2E1-1

教師なしクラスタリングによるテキストのトピック抽出

Extracting topic from texts with unsupervised clustering.

狩野 竜示[*1] 根本 啓一[*1] 大西 健司[*1]
Ryuji Kano Keiichi Nemoto Takeshi Onishi

[*1] 富士ゼロックス(株)研究技術開発本部
Research & Technology Group, Fuji Xerox Co., Ltd.

There are several kinds of methods to extract topic from text data. However, different kinds of topic are demanded depending on a situation. Some people want to know abstract topic, which is frequently mentioned in the text data, while other people want to know more concrete topic which is not so much mentioned in the text data. Generally, the resolution of topic is adjusted by clustering parameters, but it doesn't fit well with the resolution referred above. We adapted two other parameters, which are the upper limit of the number of the nodes and the threshold of jaccard index, to test how these parameters affect the topic.

1. 序論

多量のテキスト群に潜在する, テーマや話題を表す単語群をトピックと呼ぶ. これらの単語群を抽出する方法として, トピックモデリングを利用し, テキスト群に潜在する話題を単語群によって表す手法(LDA)[Blei 2003]や, 単語の共起に基づいたネットワークのクラスタリングによって, テキストから単語群を抽出する手法[樋口 2001]がある. このような, 限られた単語群によって多量のテキストに存在する話題を表現する方法は, VOC などの大量のテキストを処理する場面でも有効であると考えられる. しかし, トピックからユーザが類推する話題は, トピックに含まれる単語によって異なるため, 適切なトピックの抽出が課題となっている. 本稿では, トピックと, そこから得られる話題との関連性を明らかにするため, トピック抽出に影響をあたえるパラメータと, トピックが表す話題との関連性を検証した.

2. トピック抽出

2.1 トピック抽出の過程

今回適用したクラスタリングによるトピック抽出は以下の過程から成る. すなわち, (i)テキストを形態素解析によって単語に分解する過程と, (ii)単語が同じテキスト中に出現していることを条件に, 単語をノード, 共起関係をエッジとして共起ネットワークを生成する過程と, (iii)共起ネットワークのエッジから特徴的なエッジを選別する過程と, (iv)階層性クラスタリングにより, 共起ネットワークから単語群(トピック)を抽出する過程である.

(i)の形態素解析には kuromoji[Atilika 2014]を使用し, 名詞, 形容詞, 動詞のみを解析対象とした. (ii)では, 1 単語を 1 ノードとし, 共起回数を重みとしたエッジでノード間を結んだ. この時, 共起の定義を「同じ文中に出現していること」とした. (iii)におけるエッジ選別の指標には, jaccard 係数[Manning 2002]を採用した. (iv)のクラスタリング手法には, 高速に計算できる modularity[Aaron 2004]を採用した.

2.2 トピック抽出のパラメータ

本稿で述べるトピック抽出には, 2.1 で述べたトピック抽出過程(ii)~(iv)それぞれに対応するパラメータ 3 つが存在する.

2.2.1 ノード(単語)を選別するパラメータ

(ii)に対応するパラメータは, 共起ネットワーク生成の際に使用するノード数の上限である. ネットワーク計算の際, ノード数の増加に伴い計算時間は爆発的に増加する. そのため, ネットワーク作成前にノード数の上限値 w_{thre} を設け, 出現回数の高い単語, 上位 w_{thre} 個を選別する.

2.2.2 エッジを選別するパラメータ

(iii)に対応するパラメータは, エッジ選択の基準である jaccard 係数の閾値 j_{thre} である. これは, jaccard 係数が j_{thre} 以下のエッジを除去する閾値である. ノード a と b を結ぶエッジの jaccard 係数 j(a,b)は以下の式で表される.

$$jacc(a,b) = |e(a) \cap e(b)| / |e(a) \cup e(b)| \cdots (1)$$

ここで, e(a)はノード a に繋がる全てのエッジを指し, |e(a)|は e(a)の個数を指す. jaccard 係数はある単語 a,b が同じ文中に出現する確率を指しており, 単語同士の繋がりの強さを示している.

2.2.3 クラスタの大きさを決めるパラメータ

(iv)に対応するパラメータは, クラスタリングにおいて各クラスタの大きさを決定する閾値ΔQ_{thre}である. modularity 計算の過程では, ノード a,b 間の親和度ΔQ_{ab}を, a,b の組合せ毎に計算し, この値がΔQ_{thre}を上回った組合せを同じクラスタに含めていく. ノード a,b 間のΔQ_{ab}は以下の式で表される.

$$\Delta Q_{ab} = 1/2m - |e(a)||e(b)|/(2m)^2 \cdots (2)$$

この時, m は全エッジ数を指す.

w_{thre}はネットワーク構造におけるノードの選別, j_{thre}はエッジの選別, ΔQ_{thre}はクラスタリングの度合いに対応している.

3. トピック抽出結果

トピック抽出の検証に, 実運用されている VOC データベースから取得した VOC, 29000 件を使用した. 抽出したトピックの例を表 1〜3 に示す. 表にあるように, トピックは複数の単語の組合せから成る. 表1に, w_{thre}を 200 から 1600 に変化させた時のトピックを示す. この時, j_{thre}は 0.05, ΔQ_{thre} は, 0.0 とした. 表 2 には, j_{thre}を 0.05 から 0.1 に変化させた時のトピックを示す. この時 w_{thre}を 200, ΔQ_{thre}を 0.0 とした. 表 3 には, ΔQ_{thre}を 0.0 から -0.05 に変化させた時のトピックを示す. この時, w_{thre}を 200, j_{thre}は 0.05 とした. 各条件で, 内容が近いと思われるトピックを人為

連絡先:狩野 竜示, 富士ゼロックス(株)研究技術開発本部, kano.ryuji@fujixerox.co.jp

的に選択し，同じ行に列記した．抽出したトピックの内，パラメータ変更によって変化したトピックを選択した．

4. 考察

modularity による単語の共起ネットワークのクラスタリングにより，トピックを抽出した．この時，種々のパラメータ設定によって，トピックが示す話題がどのように変化するかを考察する．パラメータを調整すると，トピックに含まれる単語数が変化する．この変化がトピックの示す話題に与える影響には，主に 2 種類があると考えられる．例として，表 3 にある(価格,安い)というトピックが(価格,用紙,安い,購入)に変化する場合は，単語が増える事によってトピックの示す話題に具体性が増す場合と考えられる．他方，(プリント,出力) が(コピー,プリント,出力,カラー,モノクロ)に変化した場合は，インターネット接続に関する類語が増え，概念が拡大したと解釈出来る．このように，パラメータ調整によるトピックの変化には，具体性増加と概念拡大の 2 種類存在する．

4.1 ノード選別パラメータ (w_{thre})の影響

w_{thre} を大きくした時のトピックは，より具体性が増した細かい話題に言及する傾向にある．例として，表 1 の(消耗品,届く)は(消耗品,届く,自動,埼玉)に変化している．これは，w_{thre} の増加によって，今までトピック候補になかった出現頻度の低い，より具体性の高い単語がトピックに含まれるようになったからと考えられる．表 1 にある 2 種類のパラメータで抽出した各トピックを官能評価したところ，トピックが変化したものの内，具体性を増したトピックが 12 件，概念が拡大したものが 7 件であった．

4.2 エッジ選別パラメータ (j_{thre})の影響

一方 j_{thre} を上げると，共起の少ない単語同士の連結が除去されるため，トピックに含まれる単語の数は減少した．j_{thre} は単語同士の関連の強さを表しているため，この数値を上げると，互いに関連の強い単語のみがトピックとして抽出されるようになる．これは，表 2 にある(文書,FAX,受信)が(FAX,受信)に変化した事からも，話題の具体性を減少させる方向に働く．4.1，と同様にトピックの官能評価を行ったところ，j_{thre} の減少によって，具体性が増したものが 17 件，概念が拡大したものが 5 件であった．

4.3 クラスタリングのパラメータ (ΔQ_{thre})の影響

ΔQ_{thre} を小さくすると，トピックの示す話題は概念が拡大する傾向にあった．表 3 にある (接続,ネット) が(接続,LAN,ネット,無線)に変化した場合と，(プリント,出力) が(コピー,プリント,出力,カラー,モノクロ)に変化した場合は，インターネット接続，あるいはプリンターに関する類語が増え，概念が拡大したと解釈出来る．4.1，4.2 と同様に，変化したトピックの官能評価を行ったところ，具体性が増したものが 4 件，概念が拡大したものが 12 件であった．

4.4 2 種類のトピック変化

前述の通り，トピック抽出のパラメータを変化させた時，トピックの示す話題は「概念拡大」と「具体性増加」の 2 通りに変化すると考えられる．この事は概念カテゴリーという言葉を使って以下のように解釈出来る．具体性が増した例として，パラメータ調整によって「価格−高い」から「価格−用紙−高い」に変化したトピックを考える．ここで，「価格」が評価軸，「高い」は評価語，「用紙」は評価対象という概念カテゴリーに属しているとみなす．新たに加わった「用紙」という単語は，評価対象の概念カテゴリーに属し，これは変化前のトピックに無かった概念カテゴリーである．このように，既存のトピックに無い概念カテゴリーに属する単語が加わる場合を，具体性増加と捉える事が出来る．これに対して，「価格−用紙−高い」が「価格−用紙−インク−高い」に変化した場合を，概念拡大の例とする．この例で，「用紙」，「インク」は共に評価対象に該当する単語であるが，「インク」が加わる前に既に評価対象の語「用紙」がトピックに存在している．このように，トピックに含まれるべきいくつかの概念カテゴリーを仮定した時，既に単語がある概念カテゴリーに新しく単語が追加された時は，概念が拡大し，新たな概念カテゴリーに単語が追加される場合，具体性が増加すると考えられる．

表 1 w_{thre} 別抽出トピックの例

w_{thre}=200	w_{thre}=1600
(価格,安い)	(安い,価格,用紙)
(消耗品,届く)	(消耗品,自動,届く,埼玉)
(最新,バージョン)	(最新,ファームウェア)
(予算,申請)	(来期,申請,予算)

表 2 j_{thre} 別抽出トピックの例

j_{thre}=0.05	j_{thre}=0.1
(作業,報告,状況,稼働,サポート,終了)	(稼働,報告)
(文書,FAX,受信)	(FAX,受信)
(コピー,プリント,出力,カラー,モノクロ)	(カラー,モノクロ)

表 3 ΔQ_{thre} 別抽出トピックの例

ΔQ_{thre}=0.0	ΔQ_{thre}=-0.05
(価格,安い)	(価格,用紙,安い,購入)
(接続,ネット)	(接続,LAN,ネット,無線)
(プリント,出力)	(コピー,プリント,出力,カラー,モノクロ)

5. まとめ

上述の通り，パラメータ調整による，トピックの示す話題の変化には，類語増加による概念拡大と，具体性の増加の 2 種類が存在した．今回検証した 3 つのパラメータの内，ΔQ_{thre} の増加は概念拡大，w_{thre} の増加，及び j_{thre} の減少は具体性の増加をもたらす傾向がある事が判明した．そして，これらの違いは，トピック内に同一の概念カテゴリーに属する単語が存在するか否かによって生じた．今回得られたこの知見を活かし，概念拡大，具体性増加の望む方向に抽出トピックを変化させられる技術の開発を目指す．

参考文献

[樋口 2001] 樋口耕一: KH Coder, http://khc.sourceforge.net (2001)

[Aaron 2004] Aaron Clauset, M. E. J. Newman, Cristopher Moore: Finding community structure in very large networks, Phys. Rev. E 70, 066111 (2004).

[Atilika 2014] Atilika: kuromoji - Japanese morphological analyzer, http://www.atilika.org/

[Blei 2003] Blei, D., Ng, A, and Jordan, M.: Latent dirichlet allocation, The Journal of Machine Learning Research, 3, p.993-1022, 2003.

[Manning 2002] Manning, C.D. Schütze, H: Foundations of statistical natural language processing, The MIT Press, London (2002).