A-004

研磨された共起ネットワークからの内容マイニング
Content Mining on a Polished Co-occurrence Network

越智 麻生汰 † 河村 泰之 ‡ 宇野 毅明 §
Maota Ochi Yasuyuki Kawamura Takeaki Uno

1. はじめに

近年, 記憶装置の大容量化や安価な計算機の普及によって, データを収集することが容易となった. データ量が膨大となり, データ分析に必要な労力も大きくなりつつある.

テキストマイニングも例外ではなく, 単語のような大量の特徴量を扱うために, 計算時間は増加している. その上, マイニング手法を適用することで得られるデータ量も膨大なものとなり, より一層, データクリーニングや考察に費やす時間が増えている.

一方で, 巨大なデータの大まかな構造を維持しながら, 扱いやすい中規模なデータに変換する新しいマイニング技術が開発されている. その内の一つであるグラフ研磨と呼ばれる手法では, グラフの構造を明確化することで, 巨大なデータからまとまりを見つけることができる. 顧客の店舗選択モデル構築 [1] や,Twitter の話題変化抽出 [2] 等, 応用範囲は非常に広い.

本稿では, えひめ結婚支援センター「愛結び」のメール文章に現れる単語の共起ネットワークに, グラフ研磨を適用することで調査しやすい形に変換し, それで得られた語のまとまりごとに共起ネットワークを再度作成することで, お見合いの相談内容を分析する.

2. 用いるデータについて

えひめ結婚支援センターは, 少子化の主たる要因の一つである未婚化・晩婚化の対策のため, 県からの委託によって設立され, 講座や出会いの場の提供, 交際サポート等を公益目的で行っている. 「愛結び」はセンターによって運営されている情報システムであり, プロフィール登録, お見合いパートナー探し, ボランティアとのメールによる交際中のサポート等を行う.

メールは, ユーザーとボランティアの間で行われ, 主に悩み相談やスケジュール調整など, 広い範囲で活用されている. 各メールには, 発信者ユーザー id の他に, ユーザーを識別する属性が付けられている.

これらのデータはセンターから提供されたものであり, 個人を特定できないように処理されている. データの収集期間は 2011 年 10 月から 2015 年 1 月までで, 総メール数は 117,881 通となる. 相談はユーザーから発信されるので, 発言者属性がユーザーでかつ, お見合い関係のメールに限定すると, 対象となるメールは 39,733 通になった.

3. 共起ネットワーク

本文では. 共起ネットワークを用いて当該メールの話題調査を行う. 共起ネットワークとは, 単語を頂点とし,

例文1：一郎は二郎が描いた絵を三郎に贈った
例文2：三郎は二郎へ絵のお礼の手紙を描いた

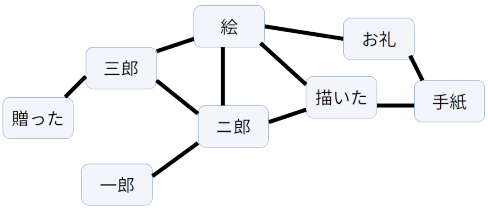

図 1: 共起ネットワークの作成例

単語間の出現パターンに基づいて頂点対を接続したグラフである. 図 1 のように, 助詞のような機能語を無視し, 隣接する 2 単語を接続したものとする. このネットワークによって, ある単語がどのような話題を構成しているか等を, 複数文章を直接読むことなく考察することができる.

今回用いるデータは, 内容が多様な文章であるため, 非常に複雑なグラフが出来上がる. そのため, 目視で全体像を把握することは困難となる.

比較的出現頻度の高い単語のみを用いてグラフを構成する方法も考えられるが, 少数ながらも有用な情報を切り落とす可能性が出てくる. また, 話題内容を限定して調査する方法も考えられるが, 文章話題の推定や, 調査話題を含む文章を決める作業という新たな問題が生じる.

4. グラフ研磨

そこで, まず共起ネットワークに対してグラフ研磨を行う. それによって類似度の高い単語がクリークとして現れる. クリークを構成する単語を抽出し, その単語を含む文章のみで再度共起ネットワークをつくることで, ある程度内容のまとまりがみられる文章ごとに話題調査を行う.

Algorithm 1 に疑似コードを記す. 元グラフの全頂点ペアに対して類似度を求め, 類似度が閾値以上であれば繋ぎ直すことを, グラフに変化がなくなるか, 実験者が指定した最大回数まで繰り返し行う. 高速な実装に関しては参考文献 [3] が詳しい.

類似度 $sim(u, v)$ を算出する式には様々な尺度があり, 今回は次のように定義される normalized PMI を用いる:

†愛媛大学 教育学部 総合人間形成課程 情報教育コース
‡愛媛大学 教育学部
§国立情報学研究所 情報学プリンシプル研究系

$$sim(u,v) = \frac{\log \frac{P(u)P(v)}{P(u,v)}}{-\log P(u,v)}.$$

$P(u)$ は素性 u の出現確率, $P(u,v)$ は素性 u と v の共起確率である. グラフにおいては u, v を頂点としたとき,

$$P(u) = |N(u)|/|V|,$$
$$P(u,v) = (|N(u)| \cap |N(v)|)/|V|$$

で定義される. これは, 単語の共起度合いを計測する指標である PMI の値域を -1.0〜1.0 になるように正規化したもので, グラフに適用した場合は, 頂点ペアで共通している近傍頂点が多いほど 1.0 に近づく.

類似関係で接続し直すことで, つながりの薄い頂点を取り除き, つながりの強い頂点を小規模のクリークにまとめることができる. これにより, 元のグラフ構造を考慮しつつ, データを見やすい形にすることができる.

共起ネットワークにおいても, 頻出する文脈パターンによってクリークやそれに近い形をなす部分や, 次数の高い語と特定話題内で頻出する語によってスター型をなす部分など, 特徴的な部分において, 頂点間の類似度が高くなるため, グラフ研磨によって話題の中核となる語を抽出できると考える.

Algorithm 1 グラフ研磨のもっとも単純な実装

```
V:頂点集合, E:辺集合, σ:類似度下限値
E' = φ; V' = φ;
for all u ∈ V do
  for all v ∈ V do
    if sim(u,v) > σ then
      E' = E' ∪ (u,v)
      V' = V' ∪ u
      V' = V' ∪ v
    end if
  end for
end for
return (V', E')
```

5. 実験と考察

グラフ研磨によってできあがるクリークは類似度の高い単語の集まりであり, 一つのクリークに含まれる単語に絞って再度共起ネットワークを構成することで, ある話題に限定した単語の関連構造を観察できる.

以下に手順を記する.

1. 正規化, 文分割, 形態素解析を行う
2. 隣接する 2 語が接続しあう共起ネットワークを作成する
3. グラフ研磨を行う
4. 手順 3. より得たグラフから極大クリークを抽出する
5. 手順 4. より得られた各クリークに対して, クリーク内の単語のうち 1 語でも含む文章のみで共起ネットワークを作成すること を行う.

2. では名詞, 動詞, 形容詞のみ対象とし,5 回以上同パターンが出現した場合に接続をしている. また, 出現頻度が高く, 意味が薄いと思われる以下の単語を除外している.

> する, いる, ある, こと, れる, なる, よう, いたす, もの, せる, ため, よい, てる, やる, あと, くださる, なれる, あう, いろいろ, どちら, みたい, そう, いつ, つもり, とこ, いま, ごと, いい, 思う

類似度 normalized PMI の閾値は 0.7 を用いた.

クリークの抽出については, サイズが小さいと, 手順 5. で取得する文章が極端に少ないことが多いので, サイズにも閾値を設けた. 本稿では, サイズが 4 以上のものを求める.

図 2 は研磨前のグラフ, 図 3 は研磨後のグラフの一部となる. グラフからサイズ 4 以上のクリークが 94 個抽出された. 表 1 に一部を掲載する.

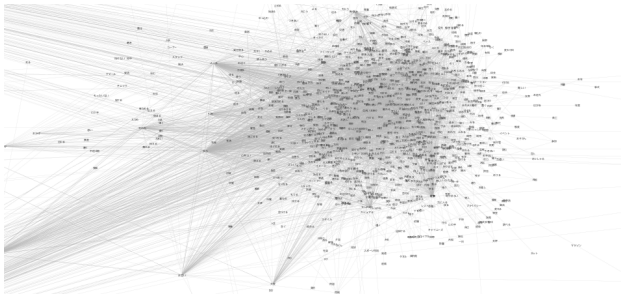

図 2: 研磨前の共起ネットワーク

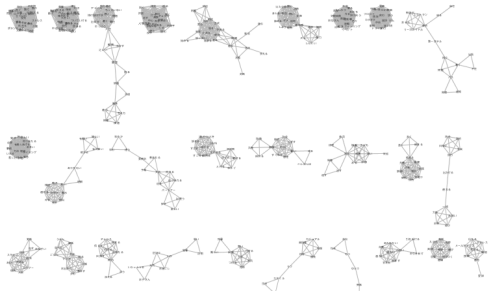

図 3: グラフ研磨を適用した共起ネットワーク

研磨前のグラフからは, 話題の中心と思われる次数の大きな部分がいくつか確認できるが, そこから文の流れ

表 1: 抽出されたクリーク一覧 (一部)
{ あたる, 乱文, 用件, 長文 }
{ いか, お過ごし, 元気, 区切り }
{ くれる, 折り返し, 折り返す, 留守電 }
{ スタート, 伺い, 解除, 開始 }
{ お年, お正月, お迎え, 迎える }
{ 回復, 帰れる, 段階, 気長 }
{ おく, 縮まる, 縮める, 置く }
{ ひく, 寝込む, 引く, 風邪 }
{ もてる, 好感, 持てる, 自信 }
{ いただき, 下す, 心遣い, 足労 }
{ 乏しい, 生かす, 豊富, 離婚 }
{ 助かる, 嬉しい, 有り難い, 有難い }
{ きく, 利く, 効く, 融通 }
{ パートナー, 出逢う, 探す, 新しい }
{ 充実, 有意義, 短い, 過ごせる }
{ 実家, 帰省, 年明け, 未定 }

をくみ取ることは困難である. 一方, 研磨後のグラフからは, 同一表現, 表記ゆれや, 同一文中に同時に使われても違和感を覚えない単語によってできているクリークが多く見られた. この中から, 一部のクリークについて手順 5. を行い, 話題の考察を行う.

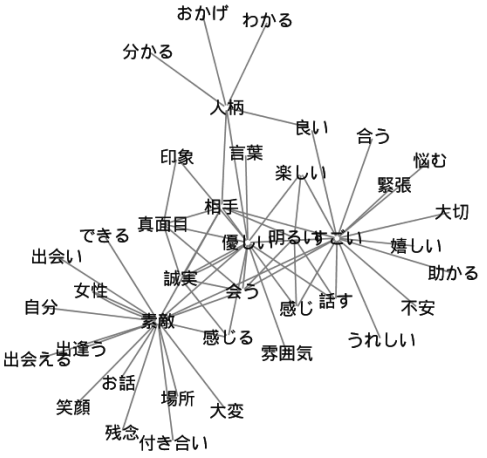

図 4: { すごい, 人柄, 優しい, 明るい, 真面目, 素敵, 誠実 } とその近傍のネットワーク

5.1. 相手方にプラスイメージをもつ話題

図 4 に { すごい, 人柄, 優しい, 明るい, 真面目, 素敵, 誠実 } のいずれかを含む文章による共起ネットワークから上記の語と近傍を抽出したものを図示する.

「優しい」,「明るい」,「すごい」の共通接続ノードに 「楽しい」,「感じ」 がある. 「優しい感じで楽しい」といったように, 一緒にいて楽しいと思われることの要因であると考えられる.

「素敵」の近傍ノードに「場所」,「お話」,「笑顔」があり, 「お話が素敵で〜」のように, 相手を素敵と感じる際の要因であることと同時に, 同研磨後クリーク中の「人柄」,「優しい」,「明るい」などを感じる際の間接的要因になっているとも思われる.

また,「優しい」,「真面目」,「誠実」 が研磨前のグラフでもクリークを成しており, 文中でも「優しく真面目な〜」,「誠実で優しく〜」といった具合に, 強調するために重ねて使われていると考えられる.

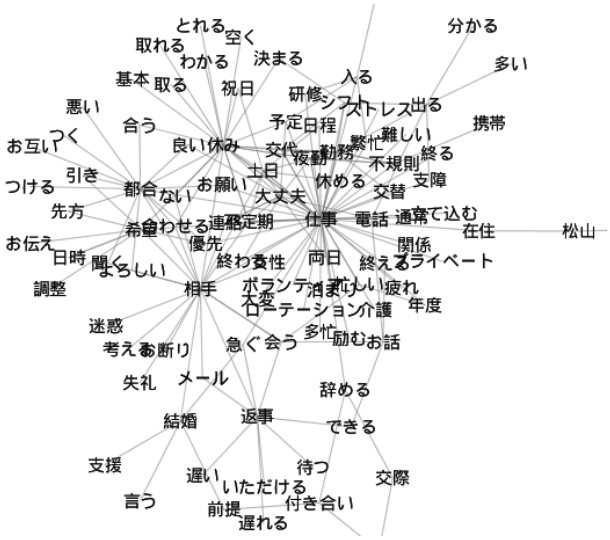

図 5: { ストレス, プライベート, ボランティア, ローテーション, 不定期, 両日, 交代, 交替, 介護, 休める, 優先, 励む, 在住, 夜勤, 年度, 急ぐ, 支障, 泊まり, 疲れ, 研修, 立て込む, 終える, 終る, 繁忙, 辞める, 通常 } とその近傍のネットワーク

5.2. 仕事にまつわる話題

図 5 に { ストレス, プライベート, ボランティア, ローテーション, 不定期, 両日, 交代, 交替, 介護, 休める, 優先, 励む, 在住, 夜勤, 年度, 急ぐ, 支障, 泊まり, 疲れ, 研修, 立て込む, 終える, 終る, 繁忙, 辞める, 通常 } のいずれかを含む文章による共起ネットワークから上記の語と近傍を抽出したものを図示する.

「仕事」を中心に,「研修」,「夜勤」など, それに関連する語が多く抽出された. 「都合」「希望」といった, 日程調整の際に用いられる単語も見られ, スケジュール調整によってお見合いがうまくいかないケースがあると思われる.

また「仕事」と「付き合い」,「交際」が「辞める」 によって接続されてかつ,「仕事」が「介護」や「松山」,「在住」とつながっており, 勤務地や介護, 仕事などと結婚の両立が, 結婚に至るまでの障害であると解釈することができる. 「相手が仕事を辞めたくなく, 付き合いを断られた」,「介護のため仕事を辞めて県外に出る」などといった文章が考えられる.

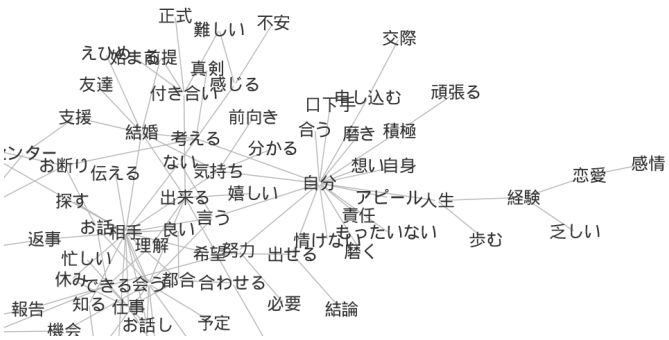

図 6: { もったいない, アピール, 人生, 出せる, 努力, 口下手, 情けない, 想い, 申し込む, 磨き, 磨く, 積極, 責任 } とその近傍のネットワークの一部

5.3. 自己評価が低いユーザーの話題

図 6 に { もったいない, アピール, 人生, 出せる, 努力, 口下手, 情けない, 想い, 申し込む, 磨き, 磨く, 積極, 責任 } のいずれかを含む文章による共起ネットワークの一部を図示する.

「自分」を中心に,「口下手」のような否定的な語が目立つ.「××な自分にはもったいない」 ,「口下手で話が弾まない」,「自分が情けない」のように, 自己評価が低い, 欠点を自認していると思われるようなユーザーがいると考えられる.

6. まとめ

本論文では共起ネットワークをグラフ研磨することで, 観察者による恣意的な文章選択や, 中～少数データの消去による情報の損失をできる限り排除しつつ, 内容が多様でデータサイズの大きなテキストから, まとまった話題ごとの調査をおこなうことができた.

一方で, クリークから抽出した単語集合の解釈が新たに必要となり, そこに意味があるかどうかは観察者の判断にゆだねられる. また, あまりにも低頻出の共起パターンを残した場合, クリークから得られる単語集合が, 非常に少数の文章からできた特異的なものになりかねない. 今後の課題として, パラメタの調整を行うために, 得られた結果の良し悪しを定量的に判断する方法の検討が挙げられる.

謝辞

本研究の一部は科研費 16K12539 の助成を受けたものである。

参考文献

[1] 中原 孝信, 羽室 行信, 宇野 毅明,「グラフ研磨手法を用いた顧客の店舗選択モデルの構築」, オペレーションズ・リサーチ, 60(2), 89–95, 2015.

[2] 前川 浩基, 内田 将史, 大内 章子, 宇野 毅明, 羽室 行信,「データ研磨手法を用いた Twitter ユーザの関係構造変化の検出」, 人工知能学会全国大会論文集, Vol28, 1–4, 2014.

[3] 宇野 毅明, 中原 孝信, 前川 浩基, 羽室 行信,「データ研磨によるクリーク列挙クラスタリング」, 情報処理学会アルゴリズム研究会報告書, 2014–AL–146(2), 1–8, 2014.