

複数関連文書間の読書支援のための類似度計算手法

田中 俊一 岡村 潤 森 辰則 中川 裕志

横浜国立大学 工学部 電子情報工学科

E-Mail: {tanashun,jun,mori}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

1 はじめに

情報検索は利用者が興味ある文書を読み始める良い端緒を与えるが、文書を読み始めた後における読書支援としては別の形態の検索システムが望まれる。例えば、利用者がある文書を読んでいたとしよう。その際に、現在読んでいる箇所について、関連する情報を補足説明として得たいことが多々ある。しかし、文書単位の検索システムでは一つの文書を選び出してくれはするものの、その中で実際に必要とする箇所までは推薦してくれない。また、読んでいる箇所に関する検索要求の構成という煩わしさを利用者に強要しなければならない。これに適するのは文書中の小単位に注目し、その単位で検索をしたり、あるいは前もって関連づけをしておくシステムである。現在読んでいる箇所に関連のある事柄だけを補足説明として読みたい場合には、このシステムの検索精度を十分に向上させ、なるべく他の文書を読まなくても済むようにできれば、元の文書に集中できるため利用者にとって有用な読書支援になる。特に 1 回の検索においてごく少数の補足説明を読むとすれば、再現率-適合率曲線において低再現率域の適合率を向上させることが重要である。

我々はマニュアルなどの文書に対する読者支援について取り組んでいるが、従来の語単位でのリンク生成に対して、文書小部分（セグメント）単位でのリンク生成を自動的に行い、ハイパーテキスト化して出力するシステムをこれまでに提案している [大森 97]。このシステムではセグメント中の語を抽出し、*tf-idf* 法、ベクトル空間モデルに基づくセグメント間の類似度計算を行なっている。

セグメント間での類似度計算には、セグメント中の語句の他に、大域的な情報、すなわち、セグメントが埋め込まれている文書とそのセグメントの間の関係を利用ができる。これが通常の情報検索との相違点であり、これを類似度計算に反映させることにより、精度向上が期待される。

我々はすでにこの種の情報として語彙連鎖を利用することを提案し、改善効果が得られることを確認している。また、セグメント内の局所情報として語の共起情報も有効であることを示している [岡村 98b]。

本稿ではさらなる精度向上を目的として、セグメント間の類似度計算に語の共起情報と語彙連鎖の効果を組み合わせる手法を提案し、それぞれ単独で用いた場合よりも精度が向上することを報告する。

2 複数文書間関連づけシステム

我々の提案する複数文書の自動関連づけシステムは、図 1 に示す 3 つのサブシステムからなっており、セグメントを単位とする *tf-idf* 法、ならびにベクトル空間モデルを基本としている。

我々の提案するシステムは、意味的にまとまりのある文書小単位を 1 つの単位としている。ここでは「章」「節」といった文書に既存の論理構造を単位とした。これをセグメントと呼ぶ。

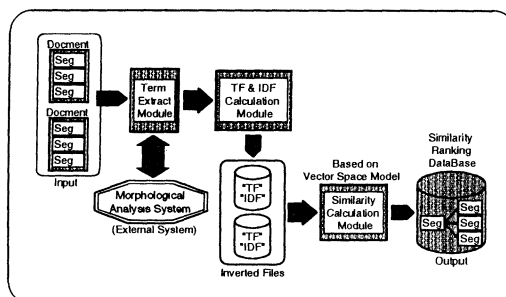


図 1: 複数文書間関連づけシステムの構成

3 語の関連を考慮した類似度計算の改善

文書単位での関連づけの場合と異なり、セグメント間の対応づけでは、セグメント内の情報だけでなくその回りのセグメント群との関係を類似度計算に反映させることができるので計算精度向上が期待できる。

我々は、類似度計算に反映させるものとして、語の関連に注目した。語の関連には一文内に生じるような局所的なものから、セグメントをまたがる大域的な話題の流れまで様々な範囲のものがある。我々はこの視点から、より高精度なセグメント対応を得るために、局所的な情報である語の共起情報と大域的な情報である語彙連鎖をセグメント間の類似度計算に反映させることを提案している。

語の共起情報は文内レベルでの語の共起を考慮する手法である。語彙連鎖の利用は、セグメント間に跨る語の出現を類似度計算に反映させる方法であり、隣接するセグメント間における語の関連を考慮していることになる。

3.1 節、3.2 節に述べる手法により、それぞれの情報を利用した類似度計算の方が *tf-idf* 法の方のみの計算よ

り精度が向上することは [HTNJ98, 岡村 98a] に述べられている。これらを組み合わせることにより両者の効果を重ねあわせ、さらに精度が向上すると期待されるので、その手法について本章で提案する。

3.1 共起情報の利用

高木らは、検索要求と文書中に同様に現れる共起単語対の重要度を増すことにより、検索精度が向上することを報告している [高木 96]。この考え方は我々の類似度計算に利用できる。すなわち、単語の共起情報によってセグメント内の単語頻度 tf を補正して類似度計算を行う。

3.2 語彙連鎖の利用

一つの文書中には、セグメント単位での $tf \cdot idf$ 法では検出されない、複数のセグメントにまたがって出現する重要な語がたびたび登場する。このような語をとらえることができれば、セグメント間の対応付けの精度向上に有用であると考えられる。我々は、この効果を類似度計算に反映させるために語彙連鎖の存在による補正を提案している。

語彙連鎖とは、文章中で語彙結束関係にある語のまとまりのことである。一般に言う語彙結束性とは、語の意味的なつながりのことであり、この意味での語彙連鎖は概念辞書等、外部知識を用いて計算される [Gre96]。しかし関連の深い文書群においては同じ語の連鎖を捉えるだけでも十分効果が得られると期待できる。複合名詞等、複数の構成素からなる語は反復する頻度が少ないので、その構成素における語彙連鎖を考える。そこで、本稿での語彙連鎖は文書中で同じ語が連続して出現している部分を指すものとする。

語彙連鎖を用い、セグメントを越えて出現する語を捉えることにより、複数のセグメントに渡る重要語を見つけることが可能となる。すなわち、ある語が語彙連鎖を形成している箇所は、その語が中心的話題となっている可能性が高い。よって語彙連鎖の形成を、あるセグメントにおける話題の中心の判定に用い、語の重要度に反映させることが出来るであろう。

語彙連鎖をセグメントの類似度計算に反映させる手法として、当該語の tf 値を補正する方法を提案している。

3.3 共起情報と語彙連鎖情報の統合

ここでは、上記二つの手法を組み合わせることによって両者を統合した改善を試みる。

共起情報と語彙連鎖は異なった性質の語の関連情報であるため独立した改善効果があると期待される。よって、各補正の効果を線形和により結合する。結合方法はマクロ結合手法と、ミクロ結合手法の2つを提案する。

3.3.1 マクロ結合手法

共起情報を用いる手法と語彙連鎖を用いる手法を独立に用いて、それぞれセグメント間類似度を先に計算し、その結果の類似度を結合させる方法である。すなわち、あるセグメントの組合せ (d_A, d_B) について、共起手法によって求められる類似度 $S_c(d_A, d_B)$ と、語彙連鎖手法による類似度 $S_l(d_A, d_B)$ を式 (1) で結合させることによって、新たな類似度 $S_t(d_A, d_B)$ を得る。

$$S_t(d_A, d_B) = S_c(d_A, d_B) + W_m \cdot S_l(d_A, d_B) \quad (1)$$

ここで、 W_m は共起手法と語彙連鎖手法のどちらに重きをおくかを決定するマクロ結合重みパラメタである。この値が大きいほど語彙連鎖の類似度を重視することになる。今回はマクロ結合重みパラメタ W_m については1として計算した。これは両手法による類似度の重みを等しくした場合である。実際には、両手法とも類似度はベクトル空間法によって求められるので、両手法を統合した類似度 $S_t(d_A, d_B)$ は、cosine 値の線形和となる。

3.3.2 ミクロ結合手法

共起情報による補正も語彙連鎖による補正も tf 値を変化させることであることに注目すると、 tf 値の補正の段階で両者を結合することが考えられる。セグメント d_i における語 t_s の tf 値を補正するとすれば、以下の式 (2) のようになる。

$$tf'(d_i, t_s) = tf(d_i, t_s) + W_r \times ((1 - W_c) \times rev_{coc}(d_i, t_s) + W_c \times rev_{lex}(d_i, t_s)) \quad (2)$$

(但し、 $0 \leq W_c \leq 1$)

ここで、 $rev_{coc}(d_i, t_s)$ は共起手法による補正值であり、 $rev_{lex}(d_i, t_s)$ は語彙連鎖手法による補正值である。 W_r は、 $tf \cdot idf$ 手法に対してどれだけ補正手法を重く見るかを決定する補正重みパラメタであり、 W_c は共起手法、語彙連鎖手法のどちらに重きをおくかというバランスを決定する結合バランスパラメタである。今回は補正重みパラメタ W_r 、結合バランスパラメタ W_c については $W_r = 2$ 、 $W_c = 0.5$ として計算した。これは、各補正值の tf に対する比率を1にするものである。

4 各類似度計算手法の評価

ここでは、

1. 共起情報を考慮した手法
2. 語彙連鎖を考慮した手法
3. 共起情報と語彙連鎖を組み合わせた手法

という三つの手法について、セグメントの全ての組合せにおける類似度計算の結果、順位づけられるセグメント対応を適合率－再現率曲線にて評価した。

また、再現率－適合率曲線の評価において、我々が提案する手法の優位性を示すために、統計的検定を行った。すなわち、ある2手法によって得られる2つの再現率－適合率値集合についてその値に差があるかどうかを検定した。今回は2つの手法の比較はある同じ再現率における適合率の比較をもって行なう。再現率－適合率の分布が正規分布とは断定できないので、ノンパラメトリック検定であるウィルコクソンの符号順位検定 [Hu193] を行った。この手法は対応のある2群の代表値の差を見るためのノンパラメトリック検定方法である。

4.1 実験

本システムは様々な種類の文書に適用可能であるが、分冊化されていることが多く複数文書を同時に読むことが一般的であるという理由により、我々はマニュアルを実例として実験を行なっている。大規模マニュアルについては、我々は既に実験を行っており、そこである程度のシステムの有効性を確認している [大森 97]。しかし大規模マニュアルにおいては完全な正解集合を人手で作るのが困難であるため、各手法の効果を考察するには、完全な正解集合を作成することができる小規模のマニュアルで実験を行う必要がある。そこで我々は、同一メーカーの3つのビデオのマニュアル A[三菱電 a], B[三菱電 b], C[三菱電 c] を用いて実験を行なった。ここでは、紙面の関係で A,B の組における結果のみを示す。各マニュアルの諸元は次の通りである。

マニュアル	A	B	C
セグメント数	33	28	32
大きさ (Kbyte)	112	102	88

A,B 両マニュアルで本システムによる関連づけを行ない、セグメント対応の評価を行なった。セグメントの全組合せ数は 924, そのうち正解である組合せの数は 65 であった。

各手法の平均適合率の比較を表 1, 2に、適合率－再現率の関係を図 2, 3に示す。表 2より、マイクロ統合手法のほうがマクロ統合手法よりも良い平均適合率が得られていることが分かる。図 2によればマクロ統合手法は、共起手法と語彙連鎖手法の中間の精度を持つ。一方マイクロ統合手法は図 3ならびに表 1, 2によると他のどの手法よりもよい結果を生んでいる。しかし、いずれの手法においても高再現率域ではほぼ同じ性能に見えるので、統計検定によりマイクロ統合手法の優位性を検証する。すなわちウィルコクソンの符号順位検定を有意水準 5% でおこなったところ、共起手法ならびに語彙連鎖手法に対して、マイクロ統合手法が勝っているという結果が得られた。

表 1: 共起手法と語彙連鎖手法の比較 (面積から導出した平均適合率)

組合せ	共起手法	語彙連鎖手法
$A \Leftrightarrow B$	0.478	0.552

表 2: マクロ統合手法とマイクロ統合手法の比較 (面積から導出した平均適合率)

組合せ	マクロ統合手法	マイクロ統合手法
$A \Leftrightarrow B$	0.523	0.565

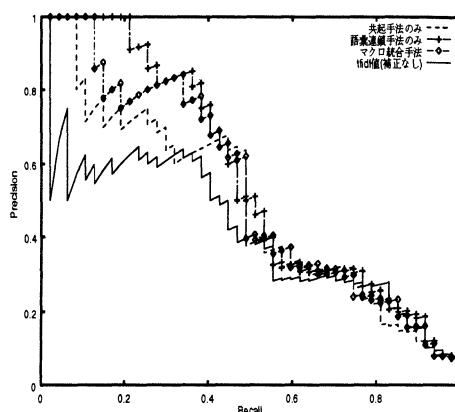


図 2: 適合率－再現率曲線 ($A \Leftrightarrow B$)

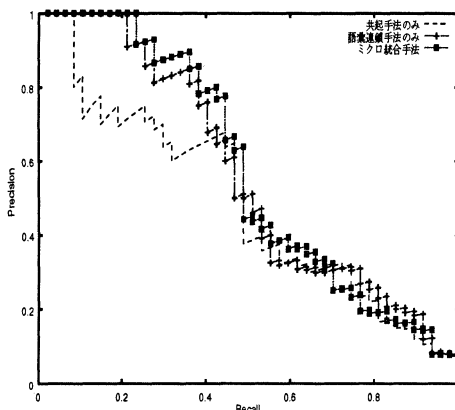


図 3: 適合率－再現率曲線 ($A \Leftrightarrow B$)

4.2 考察

図3において共起情報と語彙連鎖を用いた手法を比較すると、語彙連鎖の方が共起情報よりも再現率の低～中域において適合率が向上しており、より高精度なセグメント対応を得ていることがわかる。また、語彙連鎖とマイクロ統合手法を比較した場合、マイクロ統合手法のほうが低再現率域を中心に適合率が向上している。高再現率域においては語彙連鎖の方が若干勝っているため、類似度の低い組については語彙連鎖のみの補正を行なうということも考えられる。

マクロ統合手法では共起手法と語彙連鎖手法でそれぞれ計算された類似度の平均をとる手法のため、両手法に性能差がある今回の場合には性能の悪い手法、すなわち共起手法の影響を受け、総合的な性能向上にはならないことが確認された。一方、マイクロ統合手法では共起手法と語彙連鎖手法を直接 tf 値の補正に反映させるため、両者の効果を類似度計算に含めることが出来た。

読書支援を目的とする情報検索システムにおいては、システムは通常、検索結果を類似度の順番で提示するが、その際には、類似度の上位に正解が集まっていることが望ましい。今回得られた結果では、この点についてマイクロ統合手法の有用性が見られた。

5 おわりに

本稿では、複数マニュアルの自動関連づけシステムについて述べ、その中のセグメント間の類似度計算について、標準的な $tf \cdot idf$ 手法を補正する二つの手法である共起情報と語彙連鎖について述べた。また、その二つの手法を統合する実験を行ってその有効性を示し、さらに、ウィルコクソンの符号順位検定により有用性を検証した。

今後の課題としては、実際の読書支援の場面で本手法の有効性を検証することや、共起情報による補正と語彙連鎖による補正を結合させる際の重みパラメータについて、関連づけを行う文書によっては補正の比率を変更したほうがよい結果が得られる可能性もあるため、パラメータの調整をすることなどが考えられる。

参考文献

- [Gre96] Stephen J. Green. Using lexical chains to build hypertext links in newspaper articles. In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases, Portland, Oregon*, 1996.
- [HTNJ98] H.Nakagawa, T.Mori, N.Omori, and J.Okamura. Hypertext authoring for linking relevant segments of related in-

struction manuals. In *Proceedings of COLING-ACL '98*, pp. 929–933, 1998.

- [Hul93] David Hull. Using statistical testing in the evaluation of retrieval. In *Proceedings of SIGIR '93: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338, 1993.
- [大森 97] 大森信行, 蔵方隆宏, 岡村潤, 森辰則, 中川裕志. 情報検索手法を用いた複数文書間の関連箇所抽出 – 電子化マニュアルへの適用 –. 言語処理学会第3回年次大会, pp. 257–260, 1997.
- [岡村 98a] 岡村潤, 大森信行, 山口登志実, 森辰則, 中川裕志. 共起情報を考慮した $tf \cdot idf$ 法に基づく関連文書間の自動ハイパーテキスト化. 言語処理学会第4回年次大会, pp. 560–563, 1998.
- [岡村 98b] 岡村潤, 田中俊一, 森辰則, 中川裕志. 複数マニュアルの自動ハイパーテキスト化における類似度計算手法について. 自然言語処理研究会報告 98-NL-127, 情報処理学会, 1998.
- [高木 96] 高木徹, 木谷強. 単語共起関係を用いた文書重要度付与の検討. 情報学基礎研究会報告 96-FI-41-8, 情報処理学会, 1996.
- [三菱電 a] 三菱電機株式会社. 三菱ビデオ HV-BZ66 取扱説明書.
- [三菱電 b] 三菱電機株式会社. 三菱ビデオ HV-F93 取扱説明書.
- [三菱電 c] 三菱電機株式会社. 三菱ビデオ HV-FZ62 取扱説明書.