分割文書データのNMFによるトピック抽出

筒井 昌彦† 横井 健†
東京都立工業高等専門学校†

1 はじめに

情報化社会と呼ばれる今日、情報の整理は重要な課題となってきている。情報整理をするために文書情報に含まれる潜在的意味(トピック)に着目する手法が提案されている。トピックを抽出する手法として Non-negative Matrix Factorization (NMF) [1]が提案されているが、扱う文書量が増加すると計算に使用するメモリの不足、また計算量が膨大となる。そこで本研究では分割した文書集合に対して、NMF を適用し、それらのトピックを結合することで、大規模な文書集合におけるトピック抽出を行う。

2 文書情報に対するNMF

ある文書集合をその集合に含まれる単語数 n×文書数 m の単語文書行列 V で表現する。その要素は文書中の各単語の出現頻度とする。NMF は行列 V を(1)式のように非負値で構成された n×r 行列 W と r×m 行列 H に分解し、W の列ベクトルが、トピックを表現している。

$$V \approx WH \tag{1}$$

ただし r は任意に決定する数でここではトピック数を示す。行列 W は(2)式で示すようになり、w_{xy} はトピック R_y における単語 t_x の重みを示している。

$$W = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{cccc} R_1 & R_2 & \cdots & R_r \end{array} \\ \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1r} \\ w_{21} & w_{22} & & w_{2r} \\ \vdots & & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nr} \end{bmatrix} \end{array} \tag{2}$$

なお、W と H は(3)式を最小化するように交互最小二乗法を用いて計算する [2]。

$$f(W,H) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(V_{ij} - (WH)_{ij})^2 \tag{3}$$

3 トピックの結合方法

本研究では大規模な文書集合に対応するために、分割された文書集合D_kに NMF を適用し、それぞれの分割した文書集合から得られたトピックを結合する。文書行列D_pとD_qの結合方法の流れを以下に示す。ここでD_pとD_qに含まれるt_xの集合を(3)、(4)式に示すT_p、T_qとする。

$$T_p = \{t_x | t_x \in D_p\} \tag{3}$$

$$T_q = \{t_x | t_x \in D_q\} \tag{4}$$

1. 結合後のトピックを表現にするために用いる単語t_xの集合 T を(5)式のように定義する。なお、それぞれのトピックについて新たに追加される単語に対する重みは 0 とする。

$$T = T_p \cup T_q \tag{5}$$

2. コサイン尺度を用いて文書集合D_aとD_bの各トピックの類似度を比較する。
3. それぞれの文書集合で最も類似しており、閾値以上の類似度が得られたトピック$R_{p,a}$と$R_{q,b}$を(6)式にて結合し、新たにトピック$R_{(a+b)}$を作成する。

$$R_{(a+b)} = \frac{\alpha R_{p,a} + \beta R_{q,b}}{\alpha+\beta} \tag{6}$$

なお、α、βは(7)式の通りである。

$$\begin{cases} \alpha = \max w_{aj}^{p} - \min w_{aj}^{p} \\ \beta = \max w_{aj}^{q} - \min w_{aj}^{q} \end{cases} \tag{7}$$

4. 結合の対象とならなかったトピックはそのまま出力する。

(6)式を用いることで、差が顕著に表れている文書集合に重みを傾けることができ、トピックとして有用なものが抽出できると考えられる。

また、この結合方法では 2 つの文書行列しか結合できないため、2 つ以上の文書行列に分割した

Topic extraction from divided document data by NMF
†Masahiko tsutsui
†Takeru Yokoi
†Tokyo Metropolitan College of Technology

場合、上記の処理を繰り返し、結合をする。

4　実験と考察

本節では分割していない文書集合にNMFを適用した場合に得られたトピックと、提案手法によって得られたトピックを比較し、考察する。

4.1 実験方法

実験データとしてClutoのwebサイト[1]にて公開されているデータセットの中から計算機科学に関する文献より抽出したデータ(cacmcisi)を用いた。単語数は41,681、文書数は4,663である。

これを4つのデータに分割し、単語文書行列D_kに変換した。単語数と文書数はそれぞれ1,179×1,166、1,451×1,166、5,266×1,166 、11,199×1,165である。トピック数rを20に設定したNMFを上記4つの単語文書行列と適用した。その後、それぞれから得られたトピックにコサイン尺度を用いて結合をする。なお、類似度の閾値は0.7と設定した。

また、分割していない単語文書行列にもr=20と設定したNMFを適用し、得られたトピックを提案手法のトピックと比較する。

4.2 実験結果と考察

表1に分割していない状態(比較手法)でNMFを適用し、得られたトピックと提案手法にて得られたトピックをそれぞれ代表する単語を示す。提案手法ではトピックの結合後49個のトピックが得られた。表1ではそれぞれの中から16個のトピックを示す。

比較手法と提案手法を比較するとsystemやcatalogなどのトピックは両者に一致して抽出されている。一方、比較手法では見られなかったtitleやindexなどのトピックも提案手法では抽出した。

なお、比較手法で得られたトピックのうち約80%を提案手法でも抽出することができた。

比較手法に対して提案手法から概ね類似した結果が得られたが、比較手法では見られないトピックとしてはあまり有用ではないものを抽出してしまった。

表1 実験結果

トピック R_y	比較手法		提案手法	
	単語 t_x	重み w_{xy}	単語 t_x	重み w_{xy}
1	library	0.85255	title	0.77826
2	system	0.50926	service	0.75099
3	inform	0.32369	inform	0.72212
4	science	0.11908	cost	0.59763
5	base	0.1170	index	0.53223
6	term	0.11558	system	0.47975
7	journal	0.10836	catalog	0.47702
8	catalog	0.10822	journal	0.46966
9	search	0.09815	class	0.4623
10	provide	0.08188	search	0.4494
11	service	0.07615	research	0.44104
12	literature	0.07305	retrieve	0.42848
13	cost	0.06293	library	0.42379
14	language	0.06254	term	0.39672
15	paper	0.05848	data	0.38319
16	research	0.03918	revel	0.37078

5　まとめ

本研究では大規模文書を分割し、文書集合にNMFを適用した。その結果、従来手法での文書に対するNMFによって得られたトピックと類似したトピックを抽出することができたがトピックとして有用ではない単語も抽出していた。そのため、今後、トピック数の変えて詳細に検討する必要があると考えられる。

参考文献

[1] 柘植 覚,獅々堀 正幹,北 研二 “Non-negative Matrix Factorizationを用いた情報検索”, 情報処理学会研究報告. 情報学基礎研究会報告 2001 , pp.1-6

[2] Chih-Jen Lin, “Projected Gradient Methods for Non-negative Matrix Factorization”, Neural Computation, 19(2007), 2756-2779.

[1] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download