The University of Osaka
Institutional Knowledge Archive

Title	Wikipediaを用いた概念間の関連度測定に関する研究
Author(s)	伊藤, 雅弘
Citation	大阪大学, 2011, 博士論文
Version Type	VoR
URL	https://hdl.handle.net/11094/1157
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

【20】

氏　　名	伊　藤　雅　弘
博士の専攻分野の名称	博　士（情報科学）
学　位　記　番　号	第　２４６６３　号
学位授与年月日	平成23年3月25日
学位授与の要件	学位規則第４条第１項該当 情報科学研究科マルチメディア工学専攻
学　位　論　文　名	Wikipediaを用いた概念間の関連度測定に関する研究
論　文　審　査　委　員	（主査） 教　授　西尾章治郎 （副査） 教　授　細田　耕　　教　授　藤原　融　　教　授　薦田　憲久 准教授　原　隆浩　　情報通信研究機構グループリーダー　木俵　豊

論　文　内　容　の　要　旨

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重要な技術の１つとして，概念間の意味的関連度（Semantic Relatedness）の測定に関する研究が行われている．例えば，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．

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，また，記事（概念）同士がハイパーリンク（リンク）で互いに参照されている．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として研究者から注目を集め，概念間関連度の測定に関する研究も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．例えば，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．

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，Wikipediaの各記事に存在するリンクの共起性を解析することによって，概念間関連度を測定する手法を提案する．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ため，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．

第３章では，Wikipediaから得られる異なる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．本手法では，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．

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．さらに，それら複数の情報を機械学習手法である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．

論文審査の結果の要旨

近年、情報爆発時代に向けて、莫大なテキストデータを処理する技術が注目を集め、その重要な技術の１つとして、テキスト中に出現する概念（語）の間の関連度を求める概念間関連度の測定に関する研究が行われている。しかし、日々出現する新しい概念や専門的な概念に対応するためには、より低い解析コストで高精度な概念間関連度の測定が必要となる。そこで、大規模なWeb事典であるWikipediaを解析データとした概念間関連度の測定が注目されている。

本論文は、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を解析データとした概念間関連度の測定手法についてまとめたものである。その主要な成果を要約すると以下の通りである。

(1) Wikipediaの各記事に存在するリンクの共起性を解析することによって、概念間関連度を測定する手法を提案している。この手法では、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ため、従来手法より計算量が大幅に少ない。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、従来手法に比べて高精度な関連度測定が行える。
(2) Wikipediaから得られる異なる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ている。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ことによって情報を補い合い、より高精度な概念間関連度の測定を行う。
(3) 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している。さらに、それら複数の情報を機械学習手法である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ている。

以上のように、本論文は、膨大なテキストデータを処理するためのWikipediaを用いた概念間関連度の測定に関する先駆的な研究として、情報科学に寄与するところが大きい。よって本論文は博士（情報科学）の学位論文として価値のあるものと認める。