Aprioriアルゴリズムを用いた共起ルール抽出に関する実証研究

(株)ケイテック
丸山　優輔

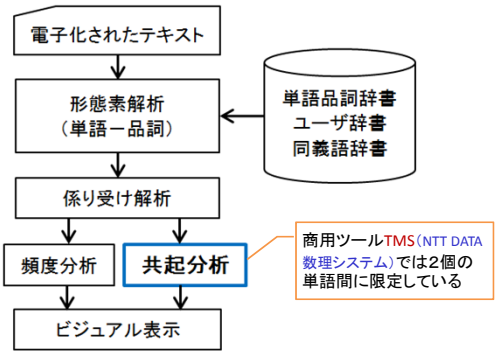
東京工芸大学工学部
コンピュータ応用学科
宇田川　佳久

目　次

1. 研究の背景と概要

- 大量のデータが日々蓄積されている
- テキストデータは人間のコミュニケーションの中核を構成する
- これまでに様々な分析技法が開発されてきた
- aprioriアルゴリズムは、「頻出する単語の集合」や「共起ルール」を効率的に発見できる
- 本文では、
  (1) N個の単語から構成される共起ルールの生成方法について実験を行い
  (2) 有意な共起ルールを生成するための方策を考察した

データマイニングの概要



形態素解析

形態素解析：

自然言語で書かれた文を形態素（意味を持つ最小単位）に分割し、それぞれの品詞を判別する処理

今回の実験では、Text Mining Studioの形態素解析ツールを採用した
（係り受け解析も行える）

形態素解析の例

お客様に有用なシステムを提供するビジネスを展開しています.

お客様に有用なシステムを提供するビジネスを展開しています.

単語ID	見出し語	原形	置換語	品詞	係り先
1	お客様に	お客様	お客様	名詞	4
2	有用な	有用	有用	名詞	3
3	システムを	システム	システム	名詞	4
4	提供する	提供	提供	名詞	5
5	ビジネスを	ビジネス	ビジネス	名詞	6
6	展開しています	展開	展開	名詞	-1

①お客様に有用なシステムを提供するビジネスを展開しています.
②お客さまに有益なシステムを提供する事業を推進しています.

同義語
「お客様」「お客さま」　　「有用」「有益」
「ビジネス」「事業」　　「展開」「推進」

単語ID	見出し語	原形	置換語	品詞	係り先
1	お客様に	お客様	お客様	名詞	4
2	有用な	有用	有用	名詞	3
3	システムを	システム	システム	名詞	4
4	提供する	提供	提供	名詞	5
5	ビジネスを	ビジネス	ビジネス	名詞	6
6	展開しています	展開	展開	名詞	-1
7	お客さまに	お客さま	お客様	名詞	10
8	有益な	有益	有用	名詞	9
9	システムを	システム	システム	名詞	10
10	提供する	提供	提供	名詞	11
11	事業を	事業	ビジネス	名詞	12
12	推進しています	推進	展開	名詞	-1

単語頻度解析
　特定の単語の発生個数

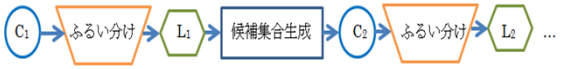
共起語
　が連続して発生する複数の語のこと
　例：コロケーション（連語関係）

現実的には N語の共起語が考えられるが、
商用の TMS では、2語の共起語しか解析しない

3. Apriori アルゴリズムとルール抽出

- Aprioriアルゴリズムは、1993年にAgrawalらが提起し、バスケット解析を現実のものとした。

No	項目
100	A C D
200	B C E
300	A B C E
400	B E



C1	支持度
A	2
B	3
C	3
D	1
E	3

L1	支持度
A	2
B	3
C	3
E	3

C2	支持度
A B	1
A C	2
A E	1
B C	2
B E	3
C E	2

L2	支持度
A C	2
B C	2
B E	3
C E	2

```
/*  APRIORI アルゴリズム  */
LS= Φ;          // 条件を満たすデータの累積
k= 1
Fk= { i | i ∈ {長さ1の頻出集合} };
repeat
      k = k+1;
      Ck = apriori_gen(Fk-1 );    // 候補の生成（要素の組合わせ）
      Tk= Φ;
      for ( each c ∈ Ck ) {
          if ( c ∈ T  &&  |t| ≧ N × minSup ) {
                LS.add(c);        // 条件を満たすデータの累積
                Fk.add(c) };      // このループで見たかったデータ
          }
      }
until Fk = Φ;
Result = LS;
```

パラメータ：閾値（支持度）

相関ルールの検出

指定された 確信度 以上の時、ルール
X ⇒ Y　を生成する。

確信度（X, Y）

$$= \frac{\text{条件 X と Y を含むすトランザクション数}}{\text{条件 X を含むトランザクション数}}$$

$$= \frac{\sigma(X \cup Y)}{\sigma(X)}$$

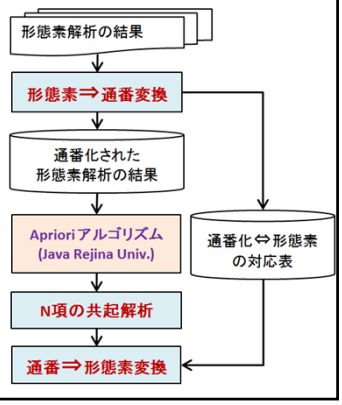
リフト（X, Y）

$$= \frac{\text{確信度（X, Y）}}{\text{条件 Y を含むトランザクション数}}$$

$$= \frac{\sigma(X \cup Y)}{\sigma(X) \cdot \sigma(Y)}$$

頻出アイテム集合の生成実験

- Canada Rejina大学が公開しているAprioriプログラム（Java）を採用した
- 本研究では、通番変換、N項の共起解析プログラムを開発した



実験対象のデータについて

- 本学の企業説明会に参加した企業から学生へのメッセージ文である.
- 単語の集合（アイテム集合） 333個
- ユニークな単語は1,450個

単語ID	見出し語	原形	置換語	品詞	係り先
1	ABCテクノロジーズは	ABCテクノロジーズ	ABCテクノロジーズ	名詞	10
2	グループ全体を	グループ全体	グループ全体	名詞	3
3	支える	支える	支える	動詞	4
4	技術スペシャリスト集団として、	技術スペシャリスト集団	技術スペシャリスト集団	名詞	10
5	高品質な	高品質	高品質	名詞	6
6	技術・サービスの	技術	技術	名詞	7
7	提供基盤となり、	提供基盤	提供基盤	名詞	10
8	特に	特に	特に	副詞	10
9	保守・運用の	保守	保守	名詞	10
10	面で、	面	面	名詞	16
11	お客様に	お客様	お客様	名詞	12

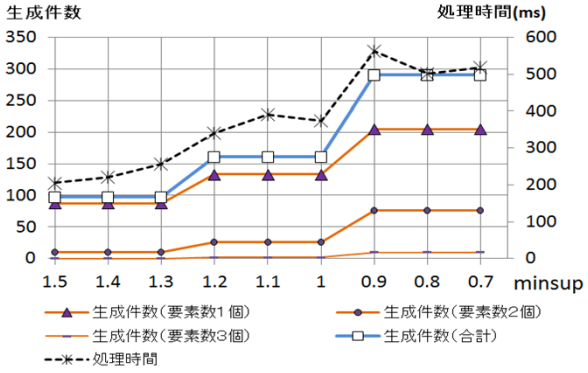
形態素 ⇒ 通番変換

テストデータ(行数:333、単語数:1450)

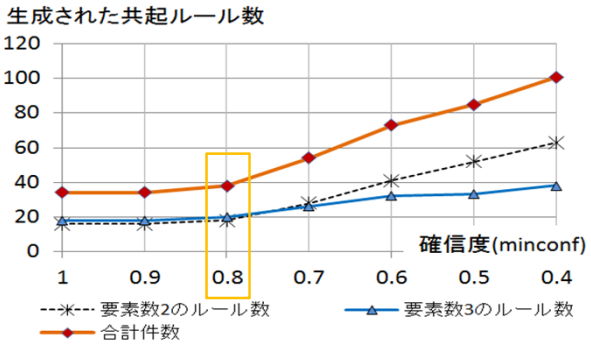
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32 33 34 35 36 37
38 39 40 41 42
43 44 45 46 47 48 49 50 51 52 53 54 55 56
24 57 58 59 60 61 62 63 64
11 15 22 65 66 67 68 69 70 71 72 73 74 75 76 77
31 78 79 80 81 82 83 84 85 86 87 88
＜中略＞
11 13 16 77 156 218 619 628 736 896 1014 1439
1440 1441 1442 1443
24 25 46 156 194 370 440 676 1257 1444 1445
1446 1447 1448 1449
13 60 218 628 1204 1293 1450

1, ABCテクノロジーズ
2, グループ全体
3, 支える
4, 技術スペシャリスト集団
5, 高品質
6, 技術
7, 提供基盤
8, 特に
9, 保守
10, 面
11, お客様
12, 安心

頻出単語の抽出実験の結果



共起ルール生成の実験の結果



{**研修、力、入れる**}の3個の要素から生成されたルール
・元々の意味を反映しているように解釈できる

研修, 力 → 入れる cf 1.0
研修, 入れる → 力 cf 1.0

{**大企業、オンリー、ワン**}の3個の要素から生成されたルール

・元々の意味を反映していないルールも生成 されている

大企業 → オンリー, ワン cf 1.0
オンリー → 大企業, ワン cf 1.0
ワン → 大企業, オンリー cf 1.0
大企業, オンリー → ワン cf 0.8
大企業, ワン → オンリー cf 0.8
オンリー, ワン → 大企業 cf 1.0

「大企業よりもオンリーワンを目指す」

考察

- 相関ルール生成では、元々の意味を反映しないルールも生成されることが判明した。

↑

ルールが組み合わせ的に生成されるため

- 頻出単語集合、共起ルールを生成した際の最少支持度と最少確信度**(パラメータ)に基準がない**

↑

行数や要素数がデータによって大きく異なるため、ある値よりも小さくすると膨大なデータが生成される

データ数(単語数)に換算して 2個あるいは 1個

おわりに

改善策としては、「品詞」や「係り先」情報をデータに含めることが考えられる

- 「品詞」情報を含める
  - 帰結部が名詞，動詞になるものは優先順位を上げる
  - 帰結部に副詞があった場合は優先順位を下げる
- 「係り先」情報を含める
  - 単語の発生順序を考慮する
    (ただし、能動態、受動態では語が反転する可能性があり、より深い構文解析が必要)

ご清聴ありがとうございました