1F-03

複数情報源による共起語グラフ可視化アプリケーションの提案

堺　初穂†　　佐賀　亮介‡

大阪府立大学大学院工学研究科‡

近年、テキストマイニングの分野で共起グラフの分析によるキーワードや話題の抽出が多くなされている。しかし著者や掲載紙などの情報源によって、同じ話題でも含まれている語が異なる可能性がある。そこでグラフを情報源ごとに生成して分析し、その結果を同時に提示することで、情報源ごとの特徴が明確になると考えられる。本論文では情報源ごとに共起グラフを作成し、重ね合わせて表示させるアプリケーションを提案する。

キーワード：共起グラフ, 可視化

1.はじめに

新聞などのメディアやブログなどの情報源には、嗜好や思想などのコンテキストに基づいたバイアスが存在しており、情報消費者は、そのバイアスを含んだ情報を獲得している。各情報源のコンテキストは、新聞記事における社説や書評を通したテキストデータにおいて主張されることが多い。たとえば、佐賀らの新聞記事の比較分析[1]では、二者間において、同じ話題について話すとしても、主張が異なっていることが示されている。このように、情報消費者は情報源特有のバイアスをもつ情報を受けてしまうため、複数の情報源から得た情報を客観的に処理するためには、情報源の特徴や差異を把握する必要がある。

そこで、本論文では、テキストマイニングの分野で使用されている共起グラフを基にした複数情報源を可視化するアプリケーションを開発する。

2.関連研究

共起グラフは、語をノード、共起関係をエッジとしたネットワークグラフであり、共起とは 2 つの語が同じ文書に同時に出現することである。共起グラフを用いた研究は数多くなされている。共起グラフ分析による話題の抽出や文書クラスタリングなどの研究がなされており[2][3], また語義曖昧性解消にも共起グラフが用いられている[4]。しかしながら、これらの手法は単一の情報源についての研究であり、複数のものを取り扱ってはいない。

本論文は、複数の情報源を対象に共起グラフを生成し、合成・統合することにより、各情報源の特徴や情報源間の差異を抽出するものである。

3. アプリケーションの構成

図 1 に、提案するアプリケーションの概要を示す。本アプリケーションは複数の情報源のデータベースと共起グラフ生成部とグラフの統合部、そして GUI からなる。共起グラフ生成部では、各情報源から共起グラフを生成し、話題を特定する。グラフの統合部では、複数の情報源から生成された共起グラフを統合し、GUI にてその結果を表示する。ユーザは、GUI を通して統合したグラフを閲覧し、情報源の特徴や差異を発見する。

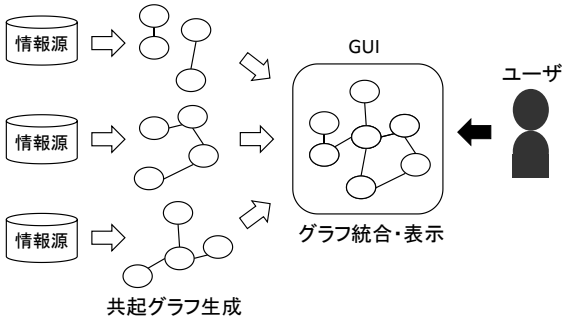

図 1　アプリケーション概要

3.1　共起グラフ生成

共起グラフを生成するために、情報源ごとにキーワードを抽出したのち、そのキーワードに基づいて共起を算出する。キーワード抽出を行うために、TF-IDF 法や出現回数を用いる。そして、抽出されたキーワード間の共起関係を情報源ごとに計算する。本研究では、共起度として Jaccard 係数を用いた。単語 *X,Y* の Jaccard 係数 *J(X,Y)*は以下の式で求められる。

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X + Y - X \cap Y|} \quad (1)$$

ここで|・|は集合数を表し, $X \cap Y$ は X と Y の単語を含む文書集合を示す。Jaccard 係数が閾地を超えた場合、エッジが存在するものとしてノード間にエッジを引く。

3.2　グラフクラスタリング

話題は、共起グラフ上において、複数の単語間でクリークを生成している箇所から暗黙的に把握することができる。これを自動化するために、クラスタリングを行う。本論文では Newman が提案した Modularity によるクラスタリング（Newman 法）[5]を行う。Newman 法の基本的な流れは以下のとおりである。

Step1:各ノードを 1 つのクラスタとし、ノードの数だけクラスタを作成する。

Step2: Modularity が最も高くなる、クラスタを統合する。

ここで、Modularity Q は以下の式で求められる。

$$Q = \sum_{i=1}^{N}(e_{ii} - a_i^2) \quad (2)$$

式(2)において、N はクラスタ数、e_{ij}は総エッジ数に対する、コミュニティ i がクラスタ j とにつながっているエッジの割合、a_iは総エッジ数に対する少なくとも片方がクラスタ i に含まれるエッジの割合である。

ただし、上記モジュラリティQは計算量が膨大であるため、Q の代わりに Q の増分 Δq を用いて最大増分となるクラスタを統合していく。このとき、クラスタ i と j を統合した時の Δq は以下の式で求められる。

$$\Delta q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j) \quad (3)$$

この Δq が負の値になるまでクラスタリングを繰り返していく。

この Newman 法により求まったクラスタ内において、頻度や TF-IDF 値などを元に複数のラベル候補を作成し、このラベル情報を話題としてノードの属性として保持させる。

3.3　グラフ統合

グラフを統合するために、各共起グラフにおけるノードとエッジを元にひな形となるグラフ（スーパーグラフ）を作成する。つまり、ある共起グラフ i がノード Vi とエッジ Ei により $Gi = G(Vi, Ei)$ と表せるとし、各共起グラフを $G1$, $G2$, …, Gn としたとき、スーパーグラフ Gs は次の式にて定義される。

$$Gs = G(\bigcup_i Vi, \bigcup_i Ei) \quad (4)$$

この Gs をひな形と、Gs 上の各要素（ノードとエッジ）が共起グラフ群において登場している頻度を保持する。話題については、先述したように、同じ話題でも語は異なる使われ方をする可能性があり、逆にいうと、同じ語でも違う話題に登場する可能性がある。そのため、グラフ統合の際に、各ノードとエッジのクラスタ情報は各々保持する。

3.4　グラフ表示とユーザ操作

グラフ表示は、GUI 上で実行されるものである。統合したグラフにおいて、ユーザにより、情報源の種類やグラフ要素の登場頻度などによりフィルタをかけることを可能とする。

統合したグラフにおける特徴は色や形などにより表現される。たとえば、情報源を特徴としたとき、異なる色でノードやエッジを描画するようにする。また、複数のノードやエッジが混在しているものは、ノードはクラスタごとに異なる色で描画する。図 2 にアプリケーションのプロトタイプ例を示す。

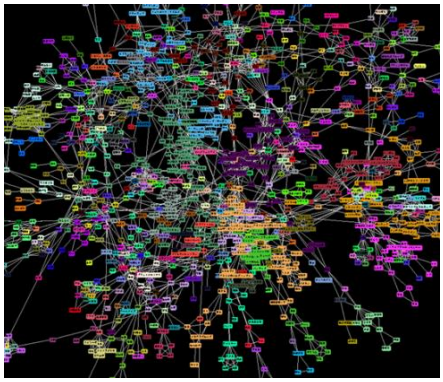

図 2　グラフ表示のイメージ

4.おわりに

本論文では、情報源ごとの違いを明確化するため、複数の情報源から共起グラフを生成し、統合、表示させるアプリケーションを提案した。

今後の課題として、より多くの情報源から共起グラフを生成し、統合したグラフから分析を行うことが考えられる。

参考文献

[1] R. Saga, H. Tsuji., Comparison Analysis for Editorials by Reversible FACT-Graph, Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2011). (2011):216-221.

[2]大澤幸生, 谷内田正彦, KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会論文誌 D 82.2 (1999): 391-400.

[3]倉由佳里,小林一郎,. 単語の共起グラフを用いた潜在的意味に基づく効果的な文書分類の検証. 人工知能学会 インタラクティブ情報アクセスと可視化マイニング研究会(第 4 回)(2013):29-33

[4]鏑木雄太,古宮嘉那子,小谷善行,. 共起語グラフのクラスタリングによる単語の多義性抽出. 言語処理学会 第 17 回年次大会 発表論文集(2011):508-511

[5] M. Newman, Fast algorithm for detecting community structure in networks. Physical review E 69.6 (2004): 066133.