情報検索手法を用いた複数文書間の関連箇所抽出法 —電子化マニュアルへの適用—

大森信行　蔵方隆宏　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

1 はじめに

近年，コンピュータに代表される電子機器，システムは飛躍的な発展を遂げ，より高度かつ複雑な処理が可能となった．これに伴いユーザも高度な知識が要求され，機器を使いこなすためには莫大な量のマニュアルを読む必要性が生じてきた．従来の紙面によるマニュアルでは説明が固定的であり，ユーザはそれぞれが必要な知識，概念の記述された項目を目次や索引で探し，読みすすめていかねばならない．また，役割に応じて複数の冊子に分かれているマニュアルも多く，逐次，参照する箇所を探していくことは容易ではない．これを助けるものとして，最近では，Microsoft Windows の Help 機能などに見られるような，語句をマウス等で指定することにより他の関連テキストを表示することができるハイパーテキストが活用されつつある．しかし，現在のところハイパーテキストの作成にはあらかじめ人間が手作業でリンク付けを行う必要があり，大規模マニュアルにおいてこの処理を行うには困難を極める．

本稿では複数のマニュアル間において，ハイパーテキストにおける参照関係であるリンクを自動的に生成するシステムを提案する．関連マニュアル間においては個々の語句に対する説明箇所の他に，一連の操作手続きなどあるまとまった文書単位での対応関係も重要である．例えば，チュートリアルにおいて例示されている操作について，それと対応する詳細記述をリファレンスマニュアルで調べる場合などが想定される．そこで本稿では図1に示すような特に節や項などあるまとまった文書単位（図1中では seg 1 など）同士が対応するハイパーテキスト化を考える．

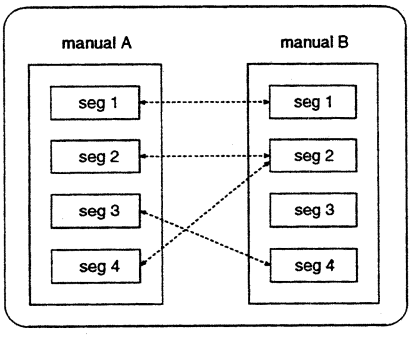

図1: システムが生成するハイパーテキストの概念図

2 関連研究

WWW の発展に伴い自動ハイパーテキスト生成は，近年注目をあびている分野である．その基礎には，重要語抽出研究，情報検索研究などが関連している．一般に自動ハイパーテキスト生成には，次の2つの課題がある．

1. いかにリンクを張るべき対象を決定，抽出するか
2. 抽出した対象をどのように関連付けるか

1. については，索引語や説明箇所の抽出について研究が行われてきた．特に前者については，重要語抽出研究として盛んに進められてきた．例えば中川らは，マニュアルにおける索引語の多くが複合語であるという事実に着目し重要語抽出を行っている [中川 96]．

2. については，シソーラス（概念間の上位下位関係）による意味的類似度を用いた手法などによって関連付けを行っている．

上記の点を踏まえて以下に，自動ハイパーテキスト生成に関連する研究について述べる．

黒橋らは，専門用語辞典を対象にハイパーテキスト生成を行った．リンクを張るべき対象は，あらかじめ与えられている索引語と，語句を定義する際の言い回しパターンをもとにテキストから抽出した語である．そして，同義語関係などから作成したシソーラスや，カテゴリ分類（人手も加わる）を用いてリンクを生成している [黒橋 92]．

また黒橋らは，文書中の重要説明箇所の特定についても研究を行っている．ここでは語に対して重要な，あるいは関連性の高い説明を使用する際，必然的にその語を繰り返し用いる必要がある，と仮定している．そして，語のテキスト中での出現密度分布を調べ，高密度な出現位置を取り出すことによって，その語の重要説明箇所を特定している [黒橋 96]．

雨宮らは，重要語抽出によるマニュアルのハイパーテキスト化を行った．ここでは，黒橋らと同様に語句を定義する際の言い回しをもとにマニュアル中の定義語を抽出し，これをキーとして文章中の参照部分と，定義部分のリンクを生成している [雨宮 96]．

以上のように，従来のハイパーテキスト生成研究では主に重要語，定義語，キーワードの説明箇所に対しリンクを生成していた．つまり，索引をたどっていく過程を電子化したものがほとんどであった．これに対し本システムでは，2つのマニュアル間において各セグメント同士のリンクを生成することにより，あるセグメント全体が表す意味内容に類似した箇所を参照することが可能となる．後に述べるように本システムは情報検索の手法を応用したものであり，非常に長い検

索要求文で文書部分を検索するのと等価になり高い精度で対応関係を見つけ出せると期待される．また，範囲を限定しない文書の関連部分を結びつける場合には，語の多義性ならびに同概念の異表記の問題がある．しかし本稿においては，同カテゴリーのマニュアルを用いることを前提としている．このため，同じ単語は同語義を表し，また同じ概念は同じ表記の語により指し示されると仮定できる．よって，シソーラスを用いなくとも精度良く対応箇所を見つけられると期待される．

3 自動ハイパーテキスト生成システム

3.1 マニュアルのハイパーテキスト化

最近の多くの機能をもつ機器やシステムでは，すべての機能を習得すること自体困難であり，適時必要な機能のみを使用すればこと足りる場合がほとんどである．このような製品ではユーザのレベルや目的によって使いわけができるように，次のような複数のマニュアルに分かれている場合が多い．

- 初心者向けチュートリアル
- リファレンスマニュアル
- 拡張，応用機能マニュアル，操作早見表

これらのうち，リファレンスマニュアルではその機器の使用法がすべてにわたって記述されているので，それ以外のマニュアルを読み進める過程でリファレンスマニュアルを参照することが多い．例えばチュートリアルマニュアルはリファレンス中の基本部分を説明したものであるから，チュートリアルマニュアルの記述は，リファレンスマニュアルの記述内容に包含されている．

そこで，我々の手法の評価実験においては，マニュアルのハイパーテキスト化のうちで最も有効と思われる，リファレンスマニュアル・その他の関連マニュアル間の対応付けに注目し，特にチュートリアルマニュアルとリファレンスマニュアル間の関連部分を自動的に対応づけることを試みた．もちろん，この手法は，他の関連マニュアルにも適用可能である．

3.2 自動ハイパーテキスト生成

本システムでは，2で述べた自動ハイパーテキスト生成における2つの課題について次のように考えている．

1. 対象は，あるまとまった文書単位(セグメント)であり，2つのマニュアル中の全文をセグメント単位に区切り，その全てを候補と考える．セグメントの単位としては，文字列の長さに基づいて機械的に区切ったものも考えられるが，ここでは意味的なまとまりを考慮し，節，項とする．

2. 関連付けについては，まず両マニュアルからそれぞれ任意の候補を選び，内容的な類似度のスコア付けを行い，値が高い組み合わせについてリンクを生成する．

1.については，HTMLなど構造をもつ記述形式になっていれば，文書構造からセグメントを認識できるため容易に自動化できる．2.については，類似度のスコア付けが問題となる．この類似度のスコア付けに情報検索で広く用いられている，$tf \cdot idf$ 法に基づくベクトル空間モデルを応用する．

情報検索は，検索要求文を満たす文書をデータベースから引き出す．このとき，検索要求文ならびに文書中の語に重みを付けた部分照合手法を用いることで，語に対する重要度を考慮している．語の重要度をスコア付けする方法としては，$tf \cdot idf$ 法が広く用いられている．さらに検索式と検索結果の適合度を示すスコアは，ベクトル空間モデルを用いることで求められる．

$tf \cdot idf$ 法 $tf(d,t)$ は，ある語 t がある文書 d 中に現れる頻度である．$idf(t)$ は，文書データベース全体においてある語 t が現れる文書の頻度に基づく値であり，次式で定義される．

$$idf(t) = \log \frac{\text{データベース中の文書数}}{\text{語 t が現れる文書数}} + 1$$

$idf(t)$ はある語 t が一部の文書に集中している度合を表しているので，$tf \cdot idf(d,t)$ はある語 t がある文書 d を弁別する能力を表している．

ベクトル空間モデル ベクトル空間モデルは，文書や検索質問を多次元空間上のベクトルとして表現し，二つのベクトルを比較することにより類似度を調べるものである．ベクトルの各次元には各検索語を，各成分には重みを割り当てる．つまり，検索式中の語の数が次元を決定する．データベース中の単語については $tf \cdot idf$ 法でのスコアを重みとする．検索語の重みについては例えば全て1としてベクトルを生成する．

ここでは，ベクトルがより同じ方向を指す文書が類似度が高いと仮定されている．

よって，検索質問と文書の類似度は2つのベクトルのなす角度によって決められ，一般的には両ベクトルのcosine値により求められる．

検索エンジンでは検索要求文と文書の類似度を計算するが，この方法を拡張するとテキストどうしの適合度を調べることができる．

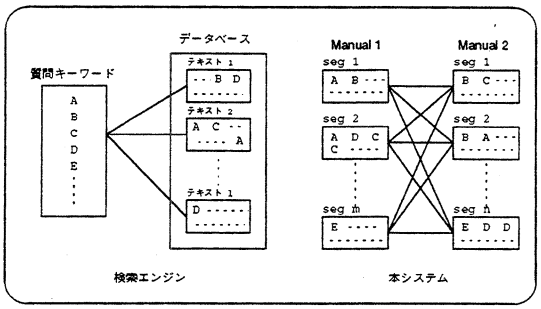

図 2: 類似度計算の比較

図2に検索エンジンと本システムの類似度計算の違いについて示す．検索エンジンでは，一つの検索質問につきデータベース中の各文書に対して重要度の順位

付けをおこなっているが，本システムではセグメントの全組合せについて類似度を求め順位付けを行い，一定基準を満たす類似度のものについてリンクを生成する．

大規模マニュアルに対してテキスト間の対応を調べる場合，キーワード数，組み合わせ数の多さゆえに計算量が大きくなることが想定される．しかし，検索エンジンのようにオンライン処理を要求されるわけではなく，本システムではオフラインで一度テキストの対応をとりリンクを生成すればよい．ここで，マニュアルの対応付けで使用するキーワードについて考える．そもそもユーザは，次のような場合に他の項目を参照することが多い．

・わからない専門用語が出現した場合
・マニュアル中のある箇所の説明だけでは操作が理解できない場合
・ある項目から派生する操作を知りたい場合

つまり用語の説明，操作説明などが参照対象となり得る．よってここでは，これらの説明の骨格をなす名詞と動詞をキーワードとして類似度計算に用いる．

4 システムの概要

本システムの入出力は，次の通りである．

入力 電子化されたマニュアル (plaintext,LaTeX,HTML)
出力 ハイパーテキスト化されたマニュアル (HTML)

なお，入力がプレーンテキスト (タグにより構造の示されていない文書) の場合，セグメントの認識ができないため，別のツールを用いてタグ付き文書に変換した後，本システムの入力とする．また現在のところ，出力はHTML形式でありこれを表示できるブラウザを用いることを前提としている．

本システムは，4つのサブシステムより構成されている．システム構成を図3に示す．

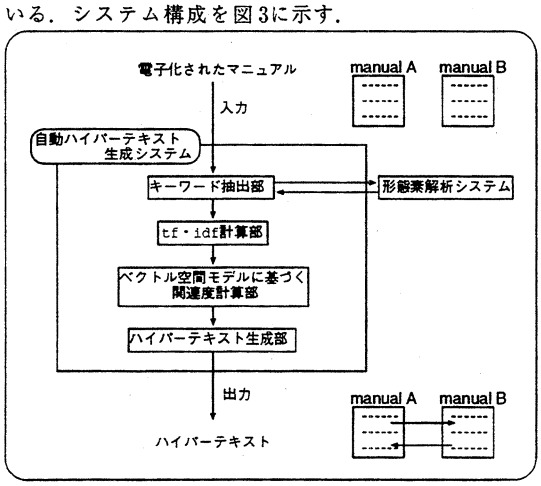

図 3: 自動ハイパーリンク生成システムの構成

キーワード抽出部 形態素解析システムを用いてテキストを単語単位に分解し，キーワードとなる語についてセグメント毎にカウントする．形態素解析システムには茶筅 1.0b4 を使用した．

$tf \cdot idf$ 計算部 カウントされたキーワードをもとに $tf \cdot idf$ 値を計算する．

ベクトル空間モデルに基づく関連度計算部 重み付けされたキーワードをもとに，各セグメント毎のベクトルを作成し，すべての組み合わせに対して *cosine* 値を求める．

ハイパーテキスト生成部 *cosine* 値の高い組み合わせに対しリンクを作成する．

ある実用ソフトウェア (APPGALLERY [日立製作所]) のチュートリアルマニュアルとリファレンスマニュアルの間で，自動ハイパーテキスト化を行なった結果を図4に示す．

画面をフレームで4分割し，左上に「オンラインヘルプ」，右上に「チュートリアル」がそれぞれ表示される．左下，右下には，それぞれのセグメントのリンク先が表示されており，いずれかをクリックすることにより，参照先がそれぞれのフレーム上部分に再表示される．その後も同様にリンク先をたどっていくことができる．

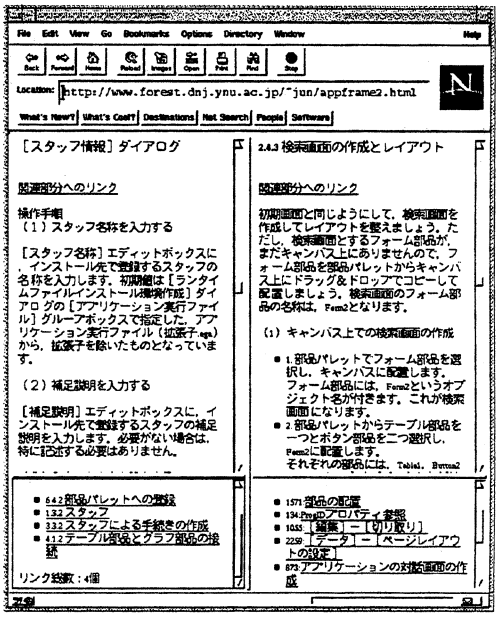

図 4: システムの利用画面

5 システム検証

5.1 評価法

情報検索で，一般的利用される再現率 (recall)，適合率 (precision) を用いてシステムの性能評価を行う．

$$\text{再現率}(recall) = \frac{\text{検索された適合対応数}}{\text{全ての適合対応数}}$$

$$\text{適合率}(precision) = \frac{\text{検索された適合対応数}}{\text{検索された対応数}}$$

再現率はある順位までに出現する正解の割合，適合率はノイズの割合をそれぞれ示す．

5.2 検証

大規模マニュアルにおいて，人手で対応関係の完全な正解を作成することは非常に困難である．例えば，APPGALLERY [日立製作所] では，チュートリアルのセグメント数65，ヘルプマニュアルに至ってはセグメント数2479であり，対応の組合せは161135通りである．人間がこの対応すべてを調べることは困難であるため，ここでは我々の手法により順位付けられた対応関係のうち上位200位までを調査して正解の分布を調べた．正解がより上位に分布していることが示されれば，本方式の有効性が近似的ながらも示されると考える．

ここでは，正解を次のように定めた．

1. チュートリアルとヘルプマニュアルで，同じ操作をしている部分がある．
2. 一方が抽象的な概念の説明であり，もう一方が具体的な操作方法の説明である．

図5に本方式で計算された対応付けの再現率，適合率を示す．順位づけされた対応の上位部分のみを対象にしているため，上位200位までに含まれる正解を近似的な正解集合と考え，上位から横軸が示す順位までを取り出した時の再現率，適合率を示している．

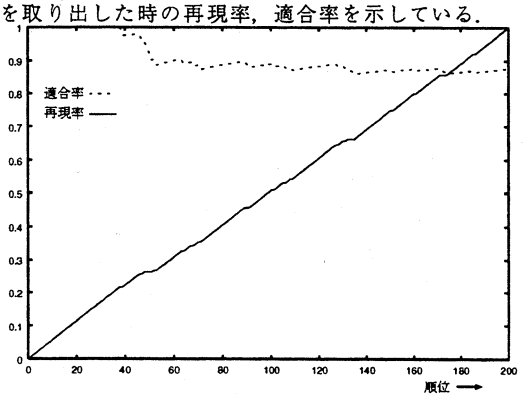

図 5: 評価実験における再現率と適合率

5.3 考察

正解集合が上位にあり，ノイズの少ない理想に近いグラフになった．

以上から，近似的ながら大規模なマニュアルに適用した場合のシステムの正当性が示された．ただし，上記のグラフによれば200位以内はほぼ正解だけで占められているので，さらに下位の分布も調べる必要がある．なおチュートリアルのセグメント数65よりも多くの正解が存在するのは，チュートリアルの1セグメントが，ヘルプマニュアルの複数のセグメントに対応している場合があるからである．実際の使用時には対応セグメントへのリンクを類似度の高い順に提示できるので，利用者に負担をかけることはない．

両セグメントに同じキーワードが何回か出現すると本システムでは類似度が大きいと判断される．しかし，その場合でもセグメント同士の内容の関連が大きいとは言えない場合があり，それらがノイズとなっている．これは，単語の出現分布のみによる本手法の限界である．

6 課題

システムの実用性を検証するために，一般的な大規模マニュアルに関して実験を行った．上位200位程度までを調査したところ、その中で正解がより上位に分布していることが分かった．

また，現在はセグメントをLaTeXのsubsubsection相当を最小単位として区切っているが，文書の大きさにばらつきが見られ，より大きな文書については対応するキーワードも多く，スコアが大きくなる傾向が見られる．大きな文書ではどの箇所が対応しているかが分かりづらい点も問題である．

キーワードについては名詞と動詞を抽出したものの，現段階ではそれらの関係については考慮していない．これについては，動詞の格情報をもとにマッチングを行うことを検討しており，一つの実験例では，上位の対応と下位の対応においてスコアに大きな差が生じるという結果が出ているため，より精度の高いリンクができると期待される．

謝辞 マニュアルを提供して下さった日立製作所に深く感謝致します．

参考文献

[中川 96] 中川 裕志，森 辰則，松崎 知美: 日本語マニュアル文における名詞間の連接情報を用いたハイパーテキスト化のための索引語の抽出，情報処理学会研究報告 96-NL-116-10，(1996).

[黒橋 92] 黒橋 禎夫，長尾 真，佐藤 理史，村上 雅彦: 専門用語の自動的ハイパーテキスト化の方法，人工知能学会誌，Vol.7，No.2，pp.336-345，(1992).

[黒橋 96] 黒橋 禎夫，白木 伸征，長尾 真: 出現密度分布を用いた語の重要説明箇所の特定，情報処理学会研究報告，96-NL-115-7，(1996).

[雨宮 96] 雨宮 秀文，森 辰則，中川 裕志: 重要語抽出による日本語マニュアルのハイパーテキスト化，言語処理学会 第2回 年次大会 発表論文集，pp.85-88，(1996).

[松本 96a] 松本 裕治, 黒橋 禎夫, 山地 治, 妙木 裕，長尾 真: “日本語形態素解析システム JUMAN version 3.1 使用説明書” 京都大学工学部 長尾研究室, 奈良先端科学技術大学院大学 松本研究室, 1996.

[松本 96b] 松本 裕治, 今一 修, 山下 達雄, 北内 啓, 今村 友明: “日本語形態素解析システム 茶筅 version 1.0b1 使用説明書” 奈良先端科学技術大学院大学 松本研究室, 1996.

[日立製作所] 日立製作所: 使ってみよう APPGALLERY，APPGALLERY オンラインヘルプ，日立製作所.