コルモゴロフ複雑性に基づく顧客要求抽出

藤原　由希子†　　五藤　智久†

†NECサービスプラットフォーム研究所

1　はじめに

サービス提案や仕様策定では、各顧客企業の特徴に応じた最適なサービスセットの提示が求められる。そのため、顧客企業内の多種多様な要求の抽出が不可欠であり、要求分析者は顧客企業のステークホルダにヒアリングを実施し、要求を収集する。しかし、一般に、顧客自身がサービス導入後を確実に推定することは難しく、要求漏れが発生しやすい。漏れを防ぐために潜在的な要求を推定することは、要求分析者の経験と勘に依存するところが大きい。過去の類似案件の参照は有効な知見を得る手段であるが、膨大かつ形式の異なる情報を再利用することは困難であった。

本稿では、過去の類似案件の情報を用いた要求抽出支援のため、コルモゴロフ複雑性に基づく顧客要求抽出法を提案する。コルモゴロフ複雑性は、情報のランダム性の指標である[1]。提案法は、過去の類似案件であるテキスト情報に対し、前処理でテキスト情報の表記を統一し、クラスタリングにより要求を抽出する。提案法は、顧客企業へのITサービス提案に適用し、実現可能性を検証した。

2　コルモゴロフ複雑性

提案法は、コルモゴロフ複雑性（Kolmogorov Complexity）に基づいて要求間の類似度を計算する。文字列xと文字列yのコルモゴロフ複雑性は、正規化圧縮距離NCD（Normalized Compression Distance）で概算する[2]。

$$NCD(x,y)=\frac{C(x\cdot y)-\min\{C(x),C(y)\}}{\max\{C(x),C(y)\}}$$

ここで、C(x)、C(y)は、それぞれ、文字列x、yの圧縮列の長さ、C(x・y)は、文字列xとyを連結させた圧縮列の長さである。NCDは、値が小さいほど2つのデータが類似していることを示す。

従来、テキスト間の類似度は、テキストに含まれる単語の出現頻度の情報を用いて計算していた。しかし、例えば、「シミュレーション」、「シュミレーション」という単語は、事前に辞書で変換しない限り、別となり類似と計算できないという問題があった。コルモゴロフ複雑性は、圧縮の際に「レーション」の一致を用いるため類似と計算できる。また、「デザインならよいが」という文Aに対し、全く同一の文B「デザインならよいが」と、異なる文C「デザインがよいなら」があった場合、従来法では、含まれる単語が同一のため、文Aと文Bの類似度は文Aと文Cの類似度と等しいと計算されるという問題があった。しかし、コルモゴロフ複雑性に基づくと、文Aと文Bの方が、文Aと文Cより類似すると計算することができる。

3　顧客要求抽出法

我々はコルモゴロフ複雑性に基づく連結クラスタリング法を提案し、商品・サービスの価値評価に適用し提案法の有効性を確認した[3]。本稿で提案する顧客要求抽出法は、基本的には同じである。まず、前処理として、過去の類似案件の表記を統一する。[3]では、変換ルールを用いて「ですます調」の「である調」への統一や「職場」と「オフィス」の統一などを行った。しかし、膨大な文書で変換ルールを完備することは困難である。そのため、本稿では、同義語辞書WordNet（version 0.91）[4]による変換を追加した。テキストをMeCab（version 0.98）[5]を用いた形態素解析によって単語に分割した後、各単語をその単語が含まれるWordNetの類義関係セット（synset）の代表語に変換した。

次に、提案法は、連結クラスタリング法を用いて要求を集約する。まず、対象集合Dと求めるクラスタ数Kが入力されると、各対象を各クラスタとした要素数Nのクラスタ集合Cを作成する。次に、対象ペア間の非類似度NCDを計算し、最も類似度の高い対象ペアd_xとd_yを探す。それから、d_x・d_yを改めてd_xとし、d_yを対象集合Dから削除する。そして、クラスタC_xとC_yを併合したクラスタを改めてC_xとし、クラスタC_yを削除する。これら対象ペア選択、対象の連結、クラスタの併合という動作を繰り返して、対象をクラスタリングする。N-K回繰り返すとクラスタ数がKとなるのでクラスタリングを終了し、クラスタ集合Cを出力する。

Requirements Extraction based on Kolmogorov Complexity
†Yukiko FUJIWARA and Tomohisa GOTOH
Service Platform Research Laboratories, NEC Corporation

クラスタリング後は、クラスタの代表を要求として抽出する。代表は、クラスタの全要素の連結と個々の対象との類似度 NCD を計算し、最も NCD の小さい対象とした。

4 検証実験

4.1 検証方法

実際のインタビュー結果の代替として、検証用データは、Web アンケート調査で収集した。2008 年度にオフィス勤務者に対し、「職場での業務効率の向上やオフィスでの働き方などについて、あなたのお考えを自由に記入ください。」と質問し、1,000 件の回答を得た。「わからない」などの無意味な回答を除くため、20 文字以上の 539 件の回答を選択した。回答は、企業の特徴ごとの要求抽出とするため、フリーアドレス適用の有無と、外出が週 2 回以上かどうかで分類した。データ数を表 1 に示す。これら 4 種のデータに対し、顧客要求抽出法を適用してそれぞれ約 20 件の要求を抽出した。

4.2 実験結果と考察

実験結果として、環境、事務、コミュニケーション（連携）、個人、マネジメント、勤務制度、その他の要求が抽出された。実験結果の一部を表 2 に示す。表で、A はフリーアドレスかつ外出少、B はフリーアドレスかつ外出多、C は固定席かつ外出少、D は固定席かつ外出多を示す。

表に示すように、企業の特徴ごとに異なる要求が抽出された。例えば、フリーアドレスでは、機器やツールの整備や紙書類の保管場所が業務効率向上に必要と認識されることが分かった。席のないフリーアドレスで、機器やツールへの依存度が高く、紙書類の置き場所が気になるという結果は妥当と考えられる。また、外出回数が多いと、TV 会議の利用や紙書類の電子化が良いと認識されており、これらの結果も妥当と考えられる。また、外出回数が少ないとフリーアドレス適用前にはコミュニケーションが問題とされず、適用後は重要性を認識することも分かった。サービス提案では、外出回数の少ない職場では、業務ノウハウ共有、外出回数が多い職場では、コミュニケーション活性化案や TV 会議を提案するとよいなどの知見が得られた。

このように、多くのテキストでなく、20 件程度の要求を読むことで知見が得られることが示された。なお、WordNet の変換で構文が不正確になる場合が多く、従来法では単語分割ミスが起きるため、提案法と従来法とは比較しなかった。

表 1 検証用データ

	外出少	外出多	計
フリーアドレス	114	94	208
固定席	213	118	331
計	327	212	539

表 2 検証実験結果

	内容	A	B	C	D
環境	機器やツールの整備が必要	○	○		
	時間短縮に TV 会議が良い		○		○
	気分転換の自分の空間必要	○			
	ロッカーへの片づけが無駄		○		
	紙書類の保管場所が必要	○	○		
	紙書類の電子化が良い		○		○
	関係者を集める場所が有効				○
	業務で異なる環境が必要				○
事務	業務ノウハウの共有が良い	○		○	
	社内手続きが増え本業圧迫		○		
連携	コミュニケーションが大切	○	○		○
	社内コミュニケーション不足	○	○		○
	仕事中の雑談はなくすべき			○	
	尻拭いが多く連携は良くない			○	

5 結論

本稿では、コルモゴロフ複雑性に基づく要求抽出法を提案した。提案法をアンケート調査結果に適用し、その実現可能性を例証した。今後、より詳細な分析を通して有効性を検証していく。

参考文献

[1] C.H. Bennett, et al., “Information Distance”, *IEEE Transactions on Information Theory*, Vol.44, No.4, pp.1407-1423, 1998.

[2] R. Cilibrasi and P. Vitányi, “Clustering by Compression”, *IEEE Transactions on Information Theory*, Vol.51, No.4, pp.1523-1545, 2005.

[3] 藤原由希子, 五藤智久, 井口浩人, “コルモゴロフ複雑性に基づく商品・サービスの価値評価”, 第 8 回情報科学技術フォーラム (FIT2009), RF-002, 2009.

[4] F. Bond, et al., “Enhancing the Japanese WordNet”, in *The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009*, 2009.

[5] http://mecab.sourceforge.net/