づくためである．また，その計算時間はデータ量に対して線形であるため，pfibf のように多量の計算が必要になることはない．

ここで，提案手法において参照先 URL が同じリンクは，たとえアンカーテキストや出現する記事が違っても，同じリンクであると見なす．たとえば図 2.1 の例では，アンカーテキスト “MS” と “Microsoft” は同じ記事 “Microsoft” を参照しており，同一のリンク（記事 “Microsoft” へのリンク）であると見なす．つまり，ここでのリンクは Wikipedia の各記事と一対一に対応しており，提案手法では特定の記事への参照に関する統計的解析を行うものである．たとえば，リンク A の Wikipedia 全体での出現回数は，リンク A が指し示す記事 A の被参照回数であり，記事 A の

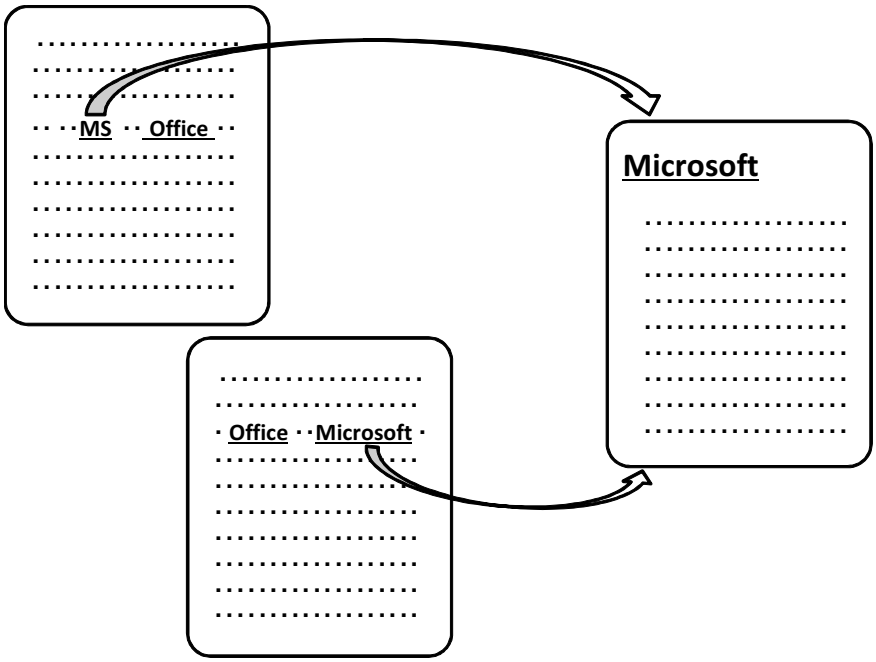

図 2.1: 同一リンクの定義

バックワードリンク数と等しくなる.

本節では，Wikipedia におけるリンクの共起性に基づいて，2 つの概念間の関連度を求める手法を提案する．以降では，まず従来研究における単語の共起性によって関連度を求める手法を解説した後，提案手法におけるリンク間の関連度の算出方法を述べる.

2.2.2 単語の共起性解析による関連度の算出

単語の共起とは，特定の範囲において，ある組の単語が同時に出現することであり，単語の共起性解析は，頻繁に共起する単語ペアは関連度が高いという考えに基づいている [42,65,73]．単語の共起性解析は，従来研究においては連想シソーラス辞書構築や概念間関連度の測定に利用されてきた．ここでは，文書コーパスにおける単語の共起性を解析することによって単語ペアの関連性を求める手法についての 2 つの先行研究を紹介する.

共起回数による単語間の関連度

共起回数から関連度を求める代表的な手法として，Cosine，相互情報量，Dice 係数がある [42, 65]．以下では，それぞれにおける関連度の計算式を示す．なお，式中の $P(x)$，$P(y)$ は単語 x と y がそれぞれ独立に出現する確率，$P(x,y)$ は x と y が同時に出現する確率とし，f_x,，f_y は x と y がそれぞれ独立に出現する回数，f_{xy} は x と y が同時に出現する回数とする．

- Cosine

$$Cosine(x,y) = \frac{f_{xy}}{\sqrt{f_x f_y}} \tag{2.1}$$

- 相互情報量

$$MI(x,y) = \log \frac{P(x,y)}{P(x)P(y)} \tag{2.2}$$

- Dice 係数

$$Dice(x,y) = \frac{2 \cdot f_{xy}}{f_x + f_y}, \quad (0 \leq Dice(x,y) \leq 1) \tag{2.3}$$

北村らは，Dice 係数の欠点は単語ペア出現回数の大小に関わらず，独立出現回数と同時出現回数の相対比により関連度が決まるという点であり，出現回数が少ない場合の信頼性の違いを考慮していないと指摘し，Dice 係数に共起回数による重み付けを行った改良版の Dice 係数を提案している [42]．以下にその式を示す．

$$\begin{aligned} IDice(x,y) &= w(f_{xy}) \frac{2 \cdot f_{xy}}{f_x + f_y}, \\ w(f_{xy}) &= \begin{cases} f_{xy} \\ \log f_{xy} \end{cases} \end{aligned} \tag{2.4}$$

二次共起（second-order co-occurrence）

Hinrich らは文書コーパスから単語の共起に基づく連想シソーラス辞書を構築し，情報検索に応用する手法を提案している [73]．具体的には，式 2.1, 式 2.2, 式 2.3, 式

2.4 に示すような共起回数だけで，ある組の関連度を算出する一次共起（first-order co-occurrence）に異論を唱え，ある組の語がどれくらい同じ語と共起しているかで関連度を算出する二次共起（second-order co-occurrence）を提案している．

この手法では，まずすべての単語を行と列においた正方行列 C を作り，その各要素 c_{ij} を単語 i と j の共起回数としている．ここで任意の単語 i における行ベクトルをシソーラスベクトル（thesaurus vector）とし，単語 i と j の関連度はそれぞれのシソーラスベクトルのコサイン類似度によって求められる．ここで，単語 i, j のコサイン類似度は，シソーラスベクトルを v_i, v_j とすると，以下のように表される．

$$\begin{aligned} cos(v_i, v_j) &= \frac{v_i \cdot v_j}{|v_i||v_j|} \\ &= \frac{\sum_{k=1}^{n} c_{ik}c_{jk}}{\sqrt{\sum_{k=1}^{n} c_{ik}^2}\sqrt{\sum_{k=1}^{n} c_{jk}^2}} \end{aligned} \tag{2.5}$$

2.2.3 提案手法

提案手法ではリンクの共起性を解析することによってリンク間（記事間）の関連度を算出する．リンクの共起とは，単語をリンクとして扱うということ以外，基本的な概念は単語の共起と同様である．つまり，リンクが共起するということは，特定の範囲においてある異なる 2 つのリンクが同時に出現するということである．リンク共起性解析では，Wikipedia 全体でのリンクの共起性を解析し，2 つのリンクの関連度，つまり Wikipedia の記事が表す 2 つの語（概念）の関連度を求める．

ところで，Wikipedia を解析する時，同じ記事内での共起をカウントすると，リンク数の多い記事の場合，非常に膨大な共起の組み合わせが存在する．そこで解析範囲を近傍のリンクに限定するウインドウを設定して，ウインドウ内のリンクだけ共起していると見なす方法が提案されている [73]．たとえば図 2.2 の，ある “○○○” という語に関する記事（図上部）から，解析対象データであるリンクを出現順で並べたデータ（図下部）を生成した例を題材に説明する．図中のアルファベットは，その記事のリンクにおけるリンク先記事を表している．この例では，ウインドウサイズが 3 の場合の解析例を示しており，解析対象データの先頭からウ

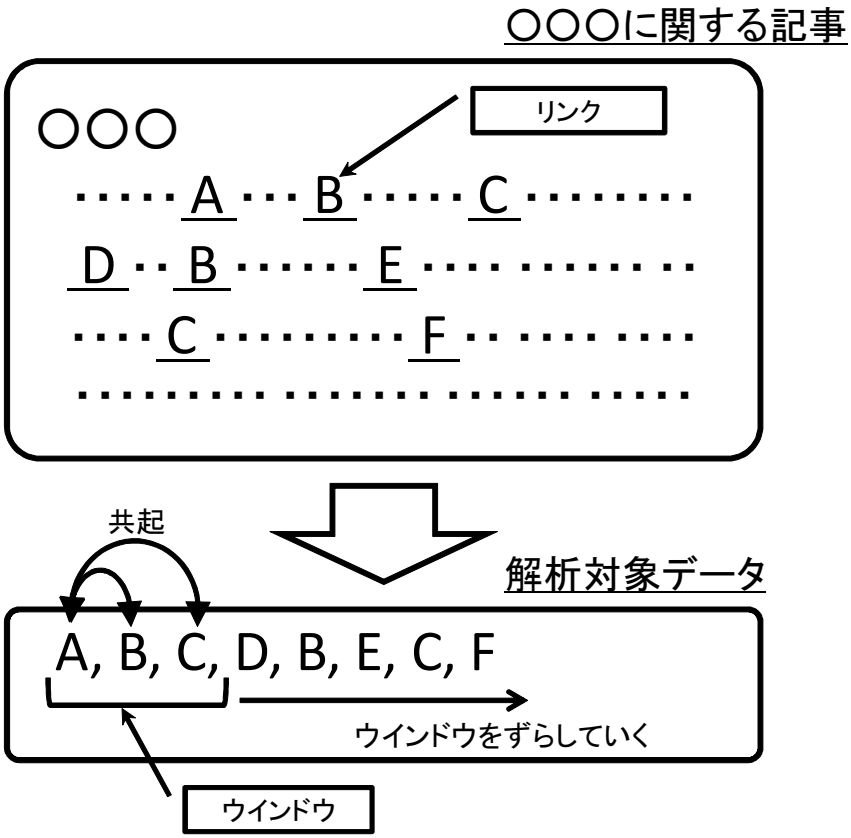

図 2.2: 記事の解析例

インドウが図に示すように 1 つずつ移動する．各ウインドウの位置で，図に示すような 2 つのリンクペアが共起していると見なされる．このように指定されたサイズのウインドウに基づいて計算された各記事のリンクペアの共起回数を，全記事で合算することによって，Wikipedia 全体におけるそれぞれのリンクペアの共起回数を算出することができる．

上述までの手法をリンクの一次共起と呼ぶ．ここで，リンクの一次共起のみを用いた場合，直接共起しないリンクペアは関連性がないと判断されるが，本来関連性はあるが直接共起しないようなリンクペアも存在する．そこで提案手法では，リンク間の関連度を算出するためにリンクの二次共起による関連度を用いる．リンクの二次共起は一次共起の問題を解決するために，たとえ直接共起しなくても，

共起特性の類似性，つまりそれぞれが共起する複数のリンクの中で，どの程度同じリンクを共有しているかで関連性を測る．実際，予備実験により，高い関連度であると考えられるリンクペアのうち，リンクの一次共起では関連性がないとされる場合でも，リンクの二次共起では高い関連度を示すものが存在することを確認している．たとえば，2 つの語「OPEC」と「Oil」は，直感的にも高い関連性を持つことが分かる．しかし，一次共起による解析の場合，関連性はゼロと算出された．一方，二次共起を適用した場合はこの 2 つの語は高い関連性を示した．具体的には，WordSimilarity-353 Test Collection [24] に含まれる 353 の単語ペアのうち，関連度が 5 以上の 244 ペアで調査した結果，一次共起では 192 ペアの関連度がゼロであった．これは，約 79%のリンクペアが一度も直接共起していないということを表している．一方，二次共起ではすべてのペアの関連度をゼロより大きな値で求めることができていた．以上の理由により，リンクの共起性解析をする際，一次共起よりも二次共起の方が，より関連度を求める方法として適していると判断した．

二次共起では，まずリンクの一次共起による関連度を求めた後，その関連度を使ってリンクの二次共起による関連度を算出する．以下では，それぞれについて解説する．

リンクの一次共起による関連度の算出

本研究においては，リンクペアの関連度を求める際に，二次共起によって共起性を求め，それを関連度としている．二次共起の計算におけるベクトル生成において，先に挙げた Hinrich らの手法では共起回数をベクトルの各要素としている．しかし共起回数だけを利用した場合，多く出現しているリンクは出現回数の低いリンクより，どのリンクとも共起する可能性が高くなる．つまり，出現回数の高いリンクほど関連度が高くなる可能性がある．たとえば，あるリンク A と B がそれぞれ 1,000 回出現していて，A と B の共起回数が 100 回であるのと，あるリンク C と D がそれぞれ 100 回出現していて，C と D の共起回数が 100 回であるのとでは同じ関連度であるとみなされる．しかしこの場合，リンク C と D はすべての出現において共起しているので，C と D の関連度の方が高いことは明白である．こ

の問題を解決するために考慮しなければならないことは，共起ペアのそれぞれの出現回数に対して共起回数が何回であるかということである．そこで，リンクの出現回数を考慮した計算方法として 2.2.2 項に挙げた 4 つの式をリンク間の一次共起による関連度として定義する．ここで，本研究ではリンクの共起による解析を行うため，語の共起性解析で示した 4 つの式の中に出現する x, y を単語ではなくリンクとして扱う．

リンクの二次共起による関連度の算出

二次共起による関連度を求める際，各リンクにおいてどのようなリンクと共起するかという，各リンクの共起特性を表すリンクベクトルを生成する．リンクベクトルは，リンクを次元，各リンクに対する重み（一次共起性）を要素とする，ベクトル空間モデルに基づく多次元ベクトルであり，リンク i の共起特性を表わすベクトル v_i は以下のように表される．

$$v_i \quad = \quad \{l_{i1}, l_{i2}, l_{i3}, \cdots, l_{in}\} \tag{2.6}$$

ここで，n は Wikipedia に存在する全記事数，l_{ij} はリンク i, j 間の重みであり，2.2.2 項で挙げた 4 つの一次共起性の計算式のいずれかによって算出される．

このように作成されたリンクベクトルを利用し，2 ベクトル間のコサイン類似度によって，それぞれのリンクの共起性パターンがどれだけ同じかという情報を元に，関連度を求めることができる．

ここで，処理の流れを図解によって示す．図 2.3 では概念ペアが与えられたときに，その概念ペアの関連度を生成されたベクトルから算出する流れを表している．まず，概念が与えられれば，その概念に対応するリンクベクトルを抽出し，そのベクトル間のコサイン類似度を求めることによって，概念ペアの関連度を求めることができる．

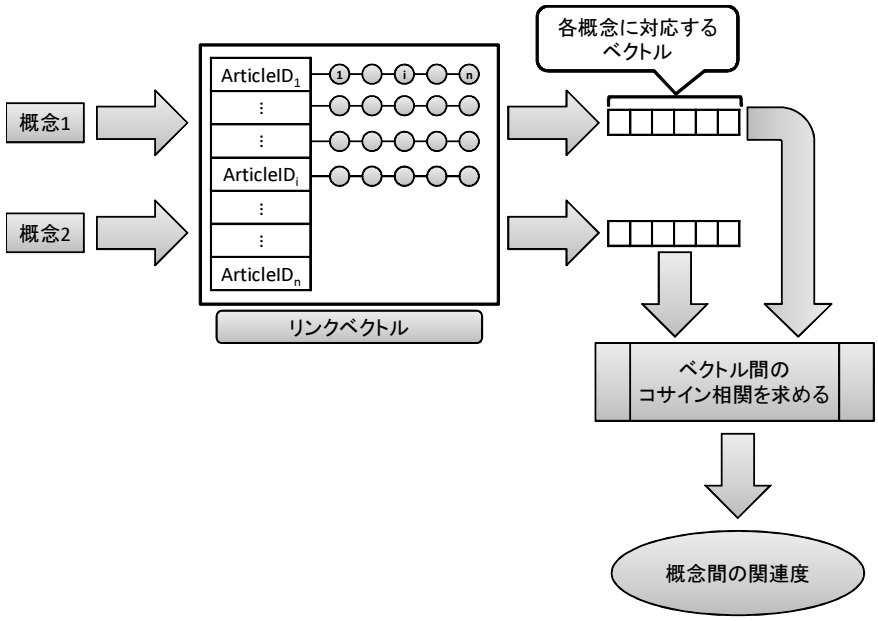

図 2.3: 処理の流れ：概念間の関連度の算出

2.3 評価実験

2.3.1 実験概要

本実験の目的は，提案手法によって測定される概念間関連度の精度と解析時間を評価することである．提案手法はリンク共起性解析であり，比較対象としては，tfidf ベースの手法と pfibf を用い，それぞれの手法によって Wikipedia から概念間関連度を測定し，解析時間と精度を比較した．なお，pfibf においてはホップ数 n を 2 に設定した．

ここで，これらの手法に加えて pfibf によってベクトル化した以下の手法 2nd pfibf も比較手法として用いる．まず，各概念（記事）を pfibf によって取得した関連度の高い関連概念 100 件とその関連度によってベクトル化する．そのベクトルどうしのコサイン類似度によって 2 つの概念の関連度を求める．本手法を用いる理由は，pfibf は記事の n-hop 先まで解析することによって，ある概念の関連概念セットを取得するためにデザインされており，関連度が強くない概念ペアを含む任意

の概念ペアの関連度測定には向いていないためである．実際に，予備実験において pfibf のみでは，測定の出来ない概念ペアが非常に多く，非常に低い精度となっている．

解析対象の Wikipedia のデータとしては，2008 年 8 月時点の英語版 Wikipedia のデータからノイズ記事を除去した，記事数約 150 万，総リンク数約 5,800 万のデータを用いた．ノイズ記事の定義は，トップページやカテゴリページなどの通常の記事ではないもの，記事内のリンク数が 5 つ以下のものである．

精度測定方法

精度に関しては，付録 A で述べる「WikiSimi Test Collection」を用いて計測した．このテストコレクションは，約 1,749 組の単語ペアを 5 人から 9 人の被験者によって関連性を主観で 10 段階評価してもらい，その平均値を関連度としている．このテストコレクション内の概念ペアに対して，その関連度を 1.3.3 項で説明した各手法や提案手法によって求め，テストコレクションにおける関連度と，実験によって測定した関連度をそれぞれ順位付けする．そして，2 つの順位の相関性を「スピアマンの順位相関係数（Spearman rank-order correlation coefficient）[74]」によって求め，概念間関連度の精度とした．スピアマンの順位相関係数とは，2 つの数値の系列における数値の大きさの順位が，どの程度近いかという順位の相関性を計算する手法であり，順位が全く同じならスピアマンの順位相関係数は最大の相関である 1 となり，順位が全く逆なら負の相関である -1 となる．

具体的には，WikiSimi Test Collection よりランダムに 100 件の概念ペアを取得し，その 100 件の概念ペアを評価セットとして用いて精度測定を行う．その評価セット作成と精度測定の一連の作業を 50 回繰り返し，その精度の平均値を求めることによって本実験の精度評価とする．この作業を行う理由としては，本実験によって得られた実験結果の精度が，テストコレクションに対して過剰適合していないことを保証するためである．

解析対象データの準備

Wikipedia のデータは GNU ライセンスに従い自由に入手し，解析可能である．Wikipedia の全データは定期的にバックアップされ，ダンプデータと呼ばれるデータとして保存されている．本実験では，ダンプデータのダウンロードサイト[1]に用意されている XML ファイルをダウンロードし，専用ツールを利用することで MySQL にデータをインポートした．いくつかツールが存在するが，本実験では Java ベースの mwdumper[2]を利用した．

MySQL に構築された Wikipedia のデータベースから，各記事のソースを解析し Wikipedia 内の記事へのリンクだけを抜き出す．この時点で，リンク元記事とリンク先記事の 2 つの列を持つリンク情報テーブルを MySQL に構築する．このテーブルから，さらにフォワードリンク数やバックワードリンク数などの必要なテーブルを作成する．次に，リンク情報テーブルに対して，フォワードリンク数やバックワードリンク数などの情報を元に，ノイズ記事を含むリンクの行を削除する処理を施したノイズ除去済みリンク情報テーブルを生成し，解析対象データとして利用した．

2.3.2 実験結果と考察

本項では，本実験の結果における解析時間と精度のそれぞれについて解説し考察する．解析には，上述の手順で構築したデータを利用して，表 2.1 に示す計算機環境を用いた．

解析に要する時間

ここで評価する解析時間は，2 つの概念ペアが与えられれば，即時にその概念ペアの関連度を算出できる状態まで Wikipedia 全体を解析するまでの時間であり，たとえば，提案手法においては，Wikipedia のすべてのリンクの共起性解析を終えるまでの時間である．表 2.2 に，提案手法におけるウインドウサイズ 2 から 4 の場合

[1] http://download.wikipedia.org
[2] http://svn.wikimedia.org/svnroot/mediawiki/trunk/mwdumper/

表 2.1: 計算機環境

項目	仕様
CPU	Intel Xeon 5160 3.0GHz × 4
メモリ	16 GB
OS	SUSE Linux Enterprise Server 10
開発言語	C++
コンパイラ	Intel C++ Compiler 9.1

表 2.2: 解析に要する時間

手法	計算時間（秒）
リンク共起性解析（ウインドウサイズ 2）	486
リンク共起性解析（ウインドウサイズ 3）	870
リンク共起性解析（ウインドウサイズ 4）	1,248
tfidf	513
pfibf	561,600

と，比較手法における解析に要する時間を示す．なお，提案手法における二次共起のベクトルの要素となる，4 つの一次共起の各計算手法 (Cosine, MI, Dice, IDice) の計算時間を計測したが，ほとんど変化がなかった．そのため，ここでは Cosine を用いた解析時間のみを示す．

提案手法の共起性解析ではウインドウサイズが 2 から 4 に増えるに伴って計算時間が増加しているが，その増加率はウインドウサイズの増加に対して 400 秒程度と線形である．共起性解析と tfidf に基づく手法の計算時間を比較すると，共起性解析のウインドウサイズ 2 においては tfidf と比べて約 0.95 倍の時間を，ウインドウサイズ 4 においては約 2.43 倍の時間を要している．

次に，共起性解析と pfibf の計算時間を比較すると，pfibf は共起性解析のウインドウサイズ 2 に対して約 1,155 倍もの時間を要している．ウインドウサイズ 4 の

処理と比較しても約 450 倍の計算時間を要している．これは，明らかに提案手法の方が計算時間において大幅に有利であることを示している．pfibf は手法の特性上，ある記事から n ホップ先の記事までのリンク構造を再帰的に計算する．pfibf に関する論文 [58] に述べられている近似手法を用いても，膨大な計算が必要である．一方，共起性解析や tfidf はリンク先を再帰的に処理しないため，少ない計算量に抑えられている．

ここで，本実験により生成されたリンクベクトルの数と次元数は，実験に用いた Wikipedia の記事数 1,542,302 と等しくなる．しかし，実際には長さがゼロのベクトルや，ゼロ要素を含むベクトルは圧縮している．圧縮後のデータ量としては，ウインドウサイズ 2 の場合，ベクトルの非ゼロ要素の最大数は 114,567 次元，平均長は約 33 次元であった．また，ウインドウサイズ 4 の場合，ベクトルの非ゼロ要素の最大数は 254,352 次元，平均長は約 104 次元であった．このような処理によって，150 万次元の正方行列を扱う場合に比べて，計算機の現実的なメモリリソース上での計算を可能にしていることが分かる．

概念間関連度測定の精度

提案手法によって測定した概念間関連度の精度として，ウインドウサイズ 2 から 5 のそれぞれにおいて，二次共起性の計算におけるベクトルの要素に 2.2.2 項で示した 4 つの一次共起の計算手法 (Cosine, MI, Dice, IDice) を用いた場合の，50 回の実験結果の平均精度を表 2.3 に示す．また，表 2.4 に比較手法における 50 回の実験における概念間関連度の平均精度を示す．

まず表 2.3 の提案手法であるリンク共起性解析の結果より，精度はウインドウサイズの違いでほとんど変化が見られなかった．一次共起性の計算手法別による精度の違いを比較すると，どのウインドウサイズにおいてもベクトル生成の際に Cosine を用いた場合に最も精度が高くなっている．特に，ウインドウサイズ 3 の時に最も精度が良い．一方，その他の一次共起性の計算手法 (MI, Dice, IDice) を用いてベクトルを生成した場合，ウインドウサイズが増えるに従って，精度が悪くなっている．このことは，リンク共起性解析においては隣り合うリンクを共起と見なすだけで十分であり，隣り合っていない離れたリンクを共起と見なすことは，

表 2.3: 概念間関連度の精度（平均値）：提案手法（リンク共起性解析）

ウインドウサイズ	手法	スピアマンの順位相関係数
2	Cosine	**0.457**
	MI	0.269
	Dice	0.431
	IDice	0.345
3	Cosine	**0.467**
	MI	0.258
	Dice	0.429
	IDice	0.341
4	Cosine	**0.462**
	MI	0.254
	Dice	0.423
	IDice	0.335

（各ウインドウサイズについて上位 1 件を強調表示）

表 2.4: 概念間関連度の精度（平均値）：比較手法

手法	スピアマンの順位相関係数
tfidf に基づく手法	0.410
pfibf	0.212
2nd pfibf	0.482

精度の低下を招くということを示唆している．結果として，本実験ではウインドウサイズが 3，一次共起性の計算手法が Cosine の精度が最も良い．以下では，提案手法と他手法との比較のための説明にこのセッティングにおける精度を用いる．

次に表 2.4 の tfidf と比較すると，提案手法であるリンク共起性解析は tfidf より

高い精度を実現している．これは，tfidf では記事内に含まれるリンクのみを利用し，記事（概念）に対する特徴ベクトルを抽出しているため，記事の質に応じて精度が低下したことに起因すると考えられる．Wikipedia において，各記事は限られたユーザによって編集されているので，記事内のリンク数や信頼性は均質でない．そのため，各記事のリンクによって得られる情報は必ずしも一般的というわけではなく，偏った内容となっている可能性がある．しかし，Wikipedia のすべての記事を通して出現する各記事へのリンク数などの統計的情報は，一部のユーザによる偏った情報ではなく，各記事に対する一般的な認識による情報となっている．つまり，ある記事へリンク付けを行うかどうかなどの，書き手によるリンク付けの偏りが存在した場合においても，リンク共起性解析は統計的情報を利用しているため，書き手によるリンク付けの偏りは平均化され，客観的情報が得られる．この理由により，tfidf よりリンク共起性解析の方が高い精度を実現したものと考えられる．

また表 2.4 の pfibf とリンク共起性解析を比較すると，提案手法は pfbif を上回る精度を実現しているが，2nd pfibf（ベクトル化した pfbif による手法）と比較すると，リンク共起性解析は低い精度となっている．しかし，リンク共起性解析における最大の精度である 0.467 は，2nd pfibf の精度である 0.482 に迫っている．前項で述べたリンク共起性解析と pfibf の計算時間が約 1,155 倍も違うことを考えると，リンク共起性解析は大幅に少ない計算量で 2nd pfibf に迫る精度を実現しているといえる．

また，平均精度だけではなく，50 回の実験の時系列での 10 回移動平均を，tfidf に基づく手法，2nd pfibf，提案手法に関して図 2.4 に示す．ここで，移動平均とは系列データを平滑化する手法であり，外れ値を排除し傾向を把握する用途に用いられる．ここでは，10 回の実験の移動平均として実験番号 1〜10 の平均値，実験番号 2〜11 の平均値，...，実験番号 41〜50 の平均値という 41 個の値を計算しグラフ化している．図 2.4 より，提案手法は，どの実験区間においても tfidf に基づく手法より優位であることがわかり，2nd pfibf に匹敵していることが分かる．

本実験結果より，Wikipedia をコーパスとしてリンクの共起性を解析を行うことは，概念間関連度の測定に有効であることが分かった．

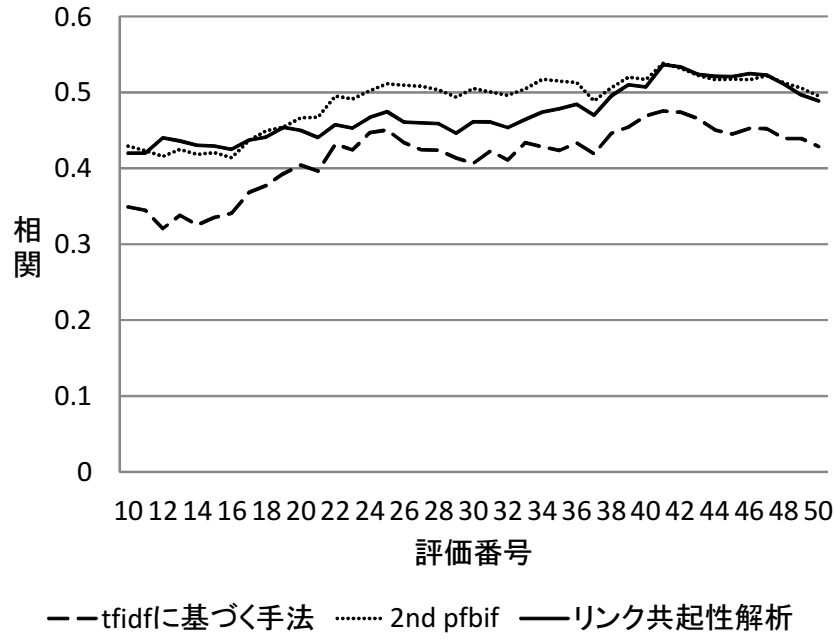

図 2.4: 他手法との比較（10 回の移動平均）

2.3.3　概念間関連度の測定例

本項では，提案手法によって求めた概念間関連度の例として，企業・製品・その他一般語に分けて，関連度の上位 30 件の概念リストを表 2.5, 表 2.6, 表 2.7 に示す．これらの表より，提案手法によって，直観的にも連想される概念が抽出できていることが分かる．

2.4　むすび

本章では，大規模な Web 事典である Wikipedia を解析し，概念間関連度を測定するスケーラビリティの高い手法として，リンクの共起性解析に基づく手法を提案した．提案手法では，記事内における記事間リンクに対して近傍での共起をカウントし，その処理を Wikipedia のすべての記事で行うことによって，2 つの記事の共起性を求める．そして，その共起性の値を用いてある記事の共起ベクトルを生成し，そのベクトルの距離をコサイン類似度によって求めることで，2 つの記事

表 2.5: 概念間関連度の例 : 企業

Rank	Microsoft	Apple Computer	Google
1	Microsoft Office	Macintosh	Froogle
2	Microsoft Windows	IMac	Google Image Search
3	Microsoft Dynamics	Mac OS X	Google Book Search
4	Windows Server System	.Mac	Google Groups
5	Windows XP	ILife	Yahoo!
6	Bill Gates	IPod	Search engine
7	Amentra	Steve Wozniak	Sergey Brin
8	Windows Vista	Xgrid Admin	Larry Page
9	Windows NT	Workgroup Manager	Google News
10	Internet Explorer	Steve Jobs	PageRank
11	Xbox	Jef Raskin	Yahoo! Search Marketing
12	IBM	Server Monitor	Google platform
13	Operating system	Proprietary software	Symantec
14	Sun Microsystems	Mac OS X v10.4	Pyra Labs
15	Windows 2000	Server Assistant	Google Scholar
16	Windows Server 2003	EMC Corporation	Gmail
17	Proprietary software	Front Row	SAP AG
18	Xbox 360	WebKit	Features of Gmail
19	Windows 95	Mac OS	Oracle Corporation
20	Paul Allen	Mike Markkula	Amazon.com
21	Windows Live Mobile	XNU	SYSTRAN
22	Windows Live Mobile Search	Fred D. Anderson	Urchin Software Corporation
23	Windows Live Writer	Batch Monitor	Novell
24	KDS	Apple Qadministrator	MSN Search
25	Silicon Graphics	Symantec	EBay
26	Windows Live Toolbar	PlainTalk	Spamdexing
27	Hewlett-Packard	IWork	W3C Markup Validation Service
28	Microsoft Visual Studio	Power Macintosh	Web crawler
29	Accenture	Universal binary	David Bret
30	ITA Software	Mac OS X Server	Inktomi

表 2.6: 概念間関連度の例 : 製品

Rank	IPod	Macintosh
1	IPod mini	Apple Computer
2	IPod photo	Microsoft Windows
3	IPod shuffle	Amiga
4	IPod nano	System 6
5	MacBook Pro	Atari ST
6	PowerBook G4	DOS
7	ITV (Apple)	Floppy disk
8	PowerBook G3	Mac OS history
9	Jerry York (businessman)	System 7 (Macintosh)
10	Digital audio player	Apple IIGS
11	Xserve	Apple II series
12	Creative Zen	Personal computer
13	Eric E. Schmidt	MS-DOS
14	Mac mini	Jerry York (businessman)
15	Iriver	Mac OS
16	MacBook	IBM PC compatible
17	IMac	Apple Lisa
18	Apple Computer	Xgrid Admin
19	Creative NOMAD	Mac OS X
20	Rio Karma	Workgroup Manager
21	AirPort	ITV (Apple)
22	ITunes Store	Server Monitor
23	Mac Pro	Commodore 64
24	ITunes	CD-ROM
25	ISight	Arthur D. Levinson
26	Jamie Kane	Computer keyboard
27	FairPlay	IBM PC
28	Los Altos School District	Macintosh Quadra
29	Iriver H10 series	Eric E. Schmidt
30	Apple Cinema Display	Universal binary

表 2.7: 概念間関連度の例 : その他一般語

Rank	Apple	Tennis	Sausage
1	Pear	Volleyball	Bacon
2	Cherry	Softball	Pork
3	Peach	Association of Tennis Professionals	Sausage making
4	Plum	Cross country running	Ham
5	Apricot	Golf	Stuffing
6	Grape	Swimming	Blood sausage
7	Granny Smith	The Championships, Wimbledon	Baked beans
8	Strawberry	Australian Open	Saveloy
9	Orchard	Football	Confit
10	Quince	Wrestling	Fried bread
11	Raspberry	Lacrosse	Horilka
12	Gooseberry	Basketball	Beef
13	Red Delicious	Tennis at the Summer Olympics	Instant mashed potato
14	Blackcurrant	Diving	Valio
15	Codling moth	Water polo	Salami
16	Golden apple	Davis Cup	Veal
17	Gymnosporangium	Table tennis	English muffin
18	Light brown apple moth	World No. 1 Tennis	Italian sausage
	Player Pre-ATP Rankings		
19	Geranyl acetate	Baseball	Hot dog
20	Banana	International Tennis Hall of Fame	Ground beef
21	Pineapple	Badminton	Onion
22	Malus sieversii	Smash (tennis)	Currywurst
23	Golden Delicious	Field hockey	Potato bread
24	Cider	Lob (tennis)	Bratwurst
25	Fruit	Gymnastics	Shine On You Crazy Diamond
26	Lemon	Volley (tennis)	PA ¢ tAc
27	Walnut	Sport rowing	Pancake
28	Cooking apple	Cheerleading	Meat
29	Malus	Drop shot	Riddles Are Abound Tonight
30	Blackberry	Tennis terminology	Toast

間，つまりそれらの記事が表す概念間の関連度を測定する．

提案手法の性能評価のために行った評価実験により，提案手法は，従来研究である tfidf に基づく手法よりも高い精度を実現し，また pfibf よりも解析時間が大幅に短いにもかかわらず，pfibf を元に特徴ベクトルを生成する手法である 2nd pfibf に迫る高い精度を実現していることが証明できた．特に，共起性解析によるリンクベクトルを生成する際の一次共起性の計算手法としては，Cosine が最も高い精度で概念間関連度を測定できることが判明した．

本研究によって，Wikipedia をコーパスとしてリンクの共起性の解析を行うことは，概念間関連度の測定に有効であることが分かった．

第3章　大域的情報と局所的情報を活用した概念間関連度

3.1　まえがき

前章では，概念間関連度を測定するための各記事の特徴情報を求める際，tfidf に基づく手法のように記事内に存在するリンク情報（その記事のフォワードリンク情報），つまり記事内の局所的情報を用いる手法の問題点を指摘し，リンク共起性解析のような Wikipedia 全体のその記事へのリンク情報（フォワードリンク情報），つまり Wikipedia 全体の大域的情報を用いることの有効性を示した．しかし，これは局所的情報である各記事のフォワードリンク情報の有効性を否定するものではない．たとえば，リンク共起性解析のような各記事のバックワードリンクに関する統計情報を Wikipedia 全体から解析する手法では，バックワードリンクが少ない場合に統計情報が少なくなり，その記事に関する特徴情報を適切に取得できない可能性がある．しかし，そのような記事でもフォワードリンクが存在する場合，その記事の特徴を端的に表現する情報源として有用な情報になると考えられる．そこで，これら 2 つの性質の異なる情報を，双方とも扱うことによって，情報を補い合い概念間関連度の測定精度向上を図れる可能性がある．

一方で，中山ら [57] の研究では，専門的な概念の記事と一般的な概念の記事でフォワードリンクとバックワードリンクの重要性に違いがあることを示している．たとえば，ドメイン特有の単語（専門用語）の場合には特にバックワードリンクが重要な役割を果たすことが予備実験によって判明している．これは，ドメイン特有の単語の場合，ドメイン内で密なリンク構造が形成されており，フォワードリンク解析では発見できなかった語彙どうしの関連情報をバックワードリンク解析から抽出できたためだと考えられる．しかし，その逆に一般的な語の場合は，様々

な分野の記事から参照されるため，バックワードリンク解析の結果が分散してしまい，精度が下がってしまうという現象がみられた．これは，バックワードリンク数の多い語（一般的な語）は，記事の内容が信頼できるため，フォワードリンクを重視して解析することが望ましい一方で，バックワードリンク解析の結果は分散してしまうため，比重を下げる必要があることを示唆している．

以上より，各記事に関してバックワードリンクの解析を行っているリンク共起性解析も各概念（記事）を特徴付ける重要な情報となるが，フォワードリンクの解析結果もまた有用な情報を持っており，両者を組み合わせることで精度向上を図ることができると考えられる．しかし，それぞれ異なる情報を扱うためには，それらの特性に応じた解析が必要となってくる．本研究では，これら 2 つの特性の異なる情報を融合した手法を提案し，概念間関連度の精度向上を図る方法を検証することを目的とする．

以下では，第 3.2 節で記事に関するリンクなどの特徴の概念間関連度測定手法に与える影響を考慮した提案手法について述べ，第 3.3 節で評価実験の結果を示す．最後に第 3.4 節で本章のまとめを行う．

3.2　大域的情報と局所的情報を融合した手法

3.2.1　概要

前節で述べたとおり，バックワードリンク数が少ない場合，Wikipedia 全体の大域的情報であるリンク共起性解析において特徴情報を十分に取れない可能性がある．その際，記事内の局所的情報である記事中のリンク情報をその記事の特徴を端的に表す情報として用いることによって，特徴情報の充実を図る．そこで，大域的情報であるリンク共起性解析による特徴と共に，各記事内における記述（ここではリンク情報）の特性も考慮するため，それら 2 種類の特徴情報を合成することによって精度向上を図る．

合成方法としては，リンク共起性解析において記事（を表すリンク）とその記事内のリンクも共起としてカウントするという方法も考えられる．しかし，この

方法では近傍リンクにおける共起と，記事とその記事中のリンクとの共起は，それぞれの総数が大幅に異なることが考えられ，それぞれの重要度の違いを考慮することが必要である．そこで，1.3.3 項で述べた tfidf に基づく手法を用いる．すでに Milne らによって，tfidf に基づく手法で生成された記事中のリンクを用いた特徴ベクトルが，その記事（概念）の特徴情報として有用であるとこが示されている．そこで，本研究では Wikipedia 全体の大域的情報としてリンク共起性解析を，記事中の局所的な情報として tfidf に基づく手法を，概念に関するそれぞれの特徴情報を算出する手法として用いる．

3.2.2 提案手法

提案手法では，リンク共起性解析によって生成される特徴ベクトルと，tfidf に基づく手法によって生成される特徴ベクトルを合成することによって，概念を表す特徴ベクトルを生成する．リンク共起性解析と tfidf に基づく手法によって生成される特徴ベクトルは，両方ともベクトル空間モデル [72] によって，Wikipedia に存在する全記事（へのリンク）を次元，その各リンクに対する重みを要素とする多次元ベクトルで表わされているため，一般的なベクトルに対する演算を適用できる．なお，ベクトルの次元数は Wikipedia に存在する全記事数である．それぞれのベクトルは，それぞれを個別の特徴ベクトルとして扱うためにベクトルの加算を行う前に正規化する必要がある．正規化は，各要素をすべての要素の合計値で除算することによって行う．合成ベクトル cv は，リンク共起性解析による特徴ベクトル lv と tfidf による特徴ベクトル tv によって，以下の式で表される．

$$cv = \alpha \cdot lv + (1 - \alpha) \cdot tv \tag{3.1}$$

α は，リンク共起性解析と tfidf に基づく手法のどちらをより重視するかのバランスを調整するパラメータである．

ここで，処理の流れを図解によって示す．図 3.1 では，ベクトルの生成の流れを示している．まず，Wikipedia の全記事を解析し，解析対象データであるリンク元記事 ID とリンク先記事 ID を 1 組としたリンク情報データを生成する．次に，そのリンク情報データに対して tfidf に基づく手法とリンク共起性解析を用いること

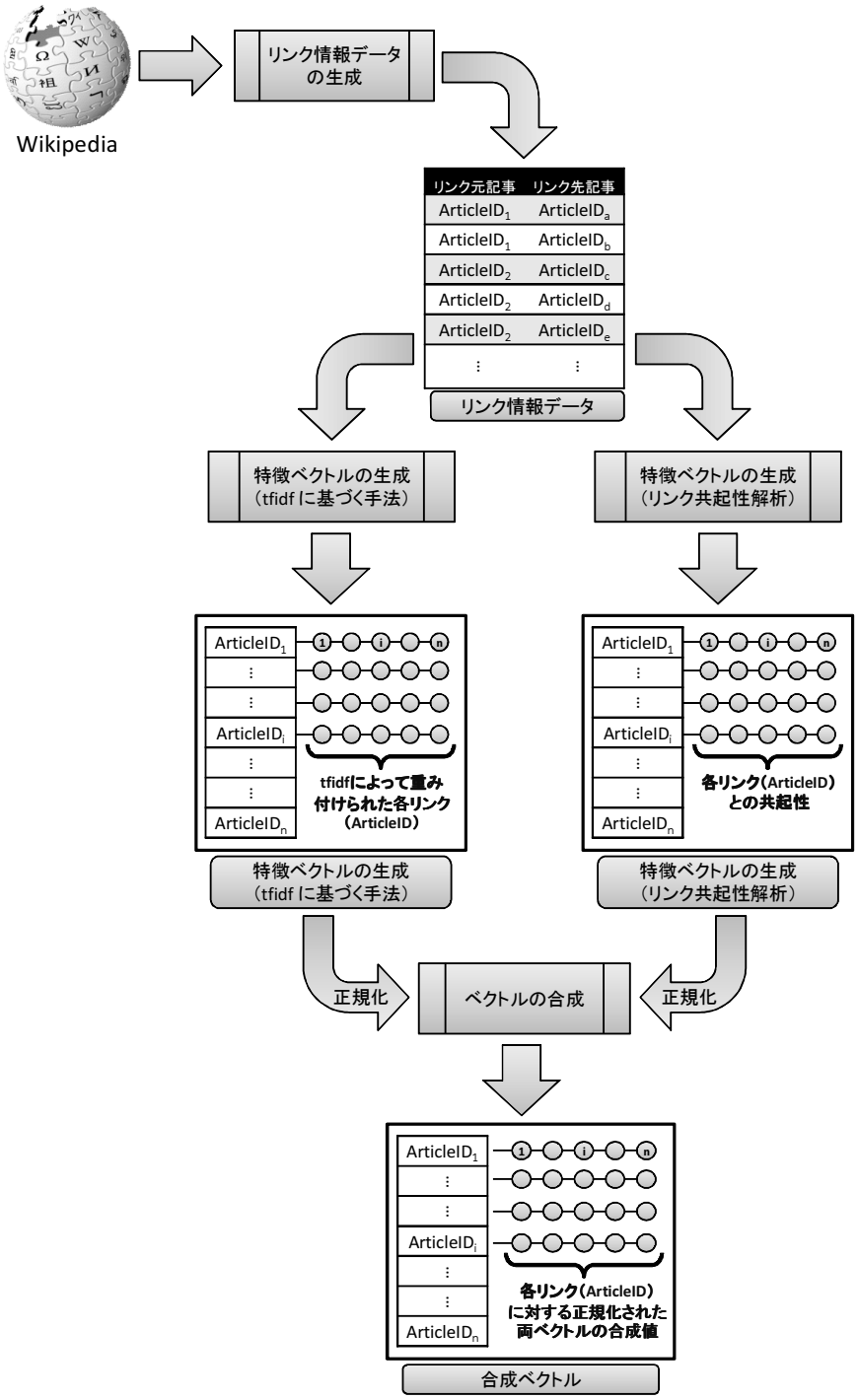

図 3.1: 処理の流れ：ベクトルの生成

によって，それぞれの特徴ベクトルを生成する．各記事（概念）の特徴ベクトルは，全記事数を n とすると n 次元のベクトルとなっている．この時点で，それぞれの手法に基づく概念間の関連度を求めるために，これらのベクトルを用いることができる．図 3.1 の下部では，リンク共起性解析と tfidf に基づく手法の融合手法に用いる合成ベクトルを生成している．合成ベクトルは，各手法によるベクトルをベクトルの全要素の合計値で除算することによって正規化し，同一概念（記事）のベクトルを加算することによって，得ることができる．

3.2.3 パラメータ α の検討

ここで，先に挙げた式 3.1 のパラメータ α の設定方法について検討する．パラメータの決定法としては，大きく 2 つに分けられる．1 つは，最適なパラメータを試行錯誤的に求め，それを固定値として設定する方法で，もう 1 つが何らかの指標をもとに，状況に応じた変動値を設定する方法である．それぞれの決定法のうち，どちらが良いかは実験的に評価する必要がある．

本研究では，固定値に関しては，評価実験において 0.1〜0.9 の 0.1 刻みの α を用いることによって，性能比較を行う．さらに，本研究では変動値の算出も試みる．初めに述べたように，提案手法は Wikipedia におけるリンクの共起性と，記事内のリンクの重要度の情報を融合することによって精度向上を図っている．それぞれの情報は，ある記事を中心に考えると，その記事へのバックワードリンクに関する情報と，フォワードリンクに関する情報ととらえることができる．つまり，それぞのれ情報はバックワードリンクとフォワードリンクの情報に依存しているということになる．たとえば，これらのリンク数が少ない場合，それぞれの特徴情報を十分に構築できないことが考えられる．そこで，フォワードリンクに比べてバックワードリンク数が多い場合はリンク共起性解析による特徴情報を，逆の場合は tfidf に基づく手法による特徴情報を優先するための動的パラメータの設定法を以下のように提案する．

- BF（Backward-Forward link） 重み付け

$$\alpha(p) = \log \frac{BL(p)}{FL(p)} \tag{3.2}$$

ここで，$BL(x)$ を記事 x のバックワードリンク数，$FL(x)$ を記事 x のフォワードリンク数とする．この式により α は，フォワードリンク数が少なくバックワードリンク数が多い場合，リンク共起性解析による解析結果を重視し，バックワードリンク数が少なくフォワードリンク数が多い場合，tfidf に基づく手法を重視する．

3.3　評価実験

3.3.1　実験概要

本実験の目的は，提案手法によって測定される概念間関連度の精度を評価することである．提案手法はリンク共起性解析と tfidf に基づく手法との融合手法であり，比較対象としては，第 2 章で述べたリンク共起性解析，tfidf に基づく手法，2nd pfibf を用い，それぞれの手法によって Wikipedia から概念間関連度を測定し，精度を比較した．なお，リンク共起性解析において，第 2 章の実験結果より，ウインドウサイズを 3，一次共起性の計算手法を Cosine，pfibf のホップ数 n は 2 に設定した．その他の解析データや，精度測定に関しては第 2 章の 2.3 節に準ずる．

リンク共起性解析による特徴ベクトルと tfidf に基づく手法による特徴ベクトルの合成に用いるパラメータ α は，最適な値を調査するために，以下の複数の値を用いる．

- 0.1 〜 0.9 の 0.1 刻みの値
- BF（Backward-Forward link） 重み付け

3.3.2　実験結果と考察

提案手法によって測定した概念間関連度の精度として，各パラメータ α において 50 回の実験結果の平均精度を表 3.1 に示す．また，表 3.2 に比較手法における 50 回の実験における概念間関連度の平均精度を示す．

表 3.1: 概念間関連度の精度（平均値）：提案手法（融合手法）

パラメータ α	スピアマンの順位相関係数
0.1	0.476
0.2	**0.482**
0.3	**0.480**
0.4	0.475
0.5	0.469
0.6	0.470
0.7	0.474
0.8	**0.484**
0.9	**0.484**
BF 重み付け	0.422

（上位 4 件を強調表示）

表 3.2: 概念間関連度の精度（平均値）：比較手法

手法	スピアマンの順位相関係数
tfidf	0.410
2nd pfibf	0.482
リンク共起性解析	0.467

まず，提案手法である融合手法は，BF 重み付けを除いて，すべてのパラメータ α においてリンク共起性解析のみの手法よりも精度向上を実現できていることが分かった．精度が向上した理由として，共起性解析の得意とする性質と，tfidf が得意とする性質の違いが影響していると考えられる．前述のとおり，共起性解析では Wikipedia 全体を通しての統計的情報を用いるため，特定の記事の質によらない一般的な語の使われ方による情報から関連性を導いている．一方，tfidf は記事

表 3.3: 解析に要する時間

手法	計算時間（秒）
提案手法 （リンク共起性解析 + tfidf）	1,386
リンク共起性解析（ウインドウサイズ 3）	870
tfidf	513
pfibf	561,600

の記述内容からその記事（概念）の特徴ベクトルを求め，その特徴ベクトルを比較することによって関連性を導いている．そのため tfidf に基づく手法の場合，精度は記事の質に左右されるが，その概念の特徴を端的に表すものとして重要な情報が含まれる場合が多い．この情報が，共起性解析での統計的情報では概念の特徴を十分に抽出できなかった場合に，概念の特徴情報を補間する役割をしていると考えられる．

また 2nd pfibf と比較すると，パラメータ α が 0.2，0.3，0.8， 0.9 においては，2nd pfibf と同等か若干上回る精度を実現している．このことは，第 2 章において調査した pfibf の解析時間が非常に高コストである点を考えた場合，非常に良い結果と言える．なぜなら，本提案手法による解析時間は，表 3.3 に示すように，リンク共起性解析と tfidf に基づく手法の解析時間の合計とほぼ同等であり，pfbif による解析時間よりも大幅に短いからである．

次に，パラメータ α における精度の違いを考察する．表 3.1 より，パラメータ α を 0.1〜0.9 まで変化させた場合，α 値の小さい 0.2，0.3 及び α 値の大きい 0.8， 0.9 の精度が良くなっている．この結果から推測できることは，パラメータ α が比較的低い時に精度が良くなる場合と，パラメータ α が比較的高い時に精度が高くなる場合があるのではないか，ということである．つまり，tfidf に基づく情報（局所的情報）を重視することで特徴情報をより精度よく取れる場合と，リンク共起性解析による情報（大域的情報）を重視することで特徴情報をより正確に取れる場合がある，ということが実験結果から推測できる．この推測を詳しく検証するために，表 3.1 のような 50 回行った精度評価実験の平均値ではなく，各回の実験結

果を個別に検証する．検証の際には，α が 0.1〜0.9 の実験結果に加えて，$\alpha = 0.0$ として tfidf に基づく手法の実験結果を，$\alpha = 1.0$ としてリンク共起性解析の実験結果を加える．また，棒グラフの横軸を順番に 0.0〜1.0 の α 値として，各実験を相関の極値に応じて以下の 2 パターンに大別した．

1. 極大値あり
   パラメータ α が 0.1〜0.9 の間に相関の最大値が存在する．2 つの情報を融合することによって概念間関連度の測定精度が向上したことを示す結果である．

2. 極大値なし
   パラメータ α が 0.0 もしくは 1.0 の時に相関の最大値が存在する．tfidf に基づく手法，もしくはリンク共起性解析によって最も概念間関連度の測定精度が良くなることを示す結果である．つまり，2 つの情報を融合しても精度が向上しなかった例である．

以上のような「極大値あり」,「極大値なし」の 2 パターンの実際のいくつかの例を，実験結果から各 5 つずつ抽出し，図 3.2，3.3 に示す．各図中には，5 種類の評価データを用いた実験結果があり，それぞれの評価データにおいて左から $\alpha = 0.0, 0.1, \cdots, 1.0$ に設定した場合の精度を示している．また，表 3.4 に 50 回の実験結果におけるパターン別の回数を示す．まず，図 3.2 に示すような，2 つの情報を融合する本提案手法が有効であることを示す「極大値あり」の実験結果が 36 回と，全体の 72%を占めている．一方で，図 3.3 に示すような，tfidf に基づく手法が有効であったことを示す「極大値なし（$\alpha = 0.0$）」の実験結果は 2 回，リンク共起性解析が有効であったことを示す「極大値なし（$\alpha = 1.0$）」の実験結果は 12 回と，ほとんどの場合でリンク共起性解析が tfidf に基づく手法より高い精度であることが分かった．このようなパターンの実験結果が，50 回の実験において複合的に起こり，結果として平均値を計算すると表 3.1 のような結果になったものと考えられる．仮に，各実験において最も精度が良くなるパラメータ α を動的に選ぶことが出来れば，平均精度が現在の結果よりさらに向上することは明白である．そのためには，パラメータ α を概念（記事）ペアに関連する情報に基づいて動的に変動させる必要がある．パラメータ α を動的に変動させる方法として，本実験では「BF 重み付け」

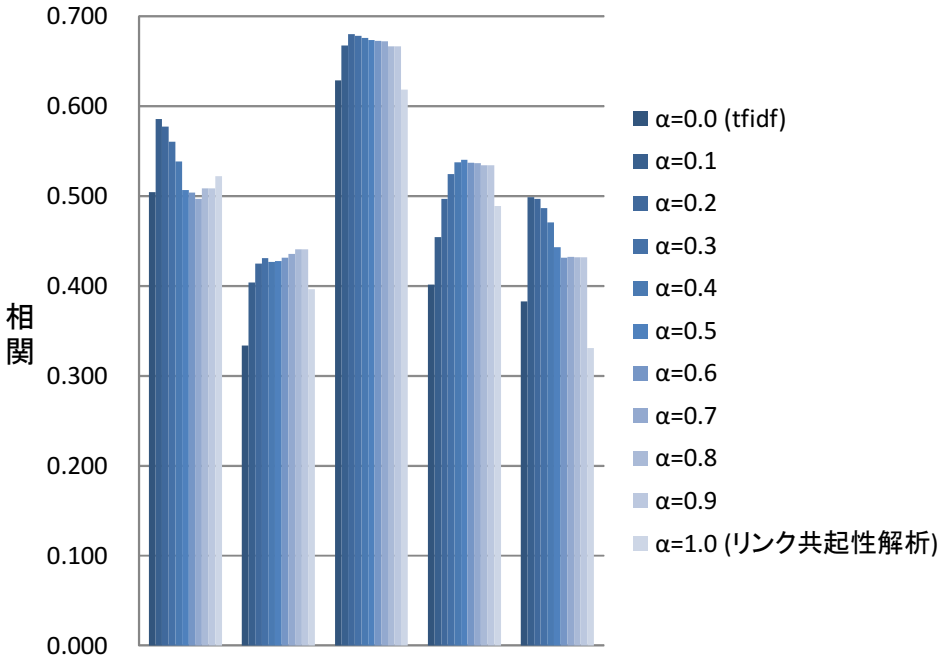

図 3.2: パラメータによる精度変化（極大値あり）

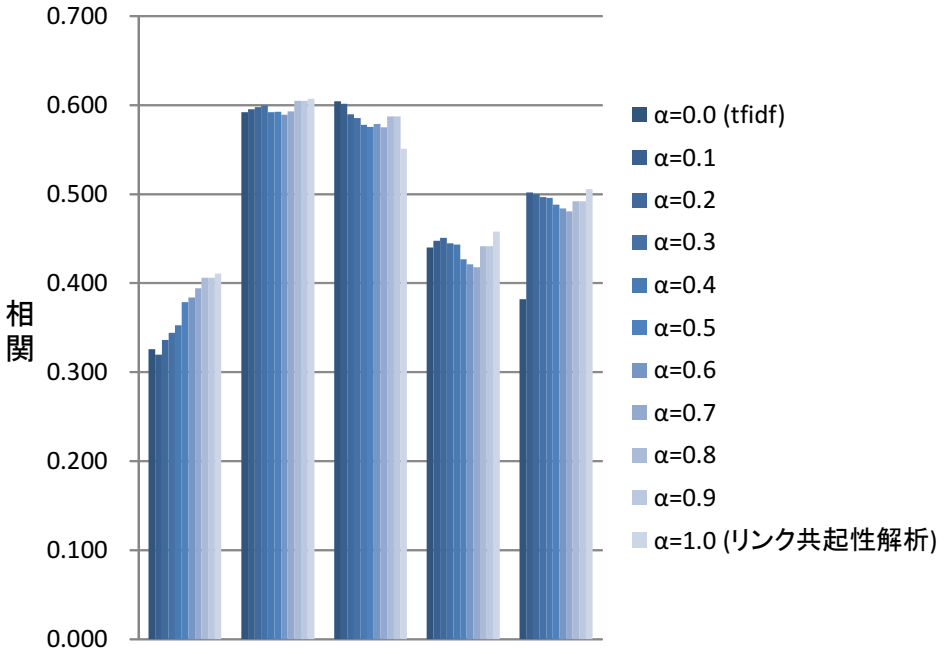

図 3.3: パラメータによる精度変化（極大値なし）

表 3.4: パターン別の回数

パターン	回数
極大値あり	36
極大値なし（$\alpha = 0.0$）	2
極大値なし（$\alpha = 1.0$）	12

を用いたが表 3.1 から分かるように，精度向上には至らなかった．この結果から，提案手法の精度をさらに向上させるためには，バックワードリンク数とフォワードリンク数の比率ではなく，記事や記事間に関するより多様な条件を考慮した適切なパラメータ α を設定する必要がある．

3.4 むすび

本章では，異なる 2 つの情報を組み合わせることによって，より高精度な概念間関連度の測定方法を提案した．リンク共起性解析のような大域的情報を用いる手法では，記事へのリンクが少ない場合，特徴情報が十分取れない可能性がある．一方で，そのような記事に対して局所的情報である記事内のリンクがある程度存在した場合，その記事の特徴を端的に表すものとしてそのような情報は依然として有用である．そこで，大域的情報である Wikipedia 全体のリンクの統計情報と，記事内のリンクの統計情報の双方を融合されることによって，概念間関連度の測定精度向上を図った．この手法では，リンク共起性解析による記事の特徴ベクトルと，記事内リンクを tfidf で重み付けすることによって求めた特徴ベクトルをパラメータ α によって合成する．

提案手法の性能評価のために行った評価実験により，提案手法の有効性が示された．具体的には，提案手法は低い解析コストにもかかわらず，パラメータ α が 0.2，0.3，0.8， 0.9 において，2nd pfibf と同等か若干上回る精度を実現していることが分かった．また，網羅的にパラメータ α を変えて実験を行った結果，各実験においてパラメータ $\alpha = 0.0$〜1.0 の変化に応じて，精度が「極大値あり」,「極大

値なし」の 2 つのパターンで変化するとこが分かった．この結果より，パラメータ α を動的に各パターンに適応させるように設定することにより，さらなる精度の向上を期待できるという知見が得られた．

第4章　SVRによる多様な情報を活用した概念間関連度

4.1　まえがき

Wikipediaには，記事間リンクやカテゴリリンク，記事の記述内容など，記事や記事間の関連性を特徴付けることが可能な様々な情報が定義されており，また，それらの情報を用いて記事間の関連性を特徴付ける新たな情報を生成可能である．これまで，Wikipediaを用いた概念間関連度の測定手法では，それらの情報の一部を単体で用いているものが多かった．具体的には，Strubeら [75] は記事が属するカテゴリに着目し，そのカテゴリ構造の情報を活用した手法を提案している．Milneら [51] は，Wikipediaの各記事の特徴を，記事内に出現する他の記事へのリンクをtfidfの考え方に基づいて重み付けしベクトル化することによって表している．中山ら [58] は，記事間のリンク構造をnホップ先の記事まで解析することによって，関連概念（記事）とその関連度を測定している．また，本論文の第2章ではWikipedia全体におけるリンクの共起性を用いて2つのリンクが表す記事（概念）の関連度を測定している．

このように，Wikipediaの様々な情報を個別に活用した各種手法が提案され，その概念間関連度の測定に対する有用性が示されてきたが，同時に，ある種の情報だけを用いる問題点も明らかになってきた．最も重要な問題は，各種情報の不足である．たとえば，カテゴリ情報を用いた手法では，記事が所属するカテゴリが少なければ，関連度測定の精度が大きく下がる．そもそも，Wikipedia全体のカテゴリ情報が記事数に対して十分でないことも問題である．また，記事内に存在するリンク情報を用いるtfidfに基づく手法や，記事へのリンク情報を用いるリンク共起性解析の場合も，それらの情報が少ない場合や情報の品質が低い場合に，精

度低下を招くという問題点がある．

一方，これまで提案されてきた手法以外に，関連度測定に貢献することが推測できる情報もある．たとえば「記事間が相互リンクの関係にあるか」や「同じカテゴリに属しているか」，バックワードリンク数，フォーワードリンク数の値などである．これらの情報は単体で用いた場合，関連度測定には不十分だが，他の情報と共に複合的に用い様々な概念間の状況を考慮できた場合，概念間関連度を測定する上で有効に働くことが予想される．このことは，第 3 章の実験結果によっても示唆されている．

そこで本章では，Wikipedia から取得可能な情報，及び既存手法などで生成可能な情報の中で，概念間関連度の測定に貢献する情報を明確にすることを目指す．概念間関連度に大きな影響を与える情報を明確にすることによって，それらの情報を複合的に用いる場合に計算量を抑えつつ高精度な概念間関連度の測定を行える可能性がある．さらに，それらの情報を機械学習手法によって学習させることによって，個々の情報を複合的に考慮した概念間関連度の測定手法を提案する．複数の情報を複合的に考慮することは，記事の様々な特性によって重視すべき情報が異なるという第 3 章で得た知見への対応となり，精度向上を期待することができる．

まず本研究では，Wikipediaから得られる様々な情報を検討するために，Wikipediaから取得，及び生成可能な情報の中で，言語に依存しない，つまり自然言語解析の必要のない情報を，概念や概念間の関連性を特徴付ける素性（Feature）の候補として提案する．そして，記事や記事間に関する複数の素性を複合的に考慮した概念間関連度の測定を行うために，機械学習手法である SVM (Support Vector Machine) の回帰問題への対応を可能とした SVR (Support Vector Regression) を用いた機械学習を行う．ここで機械学習は，入力ベクトル（一組の素性セット）とその入力に対する出力（本研究では関連度）を学習データとして，多数の学習データを与えることによって，入出力の関係を学習（近似）する．この入出力の関係を適切に学習することができれば，未知の入力ベクトルに対する出力を精度良く予測することができる．本研究においては，機械学習手法を用い，各既存手法と共にリンク数などの様々な条件下における関連度を学習する．そのため，従来手法では精

度が低下するような条件下においても，それら多様な条件を考慮することによって概念間関連度の測定精度向上が期待できる．評価実験では，概念や概念間に関する提案素性の中で関連度測定に重要な素性を，機械学習における学習の重要度を評価する尺度である F-score を用いて検証する．最後に，SVR による複数の素性を考慮した概念間関連度の測定手法について，その有効性を検証する．

以下では，第 4.2 節で概念（記事）に関する多様な情報（素性）とその重要性の検証方法について解説し，第 4.3 節で機械学習による学習方法について説明する．第 4.4 節では，重要な素性の検証結果と，機械学習に基づく手法の評価実験の結果を示す．最後に第 4.5 節で本章のまとめを行う．

4.2 概念に関する多様な情報

本節では，Wikipedia から取得可能な概念間関連度の測定に有用だと考えられる情報を網羅的に列挙する．また，どの情報が有用な情報であるかを調査するための方法を述べる．

Wikipedia は，多様な半構造化データ (Semi-structured data)[1] [1] を持っていることが大きな特徴の 1 つである．たとえば，記事タイトルやアンカーテキスト，記事テキストにおけるセクション名，テーブル，インフォボックス，他の記事へのリンクや，カテゴリリンクなどがその例である．これらの情報は，各記事がそれぞれ持っているデータであり，その記事（概念）や他の記事との関係を特徴付ける情報として利用可能であると考える．これらの情報は，以下のように大別することができる．

1. 記事間リンク
2. カテゴリリンク
3. 記事のテキスト情報

[1] 厳密には定義されていない構造を持ち，その中にテキストなどの非構造化データを含んでいる．たとえば，XML データなどは，全体は木構造のタグ構造をもつが，そのタグの中身は非構造化データであるテキストである．

記事間のリンクは，記事を特徴付ける情報として重要な役割を果たすことがこれまでの研究で明らかになっている．また，各記事からWikipediaカテゴリへのリンクや，そのカテゴリ構造も有用な情報源であり，lchなどの手法が提案されている．最後の記事のテキスト情報は，その記事が表す概念に対しての詳細な解説であるため，これまでのtfidfによる文書ベクトルを用いた手法が提案されているように，記事の特徴を表す有用な情報である．しかし，これらの素性の相互関係や重要度などは明確にされてこなかった．

4.2.1 記事（概念）間の関連性を特徴付ける素性候補

Wikipediaの記事間リンクとカテゴリリンクから生成可能な概念間の関連性を特徴付ける素性（情報）の候補として，以下のものを提案する．以下に挙げる素性は，前節で述べた3つに大別される情報のうち,「記事間リンク」「カテゴリリンク」に由来する情報のみを扱っている．3つ目の「記事のテキスト情報」を扱わない理由として，これらのテキスト情報は自然言語であり，その解析には多くの時間がかかるからである．また，リンクを用いた手法と異なり，言語に依存した手法となってしまうため，本研究では言語に非依存の概念間関連度の測定を目指すという方針のもと，自然言語解析が必要なテキスト情報は用いないこととする．

ここで，以降の素性の説明において，関連度を測定する2つの記事（概念）をp_aとp_b，記事p_xが所属するカテゴリ群をC_x，記事p_xが持つリンク群をL_x，記事p_xへのリンクをl_xと置く．

Feature 1: Same Category Check

本素性は，同じカテゴリに属する記事どうしか，という単純な情報を用いる．このような，記事が所属するWikipediaのカテゴリに基づくアプローチは，Wikipediaから概念間関連度を測定する手法の1つとして，その有用性が示されている．ここでは最も単純な発想として，同じカテゴリに所属する記事（概念）どうしは関

連性が高い，という想定の下，以下の式によって本素性を定義する．

$$Feature1 = \begin{cases} 1 & (C_a \cap C_b \neq \phi) \\ 0 & (otherwise) \end{cases} \tag{4.1}$$

Feature 2: lch

本素性は，それぞれの記事が所属するカテゴリ間の距離を用いる．1.3.3 項で述べたように，Wikipedia のカテゴリ構造を用いた手法として，Strube らが lch の有効性を示している [75]．そこで，Wikipedia のカテゴリ構造から得られる記事間の関係性を特徴付ける情報として lch を用いる． 記事 p_a, p_b 間の lch の値を式 (1.1) によって以下のように表す．

$$\begin{aligned} Feature2 &= lch(p_a, p_b) \\ &= -\log \frac{length(p_a, p_b)}{2D} \end{aligned} \tag{4.2}$$

ここで，$length(p_a, p_b)$ は，ノード p_a, p_b 間のカテゴリ構造を介した最も短いパスの長さ，D はカテゴリ構造の深さである．

Feature 3: Bidirectional Link Check

本素性は，記事どうしがリンクし合っているか，つまり双方向リンクの関係にあるか，という単純な情報を用いる．これまでの Wikipedia の記事間リンクの構造を用いる研究によって，短いホップ数でリンクされた記事どうしは高い関連性を持つことが多く，また双方向リンクが存在する場合，さらに高い関連性を持つことが分かっている．たとえば，記事「Microsoft」は記事「Microsoft Windows」へのリンクを持ち，記事「Microsoft Windows」も記事「Microsoft」へのリンクを持つ．そして，この 2 つの記事（概念）は直感的にも高い関連性を持つことが分かる．そこで，以下の式によってこの素性を定義する．

$$Feature3 = \begin{cases} 1 & (l_b \subseteq L_a \cap l_a \subseteq L_b) \\ 0 & (otherwise) \end{cases} \tag{4.3}$$

Feature 4: Inversed Link Order

本素性は，記事中におけるリンクの出現順位を情報源としても用いる．一般的に，その記事に関する概要や重要な事柄を，最初のセンテンスやパラグラフで記述することが多い．これは Wikipedia の場合も同様であることが中山ら [60] によって説明されている．つまり，Wikipedia の記事において，これらの記述の中に存在する他の記事へのリンクは，記事が表す概念と直接関係の深い記事へのリンクであることが多い．逆に，記事の下部に行くほど詳細情報や関連情報などの記述が増えていき，記事上部の記述よりも記事のタイトルに直接関係が薄い情報が増えていく傾向にあると考えられる．そこで，記事上部のリンクほどスコアが高くなるように，N（記事中の全リンク数）リンク中の n 番目のリンクのスコアを以下の式のように定義し，本素性とする．

$$Feature4 = 1 - \frac{n}{N} \tag{4.4}$$

Feature 5: 1st Link Co-occurrence

本手法は，Wikipedia における大域的な情報としてリンクの一次共起性の値を用いる．第 2 章で述べたように，Wikipedia におけるリンク共起性は，記事（概念）間関連度を測定する上で低い解析コストで効果的な情報を得られることが分かっている．そこで，以下の式に示すように，l_a と l_b の一次共起性の値を素性として用いる．

$$Feature5 = 1st\_lca(l_a, l_b) \tag{4.5}$$

ここで，関数 $1st_lca(x, y)$ はリンク x と y の，Wikipedia における一次共起性の値を返す．なお本研究では，第 2 章の実験結果に基づいて，リンク共起性解析において，ウインドウサイズを 3，一次共起性の計算方法として Cosine を用いる．

Feature 6: 2nd Link Co-occurrence

本素性は，Feature 5 と同様に Wikipedia における大域的な情報としてリンクの二次共起性の値を用いる．Feature 5 と同様，Wikipedia におけるリンクの二次共起

性も，記事（概念）間関連度を測定する上で重要な情報である．本素性では，第 2 章で述べたような ベクトル空間モデルに基づく l_a, l_b の共起ベクトルのコサイン類似度で表される，リンクの二次共起による関連度を用いる．以下に，本素性の計算式を示す.

$$\begin{aligned} Feature6 &= 2nd\_lca(l_a, l_b) \\ &= \frac{v_a \cdot v_b}{|v_a||v_b|} \end{aligned} \tag{4.6}$$

$$v_i = \{1st\_lca(l_i, l_1), 1st\_lca(l_i, l_2) \cdots, 1st\_lca(l_i, l_N)\} \tag{4.7}$$

ここで，関数 $2nd\_lca(x, y)$ は 式 (2.5) に基づき，x, y の二次共起性を返し，v_i は，リンク l_i の一次共起性による特徴ベクトルを表す．また，N は Wikipedia の全記事数とする．なお本研究では，第 2 章の実験結果に基づいて，リンク共起性解析において，ウインドウサイズを 3，一次共起性の計算方法として Cosine を用いる．

Feature 7: 1st pfibf

本素性は，記事間のリンク構造を情報源として用いる．中山らの研究によって，記事間リンクのリンク構造を解析することによって，記事間の関連性を特徴付ける有用な情報を得られることが分かっている．1.3.3 項で述べた pfibf はその代表的な手法であり，本素性は以下の式で表す pfibf による関連度を用いる．

$$Feature7 = 1st\_pfibf_n(p_a, p_b) \tag{4.8}$$

ここで，関数 $1st\_pfibf_n(x, y)$ は 式 (1.5) に基づき，記事 x, y 間のリンク構造を n ホップ先まで再帰的に解析し pfibf 値を返す．なお本研究では，中山らの研究結果 [58] に基づき，n を 2 と設定する．

Feature 8: 2nd pfibf

本素性は，Feature 7 と同様に pfbif に基づく関連度を用いるが，2.3.1 項で述べた問題を考慮し，ベクトル空間モデルによって記事 p_a, p_b の特徴ベクトルを，pfibf

によって求めた100件の関連記事（概念）とその関連度によって表し，2つのベクトルのコサイン類似度によって定義される．

$$\begin{aligned} Feature8 &= 2nd\_pfibf(p_a, p_b) \\ &= \frac{v_a \cdot v_b}{|v_a||v_b|} \end{aligned} \tag{4.9}$$

$$v_i = \{1st\_pfibf(l_i, l_1), 1st\_pfibf(l_i, l_2) \cdots, 1st\_pfibf(l_i, l_N)\} \tag{4.10}$$

ここで，関数$2nd_pfibf_n(x, y)$は式(1.5)に基づき，nホップのpfibfにより求めた結果を用いて，記事x, yの特徴ベクトルを生成し，その2つのベクトルのコサイン類似度を返す．また，v_iは，リンクl_iのpfibfによる特徴ベクトルを表し，NはWikipediaの全記事数とする．なお本研究では，中山らの研究結果に基づき，nを2と設定する．

Feature 9: tfidf of l_b in p_a

本素性は，Wikipediaにおける記事内の局所的な情報として記事中のフォワードリンクの重要度を用いる．1.3.3項で述べたように，Milneらによって記事中のリンクの重要度を関連度測定に利用するtfidfベースの手法が提案されており，その有効性が証明されている．本素性では，記事p_aに存在する記事p_bへのリンクl_bの重要度は，記事p_a, p_b間の関連度に影響を与えるという考えから，以下の式で表されるように記事p_a中に存在するリンクl_bのtfidf値を用いる．

$$Feature9 = tfidf(p_a, l_b) \tag{4.11}$$

ここで，関数$tfidf(x, y)$は式(1.2)に基づいて，記事x中のリンクyのtfidf値を返す．なお，記事p_aにリンクl_bが存在しない場合，Feature 9は0となる．

Feature 10: tfidf of l_a in p_b

本素性では，Feature 9と同様の考えのもと，Wikipediaにおける記事内の局所的な情報として記事中のフォワードリンクの重要度を用いるが，Feature 9とは逆に，

記事 p_b 中に存在するリンク l_a の tfidf 値とする．

$$Feature10 = tfidf(p_b, l_a) \tag{4.12}$$

ここで，関数 $tfidf(x, y)$ は式 (1.2) に基づいて，記事 x 中のリンク y の tfidf 値を返す．なお，記事 p_b にリンク l_a が存在しない場合，Feature 10 は 0 となる．

Feature 11: 2nd tfidf

本素性は，1.3.3 項で述べた tfidf ベースの手法であり，その概念間関連度の測定への有効性は，Milne らによって証明されている．この手法は，p_a, p_b それぞれの記事の特徴ベクトルを，記事内に存在する他の記事へのリンクを次元，その tfidf 値を要素として生成し，以下のようにそれら 2 つのベクトルのコサイン類似度によって定義される．

$$\begin{aligned} Feature11 &= 2nd\_tfidf(p_a, p_b) \\ &= \frac{v_a \cdot v_b}{|v_a||v_b|} \end{aligned} \tag{4.13}$$

$$v_i = \{tfidf(p_i, p_1), tfidf(p_i, p_2) \cdots, tfidf(p_i, p_N)\} \tag{4.14}$$

ここで，関数 $2nd\_tfidf(x, y)$ は，記事内のリンクの tfidf 値によって，ベクトル空間モデルに基づく記事 x, y の特徴ベクトルを生成し，その 2 つのベクトルのコサイン類似度を返す．また，v_i は，記事 p_i の tfidf による特徴ベクトルを表し，N は Wikipedia の全記事数とする．

Feature 12: The Number of Forward Links of p_a

本素性は，記事中のフォーワードリンク数を情報源として用いる．記事中のフォワードリンクの数は，tfidf ベースの手法や pfibf のようにフォワードリンクを情報源として用いる手法に影響を与えると考えられる．たとえば，フォワードリンクが少ない場合，その記事の特徴を表す情報源としては不十分であったり，そもそもそれほど整備が行き届いていない結果であると考えれば，リンク付けの質が悪

い可能性がある．そこで本素性を，記事 p_a のフォワードリンクの数を用いて以下の式で表す．

$$Feature12 = \log |FL(p_a)| \tag{4.15}$$

ここで，関数 $FL(x)$ は記事 x のフォワードリンクの集合を返す．

Feature 13: The Number of Forward Links of p_b

Feature 12 と同様の考えのもと，本素性を記事 p_b のフォワードリンク数を用いて以下の式で表す．

$$Feature13 = \log |FL(p_b)| \tag{4.16}$$

ここで，関数 $FL(x)$ は記事 x のフォワードリンクの集合を返す．

Feature 14: The Number of Backward Links of p_a

本素性は，記事へのバックワードリンク数を情報源として用いる．記事へのリンクの数は，リンク共起性解析や pfibf のようにバックワードリンクを情報源として用いている手法に影響を与えると考えられる．また，PageRank などの手法がバックワードリンクをその記事への投票と見なし記事の質の評価に用いていることからも分かるように，バックワードリンクは記事の特性を表す重要な情報源として用いることができる．そこで本素性を，記事 p_a のバックワードリンクの数を用いて以下の式で表す．

$$Feature14 = \log |BL(p_a)| \tag{4.17}$$

ここで，関数 $BL(x)$ は記事 x のバックワードリンクの集合を返す．

Feature 15: The Number of Backward Links of p_b

$Feature14$ と同様の考えのもと，本素性を記事 p_b のバックワードリンク数を用いて以下の式で表す．

$$Feature15 = \log |BL(p_b)| \tag{4.18}$$

ここで，関数 $BL(x)$ は記事 x のバックワードリンクの集合を返す.

Feature 16: ibf of p_a

本素性は，Feature 13, Feature 14 と同じく，記事へのバックワードリンク数を情報源として用いる．tfidf ベースの手法や pfibf では，多くの記事からリンクされている記事を，一般語を表す記事と見なしており，それらの記事は多くの記事と弱い関係を持つと判断している．そのため，このような状況を考慮するため，ibf(Inversed Backward-link Frequency) を用いている．本素性では，以下のように記事 p_a の ibf 値を用いる.

$$Feature16 = \log \frac{N}{|BL(p_a)|} \tag{4.19}$$

ここで，関数 $BL(x)$ は記事 x のバックワードリンクの集合を返し，N は Wikipedia の全記事数を表す.

Feature 17: ibf of p_b

$Feature16$ と同様の考えのもと，本素性を記事 p_b のバックワードリンク数を用いて以下の式で表す.

$$Feature17 = \log \frac{N}{|BL(p_b)|} \tag{4.20}$$

ここで，関数 $BL(x)$ は記事 x のバックワードリンクの集合を返し，N は Wikipedia の全記事数を表す.

4.2.2 F-score

F-score [14] は，各素性の機械学習における学習の重要度（貢献度）を評価する効果的な尺度であり，様々な研究で素性選択（Feature Selection）[14, 28] を行う際に用いられている [3, 11, 30, 48]．素性選択とは，機械学習において素性セットの中でより有用な部分集合だけを用いる戦略である．素性選択を行う目的としては，以下のようなものが挙げられる.

- 学習への貢献度が低い素性を除くことで，素性を生成する解析コストや学習時間を削減させる．
- 不要な学習素性を除くことで，予測精度を向上させる．
- 素性セットのうちどの素性が重要かを人間が理解しやすくなる．

また，素性選択を行う方法としては，素性の様々な部分集合を用いて機械学習を行い，探索的に精度の高くなる素性を評価する方法があるが，F-score は学習データだけを用いシンプルな計算式によって素性の重要度を計算することができる手法であり，機械学習器とは無関係に求めることができる．

入力ベクトルを x_k $(k = 1, ..., m)$ と置き，正例と負例の数をそれぞれ n_+，n_- と置くと，i 番目の素性の F-score は以下のような式で表される．

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}, \tag{4.21}$$

ここで，$\bar{x}_i$ をすべての i 番目の素性の平均値，$\bar{x}_i^{(+)}$ を正例の i 番目の素性の平均値，$\bar{x}_i^{(-)}$ を負例の i 番目の素性の平均値とする．また，$x_{k,i}^{(+)}$ を k 番目の正例の i 番目の素性の値とする．

F-score は，値が大きければ大きいほど，その素性が学習において効果的に働いている事を示しており，重要な素性と見なすことができる．本研究では，この F-score を用いることによって，4.2 節で示した記事間の関連性を特徴付けうる素性の中で，どの素性が概念間関連度の測定にとって重要かを検証する．

4.3　学習方法

本節では，4.2 節で述べた素性を機械学習させる方法について述べる．まず，本研究では，4.2 節で述べた概念間の素性とその関連度の関係を，2 クラス分類問題を扱う SVM（Support Vector Machine）を回帰問題へ適用可能とした機械学習器である SVR（Support Vector Regression）を用いて学習させ，ある概念間の素性セット（入力ベクトル）からその概念間の関連度（出力）を予測する．また，学習にお

いては機械学習器に与える入力ベクトルと出力の組である学習データを生成する必要がある．そして，学習の際には，4.3.1 項で述べるソフトマージンやカーネル関数に関するパラメータを決定する必要があり，それらのパラメータを学習データに応じて最適化することにより学習の性能を向上させることができる．以下では，それぞれについて説明する．

4.3.1 SVM / SVR

SVM は，Vapnik ら [79,80,81] によって提案された教師あり学習手法である．提案されて以来，その有用性が数多くの研究や実装により証明され，今では，クラス分類，回帰などを行う分野で非常によく利用される機械学習手法の 1 つとしてその地位を確立している．SVM は，学習データを分離する超平面を求める際のマージン最大化という戦略により高い汎化能力を持つことが大きな特徴である．また，ソフトマージンによって重なりのあるクラス分布の線形分離に対応している．以下では，まず SVM の基本戦略であるマージン最大化について解説した後，本研究で用いる回帰問題への拡張である SVR について解説する．

マージン最大化

SVM は，入力された学習データを，以下のような線形モデルを用いることによって 2 つのクラス χ_1, χ_2 に分離する．

$$y(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b \tag{4.22}$$

ここで，$\mathbf{w}$ は超平面の法線ベクトル，b はバイアスパラメータである．学習データは，$\mathbf{x}_1, \cdots, \mathbf{x}_N$ の N 個の入力ベクトルと，それぞれに対応するクラスラベル $t_1, \cdots, t_N$ からなる．なお，クラスラベル t_n は 1 か -1 で与えられ，未知のデータ点 $\mathbf{x}$ は $y(\mathbf{x})$ の符号に応じて分類される．

学習データが特徴空間で線形分離可能な場合，

$$t_n(\mathbf{w}^t\mathbf{x}_n + b) \geq 1(n = 1, \cdots N) \tag{4.23}$$

を満たし，クラスを正確に分類できる解，つまりパラメータ $\mathbf{w}$ と b の組は必ず存在する．また，一般的にこのような解は複数存在する．このように，学習データを正確に分離する解が複数存在する場合，汎化誤差が最も小さくなるような解を求めることが望ましい．SVM はマージン（分類境界と学習データの間の最短距離）という概念を用いて，このような解を求めている．SVM において，分離境界はこのマージンを最大化するものが選ばれる．この最適化問題は，

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||^2 \\ \text{制約条件} \quad : \quad & t_n y(\mathbf{x}_n) \geq 1 \quad (n = 1, \cdots N) \end{aligned} \tag{4.24}$$

という最小化問題として定式化することができる．

回帰問題への対応

ここでは，SVM を回帰問題に適用可能とした SVR について説明する．Vapnik ら [80] の提案する ϵ-SVR は，ϵ 許容誤差関数（ϵ-insensitive error function）

$$E_\epsilon(y(\mathbf{x}) - t) = \begin{cases} 0 & |y(\mathbf{x}) - t| < \epsilon \\ |y(\mathbf{x}) - t| - \epsilon & otherwise \end{cases} \tag{4.25}$$

を用いて，回帰問題を解く．ここで，ϵ 許容誤差関数は，予測値 $y(\mathbf{x})$ と実測値 t の差が $\epsilon(> 0)$ 未満の時は 0 になる．この ϵ 許容誤差関数を用いると，結局次の誤差関数を最小化する問題となる．

$$\min_{\mathbf{w},b} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N} E_\epsilon(y(\mathbf{x}_n) - t_n) \tag{4.26}$$

ここで，パラメータ C は，モデルが複雑になり汎化能力を損なう問題を考慮するためにマージン境界のある程度の誤分類を許容する，ソフトマージン法に基づくパラメータである．この最小化問題を，さらにラグランジュの未定乗数 $\alpha_n, \hat{\alpha}_n$ を導入することによって，以下の α_n と $\hat{\alpha}_n$ を最大化させるような双対問題に帰着させることができる．

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\alpha_n - \hat{\alpha}_n)(\alpha_m - \hat{\alpha}_m)\mathbf{x}_n^T\mathbf{x}_m$$

$$-\epsilon\sum_{n=1}^{N}(\alpha_n+\hat{\alpha}_n)+\sum_{n=1}^{N}(\alpha_n-\hat{\alpha}_n)t_n \tag{4.27}$$

ただし，α は以下の制約条件を満たす.

$$0\leq\alpha_n,\hat{\alpha}_n \leq C(n=1,\cdots,N) \tag{4.28}$$

$$\sum_{n=1}^{N}(\alpha_n-\hat{\alpha}_n) = 0 \tag{4.29}$$

ここで，上記は線形回帰の場合であるが，SVR は非線形回帰にも拡張することができる．まず，学習データ $\mathbf{x}$ を関数 $\phi(\mathbf{x})$ によって高次元の特徴空間へ非線形変換し，その空間において線形識別を行えばよい．しかし，これを式 (4.27) に適用した場合，高次元ベクトルの内積 $\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m)$ には膨大な計算量が必要となる．そこで，

$$k(\mathbf{x},\mathbf{x}')\equiv\phi(\mathbf{x})^T\phi(\mathbf{x}') \tag{4.30}$$

となるようなカーネル関数 k を用いる．カーネル関数としては，多項式カーネル（polynomial kernel），RBF カーネル（radial basis function kernel），シグモイドカーネル（sigmoid ernel）などが提案されているが，ここでは本実験で用いる RBF カーネルの定義を以下に示す.

$$k(\mathbf{x},\mathbf{x}')=\exp(-\gamma||\mathbf{x}-\mathbf{x}'||^2) \tag{4.31}$$

ここで，γ はカーネル関数のパラメータである.

4.3.2 学習データの生成

本研究では，機械学習器に SVR を用いて関連度を実数として予測する．機械学習器に与える学習データは，素性の個数を n とすると，各学習データに対してすべての素性 n 個（入力ベクトル）と実測値（出力）としての関連度 1 個によって $n+1$ 次元のベクトルとなる．素性の全ては，Wikipedia の記事間リンクとカテゴリリンクから生成可能なデータであり，また概念間の関連度は付録 A で述べる「WikiSimi

Test Collection」を用いる．さらに，学習データ内のすべての値は，-1から1の範囲にスケーリングを行う必要がある．機械学習により生成された学習結果であるモデルデータを用いて予測値を求める際も，同様の入力ベクトルを与えることによって，入力ベクトルに応じた予測値（関連度）を求めることができる．

4.3.3 最適パラメータの導出

SVMが高い汎化性能を持った機械学習器であることについては4.3.1項で述べたが，その能力を十分に発揮するためには，ソフトマージンのパラメータCや，カーネルのパラメータγ，ϵ-SVRにおけるパラメータϵなどの値を適切に設定する必要がある．適切なパラメータを設定することにより，より汎化能力の高い識別器を構築することができる．このようなパラメータを決定するために，学習データを用いてCross-Validation（交差検定）で予測した汎化誤差が最小になるようなパラメータを選択する方法がある．ここでCross-Validationとは，学習データをいくつかに分割し，一部を学習したモデルの評価用として用い，残りを使って学習し評価するという作業を，評価用データを変えながら分割数だけ繰り返すことによって，平均の評価値を求める評価方法である．学習データに対するCross-Validationの誤識別率や汎化誤差を最小化するパラメータを見つけるために，手作業でパラメータを少しずつ変更して探索することは可能だが，非常にコストがかかり，また最適なパラメータを見つけられるかは個人の経験と勘による．そこで，Cross-Validationの誤識別率を最小にすることを選択基準とし，自動的にSVMのパラメータを決定する方法が存在する．最も一般的な方法としては，任意の範囲を格子点探索する方法(grid search)である．たとえば，格子点探索では，探索するパラメータの値の範囲と，探索する間隔 を設定する．そして，任意の範囲，間隔でC, γを指数関数的に（たとえば，$C = 2^{-5}, 2^{-3}, ..., 2^{15}; \gamma = 2^{-5}, 2^{-3}, ..., 2^{3}$) 増加させ，それぞれの学習でのCross-Validationの 誤識別率を評価する．そして，探索したものの中で最もCross-Validationの 誤識別率が低いパラメータを採用する．

本研究においては，この方法で自動的に決定した最適パラメータを実験に用いることとする．

4.4 評価実験

本節では，提案方法の性能評価のために行った評価実験の結果を示す．以下では，まず 4.2 節で述べた様々な素性の中で，どの素性が概念間関連度に影響を与えているかを検証する．次に，これらの素性を用いた SVR に基づく概念間関連度の測定手法について，性能評価を行う．

4.4.1 重要素性の検証

検証方法

ここでは，4.2 節で述べた素性の中で，どの素性が概念間関連度にとって重要な素性かを，4.2.2 項で述べた F-score によって検証する．ここでの，素性の重要性とは，単独で概念間関連度の測定に用いた場合ではなく，その他の素性と複合的に用いた場合の，概念間関連度へ与える影響を表す．F-score を求めるためには，4.3.2 項で述べた学習データが必要となる．計算式 (4.21) からも明らかなように，各入力ベクトルのラベル値は正例か負例の 2 クラスで与えられる．しかし，本研究ではラベル値を実数の概念間関連度として与えているので，このままでは F-score の計算式に適用することができない．そこで，入力ベクトルにおけるラベル値を 2 値化することを考える．本実験では，関連度を 2 値化する閾値として 5 を設定し，「ある程度以上関連がある」グループと「それほど関連しない，もしくは全く関連がない」グループに分けることにより，F-score 値の算出を行う．表 4.1 に示すように，F-score を求めるための学習データとして，付録 A で述べる「WikiSimi Test Collection」より関連度が 5 以上の概念ペアと関連度が 5 未満の概念ペアを，それぞれランダムに 750 件取得し，合計 1,500 個の概念ペアに対する入力ベクトルを生成する．同様にこのような手順を繰り返して，50 セットの学習データを生成する．この合計 75,000 の入力ベクトルを含む 50 セットの学習データそれぞれに対して F-score を算出し，それぞれの素性ごとに 50 回の平均 F-score を求める．

表 4.1: F-score 算出のための学習データの統計

データ内容	データ数
学習データ数合計	1,500
（関連度が 5 以上の学習データ数）	(750)
（関連度が 5 未満の学習データ数）	(750)
学習データのセット数	50

F-score 算出結果

表 4.2 に，50 回の平均 F-score の算出結果を示す．まず，学習素性として用いた中で，F-score が高い上位に「2nd pfibf」や「2nd tfidf」「2nd Link Co-occurrence」「lch」が存在する．これらは，Wikipedia のデータを解析することによって概念間関連度を測定する従来手法であり，重要な情報を提供していることが改めて確認できる結果である．その他にも,「Bidirectional Link Check」が高い F-score を持っている．これは，記事どうしが双方向リンクを持っていることが，関連性が高いか低いかを大きく隔てる要素となっているという仮説が証明されていると言える．

また興味深い結果として，記事 p_b を中心とした学習素性「tfidf of l_a in p_b」の F-score は高い値であり，また「ibf of p_b」「# of Backward Links of p_b」の F-score も比較的上位に位置しているということである．一方で，記事 p_a を中心とした学習素性「tfidf of l_b in p_a」の F-score は比較的下位に位置し,「ibf of p_a」「# of Backward Links of p_a」に至っては非常に低い F-score となっている．このことは，テストコレクションに定義されている関連度が，記事 p_a から p_b への関連度と記事 p_b から p_a への関連度で同一ではなく，ある一方への方向がある可能性を示している．ここで，本実験の学習データと評価データとして用いている「WikiSimi Test Collection」は，ある概念を起点にその概念から，与えられたその他の概念群への連想する度合いを被験者がスコアリングするタスクによって構築されている．そのデータは「[連想元概念] [連想先概念] [関連度]」という 3 つ組のデータとして定義されている．つまり，本実験では p_a から p_b への連想度を予測するモデルデータを SVR に

表 4.2: 平均 F-score

順位	素性番号	素性名	F-score
1	8	2nd pfibf	0.14884
2	11	2nd tfidf	0.13945
3	3	Bidirectional Link Check	0.12054
4	9	tfidf of l_a in p_b	0.09777
5	6	2nd Link Co-occurrence	0.07312
6	2	lch	0.07303
7	17	ibf of p_b	0.06061
8	15	# of Backward Link of p_b	0.05689
9	7	pfibf	0.05102
10	5	Link Co-occurrence	0.04266
11	1	Same Category Check	0.04186
12	10	tfidf of l_b in p_a	0.03797
13	13	# of Forward Link of p_b	0.01558
14	4	Inversed Link Order	0.00558
15	12	# of Forward Link of p_a	0.00290
16	14	# of Backward Link of p_a	0.00019
17	16	ibf of p_a	0.00019

よって構築したことになる．その意味で，この方向のある連想関係を予測する上では，連想先概念，つまり本実験においては p_b を中心とした情報が重要な役割を果たす，という知見が得られたことになる．

他には，「# of Forward Links of p_a」「# of Forward Links of p_b」「Inversed Link Order」が低い F-score となっている．この結果から分かることは，本学習においてはフォワードリンクの数からは，予測精度に影響を与えるような有用な情報は得られないということであり，記事中に存在するリンクの出現位置も同様である．

4.4.2 SVRに基づく手法の性能評価

ここでは，SVRに基づく手法，つまり4.3.1項で述べたSVRを用いて4.2節で挙げた情報を学習素性として用い，概念間関連度の測定を行う手法についての性能評価を行う．本実験を行うために，まず実験における学習フェーズ，つまりモデルデータの作成を行う際の学習データと，予測フェーズ，つまりモデルデータによって関連度が未知の概念ペアの関連度を予測する際の評価データを作成する必要がある．それぞれのデータは，様々なデータ件数や関連度分布による実験を行うために,「WikiSimi Test Collection」からランダムに任意の件数，特定の関連度分布で抽出できるよう，以下の手順によって作成される．

1. 評価データとして,「WikiSimi Test Collection」から，関連度がr_1以上の概念ペアをランダムにαn件取得する．

2. 評価データとして，同様に関連度がr_1未満の概念ペアをランダムに$n-(\alpha n)$件取得する．

3. 学習データとして，評価データに含まれる概念ペアを「WikiSimi Test Collection」から除去した上で，関連度がr_2以上の概念ペアをランダムにβm件取得する．

4. 学習データとして，同様に関連度がr_2未満の概念ペアをランダムに$m-(\beta m)$件取得する．

ここで，パラメータn，m (>0)はそれぞれ作成する評価データと学習データのデータ数である．なお，本実験では第2章，第3章と同様に評価データの件数nを100件とする．また，パラメータr_1，r_2 $(0 \leq r_1, r_2 \leq 10)$はそれぞれ評価データと学習データを作成する際の関連度の閾値であり，パラメータα，β $(0 \leq \alpha, \beta \leq 1)$はそれぞれ評価データと学習データを作成する際の，関連度が閾値r_1，r_2以上のデータを取得する件数がデータ全体に占める割合である．つまり，パラメータmによって学習データ数の違いによる概念間関連度の予測性能の違いを評価することができる．また，r_1，r_2とα，βによって，評価データ中の概念ペアの関連度や学習データに含まれる概念ペアの関連度の大小のバランスを調整することができ

る．たとえば，$r_1 = 5, r_2 = 5, \alpha = 0.5, \beta = 0.5$ とすると，評価データと学習データの中に占める関連度分布がある程度均一になる．

上記の作成手順を 50 回繰り返すことにより，50 セットの評価データと学習データを作成し，これを 1 セットの実験データセットとする．つまり，1 回の実験で機械学習と評価を 50 回繰り返す．これによって，実験結果が特定の評価セットに過剰適合（over-fitting）していないことを保証する．また，学習データに評価データを含めないことで，未知の概念ペアに関しても予測可能であることを効果的に検証できる．

実際の SVR による機械学習を行うためのライブラリとしては「LIBSVM[2]」[12] を用い，4.3.1 項で述べた ϵ-SVR のカーネル関数としては RBF カーネルを用いる．学習機に与える，ソフトマージンのパラメータ C，カーネルのパラメータ γ，ϵ-SVR のパラメータ ϵ は，各実験毎に 4.3.3 項で説明した Grid Search によって最適なパラメータを算出し，そのパラメータを用いて学習を行う．

性能の評価方法は，第 2 章，第 3 章と同様，評価データに含まれる概念ペアの関連度を予測した n 件の結果と，テストコレクションで与えられている 2 つの関連度セットの順位相関を,「スピアマンの順位相関係数（Spearman rank-order correlation coefficient）」によって求め，概念間関連度の精度とする．

学習データ数による影響

ここでは，SVR に基づく手法において，学習データ数を変化させることによって，精度にどのような影響を与えるかを調査する．評価データと学習データを作成するためのパラメータとしては，以下のものを用いる．

$$
\begin{aligned}
n &= 100 \\
m &= 125, 250, 500, 750, 1000, 1500 \\
r_1 &= r_2 = 5 \\
\alpha &= \beta = 0.5
\end{aligned}
$$

[2]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

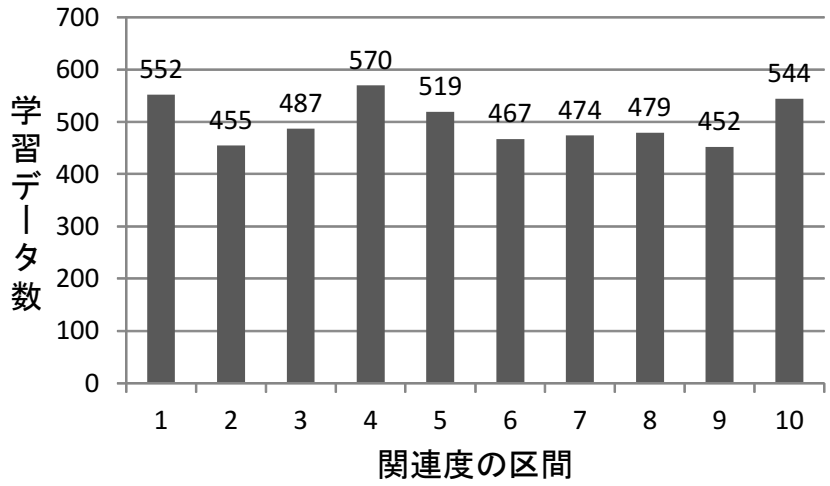

図 4.1: 評価データの関連度によるヒストグラム（学習データ数 1,500 件）

学習データに含まれる概念ペア数 m を変化させることによって，学習データ数における精度の影響を調査する．また，r_1, r_2, α, β は，データ中に含まれる概念ペアの関連度分布が均一になるように設定している．これは，関連度が低いものから高いものまで予測可能な，汎用的なモデルを構築するための措置である．ここで，作成された評価データが関連度に対してどのような分布になっているかを図 4.1 に示す．このグラフは，評価データ数が 1,500 件の場合のヒストグラムであり，グラフより評価データがそれぞれの関連度区間においてほぼ一様に存在することが分かる．また，全評価データの平均値は 5.14 となっている．

図 4.2, 図 4.3 に，学習データ数の違いによる予測精度を示す．図 4.2 は，50 回の行った実験の 10 回の移動平均を表している．本グラフより，評価データが異なる実験区間の多くで，学習データ数を増やすごとに予測精度，つまりテストコレクションから抽出した評価データとの相関が増加していることが分かる．また，図 4.3 は，50 回の実験の平均値であり，全体としても学習データを増やすことによって精度が向上している．実験結果より，多くのサンプルを学習させることによって，より精度の高い予測を行えるという知見が得られた．しかし，学習データ数の増加による，相関値の上昇率は減少しており，本実験における学習データ 1,500 でほぼ頭打ちになっている．ただし，この学習データ数が精度へ与える影響は，素

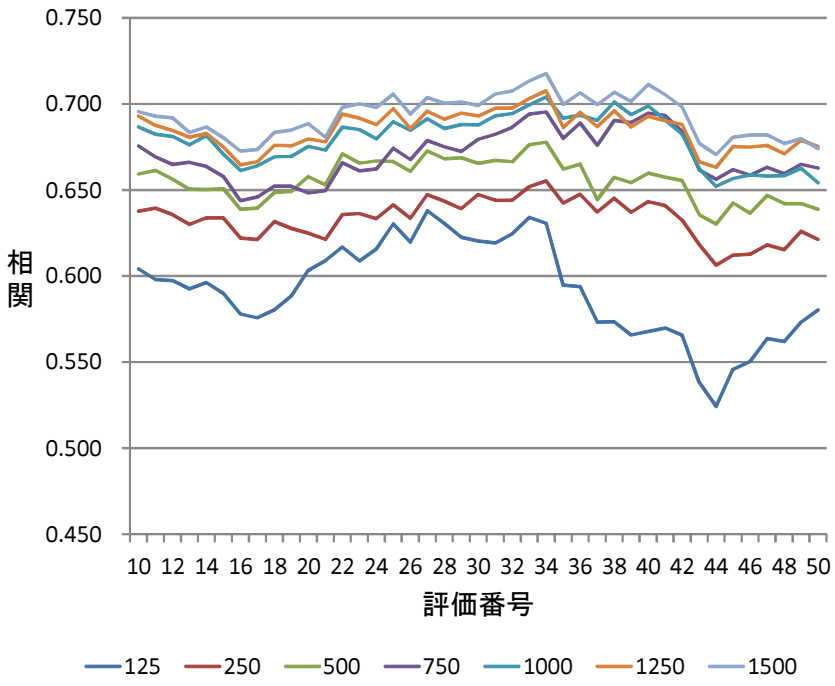

図 4.2: 異なる学習データ数での精度（10 回の移動平均）

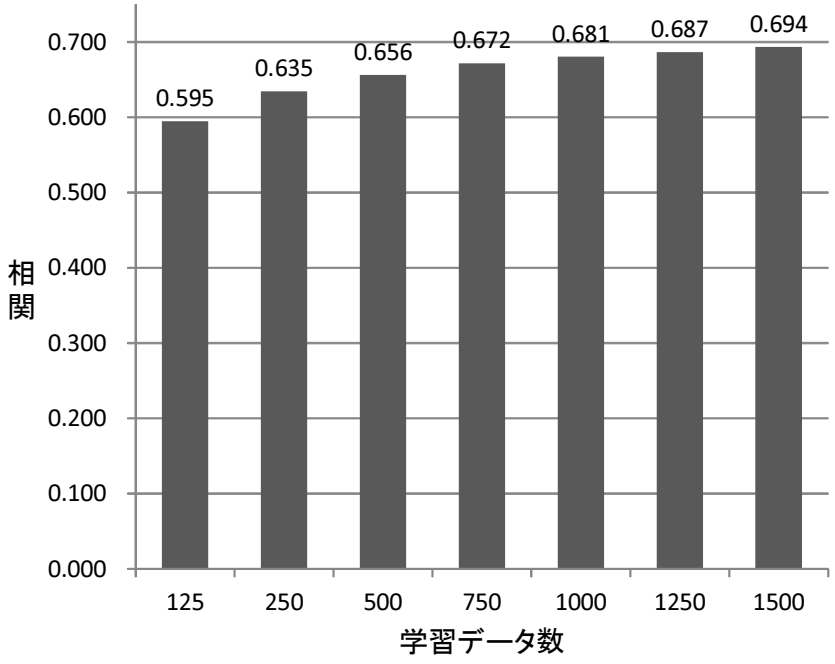

図 4.3: 異なる学習データ数での精度（平均値）

性の種類や数にも依存するため，それらを変更した際には，改めて検証する必要がある．

他手法との比較

最後に，SVR に基づく提案手法の有効性を，他手法と比較することによって示す．提案手法においては，学習データを 1,500 用い，すべての素性を学習させたものを用いる．比較手法としては，2nd tfidf，2nd pfibf 及び 第 3 章で提案した融合手法（LCA + tfidf）を用いる．なお，融合手法のパラメータ α には 第 3 章の実験結果で最も精度の良かった 0.8 を設定する．

図 4.4，図 4.5 に，提案手法を SVR-base として，他手法との比較結果を示す．図 4.4 の実験結果より，すべての実験区間において，提案手法は他のどの手法よりも大幅に高精度であることが分かる．同様に，図 4.5 の平均精度においても，他手法に比べて提案手法は大きなアドバンテージを持っていることが分かる．この結果は，既存手法による 2nd pfibf, 2nd tfidf, リンク共起性解析，lch など複数の関連度を特徴情報として考慮しつつ，様々な条件下に柔軟に対応した為だと考えることができる．つまり，双方向リンクやバックワードリンク数などの，記事から取得可能なより直接的な情報を用いることによって，各手法による関連度情報を適切なバランスパラメータによって融合できたと考えられる．これは，第 3 章の実験において課題となっていた，柔軟な合成パラメータ α の設定に対する一定の解となったと思われる．

一方で，現在の手法の計算コストを考察すると，第 2 章で議論したように，pfibf は飛び抜けて計算コストが高い．第 3 章で提案した手法において，計算量を低く抑えつつ pfibf と同等の精度を実現したが，柔軟なパラメータ設定という課題を解決することによってより高い精度が期待できることを既に述べた．本実験において，4.2 節に挙げた様々な素性を考慮することによって，柔軟なパラメータ設定によって精度向上を実現出来ているという結果が得られた．つまり，もし pfbif に関する 2 つの素性 Feature 7, Feature 8 を除いて学習させたとしても，2nd tfidf, リンク共起性解析，加えて lch といった手法が，柔軟なパラメータ設定によって融合され，2nd pfibf よりも大幅に低いコストで，2nd pfibf よりも高い精度を実現できる

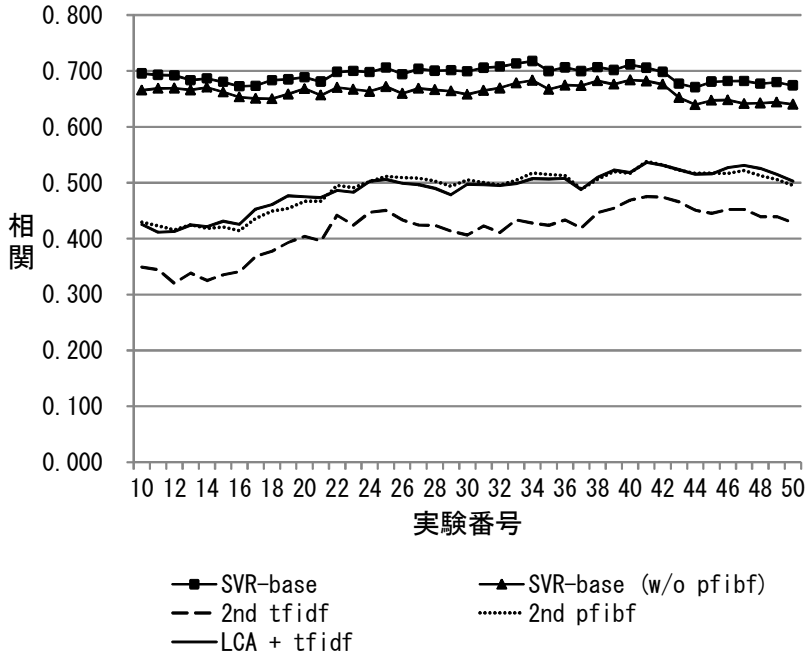

図 4.4: 他手法との比較（10 回の移動平均）

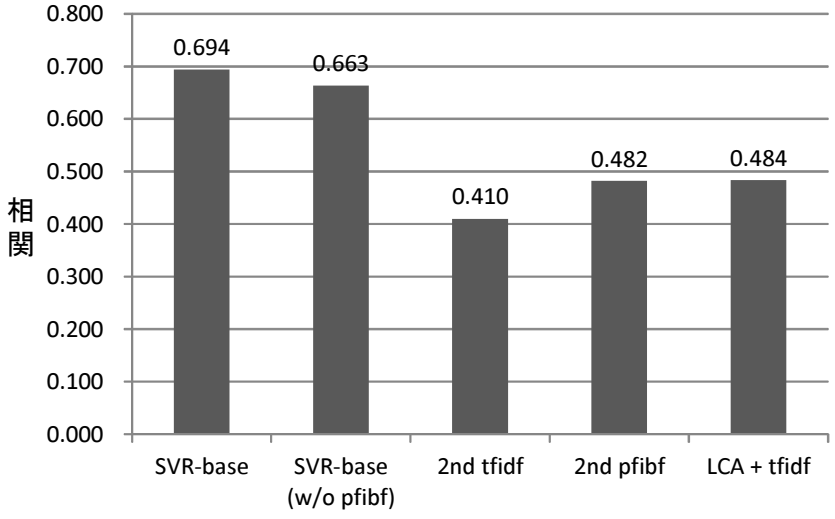

図 4.5: 他手法との比較（平均値）

ことが期待される．そこで，pfibfに関する2つの素性 Feature 7, Feature 8 を除いて学習させた結果が，実験結果の図中のSVR-base (w/o pfibf) である．図4.5より，SVR-base (w/o pfibf) は pfibfに関する素性 を用いていないにもかかわらず，どの実験区間においてもSVR-baseから若干の精度低下に抑えられ，pfibfよりも大幅に高精度な概念間関連度の測定が出来ていることが分かった．この結果は，図4.5の平均精度の結果からも明らかである．本実験により，提案手法が低い計算コストにもかかわらず高精度な概念間関連度の予測を実現していることが確認できた．

4.5 むすび

本章では，Wikipediaから取得可能な情報，及び既存手法などで生成可能な情報の中で，どの情報が概念間関連度の測定において重要なのかを検証すると共に，それらの情報を学習素性として機械学習させることによって，個々の情報を複合的に考慮した概念間関連度の測定手法を提案した．本研究では，Wikipediaから取得，及び生成可能な情報の中で，言語に依存しない，つまり自然言語解析の必要のない情報を，概念や概念間の関連性を特徴付ける素性（Feature）の候補として提案し，それらの概念間関連度の測定に与える重要性を，F-scoreを求めることによって検証した．また，それらの素性を機械学習手法であるSVMの回帰問題への対応を可能としたSVRを用いて，機械学習させることにより，記事や記事間に関する複数の素性を考慮した概念間関連度の測定を行った．

評価実験より，提案した素性の中で「2nd pfibf」や「2nd tfidf」「2nd Link Co-occurrence」「lch」などの概念間関連度の測定に関する既存手法が非常に重要であり，加えて「Bidirectional Link Check」も重要であり，「tfidf of l_a in p_b」「ibf of p_b」「# of Backward Links of p_b」などの，連想先の概念に関する記事を中心とした指標も重要であるという知見が得られた．また，SVRに基づく手法の性能評価では，学習データ数1,500において最も高い精度を達成し，他手法との比較すると，本手法は学習素性からpfibfに基づく素性を抜いてもなお，どの従来手法よりも大幅に高い精度を実現していることが分かった．この結果は，本手法は少ない計算量で非常に高精度で汎用的な概念間関連度の測定可能な手法であることを意味する．

第5章　結論

5.1　本論文のまとめ

本論文では，高い網羅性を持ち高精度な概念間関連度の測定を行うことを目的として，Wikipedia をコーパスとした解析手法を議論した．まず，第 1 章では，情報爆発時代における本研究の重要性を述べ，Wikipedia の知識抽出のためのコーパスとしての有用な特徴を明らかにした．

第 2 章では，リンクの共起性解析に基づき，高精度で計算量の少ない概念間関連度の測定方法を提案した．これまでの Wikipedia を用いた概念間関連度の測定手法において，各記事内の情報によってその記事が表す概念を特徴付ける手法の場合，精度が記事の質に影響される場合がある．また，記事間のリンク構造を解析して記事間つまり概念間の関連度を求める手法の場合，膨大な計算が発生するという問題があった．そこでこの方法では，記事内における記事間リンクに対して近傍での共起をカウントし，その処理を Wikipedia のすべての記事で行うことによって，2 つの記事の共起性を求める．そして，その共起性の値を用いてある記事の共起ベクトルを生成し，そのベクトルの距離をコサイン類似度によって求めることで，2 つの記事間，つまりそれらの記事が表す概念間の関連度を測定する．提案手法の性能評価のために行った評価実験により，提案手法は，従来研究である tfidf に基づく手法よりも高い精度を実現し，また 2nd pfibf よりも解析時間が大幅に短いにもかかわらず，pfibf に迫る高い精度を実現していることが証明できた．特に，一次共起性の計算手法としては，Cosine が最も高い精度で概念間関連度を測定できることが判明した．

第 3 章では， 大域的情報と局所的情報という性質の異なる 2 つの情報を組み合わせることによって，より高精度な概念間関連度の測定方法を提案した．リンク

共起性解析のような大域的情報を用いた手法では，記事へのリンクが少ない場合，特徴情報が十分取れない可能性がある．一方で，そのような記事に対して局所的情報である記事内のリンクがある程度存在した場合，その記事の特徴を端的に表すものとして局所的情報は依然として有用である．この方法では，Wikipedia 全体の大域的情報と記事内の局所的情報の両方を活用することによって，それぞれの情報量不足を補う．具体的には，記事の特徴を表すために，Wikipedia 全体でのリンクの共起性情報と，記事内のリンクの tfidf に基づく情報をそれぞれベクトル化し，パラメータ α によってベクトルの合成を行う．提案方法の性能評価のために行った評価実験の結果より，すべての実験において単独の情報を用いるよりも高精度に概念間関連度を測定出来ていることが分かった．具体的には，提案手法は低い解析コストにもかかわらず，パラメータ α が 0.2，0.3，0.8， 0.9 において，2nd pfibf と同等か若干上回る精度を実現していることが分かった．また，各実験においてパラメータ $\alpha = 0.0$〜1.0 の変化に応じて，精度が「極大値あり」,「極大値なし」の 2 つのパターンで変化するとこが分かった．この結果より，パラメータ α を動的に各パターンに適応させるように設定することにより，さらなる精度の向上を期待できるという知見が得られた．

第 4 章では，Wikipedia から取得可能な記事間の関連性を特徴付ける様々な情報の中で，どの情報が概念間関連度に有用であるかを検証し，さらに複数の情報を SVR によって組み合わせた高精度な概念間関連度の測定方法を提案した．これまで Wikipedia を活用して概念間関連度を測定する手法はいくつか提案されているが，いずれも単独の特徴情報を用いているものがほとんどであり，Wikipedia から取得もしくは生成可能な記事間の関係性を特徴付けることができる情報は多く存在する．そこで，Wikipedia から取得可能なそれらの情報を提示し，どの情報が概念間関連度の測定に対して有用な特徴情報であるかを，機械学習における素性の貢献度を計算する手法である F-score によって調査した．さらに，回帰問題を解くことのできる機械学習手法である SVR を用いて，それら複数の情報を学習素性として組み合わせることによって，高精度な概念間関連度の測定手法を提案した．F-score による重要情報の調査結果より，2nd tfidf, 2nd pfibf, 2nd Cooccurrence などの既存手法のベクトル比較や，双方向リンクがあるかどうか，また記事が所属するカテ

ゴリ間の距離も有用であることが分かった．また，提案手法の性能評価のために行った評価実験の結果より，すべての素性を用いた場合と，高解析コストな手法である pfibf に関する素性を除いた場合，いずれにおいても提案手法は特筆すべき精度向上を実現していることが分かった．

5.2 今後の研究課題

5.2.1 アプリケーションへの適用

本研究では，Wikipedia から測定した概念間関連度の精度をテストコレクションによって評価したが，実用においての有効性を検証するためには，実際のアプリケーションに適用した場合の評価が不可欠である．今後は，提案した概念間関連度の測定手法を文書分類や文書要約システム，情報検索システムに適用することによって，実用性を評価する予定である．

5.2.2 他プロジェクトへの適用

提案手法は Wikipedia のみならず，多数の内部リンクを持つ他の Web サイトにおいても適用可能だと考えられる．たとえば，Wikipedia の姉妹プロジェクトである Wikinews などがその適用対象である．ただし，これらのプロジェクトは，現時点では提案手法を適用できるほどの十分な量のデータがない状態である．

5.2.3 大規模オントロジの構築

現在，Wikipedia を解析することによって，関連度だけでなく，関連の種類（is-a や part-of）の抽出を行うオントロジの構築に関する研究も行われている．しかし，それらはカテゴリ構造を使うものがほとんどであり，構築できる規模は限られている．そこで，Wikpedia の記事へのリンクを含む記述文を自然言語解析することによって，概念間の関連を取得することが可能であると考える．この方法では，Wikipedia の持つ記事が表す全ての概念を対象として，オントロジを構築する

ことができる．一方で，オントロジの実用においては異なるオントロジ間のマッピングも重要な課題である．たとえば，Semantic Web などにおけるオントロジでは．各分野で専門家などの人手によって構築されたオントロジが存在する．しかし，それらの知識が個別で存在している限り，Web などの大規模で多様な情報を適切に整理することができない．そこで，Wikipedia から構築した概念構造や概念間関連度をハブ知識として用い，それぞれのオントロジを意味を考慮してマッピングすることによって，実質的な大規模オントロジを実現したい．その大規模オントロジを，将来的に知的システムの思考の知識源とすることによって，人工知能分野の発展に寄与していきたい．

付 録A　WikiSimi Test Collection

A.1　はじめに

近年，Wikipedia を用いた概念間の関連度計算手法に関する研究が行われてきた．関連度計算手法研究における評価方法は，主に主観評価，アプリケーションによる評価，テストコレクションによる評価の 3 つがある．主観評価は文献 [16,62] のように，提案する関連度計算手法を用いて抽出した関連性のある概念ペアのリストを複数の被験者に提示し，関連度を 3 段階や 10 段階で主観評価してもらい，精度を測る方法である．アプリケーションによる評価は文献 [6,78] のように，手法を語義曖昧性解消 や情報検索などのアプリケーションに適用して，その適合率や再現率を評価している．テストコレクションによる評価は文献 [25,33,75] などの研究のように，“M&C” [50] や “R&G” [69], 近年では語彙数の多さから “WordSimilarity-353 Test Collection” [24] を用いた評価が行われている．とりわけ，テストコレクションによる評価は，複数の手法を同じ基準で評価できるため，手法間の比較を行う上で重要な評価手法である．しかし，これらのテストコレクションは，Wikipedia がカバーする概念に比べて語彙の網羅性が低く，また曖昧性が解消されていないなどの問題がある．本稿では，筆者らの構築しているテストコレクション「WikiSimi Test Collection」[36] について解説する．

A.2　従来のテストコレクションの問題点

WordSimilarity-353 Test Collection [24] は Wikipedia を用いた概念間の関連度計算に関する研究 [33,75] において，もっともよく用いられているテストコレクションである．WordSimilarity-353 Test Collection は，353 組の単語を 13 人〜16 人の被

験者によって関連性を主観で 10 段階評価し，その平均を関連度としている．評価する際は，各手法によって 353 組の単語の関連度を計算し，各手法によって計算された 353 個の関連度とテストコレクションの 353 個の関連度との相関を，スピアマンの順位相関係数やピアソンの相関係数などで求めることによって，手法の精度を測る．

しかし，このテストコレクションは，Wikipedia 研究における概念間の関連度計算手法を評価する際のいくつかの問題を抱えている．まず 1 つ目に，単語ペアの数が非常に限られているということである．また，その単語も一般語に偏っている．Web や Wikipedia をコーパスとした概念間関連度の測定手法は，様々なドメインの専門語や固有名詞も含めた多くの概念を扱うことができる．実際に，概念間関連度を用いる多くのアプリケーションは，一般語よりもむしろ固有名詞間の概念間関連度を必要とする場合が多いと考えられる．しかし，WordSimilarity-353 Test Collection は 353 組の語だけしか含まれておらず，また多くの一般語を含んでいる．つまり，このテストコレクションの網羅性は概念間関連度の測定手法を評価するためには不十分である．

そして，さらに大きな問題として，定義されている単語の曖昧性が解消されておらず，評価する際に Wikipedia の概念（記事）にマッピングするという作業を行わなければならないことが挙げられる．しかし，多義語の場合，どの概念にマッピングすればよいか判断できないものもある．そのため，マッピングが正確であるという保証がない．たとえば,「stock」は市場経済における「株」という意味もあれば，店などにおける「在庫」などの意味もある．以前の研究では，このマッピング作業は研究者それぞれで独自で行っておりマッピング方法が異なっている．しかし，正当な評価を行うためには，テストコレクションにおけるそれぞれの語の意味が一意でなければならない．また，そもそもテストコレクション作成時に単一概念として被験者に提示されているわけではなく，意味の曖昧性も含んだ単なる単語として提示され，関連度を判定している点も問題である．

A.3 WikiSimi Test Collection

前節で述べた問題を解決するために，筆者は，英語版 Wikipedia の概念を基にしたテストコレクション「WikiSimi Test Collection」を構築した．Wikipedia の概念に基づいたテストコレクションを構築することによって，概念間の関連度計算手法を評価する際に，概念をテストコレクションの単語にマッピングする作業を必要とせず，より精度の高い評価が可能となると考えられる．本章では,「WikiSimi Test Collection」の構築方法と構築結果について述べる．

A.3.1 構築方法

「WikiSimi Test Collection」の構築は，複数人の英語ネイティブの被験者を雇い，以下の流れで行った．

1. Wikipedia 記事群のクリーニング
   - フォワードリンク もしくは バックワードリンク の数が 5 以下の記事をノイズ記事として除外した．
   - リストページやカテゴリページなどの通常記事以外を除外した．
2. 残った 4,800 万リンクを含む 150 万記事から，人物・地理・文化などの様々な分野の概念から，各被験者が 43 個の概念（記事）を選定した．
   (例：“Bill Gates”, “Michael Jackson”, “Indian Ocean”, “Hawaii”, “Microsoft”, “Basketball”, “Cheese”, “Dance”, “Water”)
3. 概念ペアを生成するために，被験者が各概念を表す記事内から，手動で他の概念（記事）へのリンクを選択した．加えて，関連性のある概念ペアを増やすために，それぞれの概念に対して被験者が手動で思いつく関連概念を追加した．
4. 各被験者が各概念ペアの関連度を，WordSimilarity-353 Test Collection の基準に従って 0（全く関係ない）から 10（非常に関係がある）のスコアで評価し

表 A.1: 評価基準

スコア	基準
9 - 10	強く連想される. （例：Microsoft -> Windows）
7 - 9	連想される.
5 - 7	ある程度連想される.
3 - 5	連想されるかもしれない.
2 - 3	自信がないが，連想されないと思う. （例：Microsoft -> Mac OS X）
1 - 2	基本的には連想されない.
0 - 1	絶対連想されない.（例：Cabbage -> Microsoft）

た．その際，より詳細な基準を表 A.1 のように定め，被験者に提示した．さらに，被験者のスコアリング精度をより高めるために，図 A.1 に示すようなアプリケーションを構築した．本アプリケーションは，ユーザが概念間の関連度をスコアリングする際に，それぞれの概念の意味を効率的に Wikipedia の記事から確認することができる．この作業によって，被験者がより正確な判断を行うことが期待できる．また，スコアの結果を登録する前に，スコア順に概念ペアをソートして概念ペア群のスコアリングのバランスを確認し，変更することができる.

5. それぞれの概念ペアの関連度を，すべての被験者のスコアの平均値として与えた.

A.3.2 構築結果

前節の方法で構築された本テストコレクションは，合計で 1,749 の概念ペアを含む大規模なものとなっている．各概念ペアは，5〜9 人の英語ネイティブの被験者

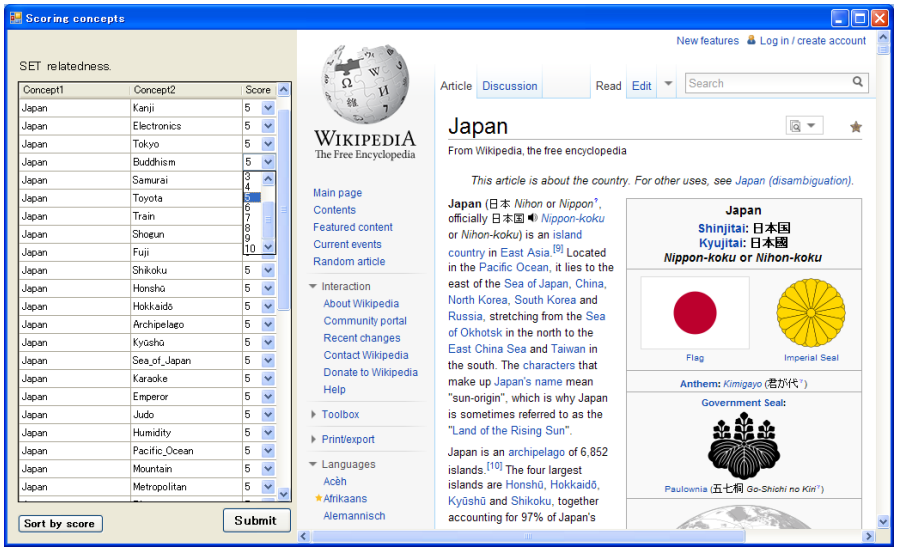

図 A.1: テストコレクション構築のためのアプリケーション

によって評価されている．構築したテストコレクションの例を表 A.2 に示す．特筆すべき点は，各ペアの関連度評価がそれぞれの概念の記事を確認した上で行われるので，確実に単一概念どうしの関連度評価が行われているということである．また，WikiSimi Test Collection の利点として，Wikipedia のリンクを利用した概念間の関連度を計算する研究との親和性が高いことが挙げられる．たとえば，文献 [58] や [33] では，取り扱っているすべての概念が Wikipedia に存在する概念（記事）であるため，評価のためのテストコレクションとして WikiSimi Test Collection をそのまま用いることができる．さらに，本テストコレクションは Wikipedia 研究だけでなく，様々な概念間関連度測定手法の評価に用いることができる．

A.4 まとめ

本付録では，Wikipedia の概念に基づく連想関係テストコレクション「WikiSimi Test Collection」について述べた．WikiSimi Test Collection は，これまでの単語を基にしたテストコレクションと違い，語彙の曖昧性を解消した Wikipedia の概念 (記事) を基にしたテストコレクションである．テストコレクションを構築する際

表 A.2: WikiSimi Test Collection の一例

概念 1	概念 2	関連度
Mobile phone	Cell phone	9.8
Thailand	Thai people	9.6
Photography	Photographic paper	8.4
Thailand	Buddhist	8.0
Photography	Photojournalism	7.6
Mobile phone	AT&T	7.2
Mobile phone	Bluetooth	6.6
Photography	Black-and-white	6.0
Mobile phone	Car phone	5.8
Thailand	Myanmar	4.6
Mobile phone	Las Vegas, Nevada	1.2
Thailand	Europe	1.2
Photography	Greek language	1.0
Photography	France	0.8
Mobile phone	Postage stamp	0.2

の被験者は，評価対象の語彙がどの概念を表しているかを明確に認識しながら関連度を評価している．このテストコレクションを用いることによって，Wikipedia に存在する多くの概念どうしの関連度をより網羅的に評価することができる．

今後は，より多くの被験者に評価してもらうことによって，テストコレクションの品質を高めるとともに，Wikipedia の概念を基にした照応関係や上位下位関係を定義したテストコレクションの構築も行っていく予定である．

謝辞

本研究全般に関して，懇切なる御指導と惜しみない御助言を頂きました大阪大学大学院情報科学研究科マルチメディア工学専攻 西尾章治郎教授に謹んで御礼申し上げます.

本研究を推進するにあたり，直接の御指導，御助言，御討論を頂きました大阪大学大学院情報科学研究科マルチメディア工学専攻 原隆浩准教授に衷心より感謝申し上げます.

本論文をまとめるにあたり，大変有益な御指導と御助言を多数賜りました大阪大学大学院情報科学研究科マルチメディア工学専攻 細田耕教授，独立行政法人情報通信研究機構 木俵豊博士に心より感謝申し上げます.

講義，学生生活を通じて，学問に取り組む姿勢をご教授頂きました大阪大学大学院情報科学研究科 マルチメディア工学専攻藤原融教授，薦田憲久教授に厚く感謝申し上げます.

本研究において，ともに研究を進め，直接の御助言，御協力，御討論を頂いた東京大学知の構造化センター 中山浩太郎助教に深く御礼申し上げます.

本研究において，多大なる御助言，御協力，御支援を頂きました大阪大学大学院情報科学研究科マルチメディア工学専攻 寺西裕一准教授，大阪大学大学院工学研究科 春本要准教授，神戸大学大学院工学研究科 寺田努准教授，大阪大学サイバーメディアセンター 義久智樹准教授，東京大学情報基盤センター 小川剛史准教授に深謝致します.

本研究において，ともに研究を進め，多大なる御協力を頂いた大阪大学大学院情報科学研究科マルチメディア工学専攻 Erdmann Maike 氏，白川真澄氏に深く御礼申し上げます.

筆者の所属する研究グループにおいて，有益な御助言を頂いた小牧大治郎氏，

岩田麻佑氏，足利えりか氏，鈴木晃祥氏，宮本大樹氏に感謝の意を表します．

本研究を進める上で惜しみない御助言，御協力，研究活動を進めるにあたっての多大なる御支援を頂いた裴明花博士，Microsoft Research Asia 荒瀬由紀博士，飯間悠樹氏に感謝の意を表します．

本研究を進めるにあたり，多くの御討論や御助言を頂きました大阪大学大学院情報科学研究科マルチメディア工学専攻 西尾研究室の諸氏に心より感謝申し上げます．

私生活において大変お世話になり，時には心の支えにもなって頂いた，これまで出会った全ての友人，知人に心より感謝申し上げます．特に，中学，高校を通して家庭教師としてお世話になった大阪府立大学人間社会学部 川部哲也博士には，私の最も困難だった時期に学業，精神両面において支えて頂きました．また，大阪府立桃谷高等学校，立命館大学，立命館大学古美術研究会，大阪大学で出会った友人達は，今までも，そしてこれからも私の人生においてかけがえのない財産となっています．

最後に，私のこれまでの人生，そして研究生活を送る上で，暖かい支援と理解を頂いた両親をはじめとする家族に心から感謝致します．

参考文献

[1] Abiteboul, S., Buneman, P., and Suciu, D.: *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000).

[2] Adler, B. T., and de Alfaro, L.: A Content-driven Reputation System for the Wikipedia, in *Proceedings of the International World Wide Web Conference (WWW 2007)*, pp. 261–270 (2007).

[3] Akay, M. F.: Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis, *Journal of Expert Systems with Applications*, Vol. 36, No. 2, pp. 3240–3247 (2009).

[4] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G.: DBpedia: A Nucleus for a Web of Open Data, in *Proceedings of the International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC 2007)*, pp. 722–735 (2007).

[5] Baeza-Yates, R. A., and Ribeiro-Neto, B. A.: *Modern Information Retrieval*, ACM Press / Addison-Wesley (1999).

[6] Banerjee, S., and Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness, in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pp. 805–810 (2003).

[7] Bray, T., Paoli, J., and Sperberg-McQueen, C. M.: Extensible Markup Language (XML), *The World Wide Web Journal*, Vol. 2, No. 4, pp. 27–66 (1997).

[8] Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S.: Web Caching and Zipf-like Distributions: Evidence and Implications, in *Proceedings of IEEE INFOCOM 1999*, pp. 126–134 (1999).

[9] Brill, E.: A Simple Rule-Based Part of Speech Tagger, in *Proceedings of the Conference on Applied Natural Language Processing (ANLP 1992)*, pp. 152–155 (1992).

[10] Budanitsky, A., and Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness, *Journal of Computational Linguistics*, Vol. 32, No. 1, pp. 13–47 (2006).

[11] Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P.: Extraction of Semantic Biomedical Relations from Text using Conditional Random Fields, *Journal of BMC Bioinformatics*, Vol. 9, (2008).

[12] Chang, C. C., and Lin, C. J.: *LIBSVM: a Library for Support Vector Machines* (2001).

[13] Chen, H., Yim, T., Fye, D., and Schatz, B. R.: Automatic Thesaurus Generation for an Electronic Community System., *Journal of the American Society for Information Science*, Vol. 46, No. 3, pp. 175–193 (1995).

[14] Chen, Y.-W., and Lin, C.-J.: Combining SVMs with Various Feature Selection Strategies, in *Feature Extraction, Foundations and Applications*, Springer (2006).

[15] Chen, Z., Liu, S., Wenyin, L., Pu, G., and Ma, W.-Y.: Building a Web Thesaurus from Web Link Structure., in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pp. 48–55 (2003).

[16] Chernov, S., Iofciu, T., Nejdl, W., and Zhou, X.: Extracting Semantics Relationships between Wikipedia Categories, in *Proceedings of the Workshop on Semantic Wikis (SemWiki 2006)* (2006).

[17] Craswell, N., Hawking, D., and Robertson, S. E.: Effective Site Finding Using Link Anchor Information, in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pp. 250–257 (2001).

[18] Crouch, C. J.: A Cluster-Based Approach to Thesaurus Construction., in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1988)*, pp. 309–320 (1988).

[19] Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data, in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pp. 708–716 (2007).

[20] Davison, B. D.: Topical Locality in the Web., in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pp. 272–279 (2000).

[21] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A.: Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407 (1990).

[22] Erdmann, M., Nakayama, K., Hara, T., and Nishio, S.: Improving the Extraction of Bilingual Terminology from Wikipedia, *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 5, No. 4 (2009).

[23] Fellbaum, C. ed.: *WordNet: An Electronic Lexical Database*, MIT Press (1998).

[24] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E.: Placing Search in Context: the Concept Revisited, *ACM Transactions on Information Systems*, Vol. 20, No. 1, pp. 116–131 (2002).

[25] Gabrilovich, E., and Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis., in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 1606–1611 (2007).

[26] Giles, J.: Internet Encyclopaedias Go Head to Head, *Nature*, Vol. 438, pp. 900–901 (2005).

[27] Grefenstette, G.: SEXTANT: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 1992)*, pp. 324–326 (1992).

[28] Guyon, I., and Elisseeff, A.: An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182 (2003).

[29] Hirst, G., and St Onge, D.: *Lexical Chains as Representation of Context for the Detection and Correction Malapropisms*, pp. 305–321, MIT Press (1998).

[30] Huang, C.-L., Chen, M.-C., and Wang, C.-J.: Credit Scoring with a Data Mining Approach based on Support Vector Machines, *Journal of Expert Systems with Applications*, Vol. 33, No. 4, pp. 847–856 (2007).

[31] Ide, N., and Véronis, J.: Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, *Computational Linguistics*, Vol. 24, No. 1, pp. 1–40 (1998).

[32] 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎：Wikipedia のリンク共起性解析によるシソーラス辞書構築のスケーラビリティ向上, 情報処理学会研究報告 (データベースシステム研究会 2007-DBS-143), 第 107 巻, pp. 539–544 (2007).

[33] Ito, M., Nakayama, K., Hara, T., and Nishio, S.: Association Thesaurus Construction Methods based on Link Co-occurrence Analysis For Wikipedia, in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2008)*, pp. 817–826 (2008).

[34] 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎：Wikipedia のリンク共起性解析によるシソーラス辞書構築, 情報処理学会論文誌:データベース, Vol. 48, No. SIG20 (TOD 36), pp. 39–49 (2008).

[35] 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎：Wikipedia からの連想シソーラス構築プロジェクト, 第 20 回セマンティックウェブとオントロジー研究会 Wikipedia ワークショップ (2009).

[36] 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎：Wikipedia の概念に基づく連想関係テストコレクション「WikiSimi3000」, 第 23 回人工知能学会全国大会 (JSAI 2009) (2009).

[37] Ito, M., Nakayama, K., Hara, T., and Nishio, S.: Semantic Relatedness Measurement based on Wikipedia Link Co-occurrence Analysis, *International Journal of Web Information Systems (IJWIS)*, Vol. 48, No. SIG11 (TOD 34), pp. 27–37 (2010).

[38] 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎：Wikipedia の多様な特徴を利用した概念間関連度と有用な特徴の調査, 電子情報通信学会データ工学研究会 (DE 2010) (2010).

[39] Jarmasz, M., and Szpakowicz, S.: Roget's Thesaurus and Semantic Similarity, in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pp. 111–120 (2003).

[40] Jiang, J. J., and Conrath, D. W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X 1997)*, pp. 19–33 (1997).

[41] Kim, S. N., and Baldwin, T.: Automatic Interpretation of Noun Compounds Using WordNet Similarity, in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pp. 945–956 (2005).

[42] 北村美穂子, 松本祐治：対訳コーパスを利用した対訳表現の自動抽出, 情報処理学会論文誌, Vol. 38, No. 4, pp. 727–736 (1997).

[43] Klein, D., and Manning, C. D.: Accurate Unlexicalized Parsing, in *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL 2003)*, pp. 423–430 (2003).

[44] Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).

[45] Leacock, C., and Chodorow, M.: *Combining Local Context and WordNet Similarity for Word Sense Identification*, pp. 265–283, MIT Press (1998).

[46] Leacock, C., Chodorow, M., and Miller, G. A.: Using Corpus Statistics and WordNet Relations for Sense Identification, *Journal of Computational Linguistics*, Vol. 24, No. 1, pp. 147–165 (1998).

[47] Lin, D.: An Information-Theoretic Definition of Similarity, in *Proceedings of the International Conference on Machine Learning (ICML 1998)*, pp. 296–304 (1998).

[48] Loong, S. N. K., and Mishra, S. K.: De Novo SVM Classification of Precursor MicroRNAs from Genomic Pseudo Hairpins using Global and Intrinsic Folding Measures, *Journal of Bioinformatics*, Vol. 23, No. 11, pp. 1321–1330 (2007).

[49] Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation, in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2007)*, pp. 196–203 (2007).

[50] Miller, G. A., and Charles, W. G.: Contextual Correlates of Semantic Similarity, *Journal of Language and Cognitive Processes*, Vol. 6, No. 1, pp. 1–28 (1991).

[51] Milne, D.: Computing Semantic Relatedness using Wikipedia Link Structure, in *Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC 2007)* (2007).

[52] Milne, D. N., Medelyan, O., and Witten, I. H.: Mining Domain-Specific Thesauri from Wikipedia: A Case Study, in *Proceedings of the International Conference on Web Intelligence (WI 2006)*, pp. 442–448 (2006).

[53] Milne, D. N., and Witten, I. H.: Learning to Link with Wikipedia, in *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2008)*, pp. 509–518 (2008).

[54] Milne, D. N., Witten, I. H., and Nichols, D. M.: A Knowledge-based Search Engine Powered by Wikipedia, in *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2007)*, pp. 445–454 (2007).

[55] Milne, D., and Witten, I. H.: An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links, in *Proceedings of the American Association for Artificial Intelligence (AAAI 2008)* (2008).

[56] Nakayama, K., Hara, T., and Nishio, S.: A Thesaurus Construction Method from Large Scale Web Dictionaries., in *Proceedings of IEEE International Conference on Advanced Information Networking and Applications (AINA 2007)*, pp. 932–939 (2007).

[57] 中山浩太郎, 原 隆浩, 西尾章治郎：Web 事典からのシソーラス辞書構築手法, 情報処理学会論文誌:データベース, Vol. 48, No. SIG11 (TOD 34), pp. 27–37 (2007).

[58] Nakayama, K., Hara, T., and Nishio, S.: Wikipedia Mining for An Association Web Thesaurus Construction, in *Proceedings of the International Conference on Web Information Systems Engineering (WISE 2007)*, pp. 322–334 (2007).

[59] Nakayama, K., Pei, M., Erdmann, M., Ito, M., shirakawa, M., Hara, T., and Nishio, S.: Wikipedia Mining –Wikipedia as a Corpus for Knowledge Extraction–, in *Proceedings of Annual Wikipedia Conference (Wikimania) (Wikimania 2008)* (2008).

[60] 中山浩太郎, 原 隆浩, 西尾章治郎：自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法, 電子情報通信学会データ工学ワークショップ (DEWS 2008) 論文集 (2008).

[61] Nakayama, K., Ito, M., Hara, T., and Nishio, S.: Wikipedia Relatedness Measurement Methods and Influential Features, in *Proceedings of the IEEE International Symposium on Mining and Web (MAW 2009)*, pp. 738–743 (2009).

[62] Ollivier, Y., and Senellart, P.: Finding Related Pages Using Green Measures: An Illustration with Wikipedia, in *Proceedings of the American Association for Artificial Intelligence (AAAI 2007)*, pp. 1427–1433 (2007).

[63] Page, L., Brin, S., Motwani, R., and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, *Technical Report, Stanford Digital Library Technologies Project* (1998).

[64] Patwardhan, S., Banerjee, S., and Pedersen, T.: SenseRelate: : TargetWord-A Generalized Framework for Word Sense Disambiguation, in *Proceedings of the American Association for Artificial Intelligence (AAAI 2005)*, pp. 1692–1693 (2005).

[65] Peat, H. J., and Willett, P.: The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval Systems., *Journal of the American Society for Information Science*, Vol. 42, No. 5, pp. 378–383 (1991).

[66] Ponzetto, S. P., and Strube, M.: Deriving a Large-Scale Taxonomy from Wikipedia, in *Proceedings of the American Association for Artificial Intelligence (AAAI 2007)*, pp. 1440–1445 (2007).

[67] Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, *Journal of Artificial Intelligence Research (JAIR)*, Vol. 11, pp. 95–130 (1999).

[68] Roget, P. M. ed.: *Roget's Thesaurus of English Words and Phrases*, Longman Group Ltd. (1852).

[69] Rubenstein, H., and Goodenough, J. B.: Contextual Correlates of Synonymy, *Journal of Communications of the ACM*, Vol. 8, No. 10, pp. 627–633 (1965).

[70] Ruiz-Casado, M., Alfonseca, E., and Castells, P.: Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets, in *Proceedings of the Atlantic Web Intelligence Conference (AWIC 2005)*, pp. 380–386 (2005).

[71] Salton, G., and McGill, M.: *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company (1983).

[72] Salton, G., Wong, A., and Yang, C. S.: A Vector Space Model for Automatic Indexing, *Journal of Communications of the ACM*, Vol. 18, No. 11, pp. 613–620 (1975).

[73] Schütze, H., and Pedersen, J. O.: A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, *Journal of Information Processing and Management*, Vol. 33, No. 3, pp. 307–318 (1997).

[74] Spearman, C.: The Proof and Measurement of Association between Two Things, *The American Journal of Psychology*, Vol. 100, No. 3/4, pp. 441–471 (1987).

[75] Strube, M., and Ponzetto, S. P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, in *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*, pp. 1419–1424 (2006).

[76] Suchanek, F. M., Kasneci, G., and Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet, *Journal of Web Semantics*, Vol. 6, No. 3, pp. 203–217 (2008).

[77] Tseng, Y. H.: Automatic Thesaurus Generation for Chinese Documents, *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 13, pp. 1130–1138 (2002).

[78] Turdakov, D., and Velikhov, P.: Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation, in *Proceedings of the Spring Young Researchers Colloquium on Databases and Information Systems (SYRCoDIS 2008)*, Vol. 355 (2008).

[79] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).

[80] Vapnik, V. N.: *The Nature of Statistical Learning Theory*, Springer, 2nd edition (1999).

[81] Vapnik, V. N.: An Overview of Statistical Learning Theory, *IEEE Transactions on Neural Networks*, Vol. 10, No. 5, pp. 988–999 (1999).

[82] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G. M., and Milios, E. E.: Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web, in *Proceedings of the ACM International Workshop on Web Information and Data Management (WIDM 2005)*, pp. 10–16 (2005).

[83] Weischedel, R. M., Meteer, M., Schwartz, R. M., Ramshaw, L. A., and Palmucci, J.: Coping with Ambiguity and Unknown Words through Probabilistic Models, *Computational Linguistics*, Vol. 19, No. 2, pp. 359–382 (1993).

[84] Zipf, G. K.: *Human Behaviour and the Principle of Least Effort*, Addison-Wesley (1949).

[85] Zobel, J., and Moffat, A.: Exploring the Similarity Space, *Journal of ACM SIGIR Forum*, Vol. 32, No. 1, pp. 18–34 (1998).