3T－5　「ディレクトリサービス」における自由形式の入力方式の検討

樋口謙一†　村田眞人†　時庭康久†　佐々木修二†　坂田真人‡
(株) 高度通信システム研究所†　東北大学‡

１．はじめに

電子メールサービス等で相手先の通信アドレスを知る手段に「ディレクトリサービス」がある。

ＯＳＩ管理におけるディレクトリサービス（以下ＤＳ）は、ＣＣＩＴＴ／Ｘ．５００で規定されており、相手先の論理的な名前は属性値アサーション（属性型と属性値の対）の列により示すことが規定されている。しかし規定された形式は我々が日常使用する宛先の記述形式とは懸け離れている。我々は論理的な名前をより自然な形式で入力できる方式について検討したので報告する。

２．入力文字列の識別上の課題と手掛かり

サービス提供のために情報を蓄積するＤＩＢは、外部からは木状に捉えられ、根(Root)の下に「地域名称」部、「組織名称」部、「個人名称」部の各エントリがある。

利用者は論理的な名前をこれらのエントリ名の列として入力するが、その形式には以下のようなレベルが考えられる。

①完全自由形式　：形式に制約を設けない入力。
②分かち書き形式　：「地域」「組織」「個人」名称部に分ける入力。
③テンプレート形式：名称ごとに属性を付す入力。

ここでは判定の容易さと入力の自然さが保てる「分かち書き」形式を採る。

「分かち書き」形式で入力された文字列は判読する上で「名称ごとの区切りが分からない」「略称の使用や記憶違い等からＤＩＢ登録名称と比較して文字の過剰、欠落、相違がある」等の課題がある。その解消のため基本的に次の３フェーズで処理を進めるものとする。

①名称の切り出し（文字列分割）
入力文字列をＤＩＴのエントリ名単位で分割する。
②名称の補正
分割された文字列は文字の過不足相違から、ＤＩＴの登録名称と必ずしも一致しない。また単位名称全体の欠落等も想定され、これらの追加修正を行う。
③照合
ＤＩＴのエントリ名と照合し、一致判定を行う。

①～③の処理で不成功と判定された場合は再度、分割補正の方法を変えて繰り返す。

なお名称には日本語文字が使用され、また当初のＤＳは企業等の組織体で適用されると考え、入力には「組織名称」部が存在し、「地域名称」部は県市町村等の行政区画の範囲で識別できるものとする。

これらの条件の下で、以下の手法を考察する。

①「接頭用語」及び「接尾用語」の活用
名称の前後に付加される定形的な語で「地域名称部」の｀県´や｀字´あるいは「組織名称部」の｀株式会社´や｀部´｀課´等（以下これらの語を「接頭用語」「接尾用語」という）がある。これらから文字列の分割や、切り出した名称の属性を判定し補正する。

②「辞書の活用」
エントリ名やこれを構成する語を収録した「辞書」を活用する。辞書はＤＩＢに登録される名称に合わせて更新されねばならないが、その維持負担は小さくする必要がある。辞書の構成の詳細は３章で述べる。

③区切り符号の活用
文字列中の空白は区切り符号と見なせる。区切り符号の規定は設けないが、入力された分は名称の区切り点の判定に積極的に活用する。

３．辞書の構成

１個の名称はいくつかの語の組合せから構成される。

入力文字列中に辞書登録名称と一致する語列が得られない場合は、名称に固有の特徴的な語を見つけ、その語を手掛かりにこれを含む名称を推定する方式を採る。

辞書はその特性を活かすために用途（「地域名称」「組織名称」）ごとに設けるものとする。

３．１「地域名称」に関する辞書

地名を収録し、どのＤＳＡにも共通的な辞書である。

登録名称は、名称の地域自身が属する階層（行政区画の階層を表す）とその上位の階層を属性として持ち、この階層の参照により、入力文字列の中で上位の地域名称

地域名称	階層	上位階層
東京	都	国
宮城	県	国
仙台	市	宮城県
山形	県	国
山形	市	山形県

図１　地域名称辞書の例

A Study of the free format input method for Directory Service
†Ken'ichi HIGUCHI, Makoto MURATA, Yasuhisa TOKINIWA, Shuji SASAKI
(Advanced Intelligent Communication System Laboratories)
‡Masato SAKATA
(Tohoku University)

が省略されている場合にも名称を生成できる。（図１）

３．２「組織名称」に関する辞書

①「辞書」登録語の種類〔型〕

辞書の登録語として次の２型を設ける。

・〔型-1〕語：エントリ名の一部として使用する語。
当初から予め収録しておく語と、運用中にＤＩＢへエントリを登録の際に追加していく語がある。
前者はどのＤＳＡにも共通の一般用語である。

・〔型-2〕語：エントリ名として実在する名称語。
エントリ名をＤＩＢに登録時、辞書へも登録する。

②登録「語」の相互関係

〔型-1〕語はこれを含む〔型-2〕語と対応付けられていなければならない。（図２）

すなわち〔型-1〕語からは、その語を含む全ての〔型-2〕語が分かり、また〔型-2〕語からはこれを構成する〔型-1〕語を知ることができる。

③「語」の属性

各〔型-1〕語は、それがいくつの〔型-2〕語すなわちエントリ名中に使用されているかを示す値「使用回数」を属性として保持する。

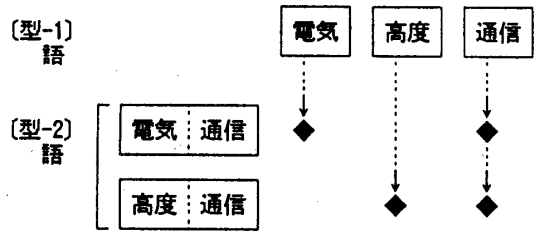

図２　辞書登録語の対応関係

④「辞書」の作成と運用維持

ＤＩＴへエントリ名を追加登録時に辞書へも登録する。

・追加するエントリ名と同一の語が辞書に無ければ、これを〔型-2〕語として新たに登録する。
・追加するエントリ名は既登録の〔型-1〕語と照合し、一致する部分は切り出してその〔型-1〕語との対応をとる。またその語の「使用回数」に１を加算する。
・どの語にも一致しない残余の部分は新たな〔型-1〕語として登録し、「使用回数」を１に設定する。

４．処理の手順

上述の考えを適用して文字列の判定を行う手順の一例を示す。入力文字列は「地域」「組織」「個人」の各名称部の順に処理する。処理の結果、他ＤＳＡ地域と判定された場合は当該ＤＳＡへの照会が伴うがここでは対象外とする。「地域名称」部が空の場合は「組織名称」部は自ＤＳＡ域内の組織として処理を進める。

４．１「地域名称」部の処理

地域の階層区分が比較的明快であり、以下による。

①「接尾用語」及び区切り符号から、辞書の照合範囲を限定するための区切り点を判定する。
②辞書により地名を切り出し、その属性を判定する。
③切り出した地名に続く入力文字列中の「接尾用語」の属性と、②の地名の階層属性を照合する。
④上位階層の地名の欠落があれば、辞書により補足する。
⑤ＤＩＴのエントリ名と照合し、当否を判定する。

４．２「組織名称」部の処理

①「接尾用語」及び区切り符号から区切り点を判定する。
②文字列の先頭から順に辞書の〔型-2〕語及び〔型-1〕語と照合する。
③〔型-2〕語と一致すればエントリ名の候補とする。
④〔型-1〕語のみに一致する場合はそれらの中でもっとも特徴的な（使用回数の少ない）語をキーにして、その語を含む〔型-2〕語を求める。（図３）
⑤その〔型-2〕語を入力文字列と比較し近似度を得る。
⑥近似度が低い場合は〔型-2〕語の次候補を採り、再度⑤を実行する。（近似度の求め方は今後の課題）
⑦全ての語を切り出し後、ＤＩＴと照合する。

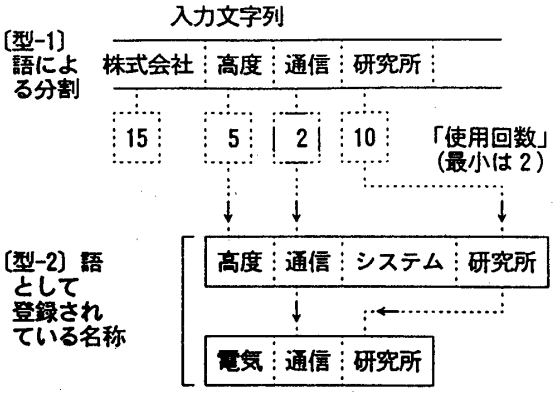

図３　〔型-2〕語によるエントリ名の推定の例

４．３「個人名称部」の処理

個人名称は「姓」のみで扱われる場合があり、入力文字列からは職名、肩書等を接頭用語、接尾用語により除去し、「姓」のみで照合するものとする。

５．おわりに

本稿は接頭接尾用語や辞書の活用により、ＤＳの入力文字列から名称を切り出し、判定を行う方式について検討した。今後はこれらの用語の設定方法や切り出しに不整合があった場合の処置方法等についてさらに精度を高める検討が課題である。

〔参考文献〕

（１）ＩＳＯ：　ＤＩＳ　９５９４　ｐａｒｔ　１−８
（２）ＣＣＩＴＴ：　Ｘ．５００シリーズ