SmallWorld による類似文書検索のための重要語選定

竹元 勇太 , 沢井 康孝 , 山本 和英
長岡技術科学大学 電気系
E-mail:{takemoto,sawai,ykaz}@nlp.nagaokaut.ac.jp

1 はじめに

インターネットの普及に伴い、WWW 上には大量の電子化文書が存在する。しかし、ユーザーが大量の情報の中から必要な情報だけを探すのは困難である。そこでユーザーが必要な情報にアクセスする手間を省く技術が求められている。その技術の一つとして類似文書検索がある。WWW 上から類似文書を検索する方法の 1 つは Google や Yahoo!などの既存の検索エンジンを使用することである。しかし検索エンジンを使うには検索対象に合ったクエリを入力することが必要となる。そこで、本稿では類似文書検索を行うためのクエリ作成に文書内の重要語を抽出することに注目した。類似文書検索における重要語とは、その文書を構成する上で必要となる語と考える。松尾ら[1]の提唱した重要語抽出法では、文書内の語を共起情報を基にグラフ(SmallWorld)化している。そしてグラフを構成する上で重要な語は、グラフを構成する基となった文書を構成する上でも重要な語であるとしている。そこで本研究では、松尾らの重要語抽出法をベースに、類似文書検索において重要となる単語を文書内から選定し検索する手法を提案する。

2 既存研究

検索クエリを使用する類似文書検索の既存研究には次のようなものがある。野口[2]は文書内全体で出現頻度上位30%の複合語を頻出語群とした。χ^2 検定を用いて頻出語群と文内共起のしやすさが大きく逸脱する単語を重要として計算している。そして、重要度の高い 5 単語を検索クエリとした。検索結果の文書と入力文書の類似度を測定し、類似度の高い順に並び替えた。これにより PageRank 方式では上位に表示されない類似ページも上位に持ってこれることを示した。しかし、クエリの個数や検索の方法を改良することが今後の課題となっている。

高木ら[3]は特許文書に対しての類似文書検索を行っている。文書内の各主題要素ごとに類似文書を検索している。これにより幅広い類似文書の検索を可能にした。

吉田ら[4]は専門分野の文書を対象に類似文書検索の研究を行っている。専門分野で特徴的な単語（未知語や複合語など）を使用することで良い成果が得られることを示している。しかし、高木らや吉田らの手法は専門分野の文書の特徴を利用しているため、Web ページなどの一般的な文書に対してそのまま適用することは難しいと考える。

3 重要語抽出法

本研究では、SmallWorld という考え方に基づいて文書内の重要語を抽出する。以下の節で SmallWorld について述べる。

3.1 文書内における SmallWorld

SmallWorld とはノードがクラスタ化されているにも関わらず、任意の 2 点間の最短パスが短いグラフのことである。文書から得られた共起グラフもこのような構造をしており、松尾らはこの SmallWorld 構造に対する貢献の高い語をキーワードとして抽出している。つまり、松尾らの抽出法では「著者の主張を表す単語」を抽出することはできる。しかしそれは、検索における網羅性や特定性という視点から考えると類似文書検索では適切な抽出法ではないと考えられる。そこで本研究では、日本語文書から共起グラフ(SmallWorld)を作成し、類似文書検索に有効な単語の抽出法を提案する。

4 提案手法

本稿で提案する手法の概要を図 4-1 に示す。クエリ作成部、類似文書検索部、類似度計算部の 3 ステップで説明を行う。

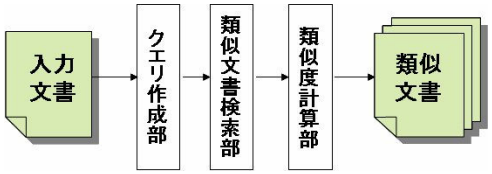

図 4-1　システムの流れ

4.1 クエリの作成部

クエリ作成部では図 4-2 の流れで入力文書の重要語を抽出し、検索クエリを作成する。

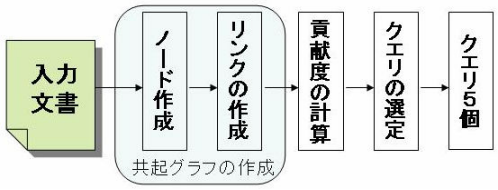

図 4-2　クエリ作成部の流れ

4.1.1 ノードの作成

文書中から検索クエリの候補となる語を抽出する。抽出する語は特定の品詞が連続した形態素(複合名詞)とした。品詞の解析には形態素解析器 ChaSen 1)を使用した。特定の品詞とは ipadic の品詞である。以下にその品詞を示す。

名詞-一般，名詞-接尾-一般，名詞-サ変接続， 名詞-形容動詞語幹，名詞-接尾-助数詞， 固有名詞-一般，固有名詞-人名，固有名詞-組織， 固有名詞-地域，記号-アルファベット，未知語

また、1 文字の形態素は対象外としている。ノードとする要素は、文書中から作成する全ての複合名詞の中で出現頻度が f_0 回以上のものとする。

4.1.2 リンクの作成

リンクの長さを計算する。リンクは関連性を表すものして考えるため、語同士の関連性を計る指標が必要である。*Jaccard* 係数は語同士の関連性を計る指標であるが、他にも共起頻度や相互情報量などの方法がある[5]。だが、共起頻度を使用した場合は出現頻度の高い語からリンクが多く張られ、相互情報量を使用した場合には出現回数が大きく、なおかつ共起頻度が高い語に対して大きな値が出やすくなる。これと比較して *Jaccard* 係数は出現頻度に関係なく共起度を算出することができることから適切な指標であると考え、本手法ではノード対 (a,b) の共起度(リンクの長さ)には *Jaccard* 係数を用いた。

$$Jaccard(a,b)=\frac{a\text{と}b\text{両方の単語を含む文の 数}}{a\text{または}b\text{の単語を含む文の数}} \quad (1)$$

入力文書から共起情報を基に *Jaccard* 係数の計算を行う。

Jaccard 係数は 0～1 の値を持ち、1 に近いほど共起しやすいことを表す。リンクの長さは *Jaccard* 係数の逆数で表し、共起しやすいものほど値が小さくなる。これは、リンクはノード間の近さを表すことが望ましいと考えたからである。そして、ひとつのノードにつきリンクの長さが小さい他のノードに対して最大 k_0 個までリンクを張る。

このようにして SmallWorld 構造を持った共起グラフを形成する。このようにして作成された共起グラフは、関連の強い単語の集まりがいくつかできると考えられる。

4.1.3 貢献度の計算

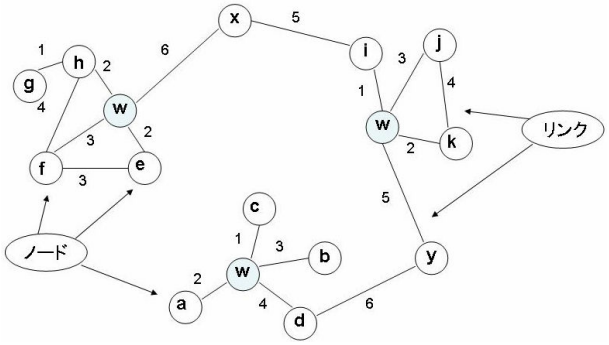

図 4-3 SmallWorld の例

検索では網羅性と特定性が重要となってくる。関連の強い単語の集まりというのは、クラスタが形成されているということである。網羅性の観点から言えば多くの異なるクラスタから単語を抽出することができれば良いということである。また、特定性の観点から言えばそのクラスタの中でもっとも重要な語を抽出できれば良いということである。このことから、本手法では図 4-3 のノード w のような単語を抽出できるような重要語抽出法を提案する。ちなみに、松尾らの手法ではノード x と y のような位置に存在する単語を重要語として抽出している。

グラフを構成する上で単語 w の貢献度 $CB_C(w)$ を計算し、貢献度 $CB_C(w)$ が高い単語 w を重要とした。貢献度とは、グラフ中の1単語が他の単語に対してどれほど関連性が高いかを表すものである。そこで、ノード a 、ノード b に対する拡張リンクの長さ $d'(a,b)$ を次のように定義する。

$$d'(a,b)\begin{cases} d(a,b) & a,b\text{間に直接リンクがある 場合} \\ w_{sum} & \text{それ以外の場合} \end{cases} \quad (2)$$

$d(a,b)$ はノード a とノード b 間のリンクの長さを表す。w_{sum} は1つのノードに対してリンクの長さが最大値のものを全てのノードの組で合計したものである。よって、w_{sum} は定数であり、図 4-3 の場合だと 64 である。

単語 w の貢献度を計算する時は、単語 w 以外の 2 単語 (a,b) 間に単語 w が存在する場合 $L'(w)$ と存在しない場合 $L'_G(w)$ を考える。$L'(w)$ は w ノード以外の全てのノードの組に対する $d'(a,b)$ の平均である。$L'_G(w)$ はノード w を取り除いたグラフにおける $d'(a,b)$ の平均である。ノード w を取り除くということは、リンクもなくなるということである。ここでノードを場所、リンクを道、リンクの値を距離として考える。すると、ノード a からノード b に移動しようする場合ノード w を介するリンク(道)がないために遠回り、または移動ができないことになる。このように w が存在する場合としない場合での移動距離が大きいほど、w の貢献度が大きいと言える。このことから、$L'_G(w)$ と $L'(w)$ の差を計算することによって単語 w の貢献度 $CB_C(w)$ を算出し、それを全てのノードに対して行う。

$$CB_C(w) = L'_G(w) - L'(w) \quad (3)$$

松尾らの抽出法では、直接リンクが張られていないノード対に対しても他のノードを介しての最短パスを計算する。このような計算を行うと、クラスタ同士を繋ぐようなノードに対して大きな貢献度 $CB(w)$ がつくようになる。本手法では、直接リンクが張られているノードにのみ注目することで、各クラスタ内から貢献度 $CB_C(w)$ の高いノードを選定することができる。また、最短パスを計算する必要がないため松尾らの手法に比べて大幅に処理時間を短縮することができる。

4.2 類似文書の検索

類似文書の検索はクエリを含む文書を抽出することで行う。以下に検索の流れを示す。

1. 5 個のクエリを含む記事を検索。
2. 検索ヒット数が 20 記事以下なら 3 へ、以上なら 5 へ。
3. 前回の検索よりクエリの個数を 1 個減らした時の組合せを考える。そして、組合せの分だけ検索をする。
4. 検索ヒット数の異なり数が 20 記事以上なら 5 へ、以下なら 3 へ。
5. 検索終了。

検索した文書数はクエリの組合せによって得られた文書の異なり数である。

4.3 類似度による並び替え

類似文書の検索によって得た類似文書候補それぞれについて入力文書との類似度を算出する。そして、類似度の高い順に並び替える。類似度計算には比較的優れた性能があり、また実装しやすいコサイン類似度 $\sigma(dx,dy)$ を採用する。

$$\sigma(dx,dy) = \frac{\sum_{i=1}^{T} x_i \times y_i}{\sqrt{\sum_{i=1}^{T} x_i \times \sum_{i=1}^{T} y_i}} \quad (4)$$

x_i, y_i はそれぞれ文書 dx, dy の複合名詞 i に対する 2 値変数(0,1)を表しており、T は入力文書及び類似記事候補中に含まれる全ての複合名詞である。

5 評価

5.1 実験データ

類似文書検索を実装し文書を入力する。本研究では 2 種類のデータに対して評価を行うことにした。実験データの詳細を以下に示す。

・共起辞書作成コーパス
　90 年～03 年 日本経済新聞
・テストデータ
　言語処理学会年次大会発表論文集 ：10 文書
　2004 年 日本経済新聞 ：30 記事
・検索記事候補
　2000～2004 年 日本経済新聞 ：193,992 記事

システムの評価は以下の部分について行った。

・クエリ作成部
・類似文書検索部

それぞれの評価において本手法と以下の手法を比較した。以下の手法は全て単語の重要度を計算するためのものであ

る。

・CB*IDF

$$CB(w)\times idf(w) \tag{5}$$

・TF*IDF

$$tf(w)\times idf(w) \tag{6}$$

・χ2*IDF

$$\left(\sum_{g\in G}\frac{\left(freq(w,g)-n_w P_g\right)}{n_w P_g}\right)\times idf(w) \tag{7}$$

$tf(w)$は入力文書中の単語 w の出現頻度，gは入力文書中で、単語の出現頻度上位30%に含まれる語，G はgの集合。$freq(w,g)$は語 w と語gの入力文書中の文内共起頻度，n_wは語 w が出現する文に含まれる語の異なり数，P_g は(語gが出現する文に含まれる語数の合計)/(文書全体に含まれる語数の合計)である。

5.2 クエリ作成部の評価

重要語抽出の精度を評価するために、入力記事から検索に必要と思われる複合名詞を全て人手で選択した。選択した単語を正解データとし、システムの出力(5 単語)に正解がいくつ含まれているかで評価を行った。図5-1と図5-2にその結果を示す。図 5-1 の縦軸は、論文 10 文書分で平均した 1 文書あたりの正解単語数である。図 5-2 の縦軸は、新聞 30 記事分で平均した 1 記事あたりの正解単語数である。

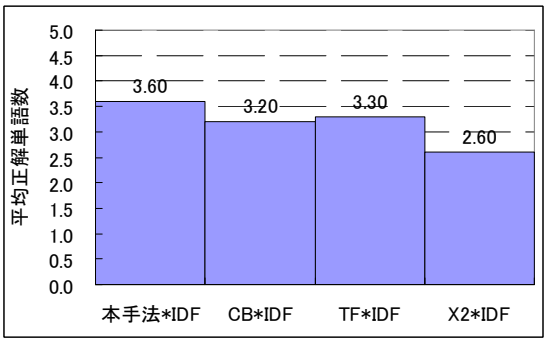

図 5-1 入力を論文とした場合の単語抽出精度の評価結果

図 5-1 の結果から、入力が論文の場合は本手法*IDF* が最も重要語抽出の精度が高く、*CB*IDF* よりも高い重要語抽出精度を実現することができた。

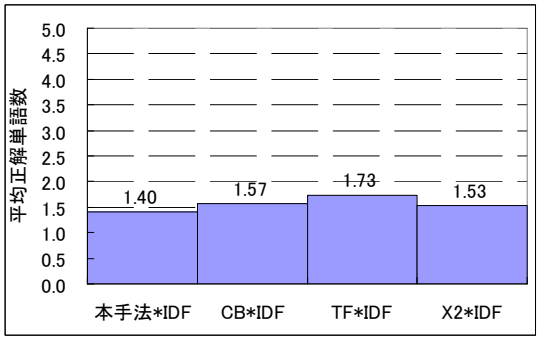

図 5-2 入力を新聞記事とした場合の
単語抽出精度の評価結果

図 5-2 から、入力文書からの重要語の抽出精度は *TF*IDF* が最も高く、本手法**IDF* が最も低いという結果になった。

このように入力データの違いによって精度に違いが出た理由の一つとしては、入力したデータの大きさに原因があると思われる。評価データとして使用した 10 本の論文と新聞 30 記事の平均文数と平均単語数を表 5-1 に示す。

表 5-1 平均文数と平均単語数

	平均文数	平均単語数
論文	319.0	975.1
新聞記事	9.2	53.4

表 5-1 から、論文の平均文数は新聞記事の約 34.7 倍、論文の平均単語数は新聞記事の約 18.3 倍であった。

松尾らの手法では、文書からグラフ構造を正確に取り出すために、文書の長さがある程度必要という欠点があるとされていた。本手法もグラフ構造の作成部分が松尾らの手法と同じであるため、この問題は発生すると考えられる。そこで、グラフ構造を補うことができれば精度向上が望めると考え、本手法に対して共起辞書を使用した。

使用方法は、まず入力文書からノードとなる単語を抽出する。それらのノード対(w_1,w_2)の共起頻度を共起辞書と入力文書から計算し、それぞれ *Jaccard* 係数を求める。それぞれの値は最大値で割り、正規化をおこなう。そして、それらを加算平均したものをノード対(w_1,w_2)の共起度とする。図 5-3 に共起辞書を使用した場合の評価結果を示す。

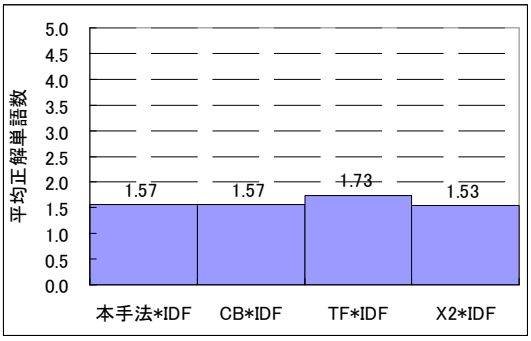

図 5-3 共起辞書を使用した場合の
単語抽出精度の評価結果

共起辞書を使用したが *TF*IDF* を超えることはできなかった。しかし、共起辞書を使用することによって本手法の精度を 0.17 上げることができた。これは共起辞書によってグラフ構造を補うことができたからである。これ以上精度を上げるためには、概念辞書などの他のグラフ構造を使用する方法も今後の課題として考えておく。

また、図 5-1 の場合は 4 つの抽出法の平均正解単語数は 3.18 だったのに対して、図 5-3 の場合には 4 つの抽出法の平均正解単語数は 1.56 であった。この結果を見ると、全体的に 1.62 も精度が落ちているため論文より新聞記事の方が問題として難しいといえる。

5.3 類似文書検索部の評価

新聞記事を入力とした場合の類似文書検索の評価を行った。本手法のみ共起辞書を使用した場合の評価を行っている。検索によって得た類似記事候補が入力文書と類似しているか評価をするため、各システムが出した類似記事候補上位 50 記事の中からランダムに 10 記事抽出した。ここで上位 50 記事とは「コサイン類似度の大きい順」ではなく、「クエリが含まれている数の多い記事の順」から50 記事のことである。これは、クエリの良し悪しに関わらず、コサイン類似度の性能によって類似文

書検索の評価結果が良くなる可能性をなくすためである。

記事の評価方法は、評価者が入力記事と類似するために必要だと考える単語の含有率を基準に、以下のように 4 段階で評価した。

1：0%～25%含まれている
2：25%～50%含まれている
3：50%～75%含まれている
4：75%以上含まれている

4 段階の評価値を 30 記事で平均し、各手法の精度とした。その結果を図 5-4 に示す。

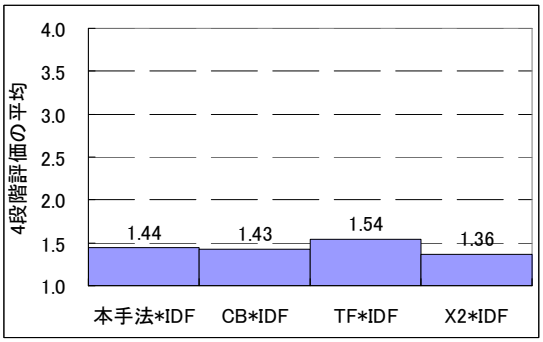

図 5-4 類似文書検索の評価

各手法で作成したクエリを使用して類似文書検索を行った結果、図 5-4 のように *TF*IDF* が最も良い精度となった。逆に最も悪い結果となったのは $\chi^2 * IDF$ であった。図 5-3 と図 5-4 の評価結果は同じ傾向が現れているため、人が検索で重要だと考える単語は、実際の検索においても重要であるということがわかった。

6 考察

本手法では、文書から共起グラフを生成し、グラフが SmallWorld 構造を持つことを利用した重要語の選定を行った。

文書から生成した共起グラフが SmallWorld 構造を作成できているかどうかを判断する基準のひとつに L と C という指標がある。L は全てのノード対の最短パスの平均、C はあるノードに直接リンクしているノード同士にリンクが張られている割合を全ノードについて平均したものである。また、リンクの張り方を *Jaccaard* 係数が大きいものからリンクを張るのではなくて、ランダムに張った場合の L と C を計算する。この場合の L と C を Lrand と Crand と呼ぶ。SmallWord 構造ができている場合、L>=Lrand、C>>Crand の式が成り立つとされている。実際に入力が論文と新聞記事ではその式が成り立つのかどうかを確認したところ、それぞれ文書数で平均した値は表 6-1 のようになった。

表 6-1 L と Lrand、C と Crand の平均

	L	Lrand	C	Crand
論文	2.31	2.18	0.37	0.26
新聞記事	2.00	1.96	0.77	0.48

表 6-1 を見る限りでは新聞記事でも SmallWorld 構造を作成できている可能性がある。確かに入力を新聞記事にした場合(図 5-2)、松尾らの手法 *CB*IDF* は *TF*IDF* に次いで良い結果を出している。もしグラフ構造を正確に取り出すことができているとすれば、問題は本手法の重要度計算方法に問題があるということになる。

本手法及び松尾らの手法には、2 つのパラメータ (f_0, k_0) がある。f_0 は単語の出現回数を表し、k_0 は 1 つのノードが張るリンクの最大値を表している。今回評価で使用したものは、入力が論文の場合は $(f_0 = 3, k_0 = 7)$、新聞記事の場合は $(f_0 = 1, k_0 = 4)$ である。この組み合わせが今回の評価で最も精度が高くなったパラメータ値である。論文では単語の出現回数が 3 回以上のものをノードとしているのに対して、新聞記事では全体的に出現頻度が低いため、同じように出現回数が 3 回以上のものをノードとした場合、重要語を取りこぼす可能性が高くなる。

また松尾らの研究ではリンクの長さを考慮していなかった。本手法ではリンクの長さを *Jaccard* 係数の逆数で示し、ノード間の近さ(共起のしやすさ)を表した。これにより、入力が論文の場合、リンクの長さを考慮しないければ平均正解単語数が 2.50、考慮した場合は図 5-1 のように 3.60 となっており、評価値を 1.1 上げることができた。

7 おわりに

本稿では、文書から生成した共起グラフ(SmallWorld 構造)を利用した類似文書検索のための重要語の選定を行った。その結果、入力が論文のように文書が比較的長いものに対しては高い精度を出すことができ、本手法の基となった手法より高い精度を得られた。しかし、入力が新聞記事のように文書の長さが比較的短いものに対しては成果を発揮することができなかった。改善方法として、共起辞書を用いるなど共起情報の補足を行い、評価値を 0.17 上げることができた。短い文書でもグラフ構造を補うことで入力を論文とした時と同様の成果が得られる可能性がある。

参考文献

[1] 松尾 豊,大澤 幸生,石塚 満:Small World 構造に基づく文書からのキーワード抽出,情報処理学会 論文誌,Vol43,No.6,pp.1825-1833,2002.

[2] 野口 光孝:文書内の重要語を検索クエリとする類似・関連 Web サイト自動検索システム,道都大学紀要 経営学部,第 5 号,pp.9-18,2006.

[3] 高木 徹,藤井 敦,石川 徹也:検索質問文書の主題分析に基づく類似文書検索,情報処理学会 研究報告,FI75-11,pp.91-98,2004.

[4] 吉田鑑地, 中村明, 川尻博光, 松本忠博, 池田尚志:専門分野の文書を対象としたキーワード抽出・類似文書検索と評価実験,言語処理学会 第 12 回年次大会,pp.660-663,2006.

[5] 徳永 健伸 .情報検索と言語処理 (言語と計算) .東京大学出版会 . p.234 .1999

[6] Watts, Duncan J. Small Worlds The Dynamics of Networks between Order and Randomness. Princeton studies in complexity. Princeton, N.J.: Princeton University Press, 1999.

[7] Milgram, S. The small world problem. Psychology Today,Vol.2,pp.60-67. 1967.

使用した言語資源及びツール

1) 形態素解析器 ChaSen, Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

2) NIKKEI, 日本経済新聞社, http://www.nikkei.co.jp/