情報検索手法を利用し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[†]

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ことが多い．ユーザがそのような分冊化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どうしが対応するハイパーテキスト化手法を提案する．この手法においては関連マニュアル間でセグメント間の関連度計算を行い，その結果に従ってハイパーテキスト化を行う．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ことにより単語頻度情報のみを利用した $tf \cdot idf$ 法と比べて，再現率，適合率が向上することを確認した．

Hypertextualization for Related Instruction Manuals Using the Techniques of Information Retrieval

NOBUYUKI OMORI,[†] JUN OKAMURA,[†] TATSUNORI MORI[†] and HIROSHI NAKAGAWA[†]

Recently manuals of products become large and are often separeted to some volumes. In reading such related manuals, we must consider the relation among segments, which contain explanations of serieses of operations. In this paper, we propose a method of hypertext generation for a set of related manuals. Our method is based on the similarity calculation between two segments. In the method word co-occurrences and case information are used for similarity ranking. Our experimental result shows the method improves both recall and precision than normal $tf \cdot idf$ method.

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．これにともな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，語句をマウスなど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ほかに，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．たとえば，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 に示すようなあるまとまった文書単位であるセグメント（図 1 中では seg 1 など）どうしが対応するハイパーテキスト化を考える．

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，項を単位とする．

† 横浜国立大学工学部電子情報工学科
Division of Electrical and Computer Engineering, Faculty of Engineering, Yokohama National University

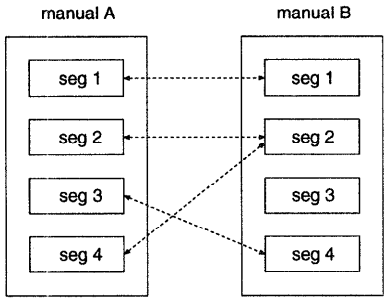


図1 システムが生成するハイパーテキストの概念図
Fig. 1 Image of hypertextualized manuals.

2. 関連研究

WWWの発展にともな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[1]．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．ハイパーテキスト化においてリンク対象となる語を決定するために索引語や説明箇所の抽出法について研究が行われてきた．文献6)にはこの分野の動向が記されている．また，用語辞典やシソーラス中の語を対象としたハイパーテキスト化も行われている[12),13)]．しかし，これらは単語を対象としたハイパーテキスト化であり，我々の考える文書間でセグメントどうしが対応するハイパーテキスト化とは異なる．

文書部分どうしを関連付ける点では，Green[5)]のlexical chainを利用した新聞記事間のハイパーテキスト化の研究が関連する．ここでは，WordNet databaseをシソーラスとして単語間の関連を調べ，記事内で意味的に関連する単語群であるlexical chainを形成する．それを利用して記事間の類似度を計算する．これに対して本研究のハイパーテキスト化はシソーラスを用いない手法である．

文書（部分）の間の類似度計算方法に，本研究では単語の重みに $tf \cdot idf$ の値を用いるベクトル空間モデル[2)]を用いている．文書間の類似度に注目する研究としては，文書をクラスタリングし，類似した複数の文書から成るグループを生成しユーザの要求に応じて提示する情報閲覧手法もある[3),8)~10)]．複数文書間のセグメントを対応付ける我々の目標において，クラスリング技術を使用することも興味深いが，単純に両文書のセグメント群を混在させてクラスタリングを行うだけでは，両文書のセグメント間での対応が必ず生成されるという保証はない．

文書中の一部分を単位として求める文書を検索する点においては，Saltonら[11)]のpassage検索が我々の研究と関連する．ここでは，検索質問と，passage（segment）の間で類似度計算を行い，値の高いpassageをユーザに提示するという研究を行っている．また，Kaszkielら[7)]は様々な文書部分の決定方法を比較し，passage検索の有効性を検証している．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点で同じである．また，ジャストシステム社のConcept Baseは検索された文書自身を容易に次回の検索質問にできるという一種の質問拡張機能を有する[14)]．この手法はある文書（部分）を検索質問とする点では我々の手法と同じである．しかし，我々の手法においては検索質問側に相当する文書部分においても語の統計情報が利用できるので，精度向上が見込まれる．これについては後述する．

高木ら[21)]は，単語の共起情報から共起に基づく重要度を算出し，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．本研究においても，1文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．

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持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，適宜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，リファレンスマニュアルや初心者向けチュートリアルなど，複数のマニュアルに分かれている場合が多い．これらのうち，リファレンスマニュアルはその機器の使用法がすべてにわたって記述されており，それ以外のマニュアルを読み進める過程で参照されることが多い．そこで本稿では，類似マニュアル間の対応付けに加えて，リファレンスマニュアルとその他の関連マニュアル間の対応付けに注目し評価実験を行った．

3.2 自動ハイパーテキスト生成

本システムでは自動ハイパーテキスト生成のために次のような関連付けを行う．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，値の降順にセグメントのタイトルを提示し，ユーザの要求に応じて実際に文書内容を表示する．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[2)]を応用しセグメント間の類似度計算を行う．さらに本システムでは格情報と語の共起情報を使って精度の向上を図っ

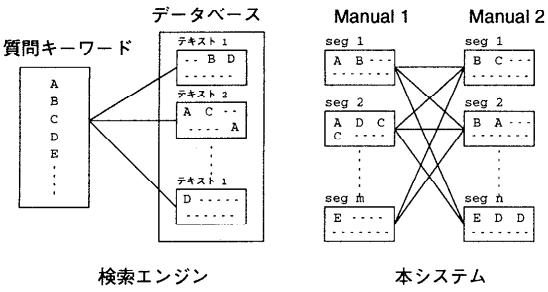


図 2 類似度計算の比較
Fig. 2 Comparison of two similarity calculation methods.

ている．

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，1つの検索要求文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行っているが，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，語数，組合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，ハイパーテキスト化においては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，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．なお，本稿で用いる「キーワード」という語は，索引語など文書の中核概念を表す語を指示するものではない．単にベクトル空間モデルにおいて次元を構成するのに利用された語を指す．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．

・分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うる．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．

以下，本システムで精度を上げるために利用している手法を説明する．

3.3 格情報，語共起情報の利用

本システムでは，操作の対応に基づいてセグメントどうしを対応付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とることができると考えられる．

ここでは，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

に反映させて類似度計算を行う手法について検討する．

3.3.1 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高木ら[21]により報告されている．我々の手法ではさらにこ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，出現するすべ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．

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

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，セグメント内のすべ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 に含まれるすべての単語について行い，tf に加算する値を得る．

次に共起重要度 cw の算出法を説明する．

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 cw を次の式で表す．

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

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ータである．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．本稿では1つの意味的なまとまりである1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，$d(d_A,t_k,p)$ は注目している文中の単語の数である．

$atf(t_k)$ 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1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．

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

$M(d_A)$ 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．

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一方，cw' は，名詞とその直後に出現する格助詞を1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．たとえば，

名詞1-格助詞1　名詞2-格助詞2 … 動詞

という文において，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1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．

3.3.2 共起情報を次元で表現する方法

共起情報を次元で表現する方法は，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）の組合せを作る．$\sum_{k=1}^{n} {}_nC_k$ 通りの組合せが作られる．

(3) キーワードの各組合せについて，$tf\cdot idf$ を計算する．

たとえば，「エンドユーザがプログラミング言語を習得する．」という句においては，「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，通常のキーワードのみの場合と同様に行った．

4. システムの概要

本システムの入出力は，次のとおりである．

入力　電子化されたマニュアル
（plain text，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 plain text（タグにより構造の示されていない文書）の場合，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．

キーワード抽出部　形態素解析システムを用いて各文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ごとにカウントする．形態素解析システムには茶筌 1.0b4 [16] を使用した．

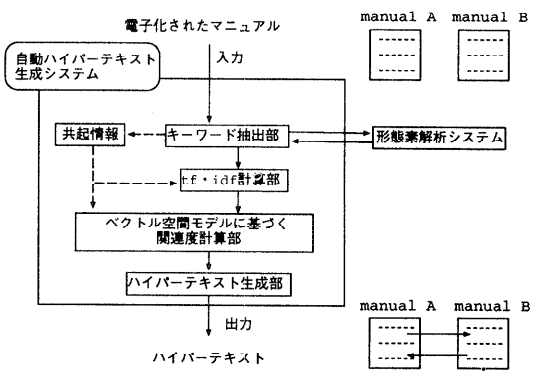


図3　自動ハイパーリンク生成システムの構成
Fig. 3　Overview of our hypertext generator.

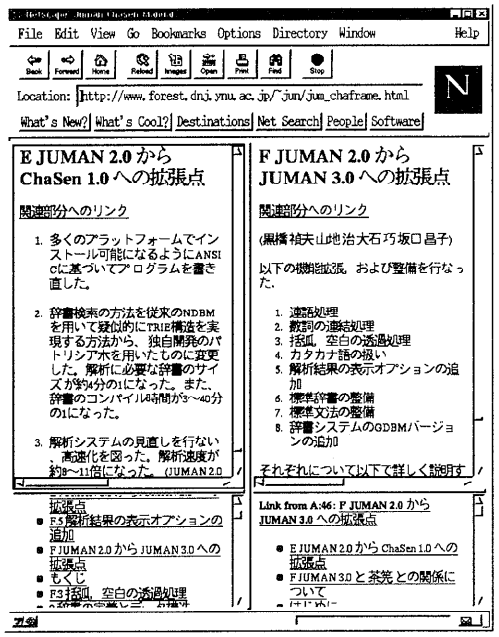


図4 システムの利用画面
Fig. 4 The use of this system.

***tf*・*idf* 計算部** カウントされたキーワードをもとに *tf*・*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,各セグメントごとのベクトルを作成し,すべての組合せに対して *cosine* 値を求める.

ハイパーテキスト生成部 *cosine* 値の降順にセグメントのタイトルを提示し,ユーザの要求に応じて実際に文書内容を表示する.

形態素解析システム JUMAN[17]と茶筌[16]のそれぞれのマニュアルの間で自動ハイパーテキスト化を行った結果を図4に示す.

画面をフレームで4分割し,左上に「JUMAN」の,右上に「茶筌」のマニュアル本文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.

5. システム評価

本章では我々の提案する手法の評価を行う.

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.

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

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

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評価

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.たとえば,APPGALLERY[15]では,チュートリアルはセグメント数65(100 kbyte),ヘルプマニュアルに至ってはセグメント数2479(2 Mbyte)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難しいので,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,大学院生2名を被験者とし,両者が次の基準に一致したと判断した対応を正解とした.

(1) 同じ操作をしている部分,または同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200位までを対象としているため,そこ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ときの再現率,適合率を示している.

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.これより,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.

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どうしの内容の関連が大きいとはいえない場合があり,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.

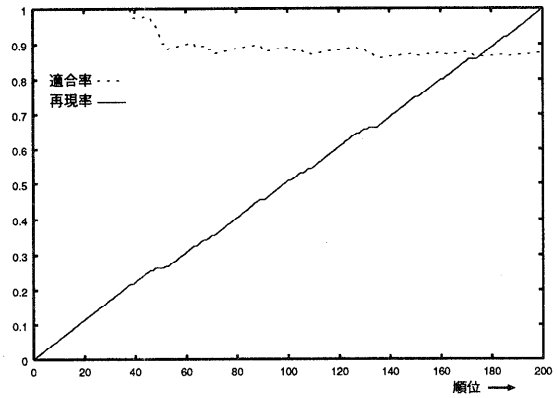


図5 大規模マニュアルにおける再現率と適合率
Fig. 5 Recall and precision of generated hyperlinks on large-scale manuals.

表1 マニュアルの組合せと正解の数
Table 1 Manual combinations and number of right correspondences of segments.

マニュアル組合せ	A⇔B	A⇔C	B⇔C
全対応の数	1056	896	924
正解数	65	60	47

5.3 対応付けの評価

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，本システムにより正しい対応付けが生成できるかを評価するために，以下の実験を行った．まず，ここでは近似として，同一メーカの3つのビデオのマニュアル[18)~20)]間で本システムによるハイパーテキスト化を行った．各マニュアルのセグメント数と大きさは，マニュアルA[20)]が32（80 kbyte），マニュアルB[18)]が33（100 kbyte），マニュアルC[19)]が28（98 kbyte）である．正解は，5.2節と同様の基準で設定した．それぞれの組合せにおける全対応の数，正解の数は**表1**に示す．

3通りのマニュアル組合せで計算された全対応付け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ときの，再現率・適合率のグラフを**図6**，**図7**，**図8**に示す．対応付けは，3.3節で述べた方法を含む，以下の4通りで行った．

(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（cw+cw' tf）
(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すべて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）

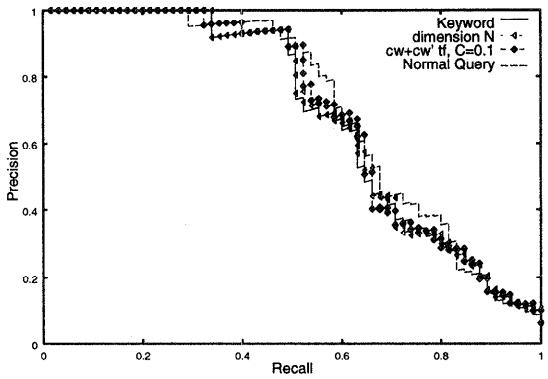


図6 全組合せにおける再現率と適合率，A⇔B
Fig. 6 Recall and precision of hyperlinks for all segments, A⇔B.

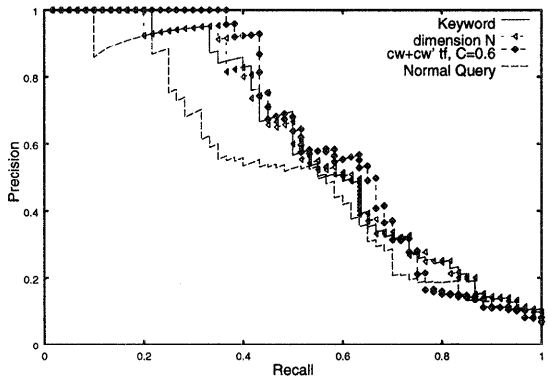


図7 全組合せにおける再現率と適合率，A⇔C
Fig. 7 Recall and precision of hyperlinks for all segments, A⇔C.

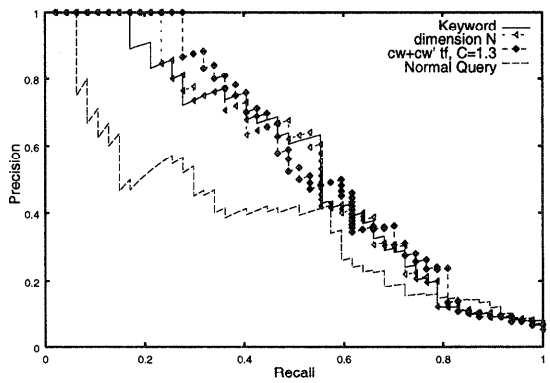


図8 全組合せにおける再現率と適合率，B⇔C
Fig. 8 Recall and precision of hyperlinks for all segments, B⇔C.

tf を補正する場合の共起重要度正規化係数 C については，適合率の11点平均の値を最も大きくする値を表している．

単語の頻度情報のみを利用する場合には，名詞と動詞の頻度情報を利用している．共起情報を次元で表現

表2 それぞれのマニュアル組合せと方法での適合率の11点平均
Table 2 11 point average of precision for each method and combination.

方法	A ⇔ B	A ⇔ C	B ⇔ C
頻度情報のみ	0.678	0.589	0.549
共起を tf	0.683	0.625	0.582
共起を次元	0.684	0.597	0.556
情報検索設定	0.692	0.532	0.395

する場合については，名詞と動詞の頻度情報に名詞のみの共起情報を利用した結果であり，名詞と動詞の共起情報は利用しない方が同じ再現率における適合率が大きかった．共起情報で tf を補正する場合については名詞と動詞の共起情報のみを利用している．

5.3.1 適合率の11点平均による評価

それぞれのマニュアル組合せでの各方法を利用したときの適合率の11点平均を**表2**に示す．これは，再現率が0.0から1.0まで0.1ごとの値の時の適合率を平均した値である．

5.3.2 適合率の統計的評価

今回，単語の頻度情報のみを使う方法と，共起情報を使う2手法の統計的比較・評価を行った．

評価対象は再現率・適合率のデータ集合であり，これについて**ウィルコクソンの符号順位検定**[4]を行った．これは，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．

3通りのマニュアル組合せの再現率・適合率データを混合したものについて，ある再現率に対応する複数の適合率の平均を計算し，再現率・適合率の1対1データ集合を作成した．これを頻度情報と共起情報の各3手法についてを作成し，実験を行った．

検定手順は次のとおりである．

(1) 前提：
- 帰無仮説 H_0：「両手法の適合率（代表値）に差はない」
- 対立仮説 H_1：「共起を使う手法の適合率（代表値）の方が大きい（片側検定）」
- 有意水準5%で片側検定を行う．

(2) ウィルコクソンの符号順位検定における N，検定統計量 T，標準正規得点 Z_0 を求める．

(3) Z_0 から，有意確率（上側確率）P_0 を求める．

(4) 帰無仮説の採否を決定する．
- $P_0 > 0.05$ のとき，帰無仮説を採択する．**「両手法の適合率に差はない」**
- $P_0 < 0.05$ のとき，帰無仮説を破棄する．**「共起を使う手法の適合率の方が大きい」**

以上の手順の検定を，単語の頻度情報のみを使う手法（Keyword）と共起情報を使う2つの手法の組合せで行った．**表3**にそれぞれの検定結果を示す．

表3 ウィルコクソンの符号順位検定結果
Table 3 Result of Wilcoxon matched-pairs signed-ranks test for each method combination.

	Keyword⇔ 共起を tf	Keyword⇔ 共起を次元
Z_0	0.49597	1.66732
P_0	0.30995 (31.0%)	0.04772 (4.8%)

5.3.3 考　　察

5.3.3.1 両文書の頻度情報を利用する効果

一方の文書ではなく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 $tf \cdot idf$ 法による重要度計算を行わず，重要度をすべて1としたときの計算結果であり，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，図6はほぼ同じものの図7，図8においてNormal Query以外の他の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．

図6，表2より，A ⇔ Bの組合せではいずれの方法で対応付けを行っても精度が高くなることが分かる．これは，ここで組み合わせた2マニュアルの構成や文章などが非常に似ており，一方にしか記述されていない説明が少なく，複雑な方法を用いなくてもセグメントの特徴を取り出すことができたためと思われる．

5.3.3.2 共起情報を利用する効果

図7，図8においては

1. 次元で表現した場合
2. tf を補正した場合

それぞれ両方の方法で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をより大きくする効果が生じるためである．

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く，精度が良いことが分かる．

表2からは，A ⇔ C，B ⇔ Cの組合せで，頻度情報のみ・情報検索設定と比べて共起情報を使うことで精度が向上していることが分かる．

次に適合率の統計的評価についての考察を行う．表3を見ると，共起情報を次元で表現する手法については，5%有意水準において単語の頻度情報のみを使う手法に対する優位性が示された．

共起情報を単語頻度 tf に反映させる手法について

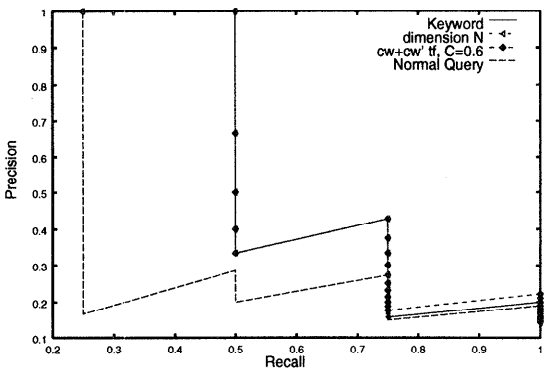


図9 特定のセグメントについての再現率と適合率
Fig. 9 Recall and precision of hyperlinks for specified segment.

は，その優位性を今回の検定によっては示すことができなかった．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るためであると考えられる．

5.3.3.3 特定セグメントについての再現率・適合率

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ため，この利用形態に合わせた評価を行う．そこで，本方式による特定のセグメントについての対応付けの精度を調べる．

例として，マニュアルAのNo23のセグメントとマニュアルC中のセグメント（複数）の対応ついて考える．**図9**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ときの，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Cの全セグメント数28であり，その中で4セグメントとの対応が正解である．

共起情報を tf の値に反映させた場合に，語の頻度情報のみの手法（Keyword）と比較すると，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語対について tf の値を補正するという方法が，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，tf の値に反映させた場合より精度向上の効果が少ない．この理由としては，新たに作った共起情報についての次元の中で，一方のセグメントでの成分が零となっているために類似度計算の際のcosine値の算出に寄与しない次元が多いためと考えられる．

6. まとめと課題

本稿では関連マニュアルのセグメントレベルでの自動ハイパーテキスト生成システムについて述べ，またそのハイパーリンクを生成する指標となる類似度計算に，情報検索で用いられる手法をベースに，操作に注目した語の共起情報を反映させる手法について述べた．

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い．その有効性を示した．

また，ビデオのマニュアルを利用した実験により，共起情報を利用することにより対応付けの精度が向上することを確認した．

本稿における自動ハイパーテキスト生成システムのユーザインタフェース部分は現在のところ1対1のマニュアルを表示する方式をとっており，これが1対 n のマニュアル提示となると，提示方法の問題が生ずる．すなわち，ユーザがあるセグメントからリンクをたどるときにリンク先セグメント群の提示方法として

- **すべてのマニュアル中から，類似度の高いセグメント群を提示**
- **ある指定されたマニュアル中から，類似度の高いセグメント群を提示**

という2通りがあり，どちらを採用するか，という問題である．この2つの提示方法についてどちらがより重要かということは，ユーザがマニュアルを読んでいる局面で生ずる検索要求によって異なると考えられ，使い分けることが必要となってくる．この「使い分け」を担当するユーザインタフェースをどのような仕様にするのかは，我々のこれからの課題である．

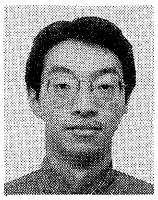
謝辞 マニュアルを提供してくださった日立製作所に深く感謝いたします．

参考文献

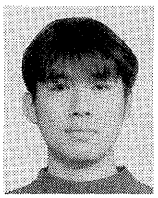
1) Basili, R., Grisoli, F. and Pazienza, M.T.: Might a semantic lexicon support hypertextual authoring?, *Proc. 4th Conference on Applied Natural Language Processing*, pp.174–179 (1994).
2) Frakes, W.B. and Baeza-Yates, R. (Eds.): *Information Retrieval – Data Structures & Algorithms*, P T R Prentice-Hall, NJ (1992).
3) Botafogo, R.A.: Cluster Analysisfor Hypertext Systems, *Proc. SIGIR '93: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.116–124 (1993).
4) Hull, D.: Using statistical testing in the evaluation of retrieval, *Proc. SIGIR '93: 16th Annual International ACM SIGIR Conference on Research and Development in Information Re-*

trieval, pp.329–338 (1993).
5) Green, S.J.: Using lexcal chains to build hypertext links in newspaper articles, *Proc. AAAI Workshop on Knowledge Discovery in Databases*, Portland, Oregon (1996).
6) Kageura, K. and Umino, B.: Methods of Automatic Term Recognition, *Terminology*, Vol.3, No.2, pp.259–289 (1996).
7) Kaszkiel, M. and Zobel, J.: Passage Retrieval Revisited, *29th Annual Meeting of the Association for Computational Linguistics*, pp.178–185 (1991).
8) Lelu, A.: Automatic Generation of "HYPERPATHS" in Information Retrieval Systems: A Stochastic and an Incremental Algorithms, *Proc. SIGIR '91: 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.326–335 (1991).
9) Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *Proc. SIGIR '92: 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.318–329 (1992).
10) Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W.: Constant Interaction-Time Scatter/Gather Browsingof Very Large Document Collections, *Proc. SIGIR '93: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.126–134 (1993).
11) Salton, G., Allan, J. and Buckley, C.: Approaches to passage retrieval in full text information systems, *Proc. SIGIR '93: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.49–58 (1993).
12) Amuba, S., Narashimhamurthi, N., O'Kane, K.C. and Turner, P.M.: Automatic Linking Of Thesauri, *Proc. SIGIR '95: 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.181–186 (1995).
13) 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).
14) 野村直之：言語工学による類似情報抽出の精度向上とその応用，「自然言語処理と検索技術」講習会資料，電子情報通信学会言語理解とコミュニケーション研究会（NLC），pp.36–56 (1997).
15) 日立製作所：使ってみよう APPGALLERY，APPGALLERY オンラインヘルプ.
16) 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 1.0b4 使用説明書，奈良先端科学技術大学院大学松本研究室 (1996).
17) 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).
18) 三菱電機：三菱ビデオ HV-BZ66 取扱説明書.
19) 三菱電機：三菱ビデオ HV-F93 取扱説明書.
20) 三菱電機：三菱ビデオ HV-FZ62 取扱説明書.
21) 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 (1996).

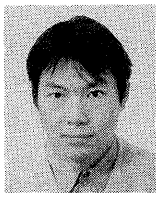
(平成 10 年 2 月 3 日受付)
(平成 11 年 3 月 5 日採録)



大森　信行（学生会員）
1973 年生．1996 年横浜国立大学工学部卒業．1998 年同大学院工学研究科博士課程前期修了．現在，NTT ヒューマンインタフェース研究所勤務.



岡村　潤
1974 年生．1997 年横浜国立大学工学部卒業．1999 年同大学院博士課程前期修了．現在，ソニー（株）勤務.



森　辰則（正会員）
1964 年生．1986 年横浜国立大学工学部卒業．1991 年同大学院工学研究科博士課程修了．工学博士．1991 年より同大学工学部勤務．現在，同助教授．計算言語学，自然言語処理システム，ディジタルドキュメント等の研究に従事.

中川　裕志（正会員）
1953 年生．1975 年東京大学工学部卒業．1980 年同大学院博士課程修了．工学博士．1980 年より同大学工学部勤務．現在，同教授．言語情報処理，マルチメディア情報検索等の研究に従事.