3N-7

語の連接情報を用いたWebからの関連文書検索

近藤　雄飛　　石塚　満

東京大学工学部電子情報学科　　東京大学情報理工学研究科

1．はじめに

近年の Web 上の文書の増加および検索エンジンの技術の発達により、何かの目的で文書を探す際、まず Web を検索するということが日常的に行われるようになってきた。我々研究者も同様であり、研究の過程である分野についてサーベイを行う際は、まず Web の検索を試みるというアプローチが一般的なっている。しかし、目的とする文書を探すための検索キーワードの選定は非常に難しく、現状では人間の直感に頼らざるを得ない。そのため、検索エンジンを用いて関連論文を検索することは時間のかかる作業となっている。本論文では、そのような検索エンジンを用いたサーベイを支援する手法を提案する。

関連文書を検索する手法としては、大竹ら[2]のように語の連接情報を用いた手法が存在している。我々は、これらの研究に習い、Web 上での語の連接情報を用いて、ある論文、あるいは論文集合に関連する文献を Web 上から検索する手法を提案する。実際のシステムでは、論文、あるいは論文集合を入力とし、関連文献を探すために適切な検索クエリーを出力とする。

本論文では、提案システムのうち、Web 上の語の連接情報を用いて、入力された論文や論文集合に特徴的な語を取り出す手法を提案する

2．専門用語抽出手法

検索クエリー抽出の対象となるのは名詞である。専門用語の辞典においても収録されている用語の大多数は名詞である。ここでは、単名詞のうちでも名詞、そして複合名詞だけを対象とした。また、本研究では近年のコンピューターの劇的な進化と普及を考慮し、非常に多数の文書媒体が存在している Web 上の情報を用いる手法である。具体的には検索エンジンで用語を検索した際のヒット数を出現頻度情報として用いる。

Document Search from the Web using word collocation
Yuhi Kondo
Department of Information and Communication Engineering, University of Tokyo
Mitsuru Ishizuka
Graduate School of Information Science and Technology, University of Tokyo

2.1　単名詞連接のスコア

特定のコーパスを想定し、連接頻度の導出法を以下のように記す。

単名詞の連接情報とそのスコアを用いる手法は、まず

```
[BC]          hit(L1)
[ABCDE]   hit(L2)     ・・・・複合名詞
[FGBCH] hit(L3)
```

C について、左方連接頻度を求める。
hit(L1)= hit(L2)= hit(L3)
　　=（BC の検索エンジンでのヒット数）

図 1：連接頻度の導出法

$$LN(N)=\sum_{i=1}^{n} hit(L_i)$$

$$RN(N)=\sum_{i=1}^{n} hit(R_i)$$

図 2：単名詞 N の左方、右方における全単名詞頻度

本研究では単名詞の連接部分に注目し、図 1 のように長い複合名詞においても N の 1 つ隣との単名詞についてのヒット数を $hit(L_i)$ として全ての単名詞、複合名詞についておこなった。

また、単名詞のスコア付けとしては図 2 のようにスコア付けを行った。

2.2　複合名詞への拡張

単名詞のスコア付けは上記のように定義できた。しかし、専門用語においては複合名詞が大多数であるため、専門用語についてもスコア付けを行う必要がある。ここでも、専門用語としての複合名詞の重要度は長さに依存しないという考え方を進めていく。

まず、単名詞 $N_1, N_2, N_3, \ldots\ldots N_L$ がこの順に連接した複合名詞を CN とする。CN のスコアとし

て各単名詞相乗平均をとることでスコアとしていく。よって、複合名詞：CN のスコアを以下のように定める。

$$LR(CN) = (\prod_{i=1}^{L}(LN(N_i)+1)(RN(N_i)+1))^{\frac{1}{2L}}$$

図 3：複合名詞CN のスコア

図 2,3 の$LR(CN)$は単名詞、複合名詞の出現頻度における情報を考慮しなかった。そこで、候補語独立のヒット数を $f(N)$ として、CN を補正する。また、一般的に Web 上の検索エンジンにおいては、ヒット数は膨大な数に及ぶ場合が多いため、Log をとって、最終的な$FCR(CN)$を定義する。

$$FCR(CN) = \log(f(CN)\times LR(CN)+1)$$

図 4：$FCR(CN)$ の定義

また、本手法では Web 上の検索エンジンにおいてのヒット数を用いるため、単名詞においては膨大なヒット数にとなる。それにより相対的に複合名詞のヒット数が少なくなる傾向にある。よって、その差を失くすためにヒット数を Jaccard 係数として計算した手法も用いた。以下に Jaccard 係数の導出法を記す。

複合名詞W_1W_2があるとすると、

$$hit_J(W_1W_2) = \frac{hit(W_1W_2)}{hit(W_1)+hit(W_2)-hit(W_1W_2)}$$

図 5：Jaccard 係数によるヒット数

3．専門用語抽出

3.1 実験方法

ある対象文書集合に対して、複合名詞を抽出する。切り出した複合名詞それぞれに対し、２語または３語と分け、上記の手法で求めたスコアを比較する。その後、複合名詞の中でスコアが最大となるような切り出し語を抽出する。

正解とする単語データをあらかじめ準備しておき、ヒット数をそのまま用いた手法と Jaccard 係数を用いた手法それぞれに対し、切り出し語が正解にあるかどうかを実験した。

実験において、文書集合は Entrez PubMed*1 から検索した医学系論文集合を用いた。また、出力データの妥当性を判別するために、医学用語集 Lexicon*2 を用いた。

3.2 実験結果

ヒット数をそのまま用いた手法を手法 A とし、Jaccard 係数を用いた手法を手法 B として実験結果を比較する。

	切り出し語	正解数	精度(%)
手法 A	461	37	8.02603
手法 B	259	23	8.880309

図 6：実験結果

あまり良い精度は出ていないが、これは元々の正解データの語数が少なかったため、本来妥当である語も正解と認識されなかったためである。

実際は本手法において下の図 7 のような本来妥当である語も抽出されている。

percussion injury	intracellular protein
breast carcinoma	cancer diagnosis
tracheal chain	respiratory disease

図 7：抽出語（不正解）

4．おわりに

今回の実験では、あまり良い精度がでなかった。しかし、関連文献検索においては実際に使う専門語自体は抽出されている。よって、抽出された専門用語を複数用いることで論文検索が可能となるシステムを構築し、精度を上げていくことが今後の課題である。

参考文献

[1] 湯本紘彰、森辰則、中川裕志：『出現頻度と連接頻度に基づく専門用語抽出』、自然言語処理、Vol.10 No.1, pp. 27 - 45, 2003

[2] 大竹清敬, 増山 繁, 山本 和英：『名詞の連接情報を用いた関連文書検索手法』, 情報処理学会論文誌, pp.2460-2467, 1999

[3] Violeta Seretan, Luka Nerima, Eric Wehrli：『Using the Web as a Corpus for the Syntactic-Based Collocation Identification 』, In Proceedings of International Conference on Language Resources and Evaluation (LREC 2004), pages 1871-1874, Lisbon, Portugal,2004

[4] 長尾真(編):『自然言語処理』,岩波書店,1996

[5] 『Google』, http://www.google.co.jp/

*1 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

*2 http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lexicon/current/web/release/