語の重要性を考慮した関連語抽出

Related Term Extraction Using Term Importance

近藤 光正　　田中 明通　　内山 匡
Mitsumasa KONDO　Akimichi TANAKA　Tadasu UCHIYAMA

日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

Related term extraction in the past methods were considered only the relativity of terms. We thought that associated terms by a lot of people which have elements that is famously , topicality , and citation etc. So we propose the method considering not only relativity of terms but also term importance. Our technique calculates the importance of the term by analyzing the structure of Wikipedia using algorithm like HITS. As a result of our experiment, the effectiveness of our technique was confirmed.

1. はじめに

Web の世界の爆発的な情報増加により，目的の情報や興味のある情報を如何に発見するかという問題は近年非常に重要な問題となってきている．本研究では，あるキーワードに関連するキーワードを提示することで，そのキーワードに関連する情報の検索を可能とする関連語抽出手法について提案する．関連語の従来研究においては，入力された検索クエリに補助キーワードを自動付与し，検索内容を絞り込むために複数語で構成される検索クエリを提示する研究等があるが，本研究は，ある単一のキーワードに対して，関連する単一のキーワードを複数個提示することで，そのキーワードに関連する情報を得ることを目的とする関連語抽出である．このような関連語抽出の関連研究においては，関連用語収集問題として佐々木ら [3] は，Web の検索エンジンのヒット数から，Jaccard 係数・χ^2 統計量を用いて用語間の関連性を算出する手法を提案している．また，中山ら [4] や伊藤ら [5] は連想シソーラスの構築問題として Wikipedia のリンク構造を解析することで，キーワード間の関連度を算出する手法を提案している．

従来研究における関連語抽出手法は，キーワード間の距離，即ちキーワード間の関連性のみを考慮する手法が主流であった．しかしながら本研究で目指す関連する情報を得るための関連語とは，そのキーワードに関して人が連想しやすいキーワードであると思われる．そのため本研究では，キーワード間の関連性だけでなく，知名度や話題性等が高く，現実世界で重要であるとされるキーワードこそが，連想されやすい関連語であると考える．そこで本研究では，キーワード間の関連性だけではなく，キーワード自体が持つ固有の重要度を考慮することで，人が連想しやすい関連語の抽出を目指す．本研究では，このようなキーワード固有の重要度をキーワード固有重要度と定義する．このようなキーワードの重要度は，従来法においては出現頻度を用いて算出する場合が多かったが，出現頻度を用いた手法では必ずしも実現できていなかった．例えば tf 法では，該当文書において頻度の高いキーワード程，その文書における重要な語であるとみなされるが，日常的に高い頻度で用いられる語の場合 tf 法は適さない．また idf 法においては，多数の文書中に存在しない語は特徴的であるとみなされるが，知名度の高い有名人であったり地名等には適さない．そして $tf \cdot idf$ 法においては，これらの乗算から各指標の問題点を軽減し各手法を単独で用いるよりも一般的に良い結果が得られるが万能ではない．情報検索等で良い結果を得ている BM25 法 [2] においても同様のことが言える．そこで本研究では，複数の文書集合から構成されるコーパス内のキーワード出現頻度に基づきキーワードの重要度とするのではなく，オンライン百科事典である Wikipedia の構造に基づくキーワード固有重要度と，検索エンジンに投入された検索クエリの投入回数に基づくキーワード固有重要度の二種類を提案する．

連絡先: kondo.mitsumasa@lab.ntt.co.jp

2. 提案手法

提案手法では，キーワードの重要性を考慮した関連語抽出手法を提案する．キーワード固有重要度の算出方法として，Wikipedia の構造の解析結果から算出する手法と，検索エンジンに投入された検索クエリの投入回数から算出する手法の二種類を提案する．そして，キーワード間の関連性の算出には，佐々木らが提案した検索エンジンのヒット数から Jaccard 係数を用いて算出する手法を用い，関連語候補集合は，Wikipedia 内における参照・被参照関係から作成する．

2.1 Wikipedia からのキーワード固有重要度算出法

Wikipedia からのキーワード固有重要度の算出方法の基本的なアイディアは，Web 文書のランキングアルゴリズムを改良して Wikipedia に適用し，Web 文書のランキングが高い程，その Web 文書の見出し語は権威のあるキーワード，即ち話題性があり実世界で重要であるとされるキーワードであると考える手法である．ベースとなる Web 文書のランキングアルゴリズムは，HITS アルゴリズム [1] を用いた．HITS アルゴリズムは，すべての Web 文書を authority（コンテンツ）と hub（リンク集）の 2 つから構成されると定義する．そして，良い hub から多数リンクされる authority 程良い authority であるという仮説と，良い authority に多数リンクしている hub 程良い hub であるという仮説の 2 つを繰り返し実行することで Web 文書のランキングを行う．しかしながら，HITS アルゴリズムは Web 世界における Web 文書のリンク構造をモデルにしたアルゴリズムのため，リンク構造が非常に密な Wikipedia にそのまま適用した場合，やや難がある．そこで，本手法では Wikipedia の特徴的な構造と密なリンク構造に対応させた Wikipedia ランキングアルゴリズムを提案する．そして，本アルゴリズムから算出した authority の値による順位を本手法が提案するスコア関数に近似させ，キーワード固有重要度を算出

する．

authority 算出の改良点 1：テキスト量の考慮

Wikipedia の見出し語は，知名度が高く話題性の高い見出し語程，テキストの記述量が多い傾向がある．そこで，authority 値の算出の際に，自文書のテキスト量が多ければ多いほどその文書は重要であるといった重み $text(k)$ を考慮する．

authority 算出の改良点 2：自リンクと被リンクの比率

一般的に Wikipedia の見出し語は，有名なキーワード程，自リンクと被リンクの数が多くなっている．しかしながら，地名やジャンル名のような広い概念を持つキーワードは，引用されやすいキーワードのため，自リンク数に比べて圧倒的に被リンクの数が多い傾向がある．通常の HITS アルゴリズムは良い hub から多数リンクされている authority は良い authority であるといった仮説を用いるが，圧倒的に被リンクが多い場合においてはこれらの仮説は成り立たないと予想される．また，その一方で，最近知名度が高くなってきている新人俳優や話題語等の見出し語は，誕生してから日が浅いため引用数は少ないが自リンクは多い傾向にある．そのため少ない被リンク数においても，authority 値を高める必要がある．これらの問題を解決するために，authority 値の算出の際に，自リンクと被リンクの比率 $flink(k)/blink(k)$ を考慮する．

authority 算出の改良点 3：明らかに authority とならない見出し語の扱い

Wikipedia の見出し語には,「～年」や「～一覧」といった明らかに authority とならない見出し語が存在する．これらの見出し語は，自リンクが非常に多く，被リンクも非常に多い場合があるためノイズになりやすい．そこで，明らかに authority とならない見出し語の authority 値は常に変更しないことで，この問題に対処する．

hub 算出の改良点 1：hub の平均的なリンクの質

Wikipedia の文書には，自リンクが多数あるが，hub として質の悪い文書がある．そこで，リンク先キーワードの authority が平均的に高い hub は重要であるといった仮説に変更することで自リンクは多いが hub として質の低い文書の hub 値を下げる重み $\frac{\sum_{k'} \log(a(k')+1)}{K'}$ を考慮する．

hub 算出の改良点 2：リダイレクトの扱い

Wikipedia には見出し語の異表記を解消するために，redirect が存在する．例えば,「イチロー」には「鈴木イチロー」,「ICHIRO」の redirect がある．redirect は異表記のキーワードを一意にまとめる効果だけでなく，キーワードの被リンク構造に大きな影響をもつため，redirect キーワードを親ノードにまとめることで，異表記のキーワードの重要度を算出し，被リンクの問題も解決する．

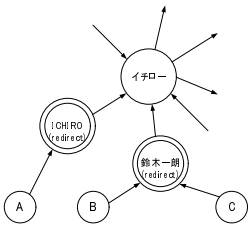

図 1: リダイレクトの例

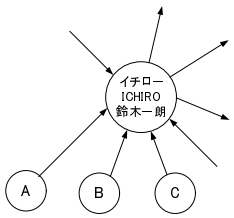

図 2: 改良後の例

Wikipedia ランキングアルゴリズム

そして，最終的な Wikipedia ランキングアルゴリズムは以下の式で定義される．

1. For all $k' \in K'$ pointing to k,

$$a(k) = \frac{\log(flink(k)+1)}{\log(blink(k)+1)} \cdot text(k) \cdot \sum_{k'} h(k') \qquad (1)$$

2. For all $k' \in K'$ pointing to by k,

$$h(k) = \frac{\sum_{k'} \log(a(k')+1)}{K'} \cdot \sum_{k'} a(k') \qquad (2)$$

以上の式を HITS アルゴリズムと同じく再帰的に繰り返す．authority の値が高いキーワード程，権威の高い内容を持つキーワードであるため，本手法では authority の値のみを用いる．そして，authority の値を降順にソートし，本研究で提案する以下のスコア関数を用いてキーワード k のキーワード固有重要度 $KIS(k)$ を算出する．このような関数を用いる理由としては，authority の値をそのまま用いると値の傾斜が激しいため，このような比較的緩やかな減衰関数を用いた．

$$KIS(k) = \exp\left(\frac{\log(y_1+1-y_0)\cdot(K-k_r+1)^a}{K^a}\right) + y_0 - 1 \qquad (3)$$

本スコア関数は，通常の exponential loss 関数と異なり，x の全体数に影響されることなく，一定の傾斜を保つ減衰曲線を描くことができ，さらに上界と下界を設定できる性質をもつ．ここで，K はキーワード k の総数，k_r は authority 値の降順にソートして算出されたキーワード k の順位，a は勾配係数，y_1 は $KIS(k)$ の上界で，y_0 は下界である．a の値が大きくなるにつれて，関数の勾配は急になる．本スコア関数の分布を図 3 に掲載する．評価実験では，$a=3$，$y_1=1$，$y_0=0.1$ の値を用いる．

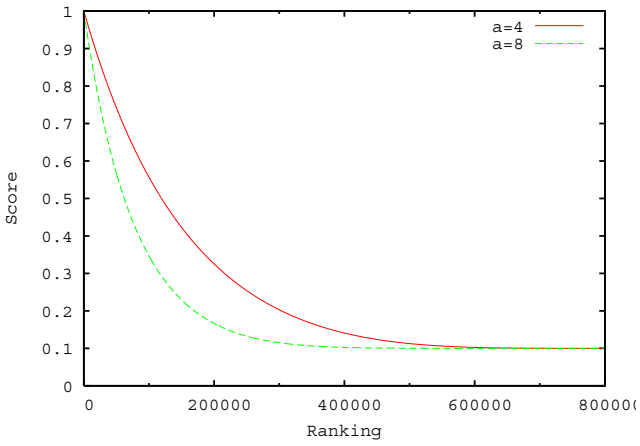

図 3: キーワード固有重要度算出で用いるスコア関数

2.2 検索エンジン上における検索クエリ投入回数からのキーワード固有重要度算出法

本節では，検索エンジン上で検索クエリ投入回数が多いキーワード程重要であると考えるキーワード固有重要度算出手法について述べる．検索エンジンの検索クエリログは，ポータルサイト goo において 2008 年 2 月から 10 月の期間の検索クエリログを用いた．検索クエリ中には，Wikipedia の見出し語として存在しないキーワードも含まれているため，Wikipedia の見出し語として存在する検索クエリのみを対象とした．そし

表 1: 評価実験で用いた手法

手法	Jaccard 係数	Wikipeida から算出した KIS	検索クエリログから算出した KIS
ベースライン	○	-	-
提案手法 1	○	○	-
提案手法 2	○	-	○
提案手法 3	○	○	○

表 2: 評価結果 (左から，○の数/△の数/×の数)

手法	上位 5 位		上位 10 位		上位 15 位		上位 20 位	
質問項目	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2
ベースライン	15/2/3	18/1/1	29/6/5	36/1/3	43/10/7	54/3/3	56/16/8	68/8/4
提案手法 1	**19/1/0**	**20/0/0**	**33/4/3**	**37/2/1**	45/11/4	**55/4/1**	57/19/4	**72/7/1**
提案手法 2	17/2/1	18/2/0	33/3/4	**37/2/1**	46/7/7	51/7/2	58/14/8	66/12/2
提案手法 3	18/1/1	19/1/0	**33/4/3**	36/3/1	**46/9/5**	54/4/2	**59/16/5**	71/8/2

て，検索クエリとして投入回数が多いキーワード順にキーワードの順位を算出し，2.1 節の 3 式で述べたキーワード固有重要度算出式を用いてキーワード固有重要度を算出する．これは，検索クエリの投入回数をそのまま用いると，Wikipedia ランキングアルゴリズムの authority 値で見られたような値の分布になるため，3 式の様なスコア関数にを用いて重要度とした方が良いためである．

2.3 最終的な関連語算出式

本手法におけるキーワード k_1 の関連語 k_2 の最終的な関連語スコアは以下の式で表される．キーワード間の関連度算出には佐々木らの実験で最も良い結果を示した Jaccard 係数を用いる[*1]．

$$R\_Term(k_1, k_2) = \frac{h(k_1 \cap k_2)}{h(k_1 \cup k_2)} \cdot KIS(k_2) \quad (4)$$

3. 評価実験

評価のベースラインとして，Jaccard 係数のみを用いた手法を用い，Wikipedia から算出したキーワード固有重要度を考慮した手法を提案手法 1 とし，検索クエリログから算出したキーワード固有重要度を考慮した手法を提案手法 2，そして上記に述べた二つのキーワード固有重要度の一対一線形和を用いた手法を提案手法 3 とする．表 1 に評価した手法一覧を掲載する．評価方法としては,「Q1. 関連する話題の検索キーワードとして適切か？」,「Q2. キーワード間の関連性は適切であるか？」の二項目によって評価を行った．本研究では，関連する話題の検索のための関連語抽出を目的としているため，キーワード間の関連性の評価 Q2 の他に，関連する話題の検索キーワードとして適切であるかの評価 Q1 を加えている．各評価項目では○適切，△どちらでもない，×不適切の 3 段階評価を行い，評価する関連語として,「直江兼続」,「バラク・オバマ」,「六本木ヒルズ」,「Linux」の 4 語をピックアップし，システム出力の上位 5 位，10 位，15 位，20 位毎に評価を行った．なお，キーワード固有重要度算出で用いるパラメータは，$y_1 = 1$，$y_0 = 0.1$，$a = 8$ を用いる．

3.1 評価結果と考察

全体の評価結果を表 2 に掲載し，各手法における「直江兼続」の関連語の上位 10 位までの出力を表 3 に掲載する．評価実験の結果，提案手法はキーワード出力の上位になればなるほど，Q1 の評価項目である関連する話題の検索キーワードとして適切である傾向があり，出力の下位になるに従ってベースライン手法と近くなる傾向がある．そして，評価項目 Q2 のキーワード間の関連性に関しては，やや提案手法が上回っているもののベースライン手法と大きな差は生じなかった．これらの理由として，Jaccard 係数のみを用いたベースライン手法は，キーワード間の関連性を正しく算出することができるが，キーワードがもつ固有の重要性を考慮していないため，Q1 の評価である関連話題の検索キーワードとしてはやや不適切な語を出力してしまう問題があるからである．表 3 の「直江兼続」の関連語結果がその様子を示している.「兼続[*2]」や「武将」といったキーワードは,「直江兼続」との関連性はあるが，関連する話題を検索するための検索キーワードとしては，やや不適切である．特に「兼続」に関しては，キーワードの共起頻度による Jaccard 係数を用いると,「兼続」の関連語は「直江兼続」以外に殆どないため，あまり重要なキーワードではないが上位に出力されてしまう．また，キーワード自体が入れ子関係になっているのも原因だと思われる．その一方で，キーワード固有重要度を用いた提案手法では，より知名度が高く話題性の高いキーワードが上位に出力される結果になっている．検索クエリログから算出したキーワード固有重要度を用いた提案手法 2 では，ドラマ名等の時代の流行的な話題に関するキーワードが上位に出力される傾向があり，Wikipedia から算出したキーワード固有重要度を用いた提案手法 1 では，時代の流行性だけに囚われない重要なキーワードが上位に出力される傾向がある．検索クエリログを用いた提案手法 2 では,「直江兼続」の重要な関連語である「上杉景勝」は，一般的な知名度があまり高くないため他の手法と比べ低い順位に位置しているが,「上杉景勝」は歴史上では比較的重要な人物である．提案手法 1 では,「上杉景勝」は上位に出力されているため，WIkipedia から算出したキーワード固有重要度は流行や一般的な知名度に捕らわれない重要度判定をしているといえるだろう．Wikipedia は各分野の準専門家達が構築した百科事典であり，そのような構造を解析することで得られた結果は，準専門家達の考えるキーワードの重要度として算出されるため，このような比較的正しいキーワードの重要度の判定がなされたのではないかと予想される．そして，双方のキーワード固有重要度の一対一線形和を用いた提案手法 3 では，それらの中間的な出力結果が得られた．線形和を用いた手法は，双方のキーワード固有重要度を考慮することで，重要でかつ検索クエリとして適切なキーワードが出力されることを期待していたが，検索クエリから算出したキー

*1 ちなみに，χ^2 統計量を用いた実験も行ったが，佐々木らの報告にもあるように，χ^2 統計量は低頻度語に高いスコアを与える傾向が見られ，Jaccard 係数を用いた手法の方が良い結果が得られた．

*2 1995 年に山形市の天文家によって発見された小惑星の名前．直江兼続にちなんで命名された．

表 3: 「直江兼続」の関連語

順位	ベースライン	Q1	Q2	提案手法 1	Q1	Q2	提案手法 2	Q1	Q2	提案手法 3	Q1	Q2
1	兼続	×	○	上杉謙信	○	○	天地人	○	○	上杉謙信	○	○
2	天地人	○	○	天地人	○	○	上杉謙信	○	○	天地人	○	○
3	上杉謙信	○	○	上杉景勝	○	○	戦国武将	○	○	上杉景勝	○	○
4	上杉景勝	○	○	石田三成	○	○	大河ドラマ	○	○	大河ドラマ	○	○
5	武将	△	○	大河ドラマ	○	○	伊達政宗	○	○	石田三成	○	○
6	戦国武将	○	○	伊達政宗	○	○	上杉景勝	○	○	伊達政宗	○	○
7	石田三成	○	○	戦国武将	○	○	武田信玄	○	○	武田信玄	○	○
8	大河ドラマ	○	○	武田信玄	○	○	石田三成	○	○	戦国武将	○	○
9	伊達政宗	○	○	武将	△	○	兼続	×	○	武将	△	○
10	武田信玄	○	○	徳川家康	○	○	妻夫木聡	○	○	徳川家康	○	○

ワード固有重要度の精度に引っ張られる形で提案手法 1 とあまり変化が見られなかった．検索クエリを用いた手法は勾配係数 a や線形和の比率を調整することによって，精度が向上する可能性も考えられるため，今後の課題としたい．

次にキーワード固有重要度を用いた各提案手法の問題点を考察する．先ほども述べたように，キーワード固有重要度を用いた手法は，上位では適切な関連語を出力するが，下位の出力では精度が落ちていることが確認できる．ベースライン手法では上位に表示される誤りの関連語が，提案手法では下位の出力で提示されることが原因として挙げられる．この問題を解決するためには，キーワード固有重要度の勾配係数 a を高く設定することで解決されるように思われるが，勾配係数 a を高く設定しすぎると，キーワードの関連性の尺度以上にキーワード固有重要度が効きすぎてしまい，精度低下を招く恐れがある．また，提案手法 2 で用いた検索クエリを用いたキーワード固有重要度は，関連性が殆ど無いサイト名や日常生活的に高い頻度で使用される生活的検索クエリが出力されるため，精度を下げる要因となった．

4. まとめと今後の予定

本研究では，キーワード間の関連性だけでなく，キーワードがもつ固有の重要性を考慮することで，関連性がありかつ関連話題として連想しやすい関連語抽出手法の提案を行った．実験の結果，Wikipedia の構造から算出したキーワード固有重要度を用いた手法はシステム出力の上位に多くの正解が見られ，さらに話題性が高く重要なキーワードが上位に位置付けられる結果を得ることができた．一般的な Web サイト上で用いられる関連話題の検索のための関連語は，少なくて 3 個，多くても 10 個あたりであるため，上位出力で比較的正しい関連語を提示できる本提案手法は有効であると考えられる．

ちなみに，本研究の成果は，2008 年 9 月末から 2009 年 3 月末まで goo ラボ*3 上にて公開実験を行った，「MyBoom サービス」[6] の機能の一部として使用した．「MyBoom サービス」では，ユーザの PC 上の Web 閲覧履歴からそのユーザの興味キーワードを抽出し，PC や携帯電話上で推薦することで，普段は検索にまで至らないが自分の興味のあるキーワードを，様々な検索システム上で検索できるシステムである．携帯電話では PC と異なり文字入力の手間がかかる問題があるが，興味キーワードの関連語を提示することで，文字入力することなく自分の興味に関連するキーワードに関する情報を探し出せる機能として実装した．図 4，5 に携帯版 MyBoom 検索画面を掲載する．

*3 http://lab.goo.ne.jp/

図 4: PC 版 MyBoom の図

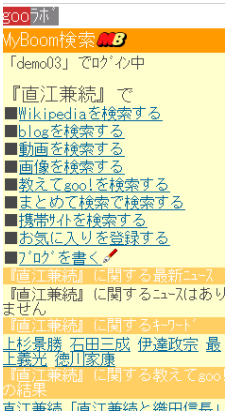

図 5: 携帯版 MyBoom の図

今後は，他の Wikipedia 研究によって構築された言語資源と連携することでよりリッチな出力をもつ関連語資源の構築や，関連語を用いた関連ニュースや動画等のアイテムを直接推薦する近傍検索的な情報推薦技術を検討する予定である．

参考文献

[1] J.Kleinberg. Authoritative sources in a hyperlinked environment. *In Proceedings 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[2] S.E.Robertson and S.walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *In Proceedings of the 17th annual international ACM SIGIR Conference (SIGIR '94)*, 1994.

[3] 佐々木靖弘, 佐藤理史, 宇津呂武仁. 関連用語収集問題とその解法. 自然言語処理, Vol. 13, No. 3, pp. 151–175, 2006.

[4] 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. SIG 11, pp. 27–36, 2007.

[5] 伊藤雅弘, 中山浩太郎, 原隆浩. 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. SIG 20, pp. 39–49, 2007.

[6] 近藤光正, 森田哲之, 田中明通, 内山匡. PC 上の Web 閲覧履歴からのクエリ抽出技術を用いたモバイル情報検索システム. 第 22 回人工知能学会全国大会, 2008.