自然言語処理 121－16
情報学基礎 47－8
（1997. 9. 12）

$tf \cdot idf$ 法を用いた関連マニュアル群のハイパーテキスト化

大森 信行　　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

近年，電子機器などの発展にともない，そのマニュアルも大規模になり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い．ユーザがそのような分冊された関連マニュアルを読み進めていく場合，一連の操作手続きなどあるまとまった文書単位 (セグメント) での対応関係に基づく参照が重要となる．そこで，本稿ではセグメント同士が対応するハイパーテキスト化を考える．関連マニュアル間でセグメント間の関連度計算を行い，その結果にしたがってハイパーテキスト化を行う．また，その関連度計算においてマニュアル文の格情報や共起情報を利用する方法について説明する．

Hypertext Generatiton from related manuals using $tf \cdot idf$ method

Nobuyuki Omori , Jun Okamura, Tatsunori Mori and Hiroshi Nakagawa
Division of Electrical and Computer Engineering, Faculty of Engineering,
Yokohama National University
{ohmori@forest,jun@forest,mori@forest,nakagawa@naklab}.dnj.ynu.ac.jp

Recently manuals of product are large and often separeted to some volumes. In reading these related manuals, we must consider the relation among some segments which contain explanations about a series of operations. In this paper, we propose a method of hypertext generation from related manuals. This method is based on the similarity between two segments. We also describe a method for similarity ranking using word co-occurrence and word dependency.

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．

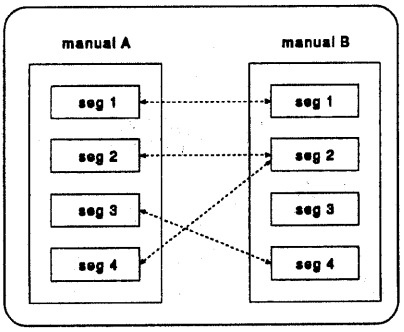

図 1: システムが生成するハイパーテキストの概念図

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].

高木らは，単語の共起情報から共起重要度を算出し，単語頻度情報と組み合わせて検索結果に対する重要度を計算している [高木 96]. 本研究においても，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，共起名詞対の重要度を反映させセグメント間の類似度を補正している．

Salton らは，検索質問と， passage という文書の一部分に対して類似度計算を行い， passage をユーザに提示するという研究を行っている [Salton 93]. 本研究においても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，セグメント間のハイパーテキスト化を行っている．本研究では，対象とする両マニュアル内の単語において頻度情報による重要度を計算している点，また語の共起関係をセグメント間の類似度計算に反映させている点で異なっている．

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の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．

検索要求文はユーザにより自由に入力できるのが通例であるからその中の検索語に関する統計情報は前もって得られないのが普通である．よって，通常は検索要求文中の検索語について重みを計算することはできず，データベース中の語についてのみ重みを計算する．一方，本システムでは，両マニュアル中の全ての語について重みを計算することが出来るため，対応箇所を見つける際の精度の向上が期待される．

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．

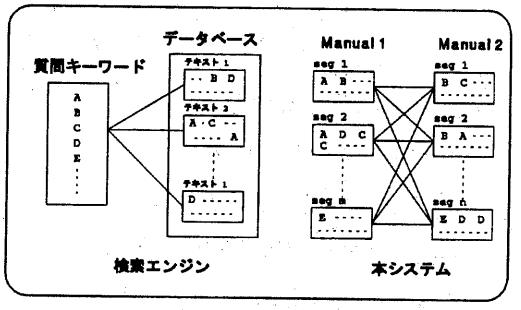

図 2: 類似度計算の比較

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．

3.3 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．基本的には操作の説明は，「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，文中に出現する名詞や動詞の間の関係を利用することで，その文の表している操作(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．

3.3.1 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。$\sum_{k=1}^{n} {}_nC_k$ 通りの組み合わせが作られる。
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。

「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、「エンドユーザ」など個々の名詞を表す次元の他に，以下のような共起情報を表す次元を考える．

> 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．ベクトル空間モデルの類似度計算は，単語のみの場合と同様に行った．

3.3.2 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]．さらに高木らの方法に加えて格情報を考慮する．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．

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．

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

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，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．例えば，

名詞1- 格助詞1　名詞2- 格助詞2 … 動詞

という動詞句において，cw' の算出では 名詞1- 格助詞1 のような続いて出現する名詞と格助詞を一つの単語と考え，　名詞1- 格助詞1　と　名詞2- 格助詞2 という 単語の対が共起していると考える．

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　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．

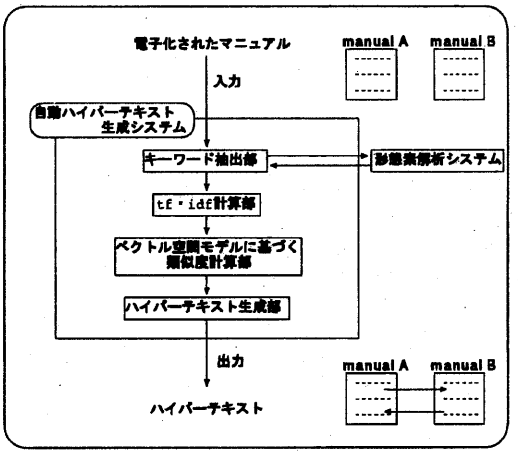

図 3: 自動ハイパーリンク生成システムの構成

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．

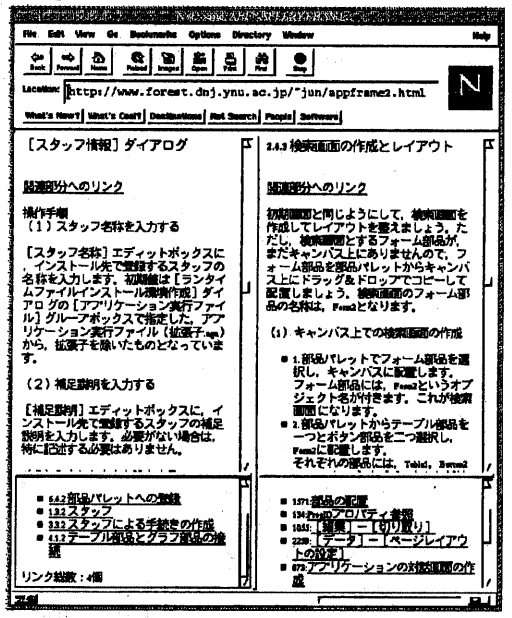

図 4: システムの利用画面

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．

5.2 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 [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. 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 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．

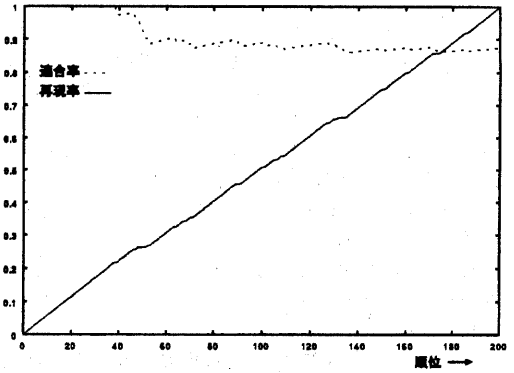

図 5: 大規模マニュアルにおける再現率と適合率

5.2.1 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．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．

5.3 対応付けの検証

対応関係の完全な正解を作成することのできるマニュアルを用いて，正しい対応付けがされているかを評価する．同一メーカーのビデオの 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，5.2 と同様の基準で 60 の対応を設定した．

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 6 に示す．対応づけは，節 3.3 で述べた方法を含む，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. 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)

tf の値を補正する場合は，共起重要度正規化係数 C を 1 として計算した．

利用する単語の情報については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．

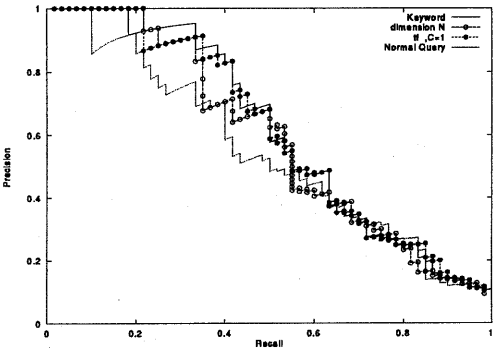

図 6: 全組合せにおける再現率と適合率

5.3.1 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．

1. 次元で表現した場合
2. tf を補正した場合

については，ともに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．

共起情報を利用した 1. と 2. の方法を比較すると，tf を補正した場合の方が適合率が大きい部分が多い．

本システムは，あるセグメントと操作手順という観点から関連のある他のセグメントとの対応付けを目的としている．そのため，特定のセグメントについて，本方式での対応の順位づけにおいて，正しい対応が上位に位置していれば，そのセグメントと関連のある段落が上位に順位付けされていることになる．

例として，マニュアル A の No22 のセグメント とマニュアル B 中のセグメント (複数) の対応ついて考える．図 7 に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，適合率のグラフを示す．ここでは，対応の数はマニュアル B の全セグメント数 27 である．

マニュアル A のセグメント 22 についての正解の数は 4 であるので，グラフは変動が大きくなっている．共起情報を tf の値に反映させた場合に，高再現率域で適合率が向上している．両セグメントに共起する単

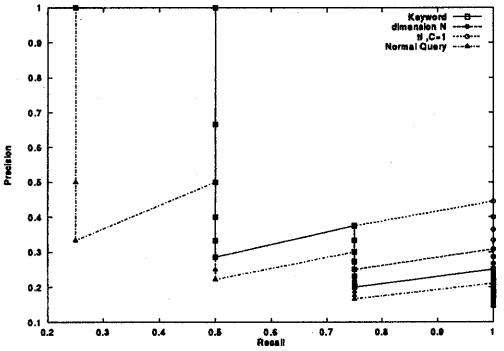

図 7: 特定のセグメントについての再現率と適合率

語対についてのみ tf の値に加算するという方法が，マニュアルという文書集合のハイパーテキスト化に有効であることを示している．

共起情報を次元で表現した場合は，単語頻度情報のみの場合と比べ適合率は向上していない．これは，次元で表現した場合は，全ての共起情報について新たな次元を作るため，類似度計算時に一方のセグメントでのみ共起している単語対に対応する次元の成分が類似度を低下させているためと考えられる．

本稿では共起重要度正規化係数は $C = 1$ として計算した．C の値を小さくすると tf を補正する効果が小さくなり，頻度情報のみを利用した場合と差がほとんどなくなる．C の値を大きくすると，誤った対応が順位づけの上位に位置し，低再現率域での適合率が低下する．これは，誤った対応をしている両セグメントで共起している単語に対して tf の補正の効果が大きくなり，そのセグメント間の類似度が大きくなるためである．

6 まとめと課題

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，ビデオのマニュアルで，共起情報の利用法として tf に補正する方法が，特定のセグメントに注目した場合に適合率の向上に有効であることを確認した．

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．

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

[Salton 93] Gerard Salton, J. Allan and Chris Buckley: Approaches to Passage Retrieval in Full Text Information Systems, 16th ACM SIGIR, pp.49-58(1993).

[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，情報処理学会研究報告 96-FI-41-8，(1996).

[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 茶筌 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.