RF-002

コルモゴロフ複雑性に基づく製品・サービスの価値評価

Product/Service Value Validation based on Kolmogorov Complexity

藤原 由希子† 五藤智久† 井口 浩人†
Yukiko Fujiwara Tomohisa Gotoh Hiroto Iguchi

1．はじめに

製品・サービスの価値は，顧客が受け入れたことの判断材料として，販売数によって事後に評価されるが，費用対効果や必要な調達資源を推定するため，事前に評価することが重要である．我々は，変化分析型決定木や変化分析型アンサンブル学習を考案し，普及初期段階から製品・サービスの将来の普及の推移を予測してきた [1,2]．しかし，これらの手法では，製品・サービスの初期段階の販売情報が必要であった．そのため，開発前に普及の推移を予測できず，製品・サービスの企画・設計に活用することはできなかった．

企画・設計した製品・サービスの価値を評価するには，製品・サービスのサンプル・イメージを顧客に提示して，インタビューやアンケートなどにより意見を収集し，要求を抽出する．しかし，複数の回答から選択する選択回答式での調査結果は，質問者が事前に想定した要求項目の分析しかできず，一方，回答者の自由な回答を引き出す自由記述式での調査結果は，多様な回答結果を整理する手間とコストがかかるという問題があった．

本稿では，製品・サービスの価値評価法として，コルモゴロフ複雑性に基づいてインタビューやアンケート調査結果を分析する新たな手法を提案する．コルモゴロフ複雑性（Kolmogorov Complexity）は，情報量に基づいた文字列のランダム性の指標である[3,4]．提案法は，前処理で文書の表記の揺れを解消してから，新たに提案する連結クラスタリング法により文書をクラスタリングし，各クラスタの代表を抽出することで意見を自動的に抽出する．検証実験として Web アンケート調査結果に適用した結果，提案法の有効性を確認できた．

本稿は，2 章で提案法を説明し，3 章で従来の文書間の類似度計算法および従来のクラスタリング法について述べる．さらに，4 章で検証実験方法について述べ，5 章で検証実験結果と考察を示し，最後に 6 章で結論をまとめる．

2．提案法

提案法は，インタビューやアンケート調査結果を入力されると，前処理を行ってから連結クラスタリング法を用いて文書を複数のクラスタに分類し，各クラスタに含まれる文書から代表文書を抽出して出力する．本章では，提案法で用いるコルモゴロフ複雑性，文書の表記揺れを解消する前処理，提案する連結クラスタリング法およびクラスタ代表の抽出法について説明する．

2.1 コルモゴロフ複雑性

コルモゴロフ複雑性は，情報量に基づいた文字列のランダム性の指標である[3,4]．ここで，文字列は，有限のバイナリ列であり，任意の有限長の対象は，バイナリ列にエンコードすることで，コルモゴロフ複雑性を計算することができる．

文字列 x のコルモゴロフ複雑性 K(x)は，万能計算機（万能チューリングマシンなど）で x を出力することができる最も短いプログラムの長さである[3]．直感的には，K(x)は，あるアルゴリズムで x を生成するのに必要な最小の情報量である．異なる万能計算機では，K(x)は異なる値となるが，差は一定の定数であることが示される．

Bennett らは，コルモゴロフ複雑性に基づいて，2 つの文字列 x と y との間の距離を考えた．まず，条件付きコルモゴロフ複雑性 K(x|y)を，y を補助入力として与えたときに x を計算する最短のプログラム長とする．K(x|y)が小さいとは，y が既知なら x を少ない情報量で生成可能という意味なので，x と y が近いと考えられる．しかし，K(x|y)は非対称なため，距離には不適当である．例えば，x が空文字列の場合，任意の y に対し，K(x|y)は小さな値となるので，直感に反して空文字列 x と長くランダムな文字列 y との距離が近いことになる．そのため，Bennett らは，

$$\max\{K(x|y)+K(y|x)\}$$

を距離として提案した[3]．これに対し，Li らは，距離が文字列 x，y の長さに応じて大きくならないよう，以下の正規化情報距離 NID（Normalized Information Distance）を提案した[5]．

$$\frac{\max\{K(x|y)+K(y|x)\}}{\max\{K(x),K(y)\}}$$

しかし，任意の文字列 x を入力として，そのコルモゴロフ複雑性 K(x)を出力する計算可能なプログラムは存在せず，K(x)は一般的には計算不能である．そのため，実問題への適用では，K(x)を概算する必要がある．コルモゴロフ複雑性の概算のため，Cilibrasi と Vitányi は，圧縮を用いることを提案した[6]．コルモゴロフ複雑性 K(x)は，文字列 x の最高の圧縮と考えられるためである．ここで，圧縮アルゴリズムとしては，圧縮前の文字列と，圧縮・展開の処理を経た文字列が完全に等しくなる可逆圧縮を用いる．可逆圧縮の例は，zgip，bzip2 などである．NIC からの導出で提案された正規化圧縮距離 NCD（Normalized Compression Distance）は以下の式で計算される[6]．

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

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 の計算式に C(y・x)が含まれないのは，zgip や bzip2 などでは，対称性 C(y・x)□C(x・y)が定義や経験から成り立つためである．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 0 以上 1+ε の値である．ε は，圧縮の不完全性に起因する値であり，理想的には，NCD は 1 以下の値となる．

† 日本電気（株），NEC Corporation

NCD は，一般的には，形態素解析を用いた従来の文書間類似度の代替手段とは言えず，DNA 配列の進化系統樹や言語の系統樹など，形態素解析を用いても共通単語の想定しにくい分野で適用されている[4,5]．しかし，提案法では，NCD を用いてインタビューやアンケート調査結果である文書間の非類似度を計算する．これは，インタビューやアンケート調査結果における意見抽出では，形態素解析で抽出される特定な単語を含むかどうかより，全体的な比較が重要と考えたためである．

本稿では，NCD の計算のための圧縮アルゴリズムとして gzip を用いた．NCD の計算では，まず，文書 x と文書 y をそれぞれファイルとし，gzip で圧縮したファイルサイズを C(x)，C(y)とする．また，文書 x と y とを連結させたファイルの圧縮ファイルサイズを C(x・y)とする．そして，コルモゴロフ複雑性を NCD の計算式で概算した．

2.2 前処理

コルモゴロフ複雑性は，計算するのに事前情報が不要なことが特徴である．しかし，実用的に考えると，文字列が類似しない同義語を認識できないことは問題である．そのため，提案法では，まず，文書の表記揺れを解消する前処理を行う．前処理では，変換ルールを予め用意しておき，全ての文書に対し，文書を変換することで，文書の表記揺れを解消した文書を作成する．

変換ルールは，例えば，「分からない→わからない」，「無い→ない」，「様々→さまざま」，「PC→パソコン」，「パーソナルコンピュータ→パソコン」，「ピーシー→パソコン」，「ケータイ→携帯」などの同義語変換や「ですます調→である調」などの文体の統一である．ここでは，文書を事前に読まなくても対象領域に関する知識から想定可能な一般的なルールを準備する．

前処理による変換の例として，元の文書集合の例を図 1，前処理後の文書集合の例を図 2 に示す．図では，1 列目の ID が文書番号，2 列目の文書内容が実際の文書である．図に示すように，元の文書集合では，「PC」と「パソコン」，「無い」と「ない」，「良い」と「よい」が混在するが，前処理後は，「パソコン」，「ない」，「よい」に統一され，文書の表記揺れが解消される．そのため，前処理後は，元の文書集合に比べ同一の単語が増え，文書の類似度が判定しやすくなることが期待できる．

2.3 連結クラスタリング法

提案法は，前処理の次に，クラスタリングを行って文書を複数に分類する．クラスタリングとは，N 件の文書を要素とする集合 D を，K 個のクラスタ C_1，C_2，・・・，C_K に分類することである．すなわち，

$$D = C_1 \cup C_2 \cup \cdots \cup C_K$$

である．クラスタリングとしては，1 件の文書が単一のクラスタに属するように分割する場合と複数のクラスタに属することを許す場合がある．単一のクラスタに帰属させる場合はハードクラスタリング，複数クラスタへ帰属させる場合はファジィクラスタリングと呼ばれる[7]．インタビューやアンケート調査結果では，明確に意見が分類できないため，ファジィクラスタリングの適用が考えられる．しかし，実際には，ファジィクラスタリングを用いると，クラスタが重なり過ぎてしまい，異なる意見の抽出が困難になる．そのため，今回は，ハードクラスタリングの場合を考えた．ファジィクラスタリングについては，5 章の考察で述べる．

ハードクラスタリングとして，本稿では，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する．連結クラスタリング法は，文書を連結しながらクラスタを併合する方法であり，図 3 の手順で実行する．

まず，文書集合 D と求めるクラスタ数 K が入力されると，各文書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，文書ペア間の非類似度 NCD を計算し，最も類似度の高い（すなわち NCD の小さい）文書ペア d_x と d_y を探す．非類似度 NCD の計算結果の例を図 4 に示す．図では，1 列目は文書番号 x，1 行目は文書番号 y である．ここで，NCD は 2.1 章で述べたように対称とみなせるため，図に示すような上三角部分を計算すればよい．上三角部分のみの計算は，計算時間の短縮のためである．

ID	文書内容
1	無い.
2	とくにない.
3	PC はデザイン・機能面がよい.
:	
N	パソコンはデザインが良い.

図 1　文書集合 D の例

ID	文書内容
1	ない.
2	とくにない.
3	パソコンはデザイン・機能面がよい.
:	
N	パソコンはデザインがよい.

図 2　前処理後の文書集合の例

Algorithm:：連結クラスタリング法

Input：　D：文書集合 $\{d_1, d_2, \ldots, d_N\}$
　　K：クラスタ数

Output：　C：クラスタ集合 $\{C_1, C_2, \ldots, C_K\}$

Begin

$C=\{\{d_1\},\{d_2\},\ldots,\{d_N\}\}$

For k=1, 2, …, N-K

$$(d_x, d_y) = \underset{d_i, d_j \in D,\ i<j}{\arg\min}\ NCD(d_i, d_j)$$

$$d_x = d_x \cdot d_y$$

$$D = D \setminus d_y$$

$$C_x = C_x \cup C_y$$

$$C = C \setminus C_y$$

End

図 3　連結クラスタリング法

それから，最も NCD の小さい文書ペアを連結した d_x・d_y を改めて d_x とし，d_y を文書集合 D から削除する．ここで，文書ペアの連結とは，文書 d_x と d_y を順に並べた文書である．文書ペアを連結したときの文書集合 D の例を図 5 に示す．図 4 で，文書 1 と文書 2 の NCD が 0.1 と最も小さく，図 3 の文書 1「ない．」と文書 2「とくにない．」が選択されたとすると，図 5 に示すように，文書 1 が連結文書「ない．とくにない．」となり，文書 2 が削除され，文書集合 D の要素が 1 つ減少する．このような文書集合の変化により，NCD の値も変化する．図 5 の文書集合に対する NCD の例を図 6 に示す．図に示すように，文書 1 に関する NCD は変化し，文書 2 に関する NCD は削除され，それ以外の NCD は変化しないこととなる．

そして連結クラスタリング法は，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．ここで，クラスタ集合 C は，変化する文書集合 D に対して元の文書を保存し，次のクラスタ代表の抽出に用いるために用意する．図 5 の例の場合には，クラスタ集合は C ={{1,2},{3},...,{N}}となる．

これら文書ペア選択，文書の連結，クラスタの併合という動作を繰り返すことにより，連結クラスタリング法は，文書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．

提案する連結クラスタリング法は，文書を連結した新たな文書 d_x・d_y を用いて非類似度 NCD を計算する．そのため，クラスタリングの過程で元の d_x および d_y に関する非類似度 NCD は不要となり，新たな d_x に関する非類似度 NCD を計算する．これに対し，従来のクラスタリング法では，3.2 章で述べるように，元の文書間の類似度に基づいてクラスタ間の類似度を計算する．したがって，クラスタリングの過程で元の文書ペア間の類似度が不要とはならない．

このように，提案する連結クラスタリング法は，NCD の計算で文書の連結を用いるだけでなく，クラスタの併合でも文書の連結を用いる方法である．

2.4 クラスタ代表の抽出法

クラスタリング結果からの意見抽出のためには，クラスタに対する何らかの要約が必要である．クラスタの要約では，一般的にはクラスタ代表を抽出する方法が用いられており，クラスタ代表は，クラスタの重心に近い要素である [7]．しかし，提案法では，NCD を用いてクラスタ代表を抽出する．各クラスタ C_x に対し，C_x に含まれる各文書 d_{xi} と C_x との NCD を計算し，最も NCD の低い文書を抽出する．ここで，NCD の計算にはクラスタの要素の連結を用いる．すなわち，$C_x=\{d_{x1},d_{x2},...,d_{xp}\}$ とすると，

$$NCD(d_{xi},C_x)=NCD(d_{xi},d_{x1}\cdot d_{x2}\cdots\cdots d_{xp})$$

で計算する．

3. 従来法

本章では，従来の文書間の類似度計算方法，従来のクラスタリング法およびそれらの問題点について述べる．

	1	2	3	…	N
1	—	0.10	0.80	…	0.55
2	—	—	0.90	…	0.85
3	—	—	—	…	0.40
:	—	—	—	—	:
N	—	—	—	—	—

図 4　NCD の計算結果の例

ID	文書内容
1	ない．とくにない．
3	パソコンはデザイン・機能面がよい．
:	
N	パソコンはデザインがよい．

図 5　連結後の文書集合 D の例

	1	3	…	N
1	—	0.75	…	0.50
3	—	—	…	0.40
:	—	—	—	:
N	—	—	—	—

図 6　連結後の NCD の計算結果の例

3.1 従来の文書間の類似度

一般的には，文書間の類似度は，形態素解析により単語を抽出してから，含まれる単語の出現頻度などに基づいて計算することが多い[8]．例えば，文書 d_i と文書 d_j との間の類似度 $s(d_i, d_j)$は，a を文書 d_i に含まれる単語数，b を文書 d_j に含まれる単語数，c を文書 d_i と文書 d_j とに共通して含まれる単語数とすると，以下のような余弦係数，Dice 係数，Jaccard 係数，重複係数のいずれかによって計算することができる．

$$\frac{c}{\sqrt{a}\sqrt{b}},\quad \frac{2c}{a+b},\quad \frac{c}{a+b-c},\quad \frac{c}{\min(a,b)}$$

または，一般的に，文書を各語の重みから構成されるベクトル（w_{i1}, w_{i2}, ・・・, w_{iM}）として表現して計算することができる．ここで，w_{im} は，i 番目の文書における m 番目の単語の重みであり，M は単語数である．重み w_{im} の例としては，tf-idf（Term Frequency-Inverse Document Frequency）

$$w_{im}=\frac{n_{im}}{\sum_{m=1}^{M}n_{im}}\log\frac{N}{|\{d:t_m\in d\}|}$$

がよく知られている[9]．ここで，n_{im} は，文書 d_i での m 番目の単語の出現回数，N は全文書数，$|\{d:t_m\in d\}|$ は，m 番目の単語を含む文書数である．文書 d_i と d_j との間の類似度は，さまざまな方法で定義できるが，例えば，

$$\frac{\sum_{k=1}^{M} w_{ik} w_{kj}}{\sqrt{\sum_{k=1}^{M} w_{ik}^2} \sqrt{\sum_{k=1}^{M} w_{kj}^2}}$$

という数式によって類似度を計算する方法がある．

新聞記事などからのトピック抽出では，トピックごとに特別な単語が利用されるために，このような形態素解析を用いた文書間の類似度が有用である．しかし，製品・サービスに関するインタビューやアンケート調査結果では，意見は平易な文章で記述されており，意見の違いが通常と異なる特別な単語には基づかない．また，インタビューやアンケート調査結果のような非形式的な文書では，文法の誤りやタイプミスが散在する．形態素解析を用いた手法では，タイプミスが異なる単語として処理され，文書の類似性が認識できない．

従来の形態素解析を用いた手法には，複合語などの問題もある．例えば，文書に「印刷物」が含まれる場合，「印刷物」という単一の単語として処理するか，「印刷」と「物」という別の単語として処理するかを事前に判断する必要がある．あるいは，文書に「印刷する」が含まれる場合，「印刷」と同じ単語として処理するか別の単語として処理するか，事前に判断する必要がある．単純に，可能な全ての組合せを単語とみなして処理すると，複合語はさまざまな単語として何度も出現して類似度への寄与が大きく，複合語以外は，類似度への寄与が小さくなるなど，複合語かどうかという単語の特徴で類似度が異なるという望ましくない結果になってしまう．

提案法で用いる NCD では，単語を抽出してから類似度を計算するのではなく，文書全体の情報を用いるため，前述したような問題はない．一方で NCD は，同義語だけでなく，「ですます調」と「である調」の混在という文体の不統一の影響も受けてしまうため，適切にクラスタリングできないという問題があった．そのため，提案法では，変換ルールを用いた前処理により，事前に文体を統一した．

3.2 従来のクラスタリング法

代表的なクラスタリング法としては，階層的クラスタリング法がある．階層的クラスタリング法は，類似の文書を併合していき，階層的なクラスタを生成する方法である[6,10]．クラスタリング結果は，図 7 に示すようなデンドログラムとなり，クラスタリングを行った後でも，任意のクラスタ数に分割することができる．

階層的クラスタリングは，(1)N 件の文書に対し，各クラスタが 1 件の文書となる N クラスタとする，(2)最も近いクラスタを併合する，(3)クラスタ数が 1 になれば，動作を終了し，そうでなければ，ステップ(2)に戻る，という手順で実行される．ステップ(2)で最も近いクラスタを選択するには，最短距離法，最長距離法，Ward 法など，さまざまな方法がある．

最短距離法は，クラスタ C_1 と C_2 とに含まれる文書のうちで，最も近いものを選んで，クラスタ C_1，C_2 の類似度とする．

$$\min\{s(d_i,d_j): d_i \in C_1, d_j \in C_2\}$$

ここで，$s(d_i,d_j)$ は，文書 d_i と文書 d_j との類似度である．

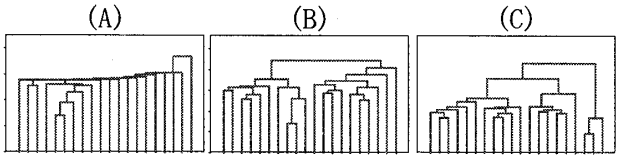

図 7　階層的クラスタリングの例．(A)最短距離法のデンドログラム，(B)最長距離法のデンドログラム，(C)Ward 法のデンドログラム

最長連結法は，逆に，クラスタ C_1 と C_2 とに含まれる文書のうちで，最も遠いものを選んで，クラスタ C_1，C_2 の類似度とする．

$$\max\{s(d_i,d_j): d_i \in C_1, d_j \in C_2\}$$

Ward 法は，クラスタ C_1，C_2 の類似度を以下の数式で計算する．

$$E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

ここで，$E(C_i) = \sum_{dj \in C_i} \|dj と C_i の重心との距離\|^2$ とする．

階層的クラスタリング法のうち，最短距離法は，クラスタ中の最も近い文書との距離を類似度とするため，1 件の文書よりも複数の文書の集まりであるクラスタの方が併合対象として選択されやすい．そのため，図 2（A）に示すように，鎖状の偏ったデンドログラムを作成し，いくつかのクラスタに分離しにくいという問題がある．これをチェイニング効果という．最長距離法は，1 件の文書とクラスタとでは，1 件の文書の方が選ばれやすいので，図 2（B）に示すように，最終的なクラスタに同じ数の文書が含まれやすいという問題がある．Ward 法は，階層的クラスタリング法の中でも，図 2（C）に示すように，最も明確なクラスタを生成しやすい手法である．

提案する連結クラスタリング法は，階層的クラスタリング法の一種であり，最短距離法と近いが，クラスタ間の類似度を連結文字列により計算する点が最短距離法と異なる．

4. 検証実験方法

提案法の有効性を確認するため，Web アンケート調査結果に対し，人手による意見の抽出と提案法および従来の文書クラスタリング法を適用して分析結果を比較した．

4.1 検証用データ

検証に用いたデータは，デジタル機器に関する調査結果であり，Web アンケートにより収集した．アンケートの対象地域は全国，男女はおよそ同数，年齢はデジタル機器のターゲットと想定される 10 代から 40 代までとした．アンケートは，2007 年 9 月 7 日に開始し，約 1 日間で 2,458 件を回収した．

アンケート内には，自由記述式の質問を 3 つ設定した．具体的な質問内容は，(A)「携帯電話が高機能化されてパソコンと同じような機能を持ちつつありますが，これについてどう思われますか？」，(B)「他のウォークマンな

どの携帯音楽プレーヤに比べ，iPod の魅力は何だと思われますか？」，(C)「iPhone について，どう思われますか？」である．ここで，（C）の iPhone は，2007 年 9 月時点では国内販売されておらず，事前評価となっている．

一般的に，自由記述式アンケートを直接実施すると，意見が少ないことが多いが，本アンケートは，[1,2]で用いたアンケートと同一であり，自由記述式の前に，携帯電話，パソコン，iPod，iPhone や個人の特性に関する選択回答式の質問も行った．このため，自由記述式として多くの意見を収集することができた．ここで，選択回答式の質問は，携帯電話やパソコンについては，所持の有無や機種，様々なサービスの利用頻度などであり，iPhone については，魅力や欠点などであり，個人の特性は，先進性やメディアへの接触頻度である．

アンケートの回答総数は 2,458 件だったが，検証においては人手による意見抽出に時間がかかりすぎて分類が困難だったため，各質問に対し，ランダムに選択した 200 件ずつの回答結果を用いた．

4.2 比較した手法

本稿では，人手による意見抽出を正解とし，提案法および従来法による分析結果を正解と比較する．

人手では，各質問に対する 200 件の回答結果を読んで類似な意見を集約していくことにより，意見を抽出した．集約の途中段階では，概ね同じだが細部が異なる意見が抽出されるため，できる限り意見の集約を続けて意見を抽出した．その結果，3 種のアンケート調査結果のそれぞれに対し，200 件の回答から約 20 件の意見を抽出した．

提案法では，3 種のアンケート調査結果のそれぞれに対し，クラスタ数 K を 20 として連結クラスタリング法を実施した後，クラスタ代表 20 件を抽出した．人手による意見抽出と比較するため，抽出したクラスタ代表 20 件だけを読むことにより，意見を抽出した．

従来法では，形態素解析を行った後，単語の Dice 係数を計算し，最短距離法，最長距離法および Ward 法でクラスタリングを実施した．形態素解析は，Mecab 0.97 を用いて実施した[11]．tf-idf を用いなかったのは，アンケート調査結果は主に平易な単語が用いられやすく，有用でなかったためである．例えば，アンケート調査結果において，「ない」は，意見を正反対に変える極めて重要な単語であったが，多くの文書に含まれ重みが小さくなってしまうなどである．従来のクラスタリング法は，STATA (version10)を用いて実施した[12]．また，連結クラスタリング法の有効性を確かめるため，NCD を用いた場合にも，最短距離法，最長距離法および Ward 法によるクラスタリングを実施した．これら従来法についても，提案法と同様に，クラスタ代表を抽出した後，クラスタ代表だけを読むことにより，意見を抽出した．

なお，提案法は，文書の表記揺れの解消も行っているが，クラスタリング性能自体を比較するため，文書の表記揺れを解消してから，人手による意見の抽出，提案法および従来法を適用した．

5. 実験結果と考察

文書ペアに対する NCD に基づく非類似度と Dice 係数に基づく非類似度の比較を図 8 に示す．図 8 において，（A）は携帯電話，（B）は iPod，（C）は iPhone の分布である．各図中には，200×199 個の点を示している．図に示すように，NCD と Dice 係数は，異なる値となる文書ペアが多かった．この結果から，NCD は Dice 係数とは異なる特徴を抽出することが分かる．また，NCD は，多くの場合，Dice 係数に基づく非類似度 より小さな値となった．例えば，Dice 係数が 0，すなわち，1-Dice 係数が 1 となり，重複単語がない文書ペアであっても，NCD は 1 以下となることがあった．

次に，（A）携帯電話についてクラスタリングした場合の各クラスタの文書数を図 9 に示す．図 9 において，(P)は提案する NCD に基づく連結クラスタリング法，（KS），（KC）および（KW）は，それぞれ，NCD に基づく最短距離法，最長距離法，Ward 法，（DS），（DC）および（DW）は，それぞれ，Dice 係数に基づく最短距離法，最長距離法，Ward 法によるクラスタリング結果である．ここで，人手による意見の抽出は，1 件の回答が複数の意見を含んでおり，明確にクラスタリングできなかったため，図示していない．また，Dice 係数に基づく最短距離法のみは，同じ類似度となるため 20 個のクラスタに分類することができず，21 個のクラスタに分類している．図に示すように，NCD でも Dice 係数でも，最短距離法は，1 つのクラスタ以外のクラスタは，1 つの文書しか含まなかった．同様の傾向は，iPod，iPhone のクラスタリング結果でも見られており，最短距離法は，チェイニング効果のため適切にクラスタリングできなかったと考えられる．したがって，以降は，最短距離法を比較の対象から除いた．ここで，連結クラスタリング法は，最初のクラスタ併合までは NCD に基づく最短距離法と同じである．しかし，以降は連結文書を作成することで最短距離法とは異なっており，連結クラスタリング法は，チェイニングを起こさず適切にクラスタリングできることが図 9 により確認できた．

人手により抽出された意見が，提案法および従来法で抽出できたかについて，（A）携帯電話の結果を図 10，（B）iPod の結果を図 11，（C）iPhone の結果を図 12 に示す．図では，左から，人手で抽出した意見，提案法（P），NCD に基づく最長距離法（KC），NCD に基づく Ward 法（KW），Dice 係数に基づく最長距離法（DC），Ward 法（DW）を示しており，クラスタの代表として各意見の内容が抽出されていれば○，抽出されていなければ空欄とした．最終行は，各クラスタリングにおける意見の抽出数を示す．

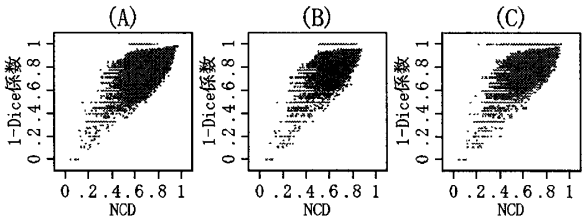

図 8　NCD と Dice 係数に基づく非類似度の分布．(A)携帯電話，(B)iPod，(C)iPhone に関する分布

図 10 に示すように，携帯電話に関しては，人手により 18 件の意見が抽出された．そのうち，提案法では 88.9%（16/18）の意見が抽出できたのに対し，NCD に基づく最長距離法では 77.8%（14/18），NCD に基づく Ward 法では 55.5%（10/18），Dice 係数に基づく最長距離法では 50.0%（9/18），Dice 係数に基づく Ward 法では 50.0%（9/18）の意見しか抽出できなかった．
また，図 11 に示すように，iPod に関しては，人手により 18 件の意見が抽出され，提案法では 66.7%（12/18）が抽出できたのに対し，NCD に基づく最長距離法では 72.2%（13/18），NCD に基づく Ward 法では 66.7%（12/18），Dice 係数に基づく最長距離法では 50.0%（918），Dice 係数に基づく Ward 法では 38.9%（7/18）の意見が抽出できた．なお，NCD に基づく最長距離法は，同一の NCD が計算されて 20 個のクラスタに分類できなかったため，21 個のクラスタに分類しており，21 個の代表からの意見抽出数を示している．

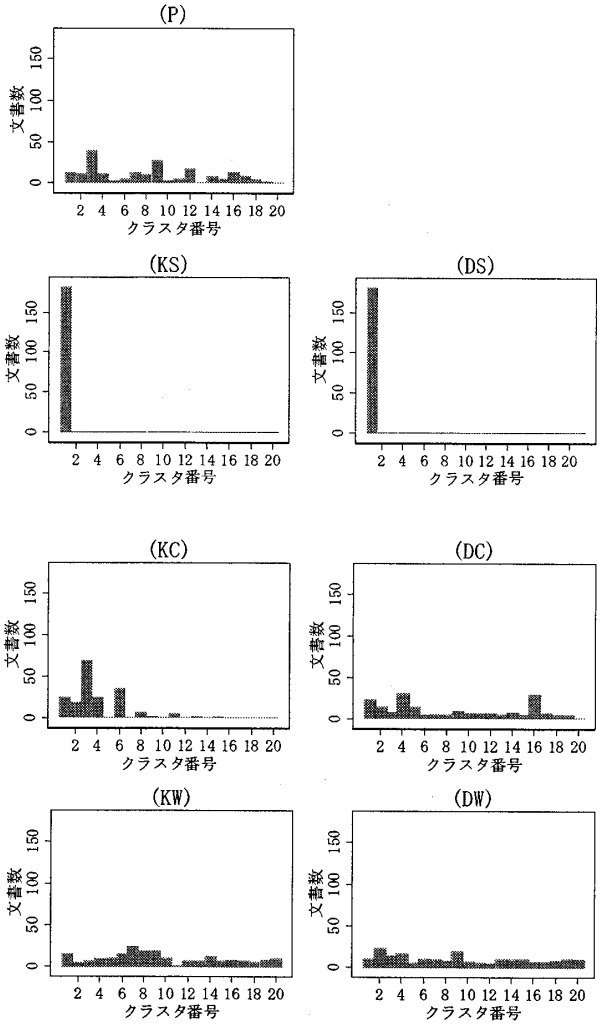

図 9　クラスタごとの文書数．（P）提案する NCD に基づく連結クラスタリング法，（KS）NCD に基づく最短距離法，（KC）NCD に基づく最長距離法，（KW）NCD に基づく Ward 法，（DS）Dice 係数に基づく最短距離法，（DC）Dice 係数に基づく最長距離法，（DW）Dice 係数に基づく Ward 法

人手で抽出した意見	P	KC	KW	DC	DW
よい・便利	○	○	○	○	○
高機能すぎる・不必要	○	○	○	○	○
機能は不十分	○	○			
小型・軽量でよい	○	○			○
電池が不安	○	○	○	○	○
料金が高い・安いならよい	○	○	○	○	○
画面が小さい・見にくい	○	○	○	○	○
操作性が悪い	○	○	○	○	
性能が低い	○	○			○
セキュリティ不安・リスクある	○	○			
故障が不安	○	○	○	○	
自分では使いこなせない	○	○	○	○	
パソコンとの使い分け必要	○	○	○		○
パソコンとのシンクロ	○				
e-mobile，au から発売を期待		○			
運転時など倫理教育必要	○				
未来情報化社会・高度 IT 社会	○				
ない・わからない			○	○	○
抽出数＝18	16	14	10	9	9

図 10　携帯電話の意見抽出結果

人手で抽出した意見	P	KC	KW	DC	DW
容量が大きい	○	○	○	○	○
デザインがよい	○	○	○	○	○
小型・薄い・軽い	○	○	○	○	○
ネット経由・パソコンと連携	○	○	○	○	○
ソフトを自由・無料で入手可能	○	○	○		
CD や MD が不必要で便利	○	○	○	○	
カスタマイズできる	○	○	○		
操作性がよい	○	○	○	○	
コストパフォーマンスがよい			○		
料金が高い		○			
消費電力が少ない					
映像を見ることができる		○		○	○
音がよい					
ブランド・知名度・話題性	○	○	○	○	○
魅力は薄れてきている	○				
アップルが好き	○	○			
アップルが好きじゃない			○		
ない・わからない	○	○	○		○
抽出数＝18	12	13	12	9	7

図 11　iPod の意見抽出結果

人手で抽出した意見	P	KC	KW	DC	DW
よい・便利	○	○		○	○
デザインがよい	○	○	○	○	
デザインが悪い					
多機能すぎる	○	○	○		
機能不十分・既存の携帯機能必要	○		○	○	
操作性がよさそう	○	○	○		
操作性が悪そう	○	○	○	○	○
料金が高そう	○	○	○	○	○
故障が不安	○	○		○	
電池が不安			○	○	
セキュリティが不安	○	○			
サイズ・重さが不安	○	○	○		
性能が不安					
他の機器がある			○		
自分では使いこなせない	○	○	○	○	○
携帯電話会社との連携が必要					
アップルに興味はあるが・話題性	○	○		○	
新しさは少ない	○				
不必要	○	○		○	○
ない・わからない	○	○	○	○	○
抽出数＝20	15	13	11	11	6

図 12　iPhone の意見抽出結果

図 12 に示すように，iPhone に関しては，人手により 20 件の意見が抽出され，提案法では 75.0%（15/20）の意見が抽出できたのに対し，NCD に基づく最長距離法では 65.0%（13/20），NCD に基づく Ward 法では 55.0%（11/20），Dice 係数に基づく最長距離法では 55.0%（11/20），Dice 係数に基づく Ward 法では 30.0%（6/20）の意見しか抽出できなかった．

今回の検証実験の結果，提案法である NCD に基づく連結クラスタリング法は，比較的多くの意見を抽出できており，提案法の有効性が確認できた．また，同一のクラスタリング法を用いると，NCD に基づく方が多くの意見を抽出できており，アンケート調査結果の分析におけるコルモゴロフ複雑性の概算である NCD の有効性が確認できた．

形態素解析を用いた類似度計算では，比較的重要度の低い記号などを除いて名詞・形容詞・動詞などに限定することがある．しかし，アンケート調査結果は，「ない」，「便利」，「賛成」などのように短文を多く含んでおり，形態素を限定すると，同じ距離が多く 20 個のクラスタに分類できないことから，形態素を限定しなかった．

本稿では，Dice 係数の実験結果を示した．これは，余弦係数，Jaccard 係数，重複係数についても検証実験を行ったところ，Dice 係数の場合が最も意見抽出数が多かったためである．Dice 係数以外の場合のそれぞれの意見抽出数は，−2〜＋1 個の範囲内であった．

クラスタリング法としては，階層的クラスタリング法以外にも，k-means 法，k-medoid 法がよく知られている．

k-means 法（ここで，k は，クラスタ数 K を示す）は，(1) K 個のベクトル $(\mathbf{c}_1,\cdot\ \cdot\ \cdot,\mathbf{c}_K)$ を初期値として生成する，(2) N 件の文書を，それぞれ，最も近いベクトル $\mathbf{c}_k$ に割り当てる，(3)ベクトル $\mathbf{c}_k$ に割り当てられた文書のベクトルの重心に更新する，(4) ベクトル $(\mathbf{c}_1,\cdot\ \cdot\ \cdot,\mathbf{c}_K)$ が変化しなくなれば処理を終了し，そうでなければステップ(2)に戻る，という手順で実行される[6,8]．k-means 法には，計算量が少ないという利点がある．しかし，クラスタ数 K やクラスタの核となるベクトルの初期値を事前に設定しなければならず，初期値に依存して局所解に陥るという欠点がある．初期値をランダムに選択すると，含まれる文書数の多い大規模クラスタに含まれる文書が多く選択されやすく，小規模クラスタが分割されにくいという問題があった．また，k-means 法は，同一サイズの球状の領域に分割されやすかった[13]．製品・サービスに対しては，多数意見・少数意見が存在する．そのため，小規模クラスタを分割しにくい k-means 法は，本稿では比較の対象としなかった．

k-medoid 法は，クラスタの 1 つの文書で代表させ，その文書との類似度を用いてクラスタリングを行う方法である[13,14]．k-medoid 法は，k-means 法において，K 個のベクトルを生成する代わりに K 個の文書を選んで，その文書ベクトルを初期値とし，各 C_k のベクトルを，

$$\arg\min_{di\in C_k} \sum_{dj\in C_k\setminus di} \left\|\mathbf{d}_i-\mathbf{d}_j\right\|$$

に更新する．k-medoid 法は，代表値を用いるので，k-means 法に比べ，外れ値の影響を受けにくく，ノイズに頑強な手法である．しかし，k-means 法と同様に，クラスタ数 K やベクトルの初期値を事前に設定しなければならず，初期値に依存して局所解に陥り，小規模クラスタが分割されにくいという問題は同様であった．そのため，k-medoid 法についても，本稿では比較の対象としなかった．

コルモゴロフ複雑性に基づくクラスタリング法として，Quartet 法が提案されており，DNA 配列，言語，音楽のクラスタリングで有効性が確認されている[15,16]．しかし，Quartet 法は，ランダム・山登りに基づいて進化を分析する手法であり，一方，製品・サービスの価値評価では進化に基づかず意見を抽出したかったため，Quartet 法は検討しなかった．

2.2 章で述べたファジィクラスタリング法は，1 件の文書に対し各クラスタへの帰属度を計算する．ファジィクラスタリング法についても，アンケート調査結果を用いて検証実験を行った．ファジィクラスタリング法としては，最も一般的なファジィ c-means 法[7]を用い，クラスタ代表は，クラスタの重心から最も近い文書とした．ファジィ c-means 法も，初期値に依存した局所解探索であるため，検証実験では，20 クラスタへの分類を 10 回試行した．その結果，20 クラスタの代表文書には重複が多く，重複を除くと，(A)携帯電話では，2〜5 件，(B)iPod の魅力では，3〜7 件，iPhone では 2〜7 件の代表文書しか抽出できず，抽出された代表文書も概ね異なっていた．このように，ファジィクラスタリング法では，クラスタが重なり過ぎ，試行のたびに抽出できる代表文書の数も代表文書自体も異なり，意見抽出が困難であることが確認された．代表文書が 20 件抽出されるよう多くのクラスタに分割したとしても，試行のたびに変化する結果の比較は困難なため，

本稿ではファジィクラスタリング法を比較の対象としなかった．

本稿では，クラスタ数 K を 20 として検証実験を行った．これは，人手による意見抽出数の約 20 件と抽出結果を比較するためである．検証実験でなく実際に意見抽出を行う場合は，事前にクラスタ数を決めることはできない．その場合は，予め NCD の閾値を決めておき，閾値より大きな NCD の文書ペアしかない場合にクラスタリングを終了する．ここで，NCD の閾値をやや高めにすることで，多めのクラスタに分割して意見抽出すればよい．本稿で比較した方法は，提案法，従来法とも階層的クラスタリング法のため，クラスタ数に結果が大きく依存しない．階層的クラスタリング法では，クラスタ数 K を 1 増やしても，K 個のクラスタのうちのいずれかが 2 分割された結果が得られるためである．したがって，異なるクラスタ数でも同様の結果が得られると考えられる．

6. 結論

本稿では，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，インタビュー・アンケート調査結果から製品・サービスを評価する手法を提案した．Web アンケート調査結果 3 種に適用し，従来法よりも多くの意見を抽出できることを確認した．また，同一のクラスタリング法を用いるとコルモゴロフ複雑性の概算である NCD に基づくほうが Dice 係数に基づくより多くの意見を抽出することができ，アンケート調査結果のような非形式的な文書におけるコルモゴロフ複雑性による非類似度計算の有効性が確認できた．

本稿では，1 件の文書を 1 つのクラスタに分類するハードクラスタリングを考えたが，アンケートの回答は，複数の意見が含まれる場合が多く，各回答を 1 つのクラスタに分類することは困難だった．それでも提案法は，人手に比べ，70%程度の意見抽出が可能だった．今後，1 件の回答を複数に分割するなどして，より意見の抽出精度を向上させたい．また，コルモゴロフ複雑性として，圧縮に基づかない手法も提案されており[16]，今後，検討していきたい．

参考文献

[1] 藤原由希子, 富沢伸行, 井口浩人, “変化分析型決定木を用いた製品・サービスの普及予測”, 第 22 回人工知能学会全国大会, 2008.

[2] 藤原由希子, 富沢伸行, 井口浩人, “閾値関数による変化分析型アンサンブル学習を用いた製品・サービス普及予測”, 第 7 回情報科学技術フォーラム(FIT), RF-006, 2008.

[3] C.H. Bennett, P. Gács, M. Li, P. Vitányi and W. H. Zurek, “Information Distance”, IEEE Transactions on Information Theory, Vol.44, No.4, pp.1407-1423, 1998.

[4] E. Keogh, S. Lonardi C. A. Ratanamahatana, L. Wei, S-H. Lee and J. Handley, “Compression-Based Data Mining of Sequential Data”, Data Mining and Knowledge Discovery, Vol. 14, No. 1, pp.99-129, 2007.

[5] M. Li, X. Chen, X. Li, B. Ma and P. Vitányi, “The Similarity Metric”, IEEE Transactions on Information Theory, Vol.50, No.12, pp.3250-3264, 2004.

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

[7] A.K. Jain, M.N. Murty and P.J. Flynn, “Data Clustering: A Review”, ACM Computing Survey, Vol. 31, No. 3, pp.264-323, 1999.

[8] 岸田和夫, “文書クラスタリングの技法：文献レビュー”, Library and Information Science, No.49, pp.33-75, 2003.

[9] 多田知道, 岩沼宏治, 鍋島英知, “イベント時系列マイニングを目的とする新聞記事からの時系列情報に基づく単語抽出”, 第 7 回情報科学技術フォーラム(FIT), F-047, 2008.

[10] 神嶌敏弘, “データマイニング分野のクラスタリング手法（1）”, 人工知能学会誌, Vol.18, No.1, pp.59-65, 2003.

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

[12] http://www.stata.com/

[13] S. Guha, R. Rastogi and K. Shim, “CURE: an Efficient Clustering Algorithm for Large Databases”, Proceedings of ACM SIGMOD International Conference on Management of Data, pp.73-84, 1998.

[14] R. T. Ng and J. Han, “CLARANS: A Method for Clustering Objects for Spatial Data Mining”, IEEE Transactions on Knowledge and Data Engineering, Vol.14, No.5, pp.1003-1016, 2002.

[15] R.Cilibrasi and P. Vitányi, “A New Quartet Tree Heuristic for Hierarchical Clustering”, Theory of Evolutionary Algorithm, 2006.

[16] C. Long, X. Zhu, M. Li and B. Ma, “Information Shared by many Objects”, Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), pp.1213-1220, 2008.