2008－NL－184 (1)
2008／3／27

# 確率伝播法を用いた文書検索用キーワードの自動抽出

三上崇志　相川勇之　平野敬　岡田康裕
三菱電機株式会社 情報技術総合研究所

カーナビや携帯電話などの電子機器の高機能化に伴い，製品の操作説明書などを電子化して機器上で検索・閲覧するニーズが高まっている．そのためキーボードがなく文字入力が困難な機器上でも，簡単に文書を検索できるインターフェースが求められている．そこで本報告では，ユーザの読み入力に応じてキーワードを自動提示するインターフェースの実現を目的として，検索対象文書からキーワードを自動抽出する方式を提案する．本方式は対象文書の論理構造を利用したブートストラップ手法により重要語句を求め，確率伝播法を用いたスコア付けにより重要語句からキーワードを高精度に抽出する．機器の操作説明書を用いた評価の結果，人間が説明書を読んで人手で抽出したキーワードの 84%を本方式により自動抽出できた．また，読みを 1 文字入力して得た上位 20 個のキーワードに対して 70%の適合率を得た．

## Automatic Extraction of Keywords for Document Retrieval with Belief Propagation

Takashi Mikami　Takeyuki Aikawa　Takashi Hirano　Yasuhiro Okada
Information Technology R&D Center, Mitsubishi Electric Corporation

Demand for retrieval and browsing of the electronic operating manuals on the multifunctional equipments such as the car navigation systems and the mobile phones is certainly booming. The interface to facilitate retrieval of the documents is required, that can work on the keyboard-less equipments whose character input is difficult. This report proposed a method for automatically extracting important phrases from the documents for retrieval, in order to achieve the interface that presents the keywords correspond to several characters input by the user. Our proposed method extracts the important phrases by the bootstrap procedure and calculates their score by belief propagation, then extracts keywords from them. We have applied our proposed method to an operating manual and have confirmed that 84 percent of keywords can be automatically extracted, compared with manually extracted keywords. We also have confirmed to achieve about 70 percent precision for 20 keywords correspond to one character input.

## 1. はじめに

携帯電話やカーナビなどの電子機器の高機能化が進み，製品の操作説明書などを電子化して機器上で検索・閲覧するニーズが高まっている．そこで我々は，キーボードがなく文字入力が困難な機器上でも，簡単な操作で操作説明書などの文書検索を可能とするため，読みを数文字入力するだけで検索用のキーワード候補を自動提示する入力サジェスト機能を開発している．

読みからキーワード候補を自動提示する機能はインターネットの全文検索サイトで実用化されているが，これらは検索クエリのログや，Web 上の膨大な文書データから得た単語出現頻度の偏りを用いる方式が主である．内容が様々で，文書量が膨大である Web においては，コンテンツからキーワード候補を抽出するより検索クエリのログから取得する方が効率的であると思われる．

しかし，機器上で操作説明書などの文書を検索する用途では，検索頻度が小さく，記憶容量も小さいため十分な検索ログを収集できない．このような場合，検索対象文書が特定できて文書量が少ないため，検索対象とする文書中からキーワード候補を抽出した方が効率的である．

本稿では，入力サジェストに適したキーワード候補を高精度に抽出するため，検索対象文書から重要語句を自動抽出する方式を提案し，その有効性を検証する．提案方式は，対象文書の論理構造を利用したブートストラップ手法および確率伝播法を用いたスコア更新によりキーワード候補を高精度に抽出する．

以降，2章で関連研究について述べ，3章で提案する方式を説明する．4章で実験による提案方式の検証を行い，5章でまとめと今後の課題を述べる．

## 2. 関連研究

Web 検索においては，入力した文字に関連した検索キーワードを提示するサービスとして「Google サジェスト」[1]や，「Yahoo! 関連検索ワード」[2]などがある．詳細なアルゴリズムは公開されていないが，これらのサイトの FAQ やヘルプの内容から，いずれも検索クエリのログに含まれる検索キーワードの統計情報を用いていると思われる．また，大塚ら[3]は統計的に偏りなくユーザを抽出し，その詳細なアクセスログを用いて関連語を抽出する方式を提案している．これらの手法は大規模な検索ログが得られる Web 検索では有効であると考えられるが，十分な検索ログを収集できない場合には適用が困難である．本研究の対象である電子機器では，ユーザが一人の場合が多く，検索頻度も小さい，さらに機器の記憶容量も小さいため，十分な検索ログの収集が難しい．

検索クエリのログを利用しないキーワード候補提示機能として，「goo サジェスト$\beta$ with ATOK」が Web 上で試験運用されている[4]．このサイトは人名やランドマークなどのカテゴリ別に，予め作成された辞書を利用して

キーワード候補を提示していると思われる．Web 検索などの広範囲に渡る一般的な検索用途に対しては，一般的な辞書を利用したキーワード提示は有効である．ただし，特定文書の検索に特化した検索キーワードを提示する目的のためには，対象文書からキーワードを自動抽出する方が，高精度にキーワードを提示できると考えられる．

このような文書からキーワードを抽出する研究として，松尾らの先行研究がある[5]．松尾らは，本研究と同様に，単語共起情報を用いて文書からネットワークを構築してキーワード抽出を行っている．しかし対象が英文であるため，日本語の解析に特有な形態素解析や複合語解析による曖昧性を考慮していない．またフォントやレイアウト情報も利用しておらず，これらの点が本研究と異なる．

## 3. 提案方式

操作説明書における章や節の見出しには，ユーザが参照する際の利便性を高めるため，機能や部品に関する短くわかりやすい表現が利用される．本研究では見出しをキーとして，以下の仮説に基づきキーワードを抽出する．

仮説 1：見出し中の語句やフォントサイズが大きい語句は重要である
仮説 2：重要語句と同じ文内に出現する語句も重要である可能性が高い

上記仮説に従い，キーワードの抽出を 2 段階の処理により実現する．まず，節の見出しを種情報とするブートストラップ手法によりキーワード候補となる重要語句を文書中から取得し，重要語句間の関連性を表すネットワーク構造を生成する．この際，フォントサイズなどのレイアウト情報と表層格や禁止語などの言語情報に従って各語句に重要性を示すスコアを付与する．

次に，節の見出しやフォントサイズが大きい語句は明らかに重要語句であるとして，これを観測情報とした確率伝播法[6]により各語句のスコアを更新する．これにより仮説 1 と仮説 2 に従った重要語句の抽出を実現する．

最後に，抽出された重要語句に読みを付与し，入力サジェスト用の辞書に格納する．実機でユーザが読みを入力した際は，これと同じ読みを持つ重要語句をスコアが高い順にキーワード候補として提示する．これによりユーザは簡単に文書検索を実現できる．

提案方式のポイントは，上記の重要語句の抽出法とスコア計算法にある．以降，3.1節で重要語句の抽出とスコア付与について述べ，3.2節で確率伝播法を用いたスコア更新について説明する．

### 3.1. 重要語句の抽出とスコア付与

ここでは，図 1 に示すアルゴリズムに従いキーワード候補となる重要語句を抽出する．

(1) レイアウト解析 / テキスト解析

まず，レイアウト解析処理により PDF 形式の文書ファイルの内容を解析し，フォント情報やページ内位置情報を持ったテキストをブロック単位（行単位）で抽出する[7]．次に，連続したテキストブロックを連結する．この際，表や図の中にある独立したテキストを過剰に連結しないよう，位置関係を考慮して連結判定を行う．その後，連結後のテキストブロックに対して形態素解析および文節内解析処理を適用し，テキストを文節単位に分割する．

(2) 語句候補抽出

テキストブロックに対して，上記で分割された文節単位を順に走査し，連続した 1 文節～5 文節からなる語句候補を抽出する．このとき，括弧表現を読み飛ばすと共に，カンマや中黒による並列表現の抽出なども行う．

(3) n 次自立語抽出

最初（n=1）では，目次情報（節の見出し）から種となる 0 次重要語句を取得する．そして 0 次重要語句の形態素解析結果から自立語を抽出して，これを 1 次自立語とする．n≧2 の場合は，次の(4)の処理で抽出された n-1 次重要語句に含まれる自立語のうち，これまでに抽出されていない自立語を n 次自立語として抽出する．

(4) n 次語句抽出

上記(3)の処理で得た n 次自立語を用いて，同じ節内のテキストブロックから n 次自立語を含む語句を探す．見つかった語句のうち，これまでに抽出されていない語句を n 次重要語句として抽出する．以下，ブートストラップ的に(3)および(4)の処理を繰り返す．

ここで n=1 の場合のみ，1 次自立語を含む語句だけではなく，フォントサイズが大きな語句を 1 次重要語句に追加する．これは，箇条書きの見出しなど，節の見出しではないが重要性の高い語句の抽出漏れを少なくするためである．

上記(3)～(4)の処理を繰り返すことで，ブートストラップ手法により重要語句のネットワークを構成する．図 2 に示す仮想的なカーナビの操作説明書に上記の処理を適用して得た重要語句のネットワーク例を図 3 に示す．

(5) 重要度スコア計算

各重要語句のスコア $S_0$ の計算は式(1)により行う．

$$\text{重要度スコア } S_0 = S_1 \times S_2 \times S_3 \times S_4 \qquad (1)$$

$S_1$：フォントスコア，$S_2$：文字数スコア，
$S_3$：表層格スコア，$S_4$：禁止語スコア，

ここで，$S_1$ は文書中の最大フォントサイズで正規化した対象語句のフォントサイズを，$S_2$ は一般的なキーワードの文字数と対象語句の文字数との差異を表す．$S_3$ は用言で始まる抽出語句に与えるペナルティであり，直前の 2 文節を走査して「が」や「を」が検出された場合は意味的なまとまりが低い語句と判定して値を下げる．$S_4$ は禁止語ペナルティである．例えば，「上記」「下図」など重要語句に含まれにくい語を登録した禁止語辞書を用いて，これらを含む語句の値を下げる．$S_1$～$S_4$ はそれぞれ 0～1 の値を取り，対象語句の重要度が高い場合，$S_0$ の値は最大 1 となる．

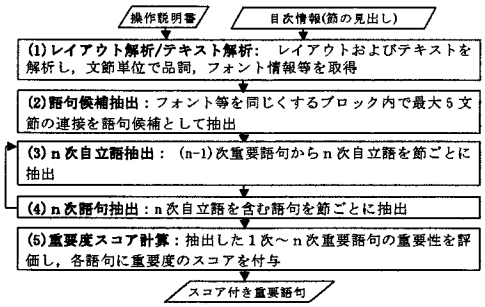


**図1　重要語句抽出アルゴリズム**

**5. 2節　観光地のルートを設定する**

各都道府県の観光地をまわるルートを設定することができます.

[1] メニューボタンを押す

[2] "観光地ルートの設定"を選択し
　　実行ボタンを押す(右図)

:

写真・文字情報のみかた

観光地についての写真情報や文字情報を表示できます.

**図2　操作説明書の例**

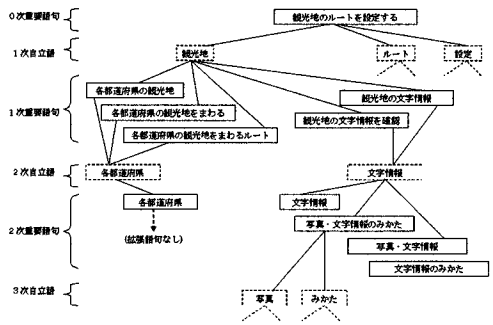


**図3　重要語句のネットワーク構造**

## 3.2.　確率伝播法を用いたスコア更新

確率伝播法は，確率ネットワークにおける周辺確率を効率よく計算するためのアルゴリズムである．確率ネットワークとは，確率変数をノードとし，そのノード間に確率変数間の依存関係に従って向きを持つアークでリンクした有向グラフである．

図4に単純な確率ネットワークの例を示す．条件付確率として $P(B|A), P(C|B), P(C|A,B)$，および事前確率として $P(A), P(B)$ が与えられれば，確率伝播法により同時確率分布 $P(A,B,C)$ を求めることができる．一部のノードの値が観測されて固定されれば，リンクを辿って周辺確率を計算していき，結果として各ノードの事後確率を求めることができる．確率伝播法は単結合ネットワークに対しては計算が収束し，真値を求めることができるが，複結合ネットワークに対しては計算が収束しない．しかし，多くの場合は近似解に収束し，収束しない場合は値が振動することが実験的に示されている[8].

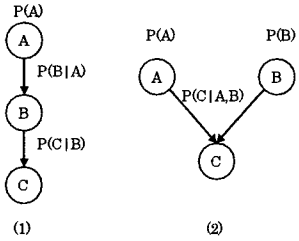


**図4　単純な確率ネットワークの 例**

本研究では，3.1節で作成した重要語句のネットワークに対して確率伝播法を適用することでスコアの精度を向上させる．ネットワーク上のノードの幾つかを重要語句として観測できれば，仮説2に従って他の重要語句のスコアを底上げできる．

確率伝播法を適用するため，重要語句のネットワークにおける各ノードに確率変数 $x$ を導入し，$x$ は次の値を取るとする．

$T$ : 重要語句

$F$ : 非重要語句

例えば $P(T)=1$ のノードは重要語句であり，$P(T)=0$ のノードは非重要語句とみなす．

今回行った実験では，親ノードに $U_1, U_2, \cdots, U_n$ を持つ子ノード $X$ の条件付確率を次のように定義した．

$$P_X(x|u_1,u_2,\cdots,u_n) = P_{U_1\to X}(x|u_1)P_{U_2\to X}(x|u_2)\cdots P_{U_n\to X}(x|u_n) \quad (2)$$

ここで，$x$ は子ノード $X$ の状態，$u_i$ は親ノード $U_i$ の状態を表す．$P_{U_i\to X}(x|u_i)$ は次式で計算する．

$$P_{U\to X}(T|T) = Sim/(Sim+1-S_0) \quad (3)$$

$$P_{U\to X}(F|T) = (1-S_0)/(Sim+1-S_0) \quad (4)$$

$$P_{U\to X}(T|F) = S_0/(Sim+S_0) \quad (5)$$

$$P_{U\to X}(F|F) = Sim/(Sim+S_0) \quad (6)$$

$Sim$ は親ノード $U$ と子ノード $X$ の類似度，$S_0$ は3.1節の方式で求めた子ノード $X$ が持つ重要度スコアである．ここで類似度 $Sim$ は二つのノードの重要語句文字列に含まれる自立語が一致する割合とした．$Sim$ の値は 0～1 の値を取り一致する自立語が多いと 1 に近い値を取る．

仮説 1 に従い，0 次重要語句に該当するノードが持つ $P(T)$ の値を 1 であるとし，これを観測値として確率伝播法により他のノードの事後確率を計算する．式(3)によれば，親ノードが重要語句である確率が高く，子ノードとの類似度が高い場合，子ノードが重要語句である確率も高くなる．また式(6)から，親ノードが非重要語句である確率が高く，類似度が高い場合，子ノードも非重要語句である確率が高くなる．親ノードと子ノードの類似度が低い場合は，式(4)，(5)が支配的になり，子ノードが重要語句であるか否かは $S_0$ に依存して決定される．伝播は子ノードから親ノードに対しても行われるため，重要語句である可能性が高くて類似度の高い子ノードを多く持つ親ノードは，重要語句である確率が高くなる．

なお，条件付確率を上記のように計算するため，全体の計算量は 1 つのノードから張られる親ノードへのリンク数について指数関数的に増える．そのため，本報告では 1 つのノードが持つリンク数に上限を設けた．リンク数上限の影響については4.3.2で述べる．

また，本研究におけるネットワークは複結合ネットワークであり，確率伝播法の計算は収束しないため，有限回の計算で打ち切っている．以下で議論していないが，今回の実験の範囲では，計算によって得られた事後確率は振動せず，一定の値に収束していた．

このようにして得られた，伝播後の各ノードが持つ事後確率 $P(T)$ を最終的な重要語句のスコア $S$ とみなす．

# 4. 実験

## 4.1. 対象データ

実験対象として， FA(Factory Automation)用装置の操作説明書（PDF 形式）を用いた．操作説明書の詳細情報は表 1の通りである．また，4.2，4.3の実験の評価用に，人手により正解データを作成した．これは表 1の操作説明書を 2 名の被験者が読み，「同説明書を検索する場合のキーワードとして利用される」と判断した重要語句の集合である．これらの重要語句は単語や複合語を含み，合計 1083 個存在する．

表 1　FA 用装置の操作説明書

| | |
|---|---|
| 言語 | 日本語 |
| ページ数 | 268 ページ |
| 文字数 | 約 17.3 万文字 |

## 4.2. 重要語句のネットワーク構造

表 1の操作説明書に対して，3.1節で示した方式によりキーワード候補となる重要語句を取得し，重要語句のネットワークを構築した．

重要語句の抽出においては正解データ 1083 語句のうち 909 語句を取得できた（再現率 84%）．抽出できなかった重要語句 16%の多くは、表のレイアウト解析の失敗や操作説明書作成者が体裁を整えるために挿入した空白文字による形態素解析失敗が原因であった．

構築されたネットワークの全ノード数，0 次重要語句数，全リンク数は表 2に示す通りである．ただし，ノードは同じ文字列表現のものも重複を許してカウントしている．これは章・節が異なれば同じ文字列も違うノードとして扱うためである．またノード中には形態素解析・複合語解析のミスにより誤抽出された重要語句も多数存在した．

表 2　ネットワーク構成

| | |
|---|---|
| 全ノード数 | 31479 |
| 0 次重要語句数 | 138 |
| 全リンク数 | 125964 |

## 4.3. 精度評価

### 4.3.1. 重要語句抽出精度

本節では提案方式の抽出精度を評価するため，実験対象の文書から人手で抽出した重要語句集合（1083 語句）と，4.2で自動抽出した重要語句集合とを比較する．

#### 4.3.1.1. 重要語句抽出精度の評価手順

人手で抽出した重要語句集合を正解として伝播前と，伝播後でのスコアを用いた再現率・適合率を求めた．ただし，確率伝播処理では 1 つのノードからのリンク数上限を親ノード数・子ノード数とも 10 とした．

#### 4.3.1.2. 重要語句抽出精度の評価結果および考察

図 5に伝播前，伝播後での再現率-適合率のグラフを示す．伝播前のグラフを見ると，スコアの値が大きい箇所（再現率が低い箇所）では適合率が上がっており，3.1節の方式によるスコア付けが妥当であることがわかる．また，伝播後は，伝播前に比べて適合率が改善している．これは 0 次重要語句を重要語句として観測したことが意図通りに伝播し，スコアの精度向上が行われたためと判断できる．

ただし，再現率が低い箇所では適合率が 0.1 程度向上するだけであり，特に再現率が 0.4 以上の範囲ではほとんど伝播の効果が得られていない．これは 0 次重要語句の数がネットワークの規模に対して非常に少なく，末端のノードへの伝播の効果が低いためと考えられる．しかし，本方式の適用目的である入力サジェスト機能では，スコアが高い順にキーワードを提示するため高スコア領域での適合率向上が重要である．この点については4.3.3で詳細に述べる．

なお，伝播後の精度をさらに高める対策として，0 次重要語句の数を増やすことが考えられるが，この数は文書の構造によって一意に決定されるため難しい．そのため 0 次重要語句の他に有効な観測方法を検討する必要がある．このような観測方法として，予めキーワードとなりそうな用語リストを与えて，これと一致するノードを $P(T)$=1 としたり，抽出された幾つかのノードに人手で $P(T)$ の値を指定するといった手法が考えられる．

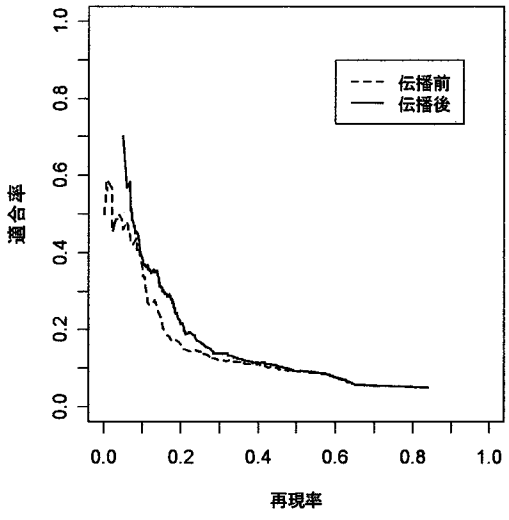


図 5　再現率・適合率

### 4.3.2.　性能評価（リンク数の影響）

3.1節で得たネットワークに対してそのまま確率伝播法を適用すると計算時間がかかる．今回の実験では各ノードからのリンク数に上限を設け，上限数以下になるようにリンクをカットした．リンク数に上限を設けたことによる抽出精度への影響を調べるため，リンク数の上限を変化させ、速度性能および抽出精度との関係を調べた．

#### 4.3.2.1.　性能評価手順

親ノードの数，子ノードの数それぞれにリンク数上限を設け，4.3.1と同様の実験により適合率の変化を検証した．親ノード数を 10，5，子ノード数を 20，10，5，として全ての組合せについて評価した．リンク数上限を超えたノードでは，伝播前のスコア $S_0$ の小さいものから順に親ノードおよび子ノードへのリンクをカットした．

#### 4.3.2.2.　性能評価結果および考察

全リンク数と確率伝播処理の計算時間の関係を表 3に示す．ここで全リンク数は，上限数によるカットを行い残ったネットワーク中のリンク数である．表 2に比べてリンク数が半分以下に減っていることが分かる．図 6はリンク数をパラメータとした再現率・適合率の測定結果を示す．図から分かるように，今回の実験範囲ではリンク数制限による影響はほとんどない．

今回用いたネットワークでは親ノードと同じ自立語を含むものを重要語句として抽出してネットワークを構築しているため，親ノード同士，子ノード同士は文字列表現が似ている．その中で伝播前のスコア $S_0$ の低いものからリンクをカットしたため，リンク数を 20 に制限した時点で偏ったノード集合が残ったのではないかと思われる．リンクのカット方法については伝播前のスコア $S_0$ による足切り以外に今後検討する必要がある．

表 3　リンクカットによる全リンク数と計算時間の違い

| 親ノード数 | 子ノード数 | 全リンク数 | 計算時間(秒) |
|---|---|---|---|
| 10 | 20 | 60918 | 25 |
| | 10 | 45395 | 21 |
| | 5 | 30916 | 17 |
| 5 | 20 | 44715 | 12 |
| | 10 | 35158 | 12 |
| | 5 | 25319 | 13 |

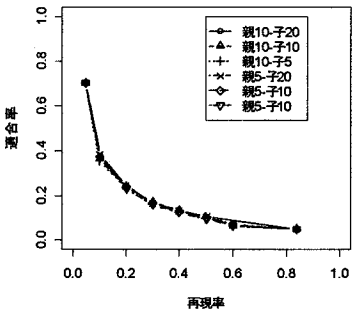


図 6　リンク数上限による比較

### 4.3.3.　キーワードの適合率評価

現在開発している入力サジェスト機能では，ユーザによって入力された「読み」に前方一致する重要語句をスコア $S$ でソートし，上位 $N$ 個をキーワード候補として提示する．ユーザは提示されたキーワードを選択し，文書検索を行う．本節では平仮名 1文字を入力した際に提示されるキーワード候補を目視確認することで，有効な検索キーワードが提示されているかを実験により検証する．

#### 4.3.3.1.　キーワードの適合率評価手順

4.2で自動抽出した重要語句集合から，撥音・拗音・小文字（“ぁ”，“ゃ”，“っ”など）を除く“あ”～“わ”の平仮名 68 種について読みが前方一致する語句を取得し，そのうちスコア $S$ の値が大きい上位 20 個をキーワード候補として選出する．平仮名ごとに選出されたキーワードを被験者が目視で確認し，検索語として適切なキーワードか否かを判断する．ここでは，キーワード集合のうち，適切と判断されたなキーワードの割合を適合率と定義した．例えば上位 10 個で適合率が 50%なら，10 個のうち 5 個が適切なキーワードで 5 個が検索には利用されない語句（ノイズ）であることを意味する．

本評価では，確率伝播前のスコア $S_0$ と伝播後のスコア $S$ を用いた場合とで適合率を比較する．また，ベースラインとしてスコアを利用せずに文字コード順に提示した場合も評価する．

なお，本実験では，キーワードの正否判断を被験者 3人によって行い，4.3.1と同様に確率伝播処理では 1 つのノードからのリンク数上限を親ノード数・子ノード数とも 10 とした．

#### 4.3.3.2. キーワードの適合率評価結果および考察

3 人の被験者による適合率の算出結果を図 7に示す．図から分かるように伝播前，伝播後の双方で，文字コード順よりも良い結果を示した．提示キーワードの個数が 2 個以上のすべての場合において，伝播前よりも伝播後の方が高い適合率を示した．伝播後の適合率は，上位 5 個で約 73%，20 個以上で約 70%となった．この結果から，確率伝播法を用いたスコア更新が，入力サジェスト機能のキーワード候補提示に対して有効であることが示された．

なお，単純に前方一致するキーワードを文字コード順に提示しても適合率が 50%を超えたが，これは，3.1節の方式において単純な形態素解析結果ではなく複合語単位で重要語句を取得している影響と考えられる．

ただし，スコアの上位 1 個のみをキーワードとして提示した場合は，伝播前の方が良い結果となっている．これは，今回の実験で 0 次重要語句を無条件にキーワードとした影響と考えられる．実際に，0 次重要語句（＝節の見出し）の中にもノイズとなる語句が混じっており，これが誤って伝播することの悪影響も存在している．

表 4 は，伝播後のスコア $S$ を用いた場合の「あ」および「い」で提示されるキーワードの例を示す．表中の「間保持」は形態素解析の失敗に起因した誤りである．このような誤りは3.1における形態素解析精度を高めることで抑制できる．

また，今回の実験では引用符などの記号や物理量の単位を含む誤った重要語句も多数キーワード候補に含まれた．これらは禁止語を設けるなどのヒューリスティクスによって除去できると考える．

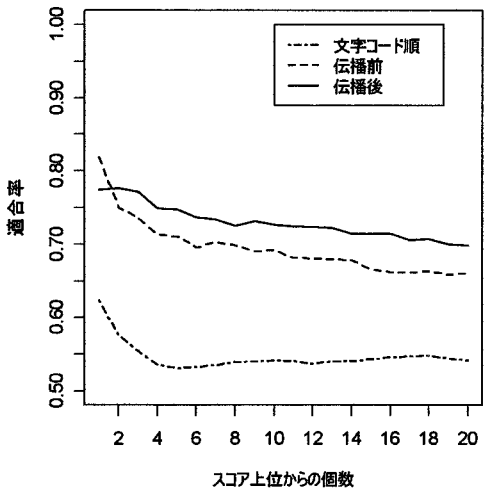


図 7　提示キーワードの適合率

表 4　提示キーワード例（上位 5 個）

| 読み「あ」を入力 | 読み「い」を入力 |
|---|---|
| アタッチメント | 一般使用 |
| アラーム情報 | インストール |
| アラーム情報の機能 | 位置補正 |
| 間保持 | 異常が発生 |
| アタッチメントの種類 | インストールと削除 |

## 5.　まとめと今後の課題

本稿では，対象文書の論理構造を利用したブートストラップ手法により重要語句の候補を求め，確率伝播法を用いたスコア更新により重要語句を高精度に抽出する方式を提案した．FA 用機器の操作説明書を用いて実験を行い，その有効性を確認できた．

今回の実験では対象文書が 1 種類であり，本方式の汎用性について評価を行えていない．今後，他の文書に対しても同様の実験を行い，評価を充実化する予定である．

確率伝播法によるスコア精度の向上では，0 次重要語句による観測しか導入できておらず，精度を向上させるための観測情報としては不十分であると考える．0 次重要語句以外に重要語句を導入するか，逆にノイズとなるようなキーワードを非重要語句として観測する，などの改良についても今後検討する予定である．また，伝播前のスコア $S_0$ の精度を上げれば伝播後のスコア $S$ の精度も向上すると考えられる．今回は伝播前のスコア付けにおいて文書中における出現頻度を考慮していないが，現在 TF-IDF のような手法を用いた，頻度情報も加味したスコア付けを検討している．

## 参考文献


[1] Google サジェスト：http://www.google.co.jp/webhp?complete=1&hl=ja

[2] Yahoo!関連検索ワード：http://developer.yahoo.co.jp/search/webunit/V1/webunitSearch.html

[3] 大塚真吾，喜連川優：大規模アクセスログ，日本データベース学会 Letters, Vol.5, No.1, pp.13-16 (2006).

[4] goo サジェスト $\beta$ with ATOK：http://suggest.search.goo.ne.jp/suggest/

[5] 松尾豊，大澤幸夫，石塚満：Small World 構造に基づく文書からのキーワード抽出，情報処理学会論文誌, Vol.43, No.6, pp.1825-1833 (2002)

[6] 麻生英樹，津田宏治，村田昇：*パターン認識と学習の統計学*，統計科学のフロンティア 6，岩波書店 (2003)．

[7] 平野敬，岡野祐一，岡田康裕，依田文夫：ページ記述言語の解析に基づく多様な文書からの構造化内容情報の抽出”，信学論 D-II, (2008 年 5 月号掲載予定)

[8] Kevin P. Murphy, Yair Weiss, Michael I. Jordan : Loopy Belief Propagation for Approximate Inference: An Empirical Study, Proceedings of Uncertainty in AI (1999)