デジタル・ドキュメント 9－3
（1997. 9. 19）

電子化マニュアルにおける自動ハイパーテキスト化手法

森 辰則　大森 信行　内間 圭介　岡村 潤　中川 裕志
横浜国立大学 工学部 電子情報工学科

梗概

昨今，ハイパーテキスト化された電子化マニュアルがいくつか見受けられるようになってきたが，参照箇所を見つけ互いに結びつける作業をすべて人手で行なうのは大変な作業を必要とする．本稿では，現在著しく発展している情報検索技術を応用し，ハイパーテキストに必要となる相互参照情報を生成する方法を2つの観点において検討する．一つは，重要語を抽出し，それについての参照関係を自動生成する手法である．今一つは，複数の関連マニュアル間において密接に関連する部分を自動抽出する手法である．

Methods of Automatic Hypertextualization for Instruction Manuals

Tatsunori MORI, Nobuyuki OHMORI, Keisuke UCHIMA,
Jun OKAMURA and Hiroshi NAKAGAWA
Division of Electrical and Computer Engineering, Yokohama National University
E-mail: {mori,ohmori,k-suke,jun}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

Abstract

Recently, instruction manuals written in the hypertext style are becoming popular. However, it is very hard to find all appropriate reference relations and generate links among them manually. In this paper, we discuss two types of method to generate hyper-links automatically. They are based on the techniques of information retrieval. Firstly, we propose a method to make hyper-links among important words, which are extracted automatically. Secondly, we describe a method to find relevant parts among several related manuals.

1 はじめに

各種情報機器に代表されるように，高機能・高性能の機器やソフトウェアが毎日のようにメーカから提供されている．しかし，それらが十分に活用されているかというと，必ずしもそうとは言えない．必要とされない機能が単に使われないというだけではなく，その存在・使用方法を知れば有用である機能も見い出せないまま死蔵されていることが非常に多い．それらが過剰に複雑化していることも一因ではあるが，それ以上に使用方法に関する情報提供の不備が大きな要因であろう．メーカが提供する使用方法に関する情報は主にマニュアルである．各メーカは非常に多くの説明項目を分類することにより，利用者のレベルや用途に分けて複数の分冊としてマニュアルを編纂している．しかし，この大部のマニュアルを渡されてもそれをはじめから読んで理解する気にはなれない．自分の持つ知識に応じて必要となる関連情報を見つけ出し，使用法を理解することは利用者にとって大変な苦痛であるからだ．

この問題を解決する方向として2つのものが考えられる．一つは，認知心理学的な視点に立ったマニュアルの内容の質的向上である．これについては，[海保 87]などに詳しく述べられている．今一つは，マニュアル利用に対して様々な手段を提供することにより利用者の利便を図ることである．前者についても議論すべきことが多々あるが，我々は工学的な立場から後者に絞って議論を進める．

様々な切口からマニュアルを利用できるようにするにはどのようにしたらよいであろうか．我々は少なくとも次の2つの技術を融合した知的マニュアルシステムが必要であると考える．一つは「多様な利用ができるマニュアルの枠組」であり，もう一つはその枠組の下での「利用者の誘導」である．これらは，ハイパーテキストと情報ナビゲーションという観点から解決策が見い出されると考える．

昨今，World Wide Web(WWW) によるハイパーテキスト型の情報提供が注目を集めている．ハイパーテキストとは文書や図などの要素の間の相互参照情報(ハイパーリンクと呼ぶ)を持つ文書であり，このリンクを仲立ちとして関連文書間の行き来ができる．マニュアルをハイパーテキスト化することにより，ある説明箇所から別の関連箇所へ自由に読み進めることができるので，マニュアルの柔軟で多様な利用のための枠組が提供される．

さらに，ハイパーテキスト化により，情報ナビゲーションのための基礎を築くことができる．ここで述べる情報ナビゲーションとは，利用者の理解にあわせて短時間で適切な情報を得られるように，読み進めるべき方向に利用者を導くことを指し示している．ハイパーテキスト化は必要な情報に到達するための「道」を切り開くものであるから，情報ナビゲーションのための基幹情報を作成することと見なせるであろう．さらに，ハイパーテキストを可視化する WWW ブラウザ(閲覧ソフトウェア)や他の情報可視化の枠組は，利用者が自分で関連項目を認識・選択しなければならないという点において受動的ではあるが，一種の情報ナビゲー

ションのためのツールと位置付けられる．もちろん，利用者の理解度に適応する高度なシステムとするためには，理解度に応じた能動的なナビゲーションが望まれるが，これについては，本稿では述べない．

このように，マニュアルをハイパーテキスト化することにより様々な利点が生じると期待されるが，これをすべて人手で行なうのは大変な作業を必要とする．さらに複数の文書間の対応づけまで行なうのは作業量からみて困難を究める．そこで我々は，この問題に対して，自動ハイパーテキスト化技術の開発という点から接近しようと試みている．具体的には，現在著しく発展している情報検索技術を応用し，ハイパーテキストに必要となる相互参照情報を生成する方法をいくつかの観点において模索している．本稿では，我々が現在試作しているこれらシステムについて述べる．

2 マニュアルにおけるハイパーテキスト化の諸相

節1で述べたように，マニュアルは利用者のレベルや用途に応じて複数の分冊に編纂されることが多い．例えば，高機能のシステム等では，全ての事柄が網羅されているリファレンスマニュアルの他に，典型的な利用方法を記したチュートリアル(入門書)を添付することが多い．このように用途が異なる文書が存在する時，ハイパーテキスト化における相互参照情情報は次の二つに大きく分けられる．

- 同一文書内の相互参照
- 複数文書間の相互参照

同じ文書内での相互参照では，ある用語の説明箇所を，その用語が現れている別の箇所が参照する場合が多い．つまり，ある用語を使用している箇所と，その用語の定義や付随する操作手順を関連付けるという，用語に注目した相互参照である．相互参照の情報が付加された用語は紙面の本におけるいわゆる索引語に対応する．

一方，複数文書間においては，上記用語の相互参照の他に，文書内で一定の大きさを持つあるまとまり，すなわち文書小単位(セグメント)の間での対応付けも必要である．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，その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．また逆にリファレンスマニュアルに現れている個々の操作について，それらを実際に利用するとどうなるかという例示をチュートリアルから得ることあろう．これは，あるセグメントに別の文書のセグメントを関連付けるという，セグメントの内容に注目した相互参照である．

以下では，上記2つの相互参照に基づき，ハイパーテキスト化を行なうためのシステムについてそれぞれ述べる．

3 重要語抽出に基づくハイパーテキスト化ツール

用語に対するハイパーリンクの設定について，その作業の自動化に関する研究が発表されている [Gre96, 雨宮 96]．しかし，相互参照情報(ハイパーリンク)の自動生成はある程度の精度で行えるものの，現状では最終的に人手による確認をせざるを得ない．本節では，この際の負担を軽減するものとして，ハイパーリンクの自動生成とその後編集を用意に行なえるようにした，タグ付け作業の支援ツールについて述べる．本ツールではまず，ハイパーテキスト化する文書から索引語を抽出し，それをもとに機械的にリンクを生成する．その後 Web ブラウザを通じたユーザインタフェースにより文書のハイパーリンクの確認と訂正を行う．これにより複数文書間にわたるリンクを含むマニュアルのハイパーテキスト化を効率よく進めることが出来る．

3.1 システム概略

本システムにおけるハイパーテキスト化は図1に示す通り次の4ステップで行なわれる．

1. ハイパーリンクを張る語の選択する．
   オフラインで HTML 化する文書から索引語候補となる重要語を抽出する．
2. 選択された語を参照・被参照箇所に分類する．
   抽出された索引語のリストと定義箇所を推定する定義パタン(後述)により，参照・被参照のタグ付けをする．
3. (利用者が)専用エディタにより索引語にふられた参照・被参照関係を後編集(ポストエディット)する．
4. 実際にハイパーリンクを生成する．

次に各ステップの説明をする．

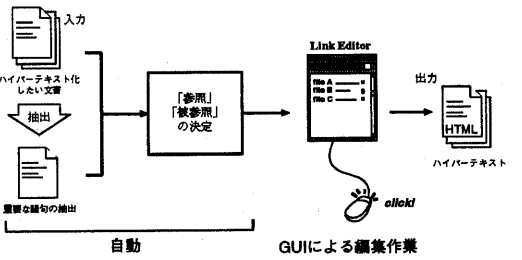

図 1: 本システムの概略

3.2 索引語の抽出

用語の相互参照を重視したハイパーテキスト化においては，ハイパーリンクを張るべき語の抽出が最大の問題となるる．本ツールでは，索引語候補を我々が既に発表している重要語抽出システム [中川 97] により選び出す．この重要語抽出手法は，マニュアル中の重要語がマニュアルの主要な概念を示す名詞を含む複合名詞であることが多いという観察に基づく．

なお，本システムでは重要語抽出アルゴリズムとは独立であり，ハイパーリンクを張るべき語のリストを入力とするために，別の手法で生成された索引語を入力とすることもできる．

3.3 ハイパーリンクの生成

抽出された索引語候補を用いれば互いに関連しあう箇所を自動的に抽出することは文字列照合のみで可能であるが，リンクの方向，すなわち，参照・被参照関係を高精度で一意に決定するのは困難である．そこで，まず参照・被参照関係を推定し，これを初期値として参照関係エディタに渡す．利用者はその関係を必要に応じて修正することにより，正しいリンクを生成できる．

3.3.1 定義パタンの使用

マニュアルでは，索引語が見出し語または定義語として使われている箇所は被参照箇所になる事が多い．そこで，「(索引語) は～である」，「(索引語) を～と呼ぶ」といった定義パタンを集め，被参照箇所の特定に使用する．パタン集は実際には Perl の正規表現に準拠したパタンを集めたファイルである．

タグ付けの際は，まず索引語を含む段落中でこのパタンに合うものを探し，見つからない場合は参照箇所とする．

3.3.2 参照箇所の間隔確保

実際に本システムのプロトタイプを利用して，得られた観察によると，おなじ語に対する参照箇所同士が近過ぎたり，被参照箇所のすぐ後ろにある参照箇所は修正段階でリンクの対象から外されることが多い．そこで，参照箇所同士および参照箇所と被参照箇所の間に一定の間隔を設けることにし，この間にある参照タグをリンクの対象から外す．実際の使用感から，デフォルトではこの間隔を 5 段落としている．この際，単に参照タグを取り外すのではなく，後で復帰可能な中間タグに置き換える．なお，被参照箇所についてはこのような間隔確保を行なわない．

3.4 Web ブラウザによるリンクエディタ

生成されたハイパーリンクのチェックは図 2に示すような Web ブラウザ上のインタフェースを用いて行なう．リンクのチェックは索引語候補ごとに行ない，チェックする索引語を選択するとタグ付けが行なわれた箇所が一文単位で表示されるようになっている．ここで，その内容を読んで，その部分を参照箇所とするか被参照箇所とするか，あるいはリンクの対象から外すかをラジオボタンの選択により決定する．ハイパーテキスト化する文書が複数ある場合でも同一ウィンドウ内で編集できる点がエディタでの作業と比較して有利である．

3.5 定義パタンの学習

現在使用している定義パタンは人手により，定義箇所になると思われるパタンを集めたものである．しかし，本ツールの使用によりユーザの編集過程を入手できるので，定義パタンの正例と負例を収集できる．これらを用いてツールの利用を通じて漸進的に定義パタンを学習することも可能である．

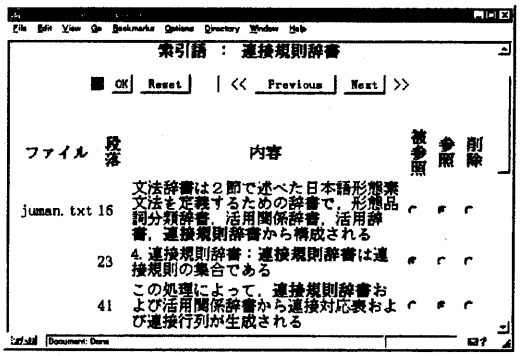

図 2: 編集画面の例

我々は本ツールにより得られた編集結果から定義パタンを学習するための予備実験を行なった．この実験では以下の情報を教師情報として，決定木獲得アルゴリズム C4.5 により，索引語の周囲の語 (文脈) からその索引語が「参照箇所」，「被参照箇所」のいずれになるかを判定するための決定木を獲得するものである．

- 索引語を含む文の情報
  文を形態素解析し，索引語の前後一定数の形態素列について，それぞれ，「形態素基本形，品詞，活用形」の組を得る．それらをならべて一つの属性ベクトルを得る．この属性ベクトルは索引語が現れた周りの文脈情報を表す．

- その文が「参照箇所」「被参照箇所」のいずれになるかの判定
  人間により行なわれた編集結果から，上記の文それぞれについて，その文に含まれる索引語が「参照箇所」であるのか「被参照箇所」であるのかの判定を得る．これは，上記属性ベクトルにより表される文脈において索引語が「参照箇所」，「被参照箇所」のいずれのクラスに分類されるかを表す．

その結果，テストに使用したマニュアルが 1 つではあるものの，交差検定 (cross-validation) による評価によれば 90% 程度の正答率で「参照」「被参照」の予測ができることが確認できた[1]．紙面の都合で実験に関する詳細については別の機会に譲る．

4 情報検索手法に基づく関連マニュアル群のハイパーテキスト化

本節では複数の関連マニュアル間において，ハイパーリンクを自動的に生成するシステムを提案する．既に述べた通り，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．そこで本節では，節や項などあるまとまった文書単位同士を対応づけるハイパーテキスト化を考える．

[1] なお，人手で調整した定義パタンにおいては，85% 程度の正答率であった．

後述す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るが，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．しかし本手法においては，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ことが期待できる．

4.1 自動ハイパーテキスト生成

本システムにおいて，我々は自動ハイパーテキスト生成について次のように考えている．

1. ハイパーリンク生成の対象は，文書小単位 (セグメント) である．2 つのマニュアル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，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，まず両マニュアルにおける任意の組合せについて類似度計算を行っておく．ハイパーリンクは利用者に提示する時に，類似度の高いものから動的に生成し，提示する．

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は，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利用する．

情報検索においては，通常，個々の文書をその中に含まれる単語の重要度で代表させる．この単語に関する重要度情報と検索要求文中の語の照合により各文書をスコア付けする．単語に重要度を与える方法としては，$tf \cdot idf$ 法が広く用いられている．さらに検索要求文と個々の文書の適合度はベクトル空間モデルを用いることで求められる．

4.1.1 $tf \cdot idf$ 法

$tf(d,t)$ は，ある語 t がある文書 d 中に現れる頻度を $M(d)$ で割った値である．$M(d)$ はセグメント内の形態素数であり，セグメント長を反映した正規化を行なっている．$idf(t)$ 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．

4.1.2 ベクトル空間モデル

ベクトル空間モデルは，文書や検索要求文を多次元空間上のベクトルとして表現し，二つのベクトルの間の角度 (のコサイン値) を比較することにより類似度を調べるものである．つまり，ある文書や検索要求文がベクトルが同じ方向を指す文書ほど類似度が高いと考える．ベクトルの各次元には各単語を，各成分には対応する単語の重要度を割り当てる．単語の重要度は $tf \cdot idf$ 法により求める[2]．

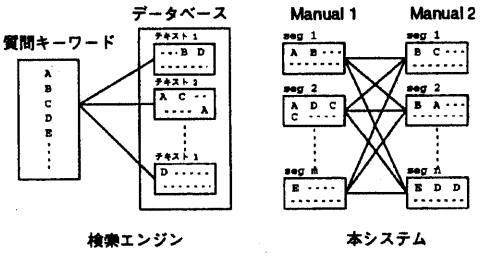

図 3: 類似度計算の比較

図 3に情報検索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う．

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，単語数，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．なお，マニュアルの対応付けで使用する単語には，用語説明，操作説明の基本単位である命題の骨格をなす名詞と動詞のみを考慮し，類似度計算を行なう．

4.1.3 格情報，共起情報の利用

本システムでは，操作の対応に基づいてセグメント同士を対応づけることを目的としている．操作の説明は，「スイッチをビデオ側に合わせる」のように

名詞 1- 格助詞 1 名詞 2- 格助詞 2 … 動詞

といった操作対象を表す名詞と操作内容を表す動詞で表される．そこで，文中の名詞や動詞の間の関係を利用することで，その文の表す操作 (の一部) に重きを置いてセグメント間の対応を取ることができると考えられる．例えば，2 セグメント内の文に同じ名詞対が共起した場合にはセグメント間の類似度に共起した名詞の重要度に応じた値を加算し補正を行うことなどが考えられる．我々は，単語の共起情報を，

1. ベクトル空間モデルの次元

2. セグメント内の単語頻度 tf

に反映させた類似度計算を行うことを考えた．

[2]通常の情報検索においては，検索要求文中の単語の頻度情報を予め知る方法はないので，検索要求文中の語の重要度については例えば全て 1 としてベクトルを生成する．

共起情報を次元で表現する方法

これは，ベクトル空間モデルで単語の重要度を表す次元とは別に，共起情報を表す新たな次元を考える方法であり，以下の計算を行う．

1. 句ごとに動詞と格情報 (格助詞とその前に位置している名詞) を取り出す。
2. 格助詞が n 個のときは、そこから 1 個以上，n 個以下を選ぶようなすべてのキーワード (名詞) の組み合わせを作る。
3. 組み合わされたキーワードの各セットについて、$tf \cdot idf$ を計算する。

例えば「エンドユーザがプログラミング言語を習得する。」という句からは、以下のような共起情報を表す 3 つの次元を新たに考える．

1．(動詞，習得)(が，エンドユーザ)
　(を，プログラミング言語)
2．(動詞，習得)(が，エンドユーザ)
3．(動詞，習得)(を，プログラミング言語)

このような共起情報を表す新たな次元についても，$tf \cdot idf$ を計算し重要度とする．

4.1.4 単語頻度 tf を補正する方法

情報検索における文書の重要度決定に，検索要求文内で共起している単語対の共起重要度を利用すると，同じ再現率に対する適合率が向上することが報告されている [高木 96]．本稿では文書間のハイパーテキスト化を考えているので，対象となる両方のマニュアルについて，出現する全ての共起単語対についての共起重要度 cw を計算し，類似度計算に反映させることを考える．さらに高木らの方法に加えて格情報を考慮する．

この手法では，2 セグメント d_A, d_B 間の類似度計算において，両セグメントに出現している共起単語対について，tf の値を次のように補正する．ある語 t_k がセグメント d_A に f 回出現した場合，新たに $tf'(d_A,t_k)$ を文書内出現頻度として語の重要度を算出する．$tf'(d_A,t_k)$ は以下の式により計算する．

$$tf'(d_A,t_k) = tf(d_A,t_k) + \sum_{t_c \in T_c(t_k,d_A,d_B)} \sum_{p=1}^{f} cw(d_A,t_k,p,t_c) + \sum_{t_c \in T_c(t_k,d_A,d_B)} \sum_{p=1}^{f} cw'(d_A,t_k,p,t_c)$$

ここで，$T_c(t_k,d_A,d_B)$ は d_A, d_B の両セグメントで t_k とある範囲内の位置で共起している単語の集合である．p は，セグメント d_A 内で，ある語 t_k が出現する場所を表しており，セグメント内の全ての出現箇所に対しての cw の和を計算している．この計算を T_c に含まれる全ての単語について行い，tf に加算する値を得る．

また，cw は，共起を調べる単語として名詞のみを考慮した共起重要度であるが，cw' は，名詞とその直後に出現する格助詞を一つの単語と考え，cw と同様に求めたものであり，格助詞と名詞の組に関する共起に着目した共起重要度である．

次に共起重要度 cw の算出法を説明する．cw' についても名詞と格助詞の組を 1 つの単語と見なす以外は算出法は同様である．まず，t_k と t_c における語間の近接出現係数 $\alpha(d_A,t_k,p,t_c)$ と共起係数 $\beta(t_k,t_c)$ を次のように定義する．

$$\alpha(d_A,t_k,p,t_c) = \frac{d(d_A,t_k,p) - dist(d_A,t_k,p,t_c)}{d(d_A,t_k,p)}$$

$$\beta(t_k,t_c) = \frac{rtf(t_k,t_c)}{atf(t_k)}$$

$d(d_A,t_k,p)$ はどれくらいの距離までを共起の範囲とするかを表すパラメタである．本稿では 1 つの意味的なまとまりである一文中の単語の共起を見ており，$\alpha(d_A,t_k,p,t_c)$ は文内に共起した単語についてのみ計算する．よって，$d(d_A,t_k,p)$ は注目している動詞句内の単語の数である．また，$dist(d_A,t_k,p,t_c)$ は，セグメント d_A で p 回めに出現した t_k について単語数で計算した t_c との距離である．$atf(t_k)$ は注目しているマニュアル内の t_k の出現総数，$rtf(t_k,t_c)$ は一文内に共起している t_k と t_c の出現総数である．

次に，t_k の共起語 t_c の近接出現共起単語の重要度 $\gamma(t_k,t_c)$ を定義する．N は各マニュアル中のセグメント数であり，$df(t_c)$ は t_c の出現する文書数である．

$$\gamma(t_k,t_c) = \tau(df(t_c)) = log(\frac{N}{df(t_c)})$$

以上で定義した，近接出現係数 $\alpha(d_A,t_k,p,t_c)$，共起係数 $\beta(t_k,t_c)$，接出現共起単語重要度 $\gamma(t_k,t_c)$ から，セグメント d_A 内の p 番めに出現する語 t_k の共起重要度を次の式で表す．

$$cw(d_A,t_k,p,t_c) = \frac{\alpha(d_A,t_k,p,t_c) \times \beta(t_k,t_c) \times \gamma(t_k,t_c) \times C}{M(d_A)}$$

$M(d_A)$ はセグメント d_A 内の形態素数であり，tf と同様の正規化を行なっている．C は共起重要度正規化係数である．この値は，大きいほど共起重要度が tf にあたえる影響が大きくなる．

4.2 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)

本システムは，図 4 に示す 4 つのサブシステムより構成されている．

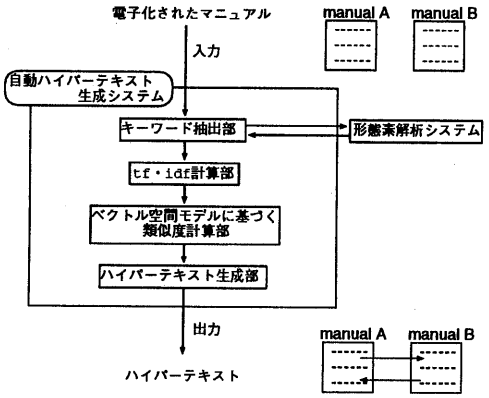

図 4: 自動ハイパーリンク生成システムの構成

一例として，図5に，ある実用ソフトウェアのチュートリアルマニュアル[日立 b]とリファレンスマニュアル[日立 a]の間で，自動ハイパーテキスト化を行なった結果を示す．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．

4.3 評価法

情報検索で，一般的に利用される再現率 (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．

4.4 大規模マニュアルによる検証

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 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．

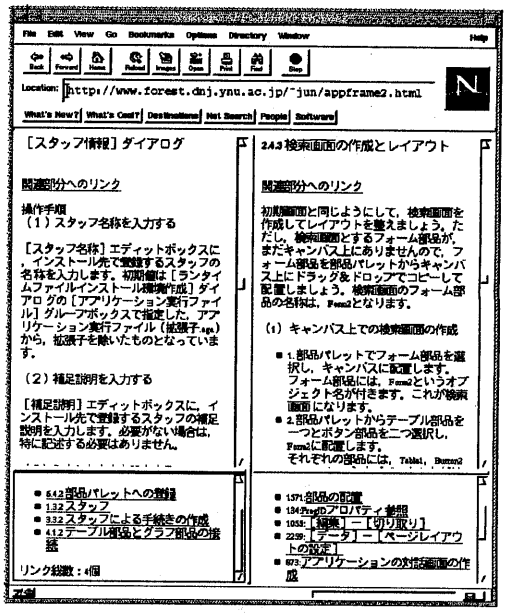

図 5: システムの利用画面

1. 同じ操作をしている部分，またはおなじ語句の説明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6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．

4.4.1 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．利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．

4.5 対応付けの検証

対応関係の完全な正解を作成可能なマニュアルを用いて，対応付けの正しさを評価する．同一メーカのビデオの2マニュアルを本システムでハイパーテキスト

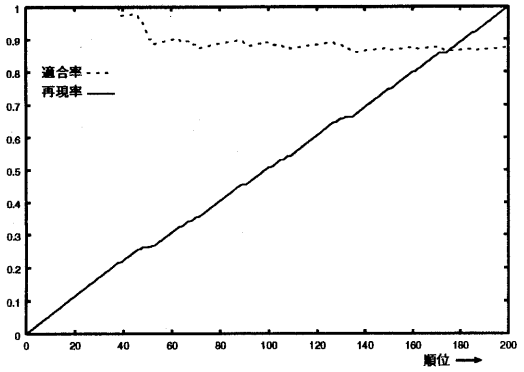

図 6: 大規模マニュアルにおける再現率と適合率

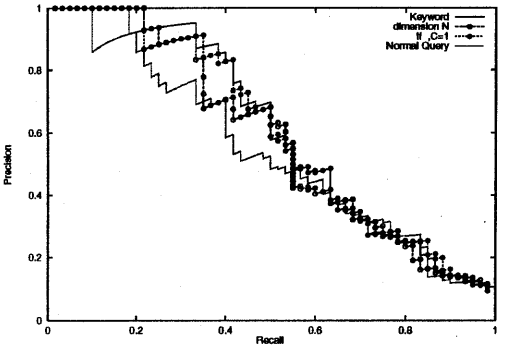

図 7: 全組合せにおける再現率と適合率

化を行った．両マニュアルのセグメント数は，マニュアル A が 31，マニュアル B が 27 である．正解は，節 4.4 と同様の基準で 60 であった．

計算された全対応付け 837 通りについて，類似度によって順位付けられた対応の上位からある順位までを選んだ時の，再現率，適合率のグラフを図 7 に示す．対応づけは，以下の 4 通りで行なった．

1. 単語の頻度情報のみ，両マニュアルの単語に $tf \cdot idf$ 計算 (図中 Keyword)
2. 単語の共起情報を次元で表現 (dimension N)
3. 共起重要度正規化係数 $C = 1$ において単語の共起情報で文書内頻度 tf の値を補正 (tf , $C = 1$)
4. 単語の頻度情報のみ，一方のマニュアルは単語の重要度を全て 1 とし，片方のマニュアルについてのみ単語に $tf \cdot idf$ 計算 (Normal Query)

単語の頻度情報のみを利用した場合には，動詞の重要度を利用せずに名詞の重要度のみを利用した場合に同再現率における適合率が高いという結果を得た．また，共起情報を次元で表現した場合は，名詞の共起情報のみを利用し動詞は利用しない場合に同再現率における適合率が最も高いという結果を得た．これらの結果と比較するために，共起情報で tf を補正する場合についても名詞の共起情報のみを利用し，tf は名詞についてのみ補正を行った．

4.5.1 考察

両文書集合に対して単語の統計的な頻度情報を計算する効果を調べる．Normal Query は，一方のマニュアルの単語には $tf \cdot idf$ による重要度計算を行わず，重要度を全て 1 とした時の計算結果であり，通常の情報検索と同じ設定である．この結果と比較すると，他の 3 通りは特に低再現率域での適合率が向上している．したがって，情報検索の場合と比較してマニュアル間のハイパーテキスト化においては，両マニュアルについて単語の統計的な情報を利用することのできる効果が現れている．

次に共起情報を考慮した場合について考察する．次元で表現した場合と tf を補正した場合いずれについても，低再現率域での適合率が向上している．これは共起情報を利用した類似度計算を行うことによって，正しい対応の類似度がより大きい値になっているためと考えられる．また両者を比較すると，tf を補正した場合の方が適合率が大きい部分が多い．

5 関連研究

一般に自動ハイパーテキスト生成は，主に，1) いかにしてリンクを張るべき対象を決めるか，2) 決定した対象を文書中の別の関連部分とどのように関連付けるか，という 2 つの部分問題から構成される．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．2. については，単語の出現頻度などに基づく統計的な類似度を用いた方法や，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．

まず，1 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) も用いる [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ることにより，その語の重要説明箇所を特定している [黒橋 96]．

中川らは，マニュアルの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た [中川 97]．

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]．

節 3で述べたシステムは，中川らのシステムと雨宮らのシステムを基礎に持つものである．定義パタンに基づく点は，黒橋らのシステムにと同じであるが，その定義パタン自身を学習してしまう点と，ハイパーリンクエディタと連動した点が新しい．

節 4で述べたシステムでは，マニュアルを操作手続きのまとまりであるセグメントに分割し，セグメント間の対応を見い出しているので，すべてのセグメントがハイパーリンクを設定すべき対象と考え，その選択については言及していない．

2. について述べる．節 4のシステムにおいて，関連する文書部分を捜し出す過程は，Salton らの Passage Retrieval の概念の応用と位置付けられるであろう．Salton らは，文書を小単位 (passage) に分割し，検索要求文と passage の間で類似度計算を行い，passage をユーザに提示することを提案している [SAB93]．しかし，情報検索においては検索要求文中の語の統計情報が，通常，前もって得ることができないのに対して，関連文書間のハイパーテキスト化では検索要求文に相当するマニュアルのセグメントに関してすでに語の統計情報が前もって得られる．これにより，類似度計算の精度が上昇することが本稿において示された．

さらに，高木らの手法 [高木 96] を参考にして，動詞句内の単語の共起情報を利用し，両セグメントで同じ共起名詞対が出現する場合には，共起した名詞の重要度を増加させる手法反映させセグメント間の類似度を補正している．ここでは高木らの手法に加えて，格助詞の情報も利用して精度の向上を図っている．

6 おわりに

本稿では，電子化マニュアルにおけるハイパーテキスト化手法について述べ，我々が提案している 2 つの試作システムについて概説した．

節 3で述べた重要語抽出に基づくハイパーテキスト化ツールは，完全なる自動ハイパーテキスト化ではなく，人間による後編集も考慮したものであった．しかし，節 3.5で述べた学習方式を組み込むことにより，システムの利用に応じて人手による訂正部分は減っていくものと思われる．より厳密な学習機構の評価ならびにその精度の向上については今後の研究の課題としたい．

節 4で述べたシステムについては，大規模マニュアルならびに小規模マニュアルを用いた検証により，ある程度の精度で関連マニュアル間の関連セグメントを結びつけることができることを示した．ただし，検証に用いたマニュアルの数が少ないので，マニュアル数を増すことにより，この検証の裏付け作業が必要となるであろう．

参考文献

[Gre96] Stephen J. Green. Using lexical chains to build hypertext links in newspaper articles. In *AAAI 96 Workshop on Internet-based Information Systems*, 1996.

[SAB93] Gerard Salton, J. Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR '93: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, 1993.

[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. 言語処理学会, 3 月 1996.

[海保 87] 海保博之, 加藤隆, 堀啓造, 原田悦子. ユーザ・読み手の心をつかむマニュアルの書き方. 共立出版, 東京, 1987.

[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. 自然言語処理研究会報告 96-NL-115-7, 情報処理学会, 1996.

[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.

[中川 97] 中川裕志, 森辰則, 松崎知美, 川上大介. 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. 情報処理学会論文誌, Vol. 38, No. 10, 1997.

[日立 a] 日立製作所. APPGALLERY オンラインヘルプ. 日立製作所.

[日立 b] 日立製作所. 使って見よう APPGALLERY. 日立製作所.