自然言語処理 117－16
（1997. 1. 21）

日本語マニュアルの内容検索システム

○松崎 知美, 三浦 健仁, 小俣 祐介, 斉藤 貴也, 山田 剛一, 森 辰則, 中川 裕志

横浜国立大学工学部

最近の電子機器やソフトウェアのマニュアルは決して読みやすいものではない。そこで、本研究では、電化製品やシステムを使っていて分からないことがあったとき、質問文を入力すれば、マニュアルの読むべき部分を示すマニュアルの内容検索システムを提案する。あらかじめマニュアルを章、節を利用してもっとも小さい節をセグメント単位に分割し、質問に対して、マニュアルの中から適切なセグメントを検索する内容検索の方法について検討した。実際に2通りの方法、一般的なベクトル空間法と複合語を考慮したtf.idf法で検索した際の結果の適合率、再現率を求めた。検索の結果最も高いスコアを得た1つのセグメントを取り出した場合で、適合率38.6%、再現率29.3%という値を得た。

Contents Retrieval System of Japanese Manual

○ Tomomi Matsuzaki, Takehito Miura, Yusuke Komata, Takaya Saitou,
Kouichi Yamada, Tatsunori Mori, Hiroshi Nakagawa

Engineering, Yokohama National University

Recently, manuals of home erectronics tool or complicated software systm are hard to read. Then, we propose a contents retrieval system of Japanese manual that when a user inputs the question, returns parts of manual to read. Firstly we divided manual into parts until the subsections. Then we examined methods to retrieve appropriate parts from one manual,corresponding to the input question. For two methods, a standard vector space method and a tf.idf method taking noun compounds into account, we calculated recall and precision. In case of retrival a part of the highest score, pricision was 38.6%, recall was 29.3%.

1 はじめに

最近の電子機器、ソフトウェアなどのマニュアルは決して読みやすいものではない。そのマニュアルに特有の言葉や、専門用語、膨大なページ数などは、マニュアルを難解なものにしている。このような状態では、使っている最中に何か分からないことがあっても、マニュアルを読んで調べるのは、なかなか手のかかるものである。電子機器やソフトウェアを使っていて分からないことがあった時、質問文を打ち込めば、マニュアルの読むべき部分を示してくれるシステムがあればユーザの助けとなるであろう。現在このようなときに助けとなるものとして、WINDOWSのオンラインHELP機能などがあげられる。分からない言葉があった時にすぐにその説明を出してくれるこれらの機能はマニュアル読者の大きな助けとなっている。しかし、このようなことができるためには、あらかじめマニュアルライターらが、多大な手間をかけてそういった機能を付与し、そのための文章を書いておかねばならない。そのような機能が付与されていないマニュアルでも、質問を打ち込まれれば、すぐに内容を検索して、読むべき部分を示すことのできる、マニュアル内容検索システムが容易に構築できることが切望されるそこで本研究ではこのようなマニュアル内容検索システムを既存のテキスト形式のマニュアルから自動構築する方法に付いて検討する。

次に、本研究に関連する情報検索の研究について概観する。

これまでの情報検索は抄録やキーワードなどの情報だけを収めた抄録型データベースの検索が中心であった。が、近年、本文をそのまま全て収めたタイプのデータベース、全文データベースが増加している。全文データベースの増加により、文書中で用いられる全ての文字や語句を検索する全文検索が行なえるようになった。全文検索のためのデータ格納の工夫や、検索アルゴリズムの開発が必要となり、高速な文書検索技術の開発が多く行なわれている。

全文検索において、全文(一つの文書全体)を検索結果として返すのみならず、文書の部分を検索結果とする検索も検討されている。長い文書においては必ずしも一つの話題についてのみ議論されるとは限らない。むしろいくつかの話題について述べられていることが多い。そこで、長い文書を文や段落、節を用いたり、目的や結論などのサブトピックスを用いて、小さな部分(passage)に分割し、その情報を文書群から文書を検索する際に利用する研究や、文書群から文書と文書の部分の両方を検索する研究が行なわれている[Salton 93, Fuller 93]。

[Callan 94]は、一定の単語数を窓長としたpassageを定義し、そのpassageから得られる類似度を手がかりにした検索実験を行なっている。文書全体とpassageのそれぞれから得られる類似度を組み合わせることにより精度向上が得られることを示した。

[Wilkinson 94]は、構造化された多くの文書群から、文書全体ならびに文書の一部を検索することについて定式化を行なっている。文書を構成する節の意味的な種類(目的、概要、要旨、題、捕捉など)に対して、重みづけした評価を行なっている。節のタイプは、人手で付与しているが、節のタイプに対応して重みづけをすることで、精度が向上することを示した。

テキスト中のある部分が何を説明しているかを調べる研究として、他に[黒橋 96]の出現密度分布を用いた語の説明箇所の特定がある。これは検索ではないが、テキスト中で複数箇所出現する語の重要な説明箇所を特定しようとする研究で、本研究で実現しようとしているシステムの立場から見て興味深いものである。

2 マニュアル内容検索システム

2.1 システムの概観

前章でも述べたように、本研究で目指すマニュアル内容検索システムとは、ユーザがマニュアルに対する質問を入力すれば、マニュアルの内容を検索し、マニュアルの読むべき部分を示すシステムである。

マニュアル内容検索システムは、対象とするマニュアルから簡単に作れることが望ましい。そこでマニュアルを用いて単語に重みを与え、これを利用して、検索を行なうことになる。検索結果は、対象マニュアルの中で質問者が調べたい内容が記述されている部分である。検索の単位をセグメントと呼ぶことにする。セグメントとしては例えば、章・節などの形式的まとまり、一定文字数ごとに分割した単位などが考えられる。

マニュアル内容検索システムは、例えば、次のよう

な応答を行なうものとする。ビデオデッキのマニュアルにおける質問と結果の表示の例である。

> 質問文：音声多重放送を録画するとどうなりますか？
> 結果表示：音声の切り換えーステレオテープや二カ国語テープの場合…
>
> 質問文：ビデオデッキを２台つないでビデオテープからビデオテープに録画するには？
> 結果表示：編集のしかた
> 録画したテープを編集すれば、さらにオリジナリティあふれるテープが作れます。
> お子様の成長記録にもう一つのアルバム…

次は、もう少し複雑なマニュアル、日本語形態素解析システムJUMANのマニュアルにおける質問と結果表示の例である。

> 質問文：連接規則辞書って３項関係は記述できないんでしょうか?
> 結果表示：連接規則は二種類の形態素の連接可能性が記述可能である。…
>
> 質問文：辞書にはどんなものがあるんですか？
> 結果表示：３.１ 辞書の概要…

さらに進んだマニュアルの内容検索システムでは、質問とユーザが適切であるとした答を事例として学習し、事例ベース推論の技術を応用して質の良い応答などを目指すことになる。

2.2 システムの動作

内容検索システムの構造を図１に示す。システムはユーザからマニュアルに対する質問文を受け付け、その質問の答となるような読むべき部分を表示する。

まず、対象マニュアルについて、以下の作業を行ない、検索のためのデータベースを用意する。

1. マニュアルを章、節といった構造を元にするなどの方法でセグメント単位に分割する。
2. マニュアルの文を形態素解析システムJUMANを用いて単語に分割する。
3. 名詞および複合名詞と名詞どうしが「の」で連接された名詞句を全て取り出す。
4. セグメントごとの名詞の出現回数を求める。
5. tf.idfによる単語の重みの計算をしておく。

このようなデータベースが実際に用意してある状態で、ユーザが質問文を打ち込むと次のような手順の末、結果を表示する。

1. 質問文を形態素解析システムJUMANにかける。
2. 名詞および複合名詞と名詞どうしが「の」で連接された名詞句を全て取り出す。
3. データベースを用いて質問文から取り出した名詞句と各セグメントのマッチングスコアを計算する。
4. 各セグメントをマッチングスコアの高い順にランキングする。
5. ランキングの順序にしたがって、セグメントを順次表示する。

質問文とマニュアルの両方とも、「の」や名詞からなる名詞句のみを取り出し、検索を行なった。

3 検索部

検索の方法としては標準的な名詞のtf.idfを用いたベクトル空間法と、複合名詞であるという情報を考慮したtf.idf法の２種類で比較検討を行なった。

3.1 標準的な名詞のtf.idfを用いたベクトル空間法

これは広く一般的に行なわれている方法である。まず個々のセグメントに出てくる名詞についてのベクトルを求める。ただし、名詞に与える重みはtf.idf重みとする。名詞NのセグメントS_jにおける重み$w(S_j, N)$の定義式は以下の通りである。

$$w(S_j, N) = tf(S_j, N) \times idf(N) \quad (1)$$

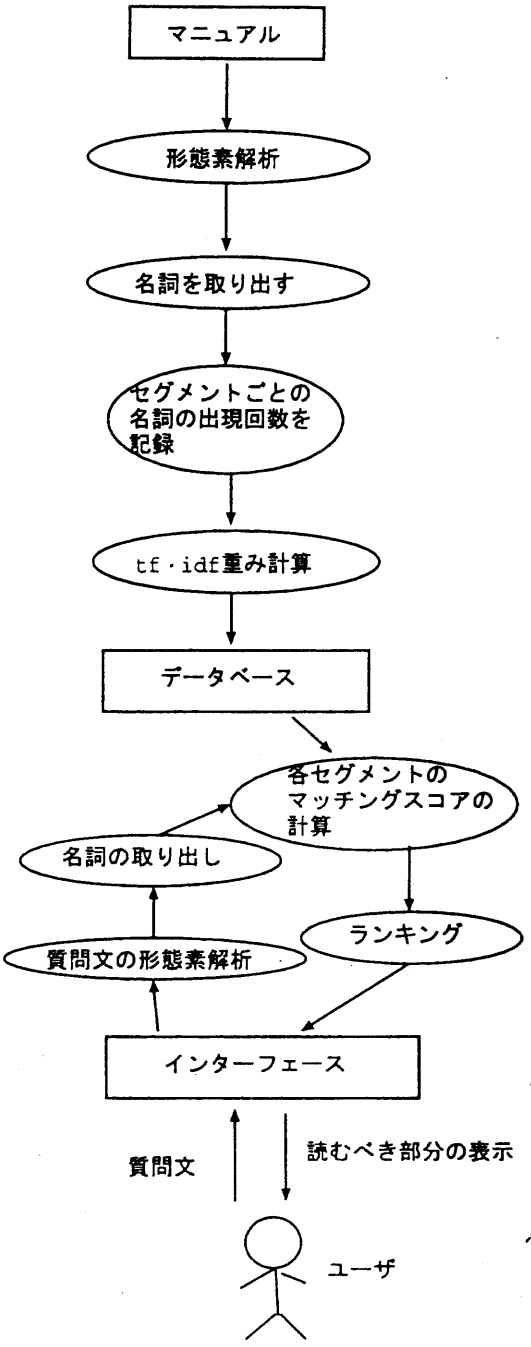

図 1: 内容検索システムの構造

$$idf(N) = \log_2\left(\frac{\#Seg}{freq(N)}\right) + 1 \qquad (2)$$

$tf(S_j, N)$：セグメント S_j における名詞 N の出現回数。
$\#Seg$：マニュアルを分割したセグメント数。
$freq(N)$：名詞 N が出現するセグメントの数。

各セグメントで出現する全ての名詞についてこの値を求める。複合語は複合語として扱わず、構成する名詞の最小単位に分割して扱う。独立して現れた名詞と、複合語として現れた名詞は同等に扱われる。あるセグメントに対するベクトルは、次のように定義される。セグメントに現れた名詞と質問文に現れた名詞をあわせた種類数を次元として持つ。ベクトルの要素が、一つ一つの名詞に対応し、その値はそれぞれのセグメントにおける名詞の tf.idf の値とする。ただしこの際、質問文に現れたが、セグメントに現れない名詞があることがある。そのような場合、その名詞に対する要素には 0 を与える。セグメントに現れた名詞と質問文に現れた名詞が併せて m 種類、名詞 $N_1, N_2, \cdots, N_m$ のとき、セグメント S_j に与えるベクトルは次の式で表される。

$$\overrightarrow{S_j} = (w(S_j, N_1), w(S_j, N_2), \cdots, w(S_j, N_m)) \qquad (3)$$

一方、質問文のベクトルは、セグメントの場合と同様、セグメントに現れた名詞と質問文に現れた名詞をあわせた種類数を次元として持つ。ベクトルの要素は一つ一つの名詞に対応し、それぞれの要素の値は、質問中に現れた名詞は 1、現れない名詞は 0 とする。先ほどと同様、セグメントに現れた名詞と質問文に現れた名詞が併せて m 種類、名詞 $N_1, N_2, \cdots, N_m$ であるとすると、質問 Q に与えるベクトルは次の式で表される。

$$\overrightarrow{Q} = (a(N_1), a(N_2), \cdots, a(N_m)) \qquad (4)$$

$a(N)$：名詞 N が質問中に現れれば 1、現れなければ 0。

この両者のベクトルの $cosine$ で、質問文 Q とセグメント S_j の類似度が求められる。式で表すと次のようになる。

$$Similarity(S_j, Q) = \frac{\overrightarrow{S_j} \cdot \overrightarrow{Q}}{|\overrightarrow{S_j}| \cdot |\overrightarrow{Q}|} \quad (5)$$

3.2 複合語のマッチングを考慮したtf.idf法

前節のベクトル空間法では名詞は単名詞に分割して計算を行ない、複合語どうしの一致については考慮しなかった。質問文の分解結果にも、もちろん検索対象のマニュアル本文中にも複合語が現れる。複合語どうしが完全に一致する事もあるが、片方が他方の複合語の一部となっていたり、部分的に一致する場合もある。このような場合に独立の名詞としてではなく、連続した名詞の一致として扱う方法について述べる。

そこでまず、質問文の分解結果のうちの一つの名詞句と、検索対象のマニュアル中の一つの名詞句のマッチングに注目する。質問文中の名詞句 $W_k^{\mathbf{Q}}(k = 1, 2, \cdots, N)$ の集合を $\mathbf{Q}$、セグメントに分割したマニュアルの j 番めのセグメント S_j 中の名詞句 $W_h^{\mathbf{S}_j}(h = 1, 2, \cdots, M)$ の集合を $\mathbf{S}_j$ とおく。

$$\mathbf{Q} = \{W_k^{\mathbf{Q}} | k = 1, 2, \cdots, N\} \quad (6)$$

$$\mathbf{S}_j = \{W_h^{\mathbf{S}_j} | h = 1, 2, \cdots, M\} \quad (7)$$

$W^{\mathbf{Q}}$、$W_h^{\mathbf{S}_j}$ のそれぞれが名詞 N_i の並びからなる名詞句であるとする。

$$W_k^{\mathbf{Q}} = /N_1/N_2/\cdots/N_n/ \quad (8)$$

$$W_h^{\mathbf{S}_j} = /N_1/N_2/\cdots/N_m/ \quad (9)$$

この2つの名詞句から一致している部分を取り出す。そのさい連続して一致している部分はひとまとまりとして取り出す。例えば次のような場合 (/ は名詞どうしの連接を表す) について考える。

$$W_k^{\mathbf{Q}} = /A/B/C/D/E/ \quad (10)$$

$$W_h^{\mathbf{S}_j} = /B/C/E/ \quad (11)$$

この場合、取り出される部分は $/B/C/$ と $/E/$ である。

なお、ひとつの複合語中に同じ名詞が2回以上出現する場合には、取り出そうとするパターンが重なる場合がある。このような場合は構成する名詞の数が多いパターンを優先して取り出す事にする。例えば、

$$W_k^{\mathbf{Q}} = /A/B/D/B/C/E/ \quad (12)$$

$$W_h^{\mathbf{S}_j} = /A/B/C/E/ \quad (13)$$

のような場合、$/A/B/$ と $/B/C/E/$ というパターンが考えられる。このような場合は $/B/C/E/$ を取り出してから、残りの $/A/B/D/$ と $/A/$ を比較し $/A/$ を取り出す。

さて、取り出したパターン (一致部分) をそれぞれ $P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_1, P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_2, \cdots, P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_r$ とする。上の例では、$P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_1 = /B/C/E/$、$P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_2 = /A/$ となる。ここで、

$$pat(W_k^{\mathbf{Q}}, W_h^{\mathbf{S}_j}) = \{P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_i | i = 1, \cdots, r\} \quad (14)$$

としたとき、セグメント S_j と質問中の名詞句 $W_k^{\mathbf{Q}}$ から得たパターンの集合を次のように表す。

$$\{P(S_j, W_k^{\mathbf{Q}})_i | i = 1, \cdots, R\} = \bigcup_h pat(W_k^{\mathbf{Q}}, W_h^{S_j}) \quad (15)$$

そして、

$$P(S_j, W_k^{\mathbf{Q}})_i = /N_1/N_2/\cdots/N_l/ \quad (16)$$

というパターンに次のような重みを与える。

$$pw(S_j, P(S_j, W_k^{\mathbf{Q}})_i) = tf(S_j, P(S_j, W_k^{\mathbf{Q}})_i) \times idf(P(S_j, W_k^{\mathbf{Q}})_i) \quad (17)$$

パターンのスコアを求める際に、パターンの tf.idf を用いている。tf.idf はパターン全体について求めている。tf はスコアを求めるセグメント内でのパターン $P(S_j, W_k^{\mathbf{Q}})_i$ の出現回数である。idf(Inverse Document Frequency) の計算式は次式を用いた。

$$idf(P(S_j, W_k^{\mathbf{Q}})_i) = (\log_2 \frac{\#Seg}{freq(P(S_j, W_k^{\mathbf{Q}})_i)}) + 1 \quad (18)$$

$\#Seg$：一つのマニュアルが分割されたセグメントの数。

$freq(P(S_j, W_k^{\mathbf{Q}})_i)$：一つのマニュアル中でパターン $P(S_j, W_k^{\mathbf{Q}})_i$ が出現するセグメントの数。

質問中の一語、$W_k^{\mathbf{Q}}$ に対するスコアを次の式で与える。

$$WScore(S_j, W_k^{\mathbf{Q}})$$

表 1: 評価に用いたマニュアルと質問数

マニュアル	size (kB)	質問数
日本語形態素解析システム JUMAN	31	20
構文解析システム SAX	29	24
家庭用ビデオデッキ	69	21
仮名漢字変換フロントエンドプロセッサ「たまご」	57	20

$$= \sum_{l=1}^{R} pw(S_j, P(S_j, W_k^{\mathbf{Q}})_l) \qquad (19)$$

質問文に対するセグメントのスコアは、質問文に出現した全ての名詞句(単名詞および複合名詞)に対するスコアを合計し、セグメントに対して正規化したものとする。

$$SScore(S_j, \mathbf{Q}) = \frac{\sum_{k=1}^{N} WScore(S_j, W_k^{\mathbf{Q}})}{\sqrt{\sum_{l=1}^{M} (pw(S_j, W_h^{\mathbf{S}_j}))^2}} \qquad (20)$$

セグメントに対する正規化は、あらかじめ、セグメントに出現する全ての名詞句を取り出しておき、その全ての名詞句を分解せずにそのままの形での tf.idf を求め、全てを二乗して足して平方根をとった値で割ることによって行なっている。

4 評価

実際に表 1に示すマニュアルについて、それぞれのマニュアルに対し 20 問程度の質問を集め、それに対する正解を人手で調べて検索システムの評価を行なった。ここで問題になるのは検索の単位となるセグメントの決め方である。これに関して、1) 章、節のうち最小の形式的構成要素 (例えばサブセクションなど) を用いる方法、2) 固定長のセグメント、具体的には 10、20、40 行の各々の長さの固定長セグメントを用いる方法、の 2 つについて評価実験を行なった。

ここで、質問に対する正解セグメントの決定法としては、あくまで質問に対して答えているものとし、関連があってもそれだけでは質問の答とならないと判断されたものは不正解とした。このため 1 問の質問に対する正解とされたセグメントの数は最も多いもので 7

表 2: 検索システムが返すセグメントの数

マニュアル	ベクトル空間法			複合語 tf.idf 法		
	min	max	平均	min	max	平均
JUMAN	1	35	22.2	1	35	22.1
SAX	1	23	9.6	1	23	8.2
ビデオ	1	24	15.9	1	24	15.2
たまご	0	43	24.5	0	43	20.5

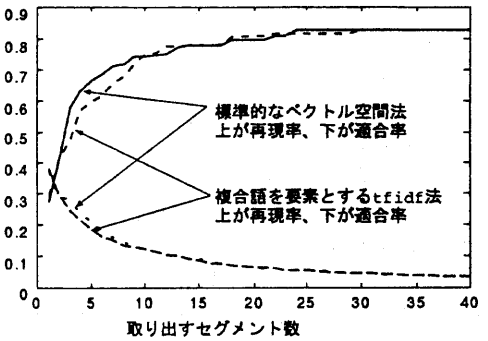

図 2: 最小形式的構成要素をセグメントとする場合で 2 つの方法でスコア上位のものから取り出した際の適合率・再現率

セグメントあった。一方、マニュアル中に適当な正解が無いと判断された質問もあった。このような質問は再現率、適合率ともに 0 として扱った。

1) 最小の形式的構成要素をセグメントとする場合

標準的なベクトル空間法と、複合語を要素とする tf.idf 法によるスコアでセグメントをランキングして上位から取り出していった際の適合率、再現率を調べて評価を行なった。4 つのマニュアル、計 85 問の質問について実験した。それぞれについてスコア 1 位のセグメントを取り出した際の適合率・再現率、スコア 2 位までのセグメントを取り出した際の適合率・再現率、…、スコア n 位までのセグメントを取り出した際の適合率・再現率を求め、質問 1 問あたりの平均値を出した。結果を図 2に示す。

ランキング 1 位のセグメントの適合率・再現率は、ベクトル空間法で 37.5%・27.4%、複合語 tf.idf 法で 38.6%・29.3% であり、複合語 tf.idf 法が勝っている。

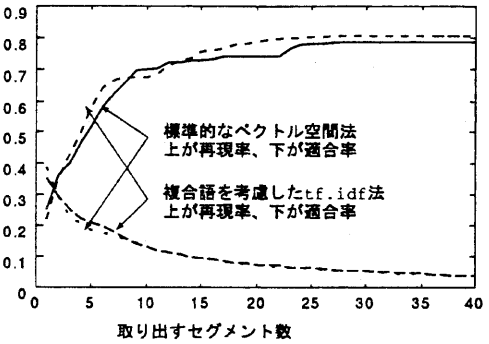

図 3: 固定長セグメント (20 行) を用いた際の適合率・再現率

しかし、グラフの再現率の 2 本の曲線を見ると、ベクトル空間法の方が取り出すセグメント数を増やした際に再現率が急激に上昇することが分かる。実際の値で見ても取り出すセグメント数を 3 セグメント以上にするとベクトル空間法が勝る。一つの質問に対して、システムが検索結果として返すセグメントの数は表 2の通りであった。(検索結果としてはスコアが 0 より大きいセグメントが全て返される。)

2) 固定長セグメントの場合

ここではマニュアルを 10 行ごと、20 行ごと、40 行ごとに、区切った際の適合率、再現率を調べた結果を述べる。ベクトル空間法で、マニュアルを 10 行ごとに区切った際で、ランキング 1 位のものを取り出した適合率 29.1%、再現率 14.6%、同様にマニュアルを 20 行ごとに区切った際で、適合率 38.7%、再現率 24.9%、40 行ごとに区切った際で適合率 36.4%、再現率 26.1% であった。同様にして複合語を考慮した tf.idf 法で、マニュアルを 10 行ごとに区切ると適合率 34.1%、再現率 17.2%、20 行ごとに区切ると適合率 35.3%、再現率 21.9%、40 行ごとに区切ると適合率 37.5%、再現率 25.2% であった。図 3はマニュアルを 20 行ごとに区切ってセグメントとした際の適合率・再現率である。章・節を利用した場合に比べて 5% 程度、再現率が劣っている。

この両方法を比較検討してみる。セグメントを小さく定めると、結果が正しければ、質問に対する答をより局所的に限定できて分かりやすいが、それだけ検索

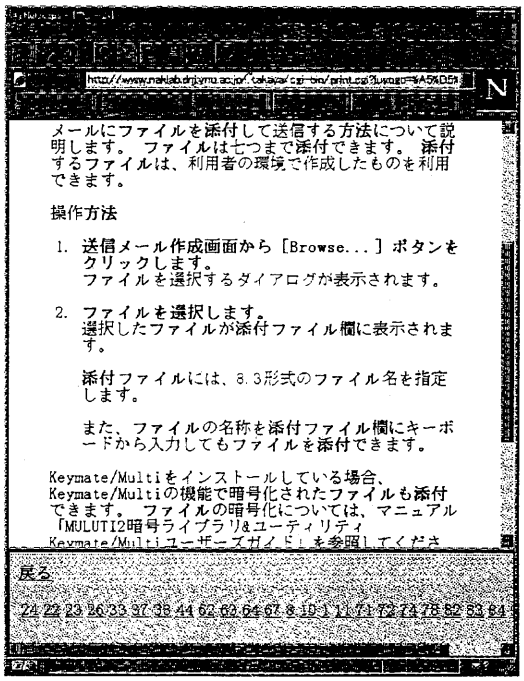

図 4: インターフェース画面 (日立製作所より提供頂いたマニュアル「Groupmax World Wide Web ユーザーズガイド」より)

の精度をあげるのが難しい。また結果には意味的なまとまりがあるとは限らないので非常に読みにくいものとなってしまうことがある。逆に、マニュアルのライターが定めた章や節には意味的なまとまりがあることが予想され、検索結果としても読みやすい。しかし、場合によってはセグメントが長いものとなってしまい、セグメントを指定されただけでは答を述べている部分を探すのに手間がかかってしまう場合がある。このような問題はセグメント表示インターフェースを工夫して解決することを試みた。以下にこれについて述べる。

5 セグメント表示インターフェース

HTML で書かれたマニュアルに対し HTML ブラウザを用いてユーザからの質問文入力と検索結果の表示を行なった。検索プログラムは cgi プログラムに

よって呼び出される形になっている。質問文はフォームを使って入力される。cgiプログラムは入力された質問文を検索プログラムに渡し、質問文中の名詞および名詞句と、検索結果としてランキングされたセグメント番号を得る。ブラウザにランキングされた順にセグメント番号を表示し、セグメント番号をクリックするとマニュアルのそのセグメントが表示される。その際にセグメントの先頭からではなく、そのセグメントではじめて出現する質問文中の名詞および名詞句の周辺部分から表示を行なう。これは質問文中の名詞が存在している文がユーザの目的の部分であると仮定して、質問文中の名詞がない文は目的の部分ではないと判断し、読む必要のない部分を省いて目的の部分から表示するためである。また、質問文中の名詞を手がかりにしてユーザが迅速に目的の部分を見つけられるように質問文中の名詞および名詞句を目立つように太文字表示してユーザの利便を図っている。例として質問文に「ファイルの添付方法」を入力した場合の表示結果を図4に示す。上のフレームではセグメント中の「ファイル」「添付」「方法」が太文字になっていて、その語が存在する文から表示されている。下のフレームではセグメント番号がランキング順に表示されている。

6 おわりに

マニュアル内容検索システムを提案し、内容検索の評価実験を行なった。表示インターフェースの改善と適合率、再現率の改善が今後の課題である。

謝辞

本研究はIPAの創造的ソフトウェアプロジェクトの援助を受けている。

実験に用いたマニュアルの一部を株式会社日立製作所より提供して頂いている。

参考文献

[WAIS 93] WAIS : WAIS Server, WAIS Workstation, WAIS Forwarder for UNIX, WAIS Inc., Technical Description Release 1.1 (1993).

[Callan 94] Callan, J. P. : Passage-Level Evidence in Document Retrieval, Proc. 17th ACM SIGIR, pp.302-310(1994).

[Wilkinson 94] Wilkinson, R. : Effectibe Retrieval of Structured Documents, Proc. 17th ACM SIGIR, pp.311-317 (1994).

[黒橋 96] 黒橋 禎夫, 白木 伸征, 長尾 眞: 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 NL 研究会 -115-7, pp.43-50(1996).

[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).

[Fuller 93] Michael Fuller, Eric Mackie, Ron Sacks-Davis, Ross Wilkinson: Structured Ansers for a Large Structured Document Collection, 16th ACM SIGIR, pp.204-213(1993)