ファジィ理論を用いたエコー情報の分類・分析システム（その１）

5S－1

内山 恵三　中村 正規
東京電力（株）システム研究所

１．はじめに

営業窓口、電話、葉書などで当社に寄せられたお客さまのご意見ご要望は、年間約１５００件以上に及んでいる。これらのお客さまの要望を分析し、過去の類似した事例を検索することにより、お客さまの要望への迅速な対応が求められている。

そのため、これらの要請（エコー情報）を効率的に分類できる、ファジィ理論を利用した自動分類方式を開発した。

本論文では、実際のデータを基に実用性を検討するための支援ツールについて報告する。

２．ファジィ理論の適用

文書を自動分類する手法は、ファジィ理論を用いた情報（文献）検索の方式[1][2]を活用する従来手法が最も一般的と考えられる。この手法は、１つの文献に共通に含まれる２つのキーワード間（キーワード対と呼ぶ）には統計的な親近関係があることに着目した方式であるが、文章の意味理解までは考慮していないため、精度が悪いという欠点がある。

本方式は、分野別にその中から名詞だけ機械的に抽出したキーワード対の親近関係から単純に分類分けするのでなく、

(1)キーワード以外に主語・目的語など文章中における用法（重要度）を一緒に抽出し[3]、用法を加味した頻度を考慮する。

(2)表記の違いを考慮する。例えば、当社の料金業務分野において「領収書」と「電気料金領収書」は同一のキーワードと見なした方が自然である。

の拡張を行い精度を向上させたものである。

A classification and analysis system for customer's opinion by fuzzy process (part 1)
Keizo Uchiyama , Masaki Nakamura
Tokyo Electric Power Company

３．分類方式[4]

(1)文書分析辞書の作成

ａ．キーワード抽出

分類がはっきりしている文書から重要キーワード抽出システム[3]を用いてキーワードを抽出する。

ｂ．同一キーワードを含む文書数の抽出

同じ用法のキーワードを含む文書数及び、キーワード対が同一である文書数をカウントする。

ｃ．キーワード間の距離の算出

Ｌ．Ｂ．Ｄｏｙｌｅが提案したキーワード間の親近性の定量化[2]を用いてキーワード間の距離を算出する。

ｄ．文書分析辞書の作成

メンバーシップ関数[2]を用いて、当該分野中のキーワード間の距離から、その分野のキーワード対の依存度（文書分析辞書）を算出する。

以上の処理を分野毎に行う。

(2)新規発生文書の自動分類

ａ．キーワード抽出

前項と同様にして、新規に発生した文書からキーワードと重要度を抽出する。

ｂ．ファジィ関係の抽出

当該分野に対する新規発生文書のファジイ関係を抽出する。新規発生文書から抽出したキーワードが当該文書分析辞書中に存在すればファジィ関係は１、存在しなければファジイ関係は０である。

ｃ．分類別依存度の算出

当該分野に対する新規発生文書の依存度を、当該文書分析辞書とファジィ関係の合成により算出する。

ｅ．曖昧度の算定

当該分野に対する新規発生文書の曖昧度を、当該分野の分類別依存度より算出する。

以上の処理を分類別に行い曖昧度が最も小さい分野が新規発生文書の分類先となる。

4．エコー情報の分類

お客さまから寄せられるご意見・ご要望の種類は、応接態度・仕事のやり方、料金、需給契約、サービス活動、広報・広聴活動、設備・工事など１２分類で管理されている。１２分類といってもお客さまのご要望は多種多様に及んでいるので、厳密には５４分野に分かれている。

エコー情報の年間の割合の例は下表の通りである。

分類名	分野名	占有率[%]	
料金	口座振替	7	２８
	検針票・領収証	6	
	その他	１５	
広報・広聴活動	印刷物でのＰＲ	9	２１
	マス媒体でのＰＲ（ﾃﾚﾋﾞ、ﾗｼﾞｵ等）	6	
	その他	6	
営業開発	電気温水器	6	１０
	その他	4	
設備・工事	計器・ＳＢ・その他内線	3	１０
	その他	6	

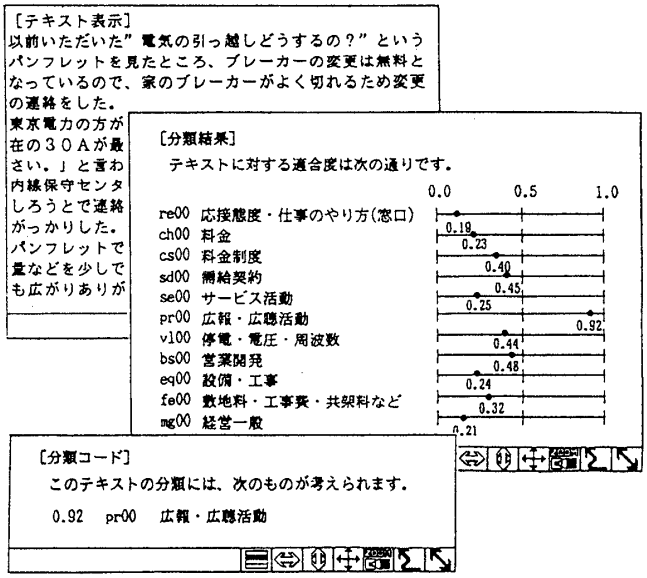
5．評価

第一ステップとして、類似性が強く分類しにくいと考えられる２分野を対象にして分類方式の有効性を評価した。５４分野の中では、マス媒体でのＰＲ（テレビ、ラジオ、新聞、映画、ビデオ等）と印刷物でのＰＲが最も似通った分野と考えられるので、自動分類を行った結果、曖昧度は０．２以下で分類でき、分類方式として有効な結果を得ている。[4]従来の重要度を加味しない方式と比較するとおよそ０．０２位精度が向上している。重要度を加味している分、キーワード数は減ったが、重要度を付加したキーワード数は用法分増えたため、結果として文書分析辞書のキーワード対の数が増えている。１つのキーワードが主語、目的語、その他の３種類の用法に使われていれば、キーワード対の数は、その組み合わせ分増える。その他のキーワード数をｎとすると、

$$4n+8$$

の数だけ増えることになる。文書分析辞書が大きくなると処理速度が遅くなるので、辞書のコンパクト化が必然的な課題である。また、似通った分野が３つ以上の場合についても分類方式の有効性を評価する必要がある。

第二ステップでは、この結果を踏まえ、実データ１２分類５４分野全部に対して分類方式の有効性を評価するためのインターフェースを開発した。分類数、分野数、重要度、キーワード数をパラメータとして、各分類、各分野の文書分析辞書、曖昧度の変化を見ることができる。以下に画面イメージを示す。



6．まとめ

エコー情報の１２分野５４分類の実データで文書分析辞書の作成など自動分類方式の過程を評価するインターフェースの開発により、類似の分野を増やした場合の自動分類の影響が文書分析辞書の中身を見ながら解析できることになる。今後、このソフトを使って文書分析辞書のコンパクト化や似通った分野への対応をはかっていく予定である。

7．参考文献

[1]森田他，「ファジィ文書検索システム～実験システムと評価～」，情報処理学会第３９回全国大会，p1067，平成元年

[2]寺野他，「ファジィシステム入門」，オーム社

[3]内山，中村，「重要キーワード抽出方式とその活用方法」，情報処理学会データベースシステム研究会，1991.7.18，p11

[4]内山，中村，「データベースにおける文書の自動分類方式」，13-117，平成５年