2P－6

イエローページサービスにおける検索知識自動生成法

別所　克人　　戸部　美春
NTT情報通信研究所

1．はじめに

業種入力により目的とする顧客情報を検索・案内するイエローページサービスにおいては、業種分類体系、案内される店・企業（以下IPという）と業種との対応関係のような検索知識が全て人手によって構築されている。人手による検索知識構築は、多大なコストを要し、構築された知識は一般的に曖昧で不完全である。また、新しい業種への対応が遅れがちになる。

本研究では、これらの問題点を解決するため、IPに関する営業内容等の説明テキストから検索知識を自動的に生成する手法について考察した。具体的には、大量の顧客情報のテキストを形態素解析により単語に分解し、得られた単語の共起性に基づく統計処理により、検索語とIP間の数値化された関連度を算出する。この関連度をもって検索知識とする一手法を提案し、その評価結果を報告する。

2．前提条件

本研究を進めるにあたり、IPの案内情報を利用し、その中の企業名と営業内容を表わす説明テキストを基に形態素解析を行なった。その結果抽出された名詞単語の集合として各IPを表現した。（表１）

表１　前提条件

IP総数	7529件
単語総数	15973語（同一単語は一つとする）

3．関連度に関する検討

3.1　アプローチ

検索語を与えた時、対応する正解IPを検索可能とするため、検索語とIPとの関係の深さに評価値を与える。これを関連度と呼ぶ。関連度を考える上では、IPを表現する単語集合と検索語との関係が表面上の関係と異なる次の条件を考慮することが重要である。

（１）与えられた検索語を持たないが、その検索語と関係が深いIPを高く評価し、
（２）与えられた検索語を持つが、その検索語と関係がないIPは低く評価する。

これを実現するため、同一IP内の単語の共起関係に基づいて単語間の関連づけを行い（想起度）、関連づけられた単語の重みを考慮した検索語とIP間の関連度を計算する手法を提案し、その有効性を評価した。

3.2　想起度

検索語（K）と単語（W）間の想起度を式（１）で定義する。

$$\frac{\text{Wの出現する正解IPの数}}{\text{Wの出現するIPの数}} \qquad (1)$$

出現件数の多い単語に限定して（496語）、Kを与えて想起度を求めた結果、商品やサービスを表わす業種相当の、Kと関係の深い単語の多くが高い想起度を示した。（想起度の上位5％以内の単語の約90％が業種相当の単語）

3.3　関連度

KとIP（A）間の関連度を、式（２）のようにA内の単語W_iとK間の想起度の平均と定義する。

$$\text{K・A間の関連度} = \frac{1}{N}\sum_{i=1}^{N}(W_i\text{・K間の想起度}) \qquad (2)$$

$$\text{A内の単語集合} \quad \{W_1, \cdots, W_N\}$$
（同一単語は一つとする）

その結果、検索語を持たないIPでも、検索語と関係の深い単語が多ければ、関連度は高くなり（図１）、検索語を持つIPでも、検索語と関係のない単語が多ければ、関連度は抑制された（図２）。

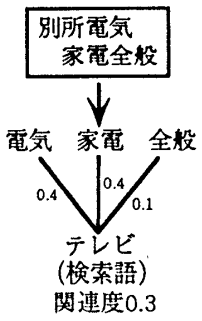

図１

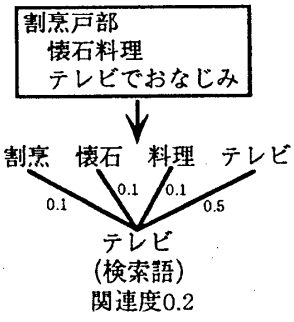

図２

An Automatic Acquisition Method of Retrieval Knowledge for a Yellow Page Service
Katsuji Bessho, Miharu Tobe
NTT Information and Communication Systems Laboratories
3-9-11 Midoricho, Musasino, Tokyo 180, Japan

いくつかの検索語について、上記の関連度計算を行い、関連度の高いIPから正解IP数だけを検索解として、再現率（＝適合率）を評価した。その結果、表2のように70%以上の再現率が得られた。

$$再現率=\frac{検索された正解IPの数}{正解IPの数} \quad (3)$$

$$適合率=\frac{検索された正解IPの数}{検索されたIPの数} \quad (4)$$

表2　関連度の定義の妥当性の評価例

検索語	再現率
レンタカー	90.7%
不動産	81.7%
宝石	80.5%
引越し	84.9%
テレビ	73.3%
冷蔵庫	72.0%

4．関連度の検索への適用

3節の関連度の計算においては、実際の検索時には未知である正解集合を用いている。実際の検索に関連度を適用するため、以下のような近似パラメータを用いることにする。

近似正解集合S：検索語が出現するIP集合

検索語Kと単語W間の近似想起度：

$$\frac{Wの出現する、SのIPの数}{Wの出現するIPの数} \quad (5)$$

検索語KとIP（A）間の近似関連度：

A内の単語とK間の近似想起度の平均

Kが与えられた時、近似正解集合を基に、各IPの近似関連度を求め、近似関連度がある閾値（今回、全IPの近似関連度の平均の3.5倍とした）以上のものを検索解集合として求めた。更に、新たに得られた検索解集合を近似正解集合として、同様の処理を繰り返し実施することにより、再現率と適合率がどのようになるか評価した。（図3）

最初の近似正解集合と真の正解集合との一致の度合によって、概ね三つのパターンに分かれた。ここで、一致の度合（一致度）は式（6）で定義した。

$$一致度=\frac{検索語が出現する正解のIPの数}{検索語が出現するか、または正解のIPの数} \quad (6)$$

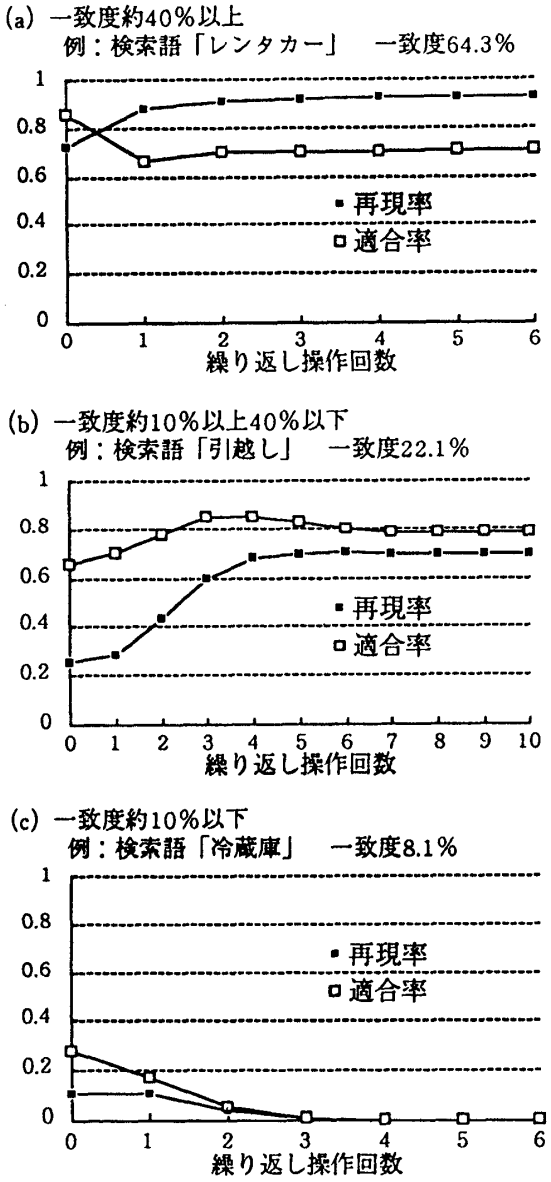

図3　検索解集合が収束するまでの再現率と適合率の遷移

パターンa、bのように、一致度がある程度以上ある場合、想起度及び関連度の近似は良いので、3節で評価した再現率、適合率に近い値まで改善される傾向が見られた。特に、一致度が高くないパターンbでは初期集合から高い改善効果（再現率最大約40%上昇、適合率最大約20%上昇）が得られた。パターンcでは、再現率、適合率の改善は見られなかった。

5．おわりに

本研究では、検索知識自動生成法として検索語とIP間の関連度算出を提案し、それが実際の検索において有効である可能性を示した。今後、評価対象の検索語を増やし評価精度を向上させると共に、検索システムに組込み、効果を確認していく予定である。