自然語のゆらぎを利用した曖昧検索方式

5N－2

唐沢　裕明　新川　晃太郎
ＮＴＴ　情報通信研究所

【1】はじめに

本稿では、データベース上の自然語で表現されたフィールドに対して、自然語解析処理によってデータ相互間のゆらぎを吸収し、その照合結果から曖昧検索用インデックスを作成する方式を提案する。その自然語表現の対象として顧客情報等に含まれる住所情報の中の建物名を用いて、作成されるインデックスの効果を示す。

【2】背景

顧客情報等における住所情報は重要な検索キーとして扱われており、近年においては地図案内システム等との連携のため、詳細な住所としての建物までを検索する要求がある。しかしながら、建物名等の自然語で表記されるデータは個々人による表現の差異が大きく、同一のデータベース上であってもデータ表現を統一することは難しい。また、検索時の入力についても、検索者はデータベース上の表記などを知っているわけではなく、その入力表現についても個々の検索者ごとに異なるため目的のレコードまで容易に到達できない。そのため、入力の試行錯誤による検索トランザクションの増加や、前方一致などの検索手法を用いることにより生じる不必要な検索結果の増大が問題とされていた。

【3】曖昧検索方式

本稿では、同一の建物であっても生ずるデータベース上の表記のゆらぎに着目し、それらの表記バリエーションを照合し、グループ化することにより作成できるインデックスを用いた検索方式を提案する。

データベース上に見られる自然語データの表現上のゆらぎを表１に示す。これらの表現上のゆらぎを自然語解析処理を用いることで、同一建物を表現する建物名をグループ化［1］する。そのグループ化処理のフローと処理例を図１に示し、以下に処理内容を述べる。

表１　データ表現のゆらぎ

ゆらぎ分類		表現のゆらぎ	
文字ゆらぎ	文字種	あさひビル	アサヒビル
		小島ビル	コジマビル
		ＳＯＮＹビル	ソニービル
	外来語表記	ヴィラ札幌	ビラ札幌
	誤記	目黒消防所	目黒消防署
単語ゆらぎ	部分欠落	銀座ソニービル	ソニービル
		中央線吉祥寺駅	吉祥寺駅
	転置	中央第１ビル	第１中央ビル
		マンション武蔵野	武蔵野マンション

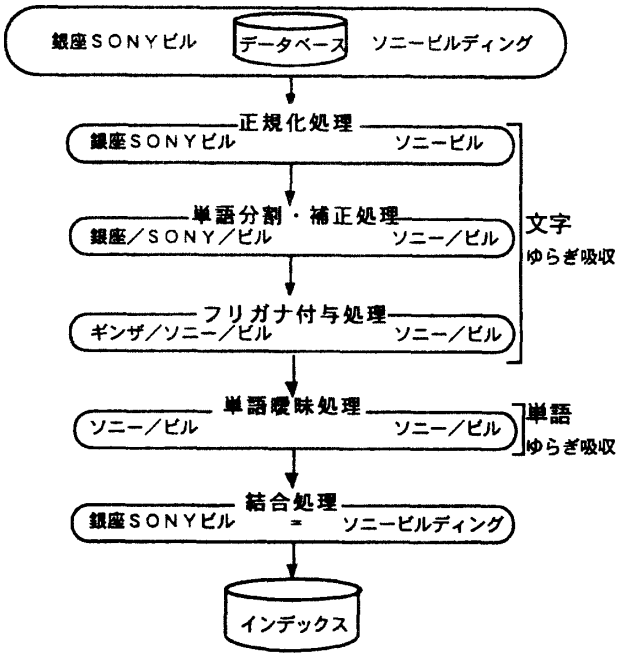

図１　グループ化処理

①正規化処理

自然語で入力された建物名に対して、統一的な表記に近づけるための正規化ルールに従った文字列上の変換を行う。正規化ルールとしては以下のようなものがある。

・カタカナ表記に統一（「あさひ」→「アサヒ」）
・建物語尾の統一（「ビルディング」→「ビル」）

②単語分割・補正処理

建物名の形態素解析処理を行い、単語分割［2］して各単語の意味を付与する。形態素解析処理中に、明らかに誤り単語として辞書登録された「消防所」などの単語を、適切な「消防署」に訂正を行う。

An Information Retrieving Method by the ambiguity of data on the database
Hiroaki KARASAWA, Kotaro SHINKAWA
NTT Information and Communication Systems Laboratories

単語分割処理結果は、単語ゆらぎの吸収を行う単語曖昧処理で用いられる。

③フリガナ付与処理

単語分割処理で分割された単語を基に、フリガナ辞書を検索することによりフリガナを付与する。このカナ表記を付与することにより、以下のような文字種のゆらぎを吸収できる。

・漢字（「小島」→「コジマ」）
・英字（「ＳＯＮＹ」→「ソニー」）
・数字（「１２３」→「イチニサン」）

④単語曖昧処理

単語分割処理で付与された各単語毎の意味とルールに基づき、建物名を構成する単語の配置を操作し、単語レベルにゆらぎのある建物名表記の統一を行う。上記の処理を図１のフローに従って行うことにより、建物名の自然語上のゆらぎが解消され、処理後のデータの完全一致を試みることで結合［３］できる。結合された建物名は同一のビルを意図した記述のグループであり、グループ化された検索キー相互にヒットするようにインデックスを構成することで曖昧検索を実現できる。

【４】評価

①曖昧対策の効果

データベース検索キー「銀座ＳＯＮＹビル」と「ソニービルディング」がグループ化され、そのグループに対して別称「銀座ソニービル」を登録した状態を図２に示す。この状態をインデックス化することにより、検索時の入力が「銀座ＳＯＮＹビル」、「ソニービルディング」、別称「銀座ソニービル」のいずれであっても、元のデータベース上のデータ「銀座ＳＯＮＹビル」と「ソニービルディング」の両者とも検索を可能にすることから検索性能の向上を期待できる。また、グループに対して別称を登録することにより、グループ単位での検索を行えるなど拡張性を持たせることができる。

上記から、データベース登録に用いられた表記の中で、潜在的に存在する自然語のゆらぎを用いて曖昧検索が実現できることが分かる。

②グループ化率

実際に東京都区内の電話帳データから数ヶ所のエリアを抜粋して、建物名のグループ化に対して定量的な評価を行った。建物名の表記を住所エリアごとにユニークにしたデータ１０，０００件に対してグループ化処理を行い、グループ内の表記ゆらぎ数とそのグループ数との関係を求めた結果を図３に示す。また、実装したグループ化処理による全建物名に対するグループ化率は２０％であり、本稿で提案した曖昧検索方式の有効性が示される。

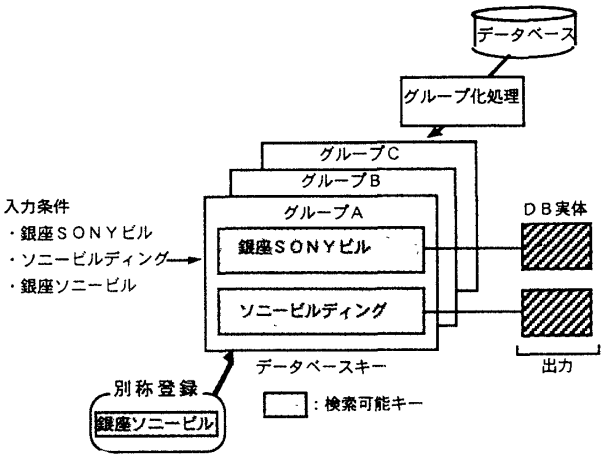

図２　曖昧検索

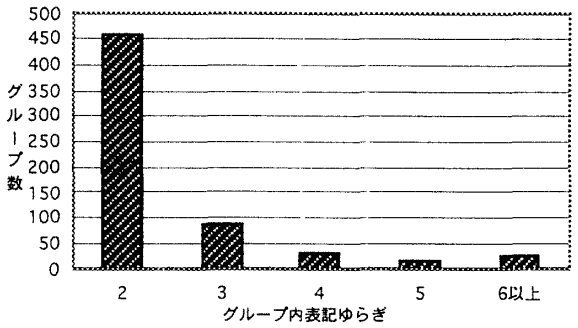

図３　グループ化評価

【５】おわりに

本稿では自然語のゆらぎを利用した曖昧検索方式を提案し、その実現性を建物名を検索キーとするデータベースのインデックスを考察し、ゆらぎの吸収性としてのグループ化率を定量的に求めた。今後の課題としては、提案を行った曖昧検索方式における検索結果の正当性の評価がある。

［参考文献］

［１］唐沢，池田：「ビル名検索方式の評価」，１９９４年信学会秋期全大，Ｄ－１３７

［２］岩瀬，大山，橋田：「企業名の普通名詞分割」，信学会論文誌，ｖｏｌ．Ｊ７０Ｄ　Ｎｏ．４，ｐｐ．８３２－８３５，１９８７年４月

［３］唐沢：「異種データベース結合方式の検討」，平成３年第４３回情処学会全大，１Ｍ－６，４－１０５