

グラフに基づくクラスタリングによる動詞類義語の獲得

竹内 孔一[†] 高橋 秀幸[†] 小林 大介^{††}

[†] 岡山大学大学院自然科学研究科 〒700-8530 岡山県岡山市津島中 3-1-1

^{††} 岡山大学工学部情報工学科 〒700-8530 岡山県岡山市津島中 3-1-1

E-mail: †koichi@cl.cs.okayama-u.ac.jp

あらまし 本研究ではクラスタリングを利用して動詞の類義語を獲得する方法について検討している。先行研究において、同時に2つの要素のクラスタを考慮しながらクラスタリングする同時共起クラスタリングを適用して、ベクトルベースのクラスタリングより精度が高いことを示した。しかし、近年ベクトルベースのクラスタリングでKernel K-meansという非線形境界でクラスタリングするより高度な手法が提案された。そこで、本報告ではKernel K-means法を我々の動詞類語獲得に適用し、先行研究における同じデータで同時共起クラスタリングとの比較を行う。この結果からKernel K-meansでのグラフ-ベクトルの等価性には限界があり、本研究が利用する2部グラフの構造は直接反映できないこと、先行研究の同時共起クラスタリングの方が高い精度で類語を獲得できることを報告する。

キーワード 動詞類語, 同時共起クラスタリング, Kernel K-means

Extraction of Verb Synonyms Using Graph-Based Clustering

Koichi TAKEUCHI[†], Hideyuki TAKAHASHI[†], and Daisuke KOBAYASHI^{††}

[†] Graduate School of Natural Science, Okayama University
Tsushimanaka 3-1-1, Okayama-shi, Okayama, 700-8530 Japan

^{††} Department of Information Technology, Faculty of Engineering, Okayama University
Tsushimanaka 3-1-1, Okayama-shi, Okayama, 700-8530 Japan

E-mail: †koichi@cl.cs.okayama-u.ac.jp

Abstract This manuscript describes evaluation results of Kernel K-means clustering approach comparing with modified Aizawa's co-clustering approach for verb synonym extraction task. Kernel K-means approach is one of the state-of-the-art vector-based clustering method which can divide vector-spaces with non-linear boundary by incorporating Kernel method. Besides the mathematical framework of Kernel K-means can cover Spectral Graph Clustering. In this manuscript, however, we reveal Aizawa's co-clustering approach overcomes Kernel K-means on the verb synonym extraction task (bi-graph clustering) in Japanese. From this results we discuss that the equivalence between graph-vector space in Kernel K-means approach can be limited, and then Kernel K-means decrease their accuracy in our verb synonym extraction.

Key words Verb synonyms, Co-clustering, Kernel K-means Clustering

1. はじめに

本研究では動詞の類義語集合をテキストから獲得し、述語項構造を整理した辞書の構築を目指している^(注1)。動詞の類義語集合とは項構造レベルでの言い換えが可能な集合を仮定している。例えば、

- 京都/ヨーロッパを 散策した/うろついた/回った/巡った

など、【地域を移動しながら回るといった意味】^(注2)を共有していると考えれば、この例は名詞と動詞の組でほとんど言い換え可能であり、これを語義と考える。この集合を獲得するタスクをここでは動詞類義語獲得と呼ぶことにする。当然、動詞は多義であるため、「巡る」にしても「薬が体内を巡る」ならば他の語義【循環】として「薬が循環する」など他のクラスタに重複して登録されるべきである。

先行研究[4]ではテキストデータを係り受け解析して名詞(助

(注1): 人手により整理した動詞項構造シソーラス(4425語(7473語義))[5]
<http://cl.cs.okayama-u.ac.jp/rsc/data/>.

(注2): 以下では語義を表現するときは □ で示す。

詞)-動詞の2部グラフを作成して(図1参照),そこから上記のような名詞と動詞が緊密に関連する部分を獲得する同時共起クラスタリング(以下CCREと呼ぶ)を提案し,ベクトルベースのシングルクラスタリングの手法(PLSIを用いた方法)より多くの類義語集合を獲得できることを示した. CCREはソフトクラスタリングで動詞,名詞とも多重にクラスタに属することを許すため,多義性に対応している.

一方で,近年ハードクラスタリングではあるが,グラフと等価なベクトル空間を作成し,Kernel法による非線形境界を利用したクラスタリング(Kernl K-means法(以下KK法))が提案された[2]. KK法の利点としてグラフベースのクラスタリングを含む形で最適化関数が一般化されているため,我々のタスクの基本データをそのまま入力することが可能である.そこで,全く同様のデータ(名詞(助詞)-動詞)というペアを入力することで直接CCRE法との比較を行うことができる.

そこで本研究では動詞類語獲得タスクに対してCCRE法とKK法を実験的に比較する.その結果,CCRE法が精度高くかつより多くの動詞類語を獲得したこと,KK法のグラフ-ベクトル空間の等価性は距離としての再定義があり,2部グラフなど構造を持ち込む場合にはグラフの構造変換が必要であることを提案する^(注3).

2. 同時共起クラスタリング

提案する同時共起クラスタリングCCREはAizawaのco-clustering[1]を元に,複数回グラフからクラスタを抽出することで動詞の多義性を考慮したソフトクラスタリングを実現する手法である.詳細は文献[4]に譲るとして,ここでは核となる考え方と特徴を示す.

まずテキストをCaboChaで係り受け解析したあと,名詞(助詞)-動詞の対になった2部グラフを構成する(図1参照).この

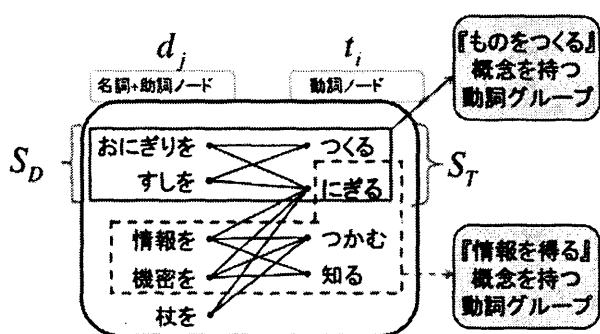


図1 係り受け解析後の名詞(助詞)-動詞対の2部グラフから類語を取り出す例

うち初期クラスタからリンクが密になっている部分を取り出す.その部分をクラスタ候補(S_T, S_D)とすると,初期クラスタ全体の情報量についてクラスタ候補(S_T, S_D)を作った方が情報量が大きいときに有効なクラスタであると判断して(S_T, S_D)をクラスタとする.具体的には下記式(1)式が正のときに(S_T, S_D)

をクラスタとする[1].

$$\delta L(S_T, S_D) = P(S_T, S_D) \log \frac{P(S_T, S_D)}{P(S_T)P(S_D)} - \sum_{t_i \in S_T} \sum_{d_j \in S_D} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)} \quad (1)$$

図1では $S_T = \{ \text{にぎる, つくる} \}$, $S_D = \{ \text{おにぎりを, すしを} \}$ をクラスタ候補として $t = \{ \text{つくる, にぎる, つかむ} \}$, $d = \{ \text{おにぎりを, すしを, 情報を, 機密を, 杖を} \}$ の要素から選んだ結果を示している. CCREではさらに,多義動詞のために,ここで得られたクラスタを削除して,繰り返しクラスタリングを行うことで今度は $S'_T = \{ \text{にぎる, つかむ, 知る} \}$, $S'_D = \{ \text{情報を, 機密を} \}$ を獲得する.つまり,図1の点線部分で囲まれた動詞グループまで獲得する.

CCREの特徴はこのクラスタ候補を初期クラスタからどうみつけるかである.基本的には名詞と動詞の全組の計算は時間がかかり過ぎるため,評価式[1]を利用して寄与度の少ないノードを削除していく.この消す順序により得られるクラスタが変わることから一種の山登り法となる.

ノードがクラスタに対してどの程度寄与しているかを評価する式として

$$\delta I(t_i, S_D) = \sum_{d_j \in S_D} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)} \quad (2)$$

$$\delta I(S_T, d_j) = \sum_{t_i \in S_T} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)} \quad (3)$$

が提案されている.これにより,基本的には t_i については式(2), d_j については式(3)の値の低いノードから削除する.

3. Kernel K-mans 法

Kernel K-means法[2]はベクトル空間上でK-meansに対してKernel法による非線形クラスタリングを導入したものである.特徴的なのはその定式化はグラフベースのクラスタリングを包含する点である^(注4).グラフとはノードとノード間を接続するエッジからなっており,エッジに対して重みを与えることでノード間の強さを表現するものである.この入力されたグラフ構造から集団としてまとめたいノード集団をエッジをカットすることでクラスタリングとして求める.よって入力データはノード,エッジとエッジの重みであり,出力としてクラスタが得られるが,内部ではベクトル空間として再認識されてKernel法を利用することで,非線形境界でクラスタを作り出す.

詳細は文献[2]に譲るとして,要点のみを説明する.グラフ構造を入力とした場合,クラスタリングは各クラスタ要素内のリンクが密になるように計算するRatio associationを考える^(注5).今,あるクラスタc内のノード集合を N_c とし,ノード

(注4): 文献[2]の他に下記の生駒日記ブログが概略の経緯を理解するのに役立つ. <http://d.hatena.ne.jp/mamoruk/20090128/pl>.

(注5): その他,クラスタ間でのリンクを減らすRatio cutなどいくつかの方法がある[2].

(注3): 残念ながら構造変換した後の計算結果は本稿記述時点で得られていない. KK法の計算が止まらなくなり原因を究明中である.

ド間のエッジを $links(N_c, N_c)$ と置くと、各クラスタ内での密度最大化は下記の式を最大化させると置くことができる。

$$RatioA(G) = \max_{N_1, \dots, N_c} \sum_{c=1}^k \frac{links(N_c, N_c)}{|N_c|}. \quad (4)$$

ここで G はクラスタリングされたグラフを示し、 k はクラスタ数、 $links(N_c, N_c)$ は与えられたクラスタ内ノード集合同士のエッジの総和を表しており、下記の式で計算する。

$$links(X, Y) = \sum_{i \in X, j \in Y} B_{i,j}. \quad (5)$$

この時、 $B_{i,j}$ はノード i から j へのエッジの重みを示しており、エッジが無ければ 0 である。ここで、クラスタ c に対してベクトル \mathbf{y}_c を仮定する。 \mathbf{y}_c とはあるクラスタ c にノード i が含まれていたとき $\mathbf{y}_c(i) = 1$ を示し、その他は 0 となるベクトルである。この \mathbf{y}_c を式 (4) に適用すると下記の式が得られる。

$$\max \left\{ \sum_{c=1}^k \frac{links(N_c, N_c)}{|N_c|} = \sum_{c=1}^k \frac{\mathbf{y}_c^T B \mathbf{y}_c}{\mathbf{y}_c^T \mathbf{y}_c} = \sum_{c=1}^k \tilde{\mathbf{y}}_c^T B \tilde{\mathbf{y}}_c \right\} \quad (6)$$

ここで B は先ほどの $B_{i,j}$ を行列で記したもので、 $\tilde{\mathbf{y}}_c = \mathbf{y}_c / (\mathbf{y}_c^T \mathbf{y}_c)^{1/2}$ である。つまり、個別のエッジの和を B で表現した形式である。

一方で、Kernel K-means の定式化を見ても、基本的なアイデアは K-means そのもので、入力データを $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ としたとき、各クラスタ $\eta_1, \eta_2, \dots, \eta_k$ を見つけるにはクラスタ平均 \mathbf{z}_c から各クラスタ内要素の距離を最小にするように求める。

$$\min \left\{ \sum_{c=1}^k \sum_{\mathbf{b}_i \in \eta_c} q_i \|\phi(\mathbf{b}_i) - \mathbf{z}_c\|^2 \right\}, \quad (7)$$

$$\mathbf{z}_c = \frac{\sum_{\mathbf{b}_i \in \eta_c} q_i \phi(\mathbf{b}_i)}{\sum_{\mathbf{b}_i \in \eta_c} q_i}. \quad (8)$$

この時、ベクトル空間を高次元空間 ϕ 上で行うと仮定して、 \mathbf{b} が $\phi(\mathbf{b})$ となっている点と各入力ベクトルに対して重み q_i を仮定している部分だけが異なる。

次に式 (9) をクラスタ毎の和から入力ベクトル毎の和に変更してみよう。クラスタ c 内での重みの和 $t_c = \sum_{\mathbf{b}_i \in \eta_c} q_i$ を仮定すると、あるベクトル i で、かつあるクラス c に対して重みをまとめた M という行列を仮定できる^(注6)。 M_{ic} は $\mathbf{b}_i \in \eta_c$ のとき $\frac{1}{t_c}$ でその他は 0 である。さらに、 $\phi(\mathbf{b}_i)$ をまとめた行列を Φ 、各ベクトルの重み q_i を対角行列でしめたものを Q とすると各クラスタ平均は $\Phi Q M M^T$ で表すことができる。よって式 (9) は下記のように変形できる。

$$\min \left\{ \sum_{i=1}^n q_i \|\Phi_i - (\Phi Q M M^T)_i\|^2 \right\}. \quad (9)$$

ここで、 i は行列の第 i 列目を表すものとする。さらに $\tilde{S} = Q^{1/2} M$ という直交行列を仮定すると全入力ベクトル

をまとめて下記のように書き換えられる。

$$\min \left\{ \|\Phi_i Q^{1/2} - \Phi Q^{1/2} \tilde{S} \tilde{S}^T\|^2 \right\}. \quad (10)$$

式 (10) は行列のノルムを表しており、フロベニウスノルムで計算すると下記の式が得られる。

$$\min \left\{ \text{trace} (Q^{1/2} \Phi^T \Phi Q^{1/2}) - \text{trace} (\tilde{S}^T Q^{1/2} \Phi^T \Phi Q^{1/2} \tilde{S}) \right\}. \quad (11)$$

式 (11) の第一項は入力ベクトルの高次元空間の内積であるが定数である。よって式 (11) 全体を最小化するには第二項を最大化することになる。よって目的関数は

$$\max_{\tilde{S}} \text{trace} (\tilde{S}^T Q^{1/2} \Phi Q^{1/2} \tilde{S}) \quad (12)$$

となる。ここでカーネル関数を $K = \Phi^T \Phi$ とおいた。

この KK 法の目的関数 (12) と先ほど求めたグラフ分割による目的関数 (6) を比較すると、カーネル関数 K が B に対応し、重み対角行列 Q が全て 1 の時 \tilde{S} が \tilde{Y} に対応する。つまりグラフ構造におけるノード間のエッジ情報がカーネル (=各ベクトル間の類似度) に対応することになる^(注7)。これにより、グラフ構造を入力としてクラスタリング結果を出力することができる。

4. 動詞類義語獲得実験

4.1 実験の枠組みと評価

動詞と名詞の係り関係から動詞類義語を獲得する。テキストデータとして日本語コーパス^(注8)の 2008 年度版 BCCWJ コアデータ白書 (以下 BKOW2008) と、毎日新聞 91 年度版から 98 年度版までの 8 年分の新聞記事 (以下、mai91-98) を利用する。これらテキストデータに対して係り受け解析を Cabocha で行った結果、名詞 (助詞付き) と動詞の共起出現回数を記録したものがクラスタリングの入力データとなる^(注9)。この入力データを CCRE 並びに、KK 法に与えてクラスタを作成させる^(注10)。出力されたクラスタは人手で動詞類語を構築した動詞項構造ソーラス^(注11)を用いて評価する。具体的には 5 階層目にある動詞集合を正解として評価する。クラスタリングにはラベルがないためどの正解動詞集合に対応しているか先験的にわからない。よって、少なくとも 2 要素以上の動詞が正解動詞集合にマッチしたとき、出力クラスタは対応する正解動詞集合のクラスタと考える。評価としては適合率と再現率を下記のように定義して利用する。

(適合率) =

$$\frac{2 \text{ 要素以上で最大で辞書動詞グループに属する要素数の総和 (延べ数)}}{\text{全出力クラスタの要素数 (延べ数)}}$$

(再現率) =

$$\frac{2 \text{ 要素以上で最大で辞書動詞グループに属する要素数の総和 (種類数)}}{(\text{辞書の要素数} - \text{コーパスに未出現の動詞数}) (\text{種類数})}$$

以下、実験結果について報告する。

(注7)：文献 [2] では ratio cut など他の場合についても同様に対応することを示している。

(注8)：<http://www.tokuteicorpus.jp/>.

(注9)：具体的には [壁を-走る 10] のように名詞 (助詞)-動詞と頻度の重み付きの 2 部グラフとなる。

(注10)：KK 法のツールが公開されていて、<http://www.cs.utexas.edu/~inderjit/software.shtml> にある Graclus を利用した。

(注11)：<http://cl.cs.okayama-u.ac.jp/rsc/data/>.

(注6)： M は $n \times k$ の行列である。

表1 CCRE法とKK法との比較(BKOW2008)

	重み	適合率	再現率
CCRE法	1	0.324(2042/6295)	0.118(960/8120)
KK法(1100)	1	0.154(685/4453)	0.084(685/8120)
KK法(300)	回数	0.159(708/4462)	0.087(708/8120)

4.2 動詞類語獲得実験結果

コーパスBKOW2008とmai91-98に対してCCRE法でクラスタリングを行った場合とKK法で行った場合の適合率と再現率をそれぞれ表1と表2に示す。CCREではエッジの重み(=名詞(助詞)-動詞の出現回数)を1にして計算を行う方が精度がよい。そこで上記表1と表2ではKK法でもエッジの重みを1とした場合と回数を用いた場合の結果を示している(注12)。また、KK法はK-meansであるためクラスタ数Kをあらかじめ与えなくてはならない。そこでクラスタ数を50から4000まで変化させてその中でもっともF値(注13)が高かったものを示している。表1ではクラスタ数Kは1100と300のとき、もっとも良かったことを示している。

表2 CCRE法とKK法との比較(mai91-98)

	重み	適合率	再現率
CCRE法	1	0.392(1900/4843)	0.123(1000/8151)
KK法(3700)	1	0.176(790/4493)	0.097(788/8151)
KK法(900)	回数	0.190(860/4538)	0.105(859/8151)

表1および表2内の適合率内の丸括弧内の分母はクラスタ数ではなく、出力された全要素数である。また適合率内の丸括弧内の分母は動詞項構造シソーラスにおける各正解動詞集合内の全動詞要素の数を示している。

表1と表2からCCRE法とKK法のクラスタ精度を比較すると、どちらのコーパスでも適合率、再現率ともに提案手法であるCCRE法の法が良いことがわかる。CCRE法はソフトクラスタリングで重複してクラスタを出す一方でKK法はハードクラスタリングであるため、適合率での値の違いは直接比較することは難しいが、それでも出力したクラスタの精度が倍異なるというのは大きな利用する側からみれば大きな差であると考えられる。さらに再現率でもCCRE法の方がKK法より高い値を示していることから、CCRE法は重複したクラスタを出しつつ、獲得すべきクラスタをKK法より多く出力していることを示している。

また、コーパス間の違いで比較すると、BKOW2008は大きさとしては新聞記事2年分程度であるにも関わらず再現率が新聞記事8年分(mai91-98)と大きく差が無い。これは単一の新聞より動詞のバラエティが大きいことを示しており、辞書構築のためのコーパスとしては良い性質であることがわかる。

一方、KK法同士で比較するとエッジの重みを考慮した方が適合率、再現率ともに上昇する。このことから重みを考慮した

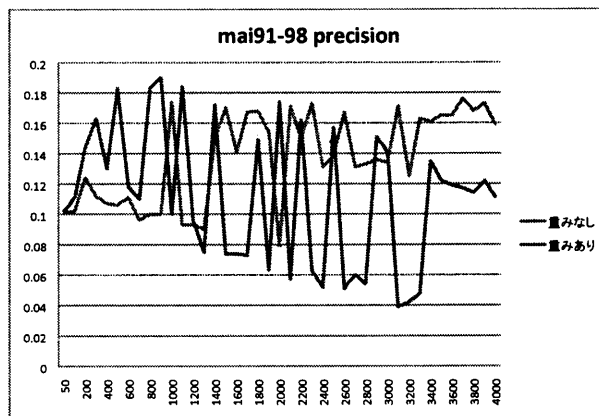


図2 KK法におけるクラスタ数を変えた適合率(50から4000まで(100刻み))

クラスタリングが有効であることがわかる。ただし、KK法のポイントである最適なクラスタ数Kの値が重み有り無しで大きく違っており信頼性に欠ける。実際にKK法を利用して動詞を獲得場合には正解データには頼れないので、なんらかの目安が必要となる。今回、上記表2を獲得するためにKを変えて適合率(図2)、再現率(図3)、F値(図4)を求めたが、これら図から明らかなように重み1、重みありのどちらの場合も単調性がなく、最適なKを得る目安はこのタスクでは得られなかった。

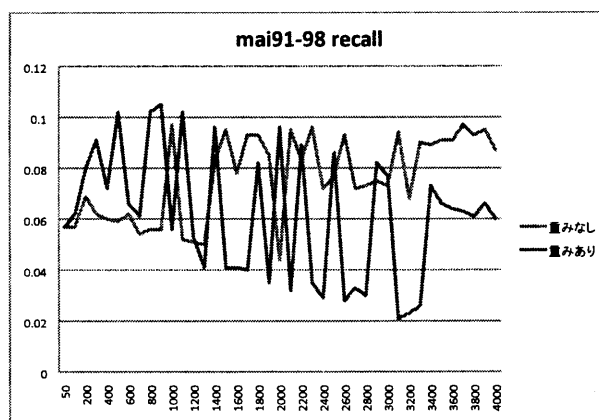


図3 KK法におけるクラスタ数を変えた再現率(50から4000まで(100刻み))

上記の結果から動詞類義語獲得は提案手法であるCCRE法が今のところKK法に比べて有益であることがわかる。今後日本語コーパスに含まれる他のコーパスについても比較する予定である。次節ではKK法について本タスクとの相違点についてモデルそのものから考察を加える。

5. 考察

本節ではKK法の定式化をもとに我々のタスクはどう適応されるべきか、どのような応用が可能であるかを検討し、よりよいクラスタリング結果の獲得について考察する。

類語が獲得できるポイントは同じ名詞に対して異なる動詞が係り関係にある場合である(図1における「つくる」「にぎる」

(注12): ここでのCCREの収束方法は文献[6]に示す最適化を行っている。
 (注13): ここでは適合率pと再現率rから調和平均(F値)を $f = \frac{2pr}{p+r}$ とした。

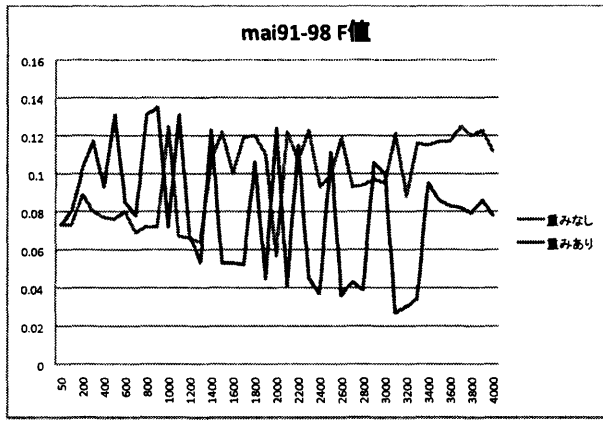


図 4 KK 法におけるクラスタ数を変えた F 値 (50 から 4000 まで (100 刻み))

に対する「おにぎりを」や「すしを」), この構造は KK 法ではノード間の接続行列 B に集約されさらにカーネル K としてノード間の近さに置き換えられることからこうした特徴は KK 法でも取り込めている。

ところが KK 法の出力結果を観測すると名詞だけのクラスタが存在した。つまり、2 部グラフとして入力しても、動詞と名詞を同時に含むようなクラスタが必ずしも生成されるわけではないことがわかる。一方で、元々の入力データは動詞と名詞の 2 部グラフという構造をもったグラフデータであり、動詞と名詞を一つのグループとしてクラスタリングすることが目的であった。つまり、動詞ノードと名詞ノードを同時に複数含むクラスタを作成するとして制約を KK 法に別途教える必要がある。

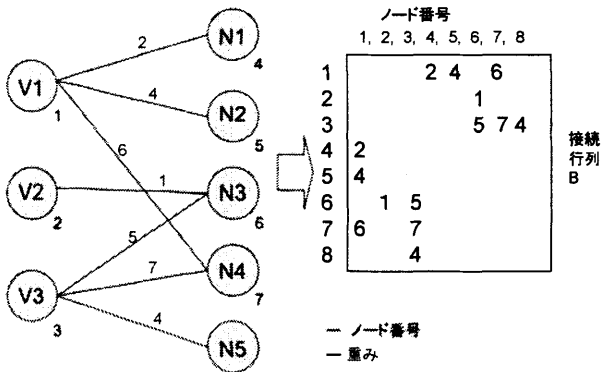


図 5 KK 法に対してグラフ構造を入力する方法

しかし、動詞と名詞を最大限残しながらクラスタリングするというものは KK 法では取り込められないと考えられる。KK 法にある種の制約を入れる方法として近年、cannot-link, must-link による半教師ありのクラスタリング法が文献 [3] で提案されている。しかしながらこの制約はあるノード間が同じクラスタに属すべきか属さないかを指定するのであって、2 部グラフとしての制約を記述することは可能でないように思える。2 部グラフを扱うにはノードからスタートして奇数番目のエッジ先のノード集合と偶数番目のエッジ先のノード集合という構造が必要となる。

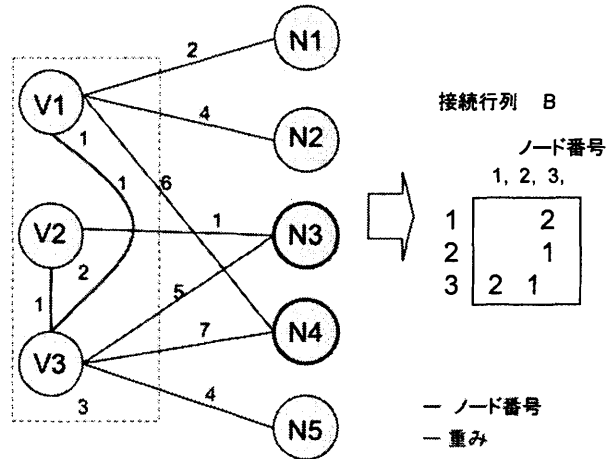


図 6 KK 法に対して動詞のみのノードを与える方法

そこで制約ではなくもう一度 KK 法の定式化に振り返って他の適用方法を考察してみたい。結局グラフ構造はノード間近接行列 B がそのままノード間の近さの定義を与えることから、ノードとして直接比較したい動詞のノードのみを与える方が KK 法の特性に合致しているように考えられる。具体的には名詞へのリンクを元に動詞間近さを定義しなおし、それを動詞ノード間のエッジ情報として入力するというものである。今回入力した方法を図 5 に、新たな動詞ノードのみの入力の様子を図 6 に示す。現段階では新たな方法で入力した場合 KK 法のツールが収束せず結果を得られていないため比較が行えていない。このあたりは今後の課題としたい。

6. まとめ

本稿ではクラスタリングを利用してテキストデータから動詞類義語集合を獲得する手法について比較による検討を行った。日本語コーパス 2008 年度版コアデータ (白書) と毎日新聞 8 年分に対する動詞類語獲得実験結果から、提案手法である CCRE 法が KK 法より適合率、再現率とも高い値を示すことを明らかにした。また KK 法の特から本研究で利用してきた 2 部グラフはそのままでは適用できないことを考察し、新たな入力方法について提案した。現在 CCRE 法では半教師ありクラスタリング法によりさらに高い精度のクラスタリング結果が得られる見通しを得ている [6]。一方で KK 法の半教師あり版が提案されており [3]、将来において比較を行いたい。

7. 謝 辞

本研究の遂行に当たって文部科学省科学研究費補助金特定領域研究「日本語コーパス」(代表) の支援を受けた。また、毎日新聞社様には新聞記事の研究利用の許諾をいただいた。ここに記して深く感謝する。

文 献

[1] Aizawa, A. : A method of Cluster-Based Indexing of Textual Data, *Proceedings of COLING 2002*, pp. 1-7 (2002).
 [2] Dhillon, I. S., Guan, Y. and Kulis, B. : Weighted Graph Cuts without Eigenvectors: A Multilevel Approach, *IEEE Transactions on Pattern Analysis and Machine Intelli-*

- gence, Vol. 29, No. 11, pp. 1944-1957 (2007).
- [3] Kulis, B., Basu, S., Dhillon, I. and Mooney, R. : Semi-supervised graph clustering: a kernel approach, *Machine Learning*, Vol. 74, No. 1, pp. 1-22 (2009).
 - [4] Takeuchi, K. and Takahashi, H. : Co-clustering with Recursive Elimination for Verb Synonym Extraction from Large Text Corpus, *IEICE Transactions on Information and Systems*, Vol. E92-D, pp. 2334-2340 (2009).
 - [5] 竹内孔一, 乾健太郎, 竹内奈央, 藤田 篤: 意味の包含関係に基づく動詞項構造の細分類, 言語処理学会第 14 回年次大会発表論文集, pp. 1037-1040 (2008).
 - [6] 竹内孔一, 高橋秀幸: 動詞辞書構築を目的とした同時クラスタリングの精度向上, 特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集, pp. 187-190 (2010).