LD-9　自然言語処理技術と構造化パターンマイニングを用いたFAQ生成支援システム

FAQ Generation Support System Using Structured Association Pattern Mining and Natural Language Processing

松澤 裕史†
Hirofumi MATSUZAWA

1．まえがき

企業では、コールセンターなどにおいて顧客からの問い合わせに対して電話応対を行っている。通常、コールセンターでは、顧客との電話応対後、その内容を要約してデータベースへ入力している。集められた問い合わせ内容は、分析担当者が経験や勘に基づいて、応対記録文書を読みながら分析し、FAQ（FAQ：Frequently Asked Question）の把握に努めてきた。数多く寄せられる問い合わせからFAQを迅速に発見できれば、コールテーカーに周知して電話応答の所要時間を減少させることができる。また、Web上に公開することで、電話件数の減少によるコスト削減も可能であり、顧客も電話せずに自力で問題を解消できることができる。問い合わせ内容から製品の問題を早期に発見できれば、新製品での改善に繋げることもできる。しかし、担当者が全ての問い合わせ記録を読むことは現実的ではない。実際に、月に十数万通の電話がある企業もあり、人手により頻度の高い問い合わせを全て把握することは不可能である。

近年、テキストマイニング技術がコールセンターの分析に用いられている[1][2]。テキストマイニングシステムを用いて、キーワードの出現頻度、係り受けや共起からコールを分析することで、どの製品に問い合わせが多かったのかを把握することができる。しかしながら、FAQの具体的内容までを把握することは困難である。本稿では、「○○．を△△したら、◇◇が□□した」などの文意を捉えることを考慮したFAQ作成支援システムを紹介する。このシステムは、データマイニング技術を応用して開発したプロトタイプである。実際のコールセンター業務に適用して、成果を上げることができた。本システムで用いられたデータマイニングのアルゴリズムは[3]にあるので、本稿では紹介しない。

2．FAQ作成支援システム

2.1　コールデータ

我々が対象とするコールセンターのデータベースから取得するデータについて説明する。データベースは、数十万件のコールデータの集合である。1件のコールデータは、識別子、タイトル、定型データ、文書データから構成される。

定型データとは、システムによって規定されたカテゴリーとその値からなり、通常、複数個のカテゴリーから構成される。例えば、PCのコールセンターであれば、マシンタイプやパーツナンバーなどが規定されている。日付データもこの定型データに含まれる。

文書データは、通常のテキストからなる。ただし、システムによっては、自動的に挿入されるようなテンプレートが含まれていることがある。また、コールテーカーに対する指導、あるいは、システムの機能として、入力された文章の質問部分と回答部分を分離可能な場合がある。このような場合には、質問部分だけを処理対象とできるように分離する。これは、質問部分だけからよくある質問を発見するためである。

2.2　構造化パターンとそのマイニング

自然言語処理技術を用いたテキストマイニング技術[1]では、構文木から単語間の係り受け関係を抽出してデータの分析を行っている。本システムは、構文木を直接扱わずに、2段の木構造に簡略化することで、データマイニング技術への適用を行った。図1に、実際の入力例とその処理の例を示す。図1にあるように、構文木から動詞とその動詞に直接、または、間接的に係り受け関係のある動詞以外の単語をグループとするような2段階の構造を構築している。以下、これを構造化データと呼ぶ。

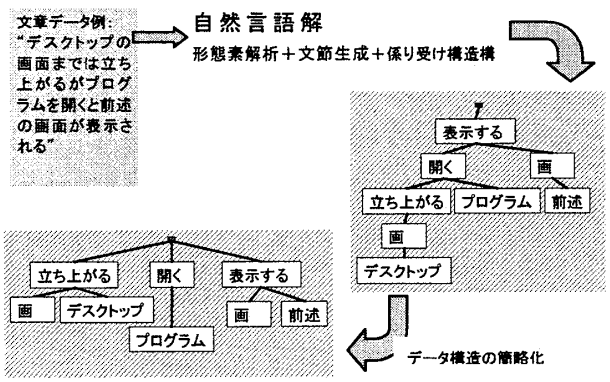

図1：自然言語処理による構造化データの生成

我々は、大量の構造化データから頻出するパターンを高速に発見する技術を開発した[3]。この技術を用いることにより、図2に示す3件の構造化データから、図2右に示す頻出構造化パターンを発見することができる。この例では、最小サポートを2とした。ここで、大文字英字は動詞に相当し、小文字は他の単語である。例えば、図2における3単語のパターン（右下）で、もし、Aが“教える”、Cが“増設する”、sが“メモリ”であれば、元の文書が「メモリの増設について教える」ということであることが想像される。共起単語だけのパターンよりも文意のわかりやすいパターンになっている。

†日本アイ・ビー・エム（株）東京基礎研究所

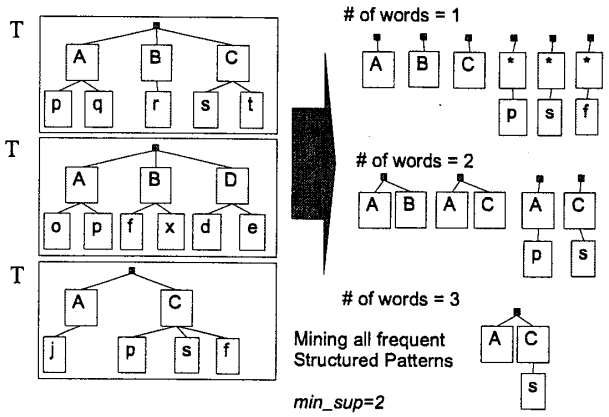

図２：構造化パターンのマイニング

2.3 FAQ作成支援システムの構成

本システムでは、以下の処理を順次適用していくことで、FAQの作成を行う。

1. データ前処理
2. 自然言語処理
3. データ選択処理
4. 頻出構造化パターンのマイニング
5. パターンブラウザを用いた元文書の抽出

2.3.1 データ前処理

データ前処理では、コールセンターのデータベースからデータを取得して、クレンジングを行う。文書データ中に含まれるテンプレートの除去も行う。このテンプレートを除去しないと、テンプレートが頻出するパターンとして発見されてしまう。また、質問部分と回答部分の分離が可能であれば、質問部分だけを対象となるように処理する。

2.3.2 自然言語処理

取り出した文書の全ての文に対して、自然言語処理を行い、図１にあるような構造化データを作成する。

2.3.3 データ選択処理

次工程で発見する構造化パターンは、全ての頻出するパターンである。頻出するパターンの数は非常に多く、多岐に渡る内容が含まれている場合、具体的な問題の発見が困難になるため、データセットを絞り込む必要がある。絞り込みは、定型データの特定カテゴリー（例えば“ハードウェア”）から単一または複数の値（例えば“モデム”）を選択する。日付による絞り込みも可能である。

2.3.4 頻出構造化パターンのマイニング

絞り込まれたデータについて、[3] のマイニングアルゴリズムを用いて、頻出する構造化パターンを発見する。具体的には、図２の構造化パターンが発見される。

2.3.5 パターンブラウザ

頻出するパターンだけで、問題が特定可能な場合が稀にあるが、実際には、元の文書を読まなければ、問題が特定できない。そこで、パターンブラウザを用いて、発見されたパターンを含む元の文書群を閲覧することで、問題を特定し、FAQ作成を行う。元の文書を読むという作業が生じているものの、特定の問題だけに絞り込まれているので、これらの文書群からFAQの作成を容易に行うことができる。また、データマイニング一般の問題であるが、発見されるパターンは非常に数が多く、面白いパターンを見つけることは容易ではない。そこで、このパターンブラウザは、単語数、動詞の数、特定の単語が含まれているかなど、複数の指標からパターンをフィルタリングするための機能を提供し、選択したパターンの元文書をインタラクティブに提示する機能を提供している。

3. 適用事例

我々は、構築したプロトタイプを用いて、実際のコールセンターのデータを用いて、特定の製品に対する１か月分の17万件のコールデータからFAQの作成作業を行った。何度かの試みにより、パターンとして着目すべきものは、単語数が３個以上であり、動詞以外の単語を含むパターンが有益であることがわかっていたため、パターンブラウザのフィルタリング機能を用いて、有益なパターンを複数取り出した。ここで、有益かどうかは、実際のコールセンターの分析担当者により判断された。

最終的に、一つの製品に対して、複数の構造化パターンが取り出され、その元文書を担当者が読み通し、最終的に19個のFAQ候補が発見された。そのうち、2個は、既知であったため、最終的に17個のFAQを生成することができた。この作業に要した時間は、22人日であった。この作業は、期末商戦に向けてのFAQ生成が目的であったため、直前の迅速なFAQ生成に貢献することができた。

4. おわりに

自然言語処理技術と構造化データマイニング技術を用いたFAQ作成支援システムについて紹介した。従来のデータマイニング技術でも見られるように、パターンを発見するだけでは、発見された大量のパターンから、どのパターンが有益であるかを判断することが困難である。我々は、パターンブラウザを構築することで、パターンをフィルタにかけて絞込み、元文書を提示することでFAQ作成支援を行った。データマイニングでは、データの集計結果を示して検討を行うが、テキストデータの場合には、集計結果だけでなく、該当する元文書の内容を読むことで、知見を得ることがある。この点は、データマイニングとテキストマイニングの差異として非常に興味深い。

参考文献

[1] 那須川哲哉、諸橋正幸、長野徹：テキストマイニング—膨大な文書データの自動分析による知識発見—，情報処理，Vol.40, No.4, pp. 358-364, 1999.

[2] 市川由美、中山康子、赤羽俊男、三好みよ子：関口寿一、藤原康祐：日報分析システムの開発、電子情報通信学会 技術研究報告 NLC2000-26, pp.31-38, 2000.

[3] 松澤裕史：大規模データベースからの頻出構造化パターンの抽出，情報処理学会論文誌：データベース，Vol.42, No. SIG8(TOD10), pp.21-35, 2001.