データベースシステム 99－9
（1994．7．20）

統計量を用いた電話帳データベース検索方式

永吉 剛　岩瀬 成人
ＮＴＴ情報通信網研究所

曖昧な問い合わせ情報を対象とする電話帳データベースの検索方式について報告する。通信先の掲載情報を検索する際、通信先の登録データと利用者が入力する条件の間には表記あるいは単語、意味レベルで様々な不一致が生じる．従来の検索システムは検索結果の検索条件に対する適合性を判断できないため、利用者が適合性を判断し、必要に応じて条件を修正し再検索していた．本稿は利用者による条件修正の成功率が低いことを実験によって示す．これを勘案して検索条件がデータベース上に持つ情報量を定義し、検索結果の適合性を確率的に推定する手法を提案する．本手法により、曖昧な問い合わせに対する検索システムの検索成功率は従来の6倍となり、大きな効果があることを示す．

A Probabilistic Information Retrieval Method for Telephone Directory System

Takeshi Nagayoshi and Shigehito Iwase
NTT Network Information Systems Laboratories

This paper proposes a probabilistic retrieval method for a telephone directory database system. There are differences in syntax, word, and semantics between query expressions and record values stored in the system, which result in mis-retrieval. Formerly a retrieval system could not evaluate the relevancy of retrieved data so that a human operator had to evaluate the relevancy. Experiments on operator's behavior of try-and-error modification of query resulted that an operator hardly succeed in finding correct modification. In order to solve this problem, the paper define information content of query and proposes a probabilistic method to estimate the relevancy of retrieved data. By applying the proposed method for the queries which have partial differences, the success rate of retrieval is improved as 6 times high as that obtained before.

1．まえがき

通信先の番号を調べるために、通信端末から直接電話帳データベース検索システムにアクセスし名義と住所から検索できるサービス（ＡＮＧＥＬＬＩＮＥ）が提供されている．使いやすいインタフェースと通信機能を持つパーソナルコンピュータや携帯情報端末の普及に伴い、今後このようなサービスは盛んに利用されると考えられる．

電話帳データベース検索サービスは、データやシステムに関して専門的な知識を持たない利用者からの多様な言語表現による問い合わせを受ける．このため目的の掲載データの名義、住所、職業といった検索キーに対して入力される検索条件は、表記や単語、意味レベルでの様々な不一致表現を生じ、一回の条件入力では検索できないことが多い．本稿では問い合わせ情報に含まれるこのような不一致表現を、問い合わせの”曖昧さ”と呼ぶ．

曖昧さの解消は、検索システム側での処理と、利用者による試行錯誤的な条件修正によって相補的に行われる．本稿ではまず利用者の条件修正行動が非効率的で失敗する可能性が高く、利用者に対する負担が大きいことを実験によって示す．利用者が試行錯誤的な条件修正を行わなければなら

表１．電話帳データベース検索システムの問い合わせにみられる曖昧さ

	曖昧さ	検索キー	問い合わせ例
仮名表記	清音と濁音	ヤマザキ	ヤマサキ
	長音、二重母音	ニューオータニ	ニューオオタニ、ニュウオウタニ
	ジとヂ、ズとヅ	イズミ	イヅミ
	音の変化	アマミヤ	アメミヤ
	漢字の読み	タイラバヤシ	ヒラバヤシ
	アルファベット	エヌエイチケイ（ＮＨＫ）	エヌエッチケー
	外来語	ヴァージンアトランティック航空	バージンアトランチック航空
	数字	イチマルキュウ（１０９）	イチレイキュウ、ワンオーナイン
名義	単語の欠落	国立近代美術館	近代美術館
		高橋克己法律事務所	高橋法律事務所
	単語順序の変化	鮨銀	銀鮨
		電子通信情報学会	電子情報通信学会
	単語の変化	高橋法律事務所	高橋弁護士事務所
		武蔵野医院	武蔵野病院、武蔵野内科
	通称	日本電信電話株式会社	ＮＴＴ，電電公社
		通商産業省	通産省
	読み違い（視覚的類似）	永吉	吉永
	聞き違い（聴覚的類似）	ロンロン	ロンドン
	勘違い	山田さん	田中さん（山田さんに顔が似ている）
住所	近隣関係	上野３丁目	外神田
	近くの駅名	上野３丁目	秋葉原
	通称	山下町	中華街
	複合語地名	日本橋人形町	日本橋
		南麻布	西麻布
	聞き違い	タマチ	タナシ
職業	包含関係にある分類名	日本料理	てんぷら料理
	類似している分類名	ワープロ教室	コンピュータ学校
	通信先の多義性	レストラン（絵を展示していて食事もできる）	貸ギャラリー

ない理由は検索結果の適合性の判断が利用者にしかできない点にある．次に検索結果の検索条件に対する適合性の確率的推定手法を提案する．最後に本手法を用いた検索システムを作成し、動作例とその有効性を示す．

2．問い合わせ情報の曖昧さ

曖昧さの分類

電話帳データベース検索サービスの特徴は、掲載データとサービス対象利用者が大量かつ多様な点にある．電話帳に登録されている掲載データは全国で約５千万件にのぼり、これらのデータの名義、住所、職業名を構成する単語には一般の単語に加えて、あらゆる人名、地名、造語、外来語が含まれている．命名や業種に関する社会動向を即座に反映するため、データが更新されるペースも速い．電話番号の検索という利用目的の性格上データベースへの問い合わせ数は大量であるが、データやシステムに関する専門的な知識を利用者に求めることはできない．このため、特定の掲載データの検索キーに対して入力される検索条件にはきわめて多様な曖昧さが生じる．

表１に仮名入力によって電話帳データベースを検索する際に見られる代表的な曖昧さを事例とともに分類して示す[１]．

従来の曖昧処理

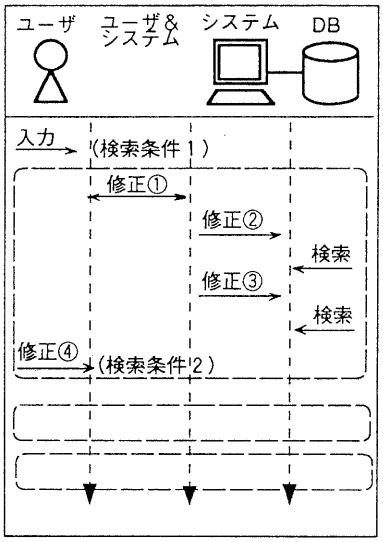
このような曖昧さの解消は、検索システムと利用者双方によってなされている．

問い合わせ前段階では別名（エイリアス）の登録が行われる．規則性のない略称や通称名は経験をもとに登録するが、複合語名義の意味系列の規則的パタンを利用して自動的に別名を派生する手法もある[２]．

問い合わせ過程では検索条件の修正が段階的になされる（図１）．

第１段階では、登録されていない俗地名や職業類語に対して、住所辞書や職業辞書（シソーラス）等を用いたガイダンスによる対話的な利用者誘導が行われる．

第２段階では、仮名表記の正規化処理が行われる．また名義の語尾には曖昧さを生じやすいため、長い入力文字列は末尾を切り捨てて前方一致検索が行われる．



修正例
①：住所、職業の誘導
②：仮名表記正規化
②：住所エリア拡大、末尾切り捨て
④：利用者が自ら行う修正

図１ 問い合わせ過程の条件修正

第３段階の修正は、検索条件を満たす検索解が存在しない（検索解なし）場合に行われる．検索条件の住所エリアの拡大や、名義の末尾をさらに切り捨てる等の処理が適用できる．

第４段階では、まず利用者が検索解に目的の掲載データ（適合データ）が含まれているか否かを判断する．適合データがなかった場合には利用者が自ら条件を修正する．登録されていない別名や引っ越し先の住所といった利用者に固有の情報を用いた修正、あるいは条件の一部削除やガイダンスで提示された類似する職業、近隣住所への修正が行われる．

システムによる従来の処理はあらかじめ用意された曖昧さに関する知識を用いるものであり、第２段階までのものが多い．特殊な表記や、社会的に見て規則性のない曖昧さは知識表現、知識獲得が困難であるためその適用範囲は限定されている．

一方第３段階以降の条件修正は毎回の検索結果に対応するため自由度が高い．再検索の必要性の有無に関する判断には以下のようなレベルがある．

(1) 検索解の数の判断
(2) 適合データの有無の判断
(3) 検索条件の内容と検索解の対応関係の判断

従来の検索システムは第３段階の修正フェーズにおいて、(1)のレベルの判断しか行っていない．

しかし人間は「目的のデータは何か」の表現は曖昧でも、「これは目的のデータか」の判断は容易にできる．そのため(2)のレベルの判断は利用者が行っている．(2)の判断が可能であれば、一切知識を持たなくても試行錯誤的な条件修正で曖昧さを解消することも可能である．(3)のレベルの判断はデータやシステムに専門的知識を有する利用者でなければ難しいと考えられる．

3．利用者検索行動の実験的分析

実験

一般の利用者の検索行動については詳細には調べられていない．利用者による条件修正の有効性を調べるため実験を行った．東京２３区の職業別電話帳（タウンページ）の全掲載データ94万件の検索システムを用いた．実験システムには名義、職業、住所（区、町）を仮名入力し、濁音と長音は正規化を行う．住所と職業は入力文字列に部分一致する候補がすべてガイダンス提示され、利用者が選択する．町が入力された場合は区を自動的に決定する．名義は前方一致検索を行う．解過多制限は無く、システムによる再検索は行わない．

４名の被験者が同一環境下で５問の曖昧な問い合わせについて検索した．なお、キーボード入力は全員支障なく行える．使用した問い合わせ文と対応する掲載データの検索キーを表２に示す．問い合わせ情報がある程度曖昧なものであることは事前に説明した．目的の掲載データが得られず、別の手段（１０４、電話帳冊子）の方がよいと感じた場合はそこで検索を中止するように指示した．

実験結果

最初の検索条件

職業名は問い合わせ文に明示的に指定していない（Q３「出版社」は複数の分類名に分かれている）．しかし延べ20例（４名×５問）の問い合わせ事例中で最初に入力した検索条件に職業を入力しなかった事例は3例のみであった．一般に人間相手のコミュニケーションでは、曖昧な問い合わせほど相手に多くの情報を提示しようとする．同様にデータやシステムに関して専門的な知識のない利用者は、最初の条件入力で、なるべく多くの情報を入力する傾向にある．表３に最初の検索条件の例を示す．

条件修正行動の特徴

５問の問い合わせ中、１回の条件入力で検索成功した（目的の掲載データを検索できた）回数、１回以上条件を修正した結果検索成功した回数、１問あたりの条件修正の合計回数を表４に示す．全体で40回の条件修正行動を観察した．

条件修正の回数は被験者によって個人差が大きい．40回のうち検索条件の一部削除による修正を24回観察した．ガイダンスで提示された候補への置換は職業で5回、町で2回みられた（例：高円寺北→高円寺南、出版社一般→出版社学習参考図書）．名義の置換は2例あった（例：キンダイビジュツカン→ニホンキンダイビジュツカン）．また、一度入力した条件をその後の再び入力する冗長な検索行動を9回観察した．

表３の検索条件は下線の部分を削除するだけでほぼ目的の掲載データのみを検索することが可能である．すなわち特別な知識を用いず試行錯誤的な条件修正のみで検索成功できる．しかし全被験

表２　実験に用いた問い合わせ文と対応するデータ

		目的の掲載データの検索キー		
	問い合わせ文	名義	住所	職業
Q1	「高円寺のロータスＯＡ学院」	ロータスオーエーガクイン	杉並区 高円寺南	ワープロ学校
Q2	「六本木の全日空ホテル」	ゼンニックウエンタープライズ／ トウキョウゼンニックウホテル	港区 赤坂	ホテル業
Q3	「半蔵門駅から歩いてちょっと行ったとことにあるPHP研究所という出版社」	ビーエッチビーケンキュウショ	千代田区 三番町	出版社一般
Q4	「外神田のモダン出版社」	モダンシュッパン	台東区 上野	出版社一般
Q5	「北の丸の近代美術館」	コクリツキンダイビジュツカン	千代田区 北の丸公園	美術館

表3　最初に入力された検索条件の例

	入力例
Q1	ロータスオーエイガクイン
	杉並区 高円寺北
	コンピュータ学校
Q2	ゼンニックウホテル
	港区 六本木
	ホテル業
Q3	ビーエイチビーケンキュウショ
	千代田区
	出版社一般
Q4	モダンシュッパン
	千代田区 外神田
	出版社一般
Q5	キンダイビジュツカン
	千代田区 北の丸公園
	美術館

表4 被験者の検索成功回数、及び条件修正回数

被験者	A	B	C	D
1回で検索成功	0	0	0	1
修正後検索成功	3	2	1	0
合計修正回数	18	14	5	3

者の修正回数に対する修正後の成功回数の比率は平均で15%と低く、その内容には冗長な修正が多く含まれることがわかった．利用者が試行錯誤的に条件を修正して検索成功するためには、論理的かつ効率的な判断のもとに根気強く条件修正を繰り返さなければならない．一般の利用者にとってこの要求を満たすことは大きな負担となる[3]．

4．統計量を用いた適合性の推定

利用者が負担をともなう試行錯誤的な条件修正操作を行わなければならない理由は、検索結果の適合性の判断が利用者にしかできない点にある．

本章では利用者が最初の検索条件に多くの情報を入力する傾向にあることに着目し、検索条件を部分的に削除した修正条件のモデル化を行う．次に修正条件がデータベース内に持つ情報量を定義ずる．これらの議論をもとに、検索条件に対する検索結果の適合性を、条件をみたすデータの数から確率的に推定する手法を提案する．ここで曖昧さに関する特別な知識は用いない．

掲載データと修正条件のモデル化

X_A、Y_A、Z_Aをある掲載データAの名義，住所、職業検索キーとし、同時にその条件をみたす検索解の集合を表わすものとする、X_Aはi_A文字とする．X_Aの先頭$i\,(0 \le i \le i_A)$文字で検索した検索解集合は、以下のような階層的包含関係にある．

$$X_A(0) \supseteq \cdots \supseteq X_A(i) \supseteq \cdots \supseteq X_A(i_A) \tag{1}$$

住所条件についても「区＋町」、「区（町を削除）」、「住所条件なし（区、町を削除）」をそれぞれ$Y_A(2)$、$Y_A(1)$、$Y_A(0)$とすると以下の階層的包含関係がある．

$$Y_A(0) \supseteq Y_A(1) \supseteq Y_A(2) \tag{2}$$

職業も同様に「職業を入力」を$Z_A(1)$、「職業なし（削除）」を$Z_A(0)$とする．

$$Z_A(0) \supseteq Z_A(1) \tag{3}$$

Aを検索解集合に含む条件C_Aは次のようになる．

$$C_A(i,j,k) = X_A(i) \cap Y_A(j) \cap Z_A(k)$$
$$\cdots\ (0 \le i \le i_A, 0 \le j \le 2, 0 \le k \le 1) \tag{4}$$

この関係を図2に示す．ここで$X(i), Y(j), Z(k)$は

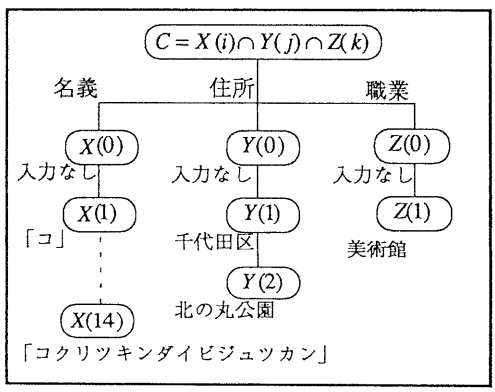


図2 掲載データの構造

以下のように表わせる．

$$X(i)=C(i,0,0)$$
$$Y(j)=C(0,j,0)$$
$$Z(k)=C(0,0,k) \quad (5)$$

同様に、ある検索条件C_qがi_q文字の名義X_q、j_q階層の住所条件Y_q、k_q階層の職業条件Z_qによって構成される場合、C_qを部分的に削除してできる修正条件Cは次のようになる．

$$C(i,j,k)=X_q(i)\cap Y_q(j)\cap Z_q(k)$$
$$\ldots\ (0\le i\le i_q, 0\le j\le j_q\le 2, 0\le k\le k_q\le 1) \quad (6)$$

(6)よりCの組み合わせの総数は以下の範囲となる．

$$h_q=(i_q+1)(j_q+1)(k_q+1)\le 6(i_q+1) \quad (7)$$

Cをみたす検索解の数を$n(C;i,j,k)$とすると、

$$i\le i' \Rightarrow n(C;i,j,k)\ge n(C;i',j,k) \quad (8)$$

が成り立つ．この関係はj,kについても同様である．

検索条件C_qでAを検索できなかった場合、C_qの部分的削除による修正条件は検索結果によって次のいずれかに分類できる．

a. 検索解なし
b. 目的の掲載データ以外の検索解：非適合条件
c. 目的の適合データを含む検索解：適合条件

最適な、すなわち最も適合性の高い修正条件は、検索解の数が最小となるような適合条件である．

検索条件の情報量

データベースに登録されているデータの数をN、ある検索条件Cをみたす検索解の数をnとすると、検索解を1件以上出力する検索条件Cをみたす掲載データがデータベース中に存在する確率$P(C)$は

$$P(C)=\frac{n}{N} \quad (9)$$

したがってデータベース内にCが有する情報量は

$$I(C)=\log_2\frac{1}{P(C)}=-\log_2\frac{n}{N}\ \text{bit} \quad (10)$$

で表わすことができる．実験に用いた94万件のデータベースから1件のデータを絞り込むために要する情報量は19.8bitである．検索解を1件以上出力するような検索条件の持つ情報量はこの値を超えない．この値を臨界情報量と呼ぶことにする．

$C=X\cap Y\cap Z$より$I(C)$はX、Y、Zの結合情報量であり、X、Y、Zはそれぞれ名義、住所、職業の固有情報量とみることができる．

$$I(C)=I(X,Y,Z) \quad (11)$$

各項目について条件を入力しない場合の情報量は0である．

$$I[C(0,0,0)]=-\log_2\frac{N}{N}=0 \quad (12)$$

また$C(i,j,k)$の平均情報量（エントロピー）を以下の式で算出できる．

$$H[C(i,j,k)]=\sum P[C(i,j,k)]\log_2\frac{1}{P[C(i,j,k)]} \quad (13)$$

検索条件の適合情報量

実在する掲載の各検索キーは互いに関連しあっている．例えば名義が「コクリツキンダイビジュツカン」なる掲載データの住所は100%千代田区北の丸公園であり、名義の先頭3文字に”セブン”とつく掲載データの職業は53%が”コンビニエンスストア”に集中している．

逆に実在する掲載データを考慮せず無作為に名義$X(i)$、住所$Y(j)$、職業$Z(k)$を選んだ場合、その検索条件$C_{NR}(i,j,k)$をみたす掲載データが存在する確率は$X(i),Y(j),Z(k)$が互いに独立に分布する場合の確率となるはずである．

$$P[C_{NR}(i,j,k)]=P[X(i)]P[Y(j)]P[Z(k)] \quad (14)$$

$C_{NR}(i,j,k)$の平均情報量は次のようになる．

$$H[C_{NR}(i,j,k)]=H[X(i)]+H[Y(j)]+H[Z(k)] \quad (15)$$

表現が曖昧であっても最初の検索条件は目的の掲載データに関する情報を部分的に含んでいる．その情報を残し曖昧な部分を削除した適合条件の

名義、住所、職業は、実在する掲載データの情報を反映して互いに関連している．関連性の強さは条件に含まれる適合データの情報量に応じて強くなる．

一方非適合条件をみたす検索解は利用者の意図とは無関係な、偶然に検索条件をみたす掲載データの集合である．そのため非適合条件をみたす検索解の存在する確率は統計的に独立な名義、住所、職業の組み合わせをみたす検索解の数に近いと考えられる．

したがって1件以上の検索解を出力する修正条件$C(i,j,k)$の適合性を、$X(i),Y(j),Z(k)$が統計的に独立に分布していると仮定した場合に検索解が存在する確率と、$P[C(i,j,k)]$との比によって定義できる．これを$P_R[C(i,j,k)]$とする．

$$P_R[C(i,j,k)]=\frac{P[C(i,j,k)]}{P[X(i)]P[Y(j)]P[Z(k)]} \tag{16}$$

この対数をとると以下のようになる．

$$\begin{aligned} R[C(i,j,k)] &= \log_2 P_R[C(i,j,k)] \\ &= I[X(i)]+I[Y(j)]+I[Z(k)] \\ &\quad -I[C(i,j,k)] \end{aligned} \tag{17}$$

$R[C(i,j,k)]$を適合情報量と呼び、$C(i,j,k)$の適合性の確率的な推定値として用いる．適合情報量は修正条件の検索解の数から算出できる．

東京23区のタウンページの全データについて、住所、名義、職業が互いに独立と仮定した場合に与える平均情報量$H[C_{NR}(i,j,k)]$と、実際の平均結合情報量$H[C(i,j,k)]$を計算した結果を図3に示す．i,j,kが大きくなると$H[C_{NR}(i,j,k)]$は臨界情報量19.8bitを超える．これは無作為に多くの情報を入力した場合、検索解なしの可能性が高くなることを意味している．一方掲載データを出力する検索条件の平均情報量は臨界情報量を超えない．適合情報量は両者の差であり、この値が大きい条件ほど検索解の数は少なく適合性が高い．

また、適合情報量は次のように展開できる．

$$\begin{aligned} R[C(i,j,k)] &= I[X(i);Y(j),Z(k)]+I[Y(j);Z(k)] \\ &= I[Y(j);Z(k),X(i)]+I[Z(k);X(i)] \\ &= I[Z(k);X(i),Y(j)]+I[X(i);Y(j)] \end{aligned} \tag{18}$$

ここで$I[X(i);Y(j)]$とは$X(i)$と$Y(j)$の相互情報量であり、これは二つの確率事象の確率的な依存関係を表わす統計量である[4]．この関係を用いて検索条件と検索結果のより詳細な関係について判断することも可能である．

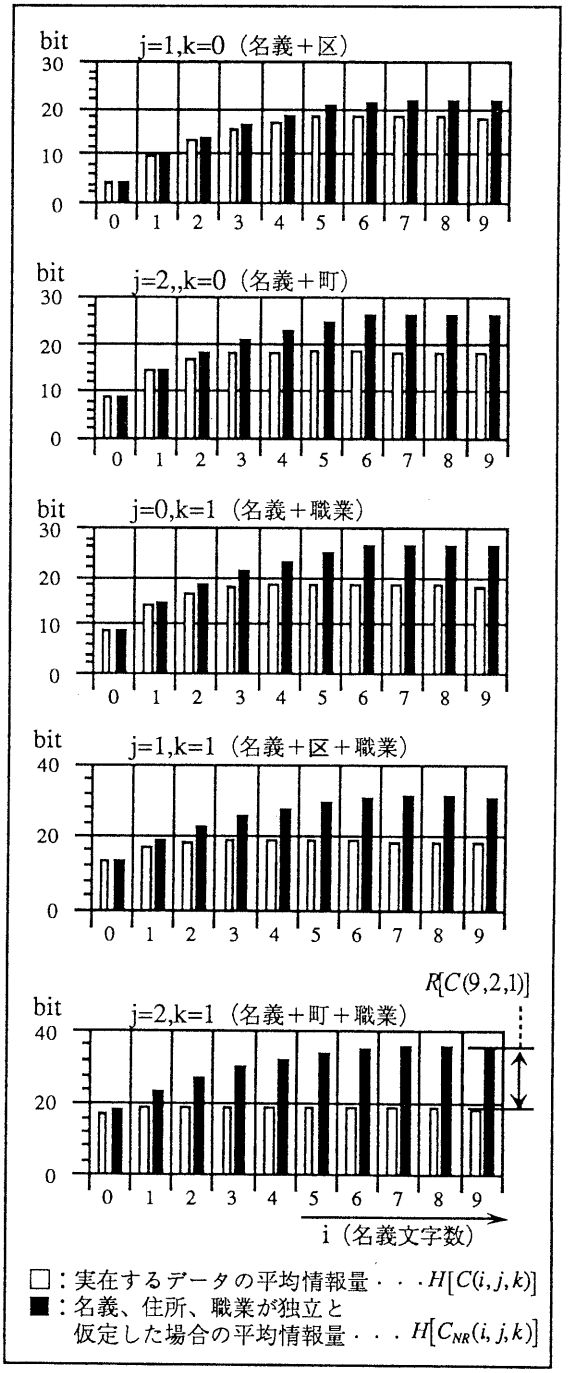


図3 電話帳データベース掲載データの平均情報量

5．適合情報量を用いた検索システム

適合情報量を用いて検索解の適合性を判断する検索システムを作成した．本システムでは最初に入力した検索条件の部分的削除によって可能な全ての修正条件を生成し、検索解の数から算出した適合情報量の値が最大の修正条件の検索解を出力する．利用者によって検索結果が不適合と判断された場合には適合情報量が大きい順に別の検索解を出力する．

実験で被験者が最初に入力した検索条を本システムに入力した．Q5の問い合わせ例

名義：キンダイビジュツカン

住所：千代田区　北の丸公園

職業：美術館

は、66個の修正条件の候補を生じ、そのうち検索解を1件以上出力する修正条件は20件あった．この20件の修正条件の検索解の数、適合データの有無、及び適合情報量を表5に示す．適合データを含みかつ検索解の数が最小の条件の適合情報量は8.32bitと特に高い値を示している．この値から、そのような問い合わせが偶然入力される確率が$1/2^{8.32}$であることがわかる．

表5　適合情報量の計算例

(i,j,k)	検索解数	適合データ	適合情報量
1,0,1	1		-1.54
1,2,0	1		-0.63
8,0,0	2		0.00
7,0,0	2		0.00
0,2,1	2	○	8.32
6,0,0	3		0.00
5,0,0	3		0.00
0,1,1	13	○	0.66
4,1,0	18		-0.28
3,1,0	18		-0.28
4,0,0	19		0.00
0,2,0	56	○	0.00
0,0,1	105	○	0.00
3,0,0	271		0.00
2,1,0	283		0.59
1,1,0	2125		0.05
2,0,0	2397		0.00
1,0,0	26212		0.00
0,1,0	73549	○	0.00
0,0,0	942837	○	0.00

5問の入力例のうちQ3を除く4問は1回目に適合データのみの検索解を出力した．Q3は2回目の検索解で適合データと1件の不適合データを出力した．適合データを検索するまでの条件入力回数の逆数を検索成功率とし、検索失敗の場合は0とすると、実験での利用者による平均検索成功率は0.15である．この値に対して本方式を用いた場合の検索成功率は0.9となり、6倍の向上が得られ、本方式の有効性が確認できた．

6．おわりに

曖昧な問い合わせ情報を対象とする電話帳データベース検索方式を提案しその有効性を示した．曖昧さに関してシステム及び利用者があらかじめ有する知識は限定されている．利用者の検索行動は成功率が低く、非効率的であることをまず実験によって示した．この実験結果をもとに、検索条件をみたすデータの数から、検索条件と検索結果の適合性を確率的に推定する手法を提案した．そのために検索条件を部分的に削除した修正条件のモデル化を行い、修正条件がデータベース内に持つ情報量を定義した．本手法には問い合わせの曖昧さに関する特別な知識を用いないという利点がある．

本方式では適合情報量を算出するために検索解の数を用いるが、データそのものは見ていない．利用頻度の多いサービスに適用するためには、少ない検索回数で検索解の数のみを得られるようなデータ構成に関する検討が必要である．

また本手法は前方一致検索を前提とするため、名義の先頭部分の曖昧さに対応できないことが予想される．名義の語構造の統計的性質を考慮した検索方式が今後の課題である．

参考文献

[1]宮部、大山、本郷：”名義検索システム”、情処学論、24,44,pp.223-228

[2]岩瀬、大山、橋田：”企業名の普通名詞分割”、信学論（D）、J70-D,4,pp.832-835

[3]永吉、岩瀬：オペレータの検索行動に基づく誘導方式の基本検討”、情処学会第47回全大、5S-7

[4]滝：情報論Ⅰ、岩波書店、p33