E-033

ウェブを利用した関連用語収集

Automatic Collection of Related Terms from the Web

小原 恭介† 山田 剛一† 絹川 博之† 中川 裕志‡
Kyosuke Kohara Koichi Yamada Hiroshi Kinukawa Hiroshi Nakagawa

1. はじめに

ある用語の関連用語は、検索質問拡張などの自然言語処理において利用されることがあり、これを自動的に獲得できるシステムが実現できれば利用価値は高い。そこで本研究では、言語資源として Web を利用した、関連用語の収集手法の構築を行なう。関連研究として佐藤、佐々木らの研究[1][2]がある。佐藤らは、語の Web 検索ヒット数を利用した共起統計による語の関連度の指標を提案し、8 割程度の精度で関連用語を得ている。

本稿では、Web 検索結果の類似性に着目した語の関連度計算手法と、それを利用した関連用語の収集システムを提案する。

2. 語の関連度を測る手法

本節では、ある用語 s と w が与えられた時に、Web を利用して s と w の関連度 *R(s,w)* を測る手法について述べる。

我々は、「s」、「w」それぞれをクエリとしたときに検索されるページ集合が類似しているほど、s と w の関連度が高いと仮定した。そこで、この検索結果の類似性に着目した関連度 *R(s,w)* を以下の手順によって測る。

(1) 「s」「w」をクエリとして、検索エンジンで、それぞれ N_1 ページの検索結果を得る。ここで、N_1 は可変なパラメータである。

(2) (1)で得られたページ群を、それぞれ茶筌[3]で形態素解析し、名詞、未知語のみ抽出する。ここで得られる語集合を *T* とする。

$T=\{t_1, t_2, t_3, \ldots, t_n\}$

T: 単語集合　t_i: 単語　*n*: 出現単語の異なり数

(3) s と w それぞれの検索結果に対して語集合 *T* に対応する総出現頻度ベクトル *CF(s),CF(w)* を作成する。

$CF(s)=(cs_1, cs_2, cs_3, \ldots, cs_n)$

$CF(w)=(cw_1, cw_2, cw_3, \ldots, cw_n)$

cW_i: 語「*W*」の検索結果ページ中に含まれる t_i の総出現頻度　*W*：*s* または *w*

(4) (3)で得られた頻度ベクトル間の類似度 $\sigma(CF(s),CF(w))$ を Jaccard 係数を用いて(式 1)で計算する。

$$\sigma(CF(s),CF(w))=\frac{\sum_{i=1}^{n} cs_i \cdot cw_i}{\sum_{i=1}^{n} cs_i^{\,2}+\sum_{i=1}^{n} cw_i^{\,2}-\sum_{i=1}^{n} cs_i \cdot cw_i} \quad \text{(式 1)}$$

(5) (4)で得られた $\sigma(CF(s),CF(w))$ を語 s,w の関連度 *R(s,w)* とする。

$R(s,w)=\sigma(CF(s),CF(w))$

以上の手順を図 1 に示す。

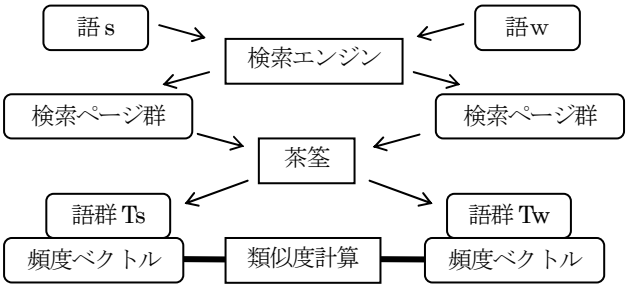

図 1．関連度計算手法

3. 関連用語収集システム

関連用語収集システムは、入力としてある用語 s を受け取り、それに関連性が深い用語群を出力する。このシステムは大きく分けると、以下の 3 ステップから構成される。

Step1: 関連ページ収集
Step2: 関連候補語抽出
Step3: 関連度計算

このシステムの処理手順を図 2 に示す。

以下の節では各ステップについて説明する。

3.1 関連ページ収集

ここでは、入力用語 s に関連するページを Web 検索エンジンを利用して収集する。今回構築したシステムでは、検索エンジンに Google[4]を利用し、「s」をクエリとした検索結果から PDF や PowerPoint ファイル等を除外した上位 N_2 ページの html のみを取得した。ここで、N_2 はシステムの入力パラメータである。

3.2 関連候補語抽出

このステップでは、収集したページ中に含まれる語から、関連候補となる語を抽出する。

まず、収集したページからタグを除去した文字列を茶筌に渡し、名詞類のみを抽出する。名詞類とは名詞や未知語、またはこれらが連接した複合名詞である。この処理で得られた名詞類が関連候補語となる。

実際に関連用語であるかどうかの判定は次のステップで行なうため、ここでは再現率の観点からなるべく多くの語を抽出したい。しかし、これらの語を全て関連候補語として認定すると関連度の計算量が増大するため、今回のシステムでは総出現頻度が 5 以上の名詞類のみを対象とした。

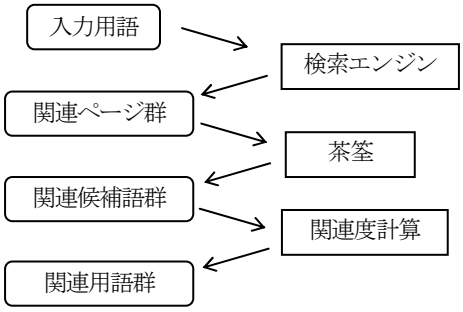

図 2．関連用語収集システム

† 東京電機大学大学院　工学研究科　情報メディア学専攻
‡ 東京大学　情報基盤センター

3.3　関連度計算

最後に、抽出した関連候補語と入力用語 s の組み合わせで、2章で提案した手法を用いて関連度を計算する。

例として、「自然言語処理」を入力用語とした時に得られる関連候補語１８０語のうち、関連度の上位３０語を表１に示す。

4.　評価実験

構築したシステムに次の５語を入力用語として与え、関連用語の取得実験を行なった。今回は、検索結果を 100 ページ利用し、関連度が 0.25 以上の語を関連用語として認定した。

入力用語：「自然言語処理」「セキュリティ」「ネットワーク」「プログラミング」「フラクタル」

4.1　評価手法

評価は、出力された関連用語群に対して実際に関連用語として正解であるかの評価を人手で行い、(式 2)で定義した精度によって行なった。実験結果は表２のようになった。

$$精度 = \frac{システムが出力した正解関連用語数}{システムが出力した関連用語数} \quad (式2)$$

表１．「自然言語処理」の関連候補語

用語	頻度	関連度	用語	頻度	関連度
言語処理	14	0.738	言語データ	6	0.336
文脈解析	22	0.567	対訳コーパス	6	0.335
音声言語処理	13	0.498	トップダウン法	9	0.319
形態素解析システム	12	0.453	ボトムアップ法	9	0.303
意味解析	36	0.425	事例ベース	6	0.300
格フレーム	6	0.399	言語理論	6	0.297
言語処理学会	13	0.388	機械翻訳	33	0.295
情報抽出	10	0.371	知識発見	6	0.293
対訳コーパス	6	0.364	形態素	13	0.288
形態素解析	55	0.357	音声言語	6	0.280
対話システム	5	0.350	処理技術	36	0.279
処理過程	10	0.342	構文解析	54	0.268
機械学習	5	0.341	研究機関	9	0.256
機械翻訳システム	27	0.340	プログラミング	6	0.254
処理学講座	11	0.337	宮崎研究室	6	0.254

表２．関連用語の収集精度

入力用語	候補語数	関連語数	正解関連語数	精度
自然言語処理	180	37	32	0.86
セキュリティ	253	85	45	0.52
ネットワーク	101	34	15	0.44
プログラミング	191	29	24	0.83
フラクタル	114	11	11	1.00

4.2　考察

(1) 提案関連度について

表１より、「格フレーム」や「機械学習」などは出現頻度が低いにも関わらず正しく関連用語として判断できていることがわかる。逆に、「コンピュータ」や「www」などの語は出現頻度は高いが関連度としては低くなった。このことから、提案した関連度は出現頻度だけでは捉えることのできない、入力用語との関連性を反映したスコアになっている。

(2) 入力用語の性質による傾向

今回実験した入力用語のうち「ネットワーク」と「セキュリティ」では、精度が比較的低かった。この原因としては、「ネットワーク」のような一般的に利用されている語や多義語などでは、検索されるページのジャンルにばらつきが多く、候補語との検索結果の類似度が全体的に低くなっていることが考えられる。これを改善するには、単に入力用語 s で検索を行うのではなく、他の語との AND 検索を行うことにより、出現するページのジャンルを絞り込む方法が考えられる。

4.3　関連研究との比較

作成したシステムの関連度計算を佐々木らの提案する指標[2]で行ったところ、上位 30 語中 22 語が関連用語であった。我々の手法では表１より、上位 30 語中 23 語が関連用語であるので、ほぼ精度は同一であることがわかる。しかし、用語ごとに関連度を見ていくと一部の語ではかなりの差がみられた。例えば、「形態素解析システム」、「格フレーム」は佐々木らの指標では関連度がかなり低い。この違いは、佐々木らの研究では語の専門用語性も判定するために、用語の表す概念の大きさを考慮にいれた指標を用いていることが影響している。つまり、用語の検索ヒット数が極端に小さいか大きい語では、関連度が小さくなる。一方我々が提案する関連度は、用語の表す概念の大きさは考慮していないため、単独の検索ヒット数が小さい語でも検索結果の類似性が高ければ、高い関連度となることに違いがある。

5.　おわりに

本稿では、提案した関連度を用いたシステムを作成し、関連用語の収集実験を行なった。実験結果では、検索されたページのジャンルにばらつきがない語の場合、適切な関連用語群を得ることができた。今後の課題としては、さらなる精度向上があげられる。考察でも述べたように、入力用語が一般語や多義語の場合、提案した手法では正確に関連用語を判定できないため、関連度計算手法の改善が必要である。また、システムとしては、他の指標と組み合わせて関連用語を選別することで、より関連語出力の精度を高められるのではないかと考えている。

参考文献

[1] 佐藤理史，佐々木靖弘: “ウェブを利用した関連用語の自動収集” 情報処理学会研究報告 NL-153-8，pp.57-64 2003

[2] 佐々木靖弘，佐藤理史，宇都呂武仁: “用語間の関連度を測る指標の提案” 言語処理学会第１０回年次大会 pp.25-28 2004

[3] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸: “形態素解析システム茶筌.” http://chasen.naist.jp/，2000

[4] Google，http://www.google.co.jp/