7B-02

ユーザ制約付き独立話題分析における制約の簡略化

西垣貴央† 山本健太† 小野田崇†
†青山学院大学理工学部

1 はじめに

本稿では話題抽出手法の一つで，話題間の関係に着目した方法である，独立話題分析 [1] について考える．独立話題分析では，信号処理の分野で使用される独立成分分析を用いて独立性の高い話題を求めている．ここで独立性が高い話題とは，話題間の相互情報量が小さい話題を示している．独立性が高い話題を求める利点として，より多くの情報量を持つ要約の作成が，容易にできる可能性が高いことが挙げられる．しかし，この独立話題分析で得られる話題は，独立性にのみ基づいて話題を抽出するため，ユーザの望む話題と異なる場合が存在する．そこで，独立話題分析で得られた話題へユーザ制約を与え，ユーザ制約を満たしてかつ独立性の高い話題を求めるユーザ制約付き独立話題分析 [2] が提案されている．本稿ではこのユーザ制約付き独立話題分析におけるユーザ制約を満たすアルゴリズムを改良することで，ユーザの負担を減らし，かつより独立性の高い話題を抽出できる方法を提案する．

2 ユーザ制約付き独立話題分析（Separate Link 制約の場合）

本章では，ユーザ制約付き独立話題分析における Separate Link 制約の場合について紹介する．Separete Link 制約の例を，LA Times のデータ [3]（新聞データで文書数は 6279，単語数は 31472 の文書データ）に対して独立話題分析を行って得た話題の重要単語を示した表 1 を用いて説明する．表 1 を見てユーザが，話題 5 を 2 個の話題に分離して，合計 8 個の話題を得たいと考える場合がある．このように，ユーザがある 1 個の話題を 2 個の話題に分離したい場合の制約を Separate Link 制約といい，この Separate Link 制約を満たしてかつ独立性の高い話題を得る方法が提案されている．以下にアルゴリズムを示す．

表 1: LA Times に独立話題分析を適用して得られた 7 個の話題を構成する重要単語

話題	重要度が高い単語			
	$w=1$	$w=2$	$w=3$	$w=4$
1	million	earn	quarter	revenu
2	scor	game	lead	rebound
3	soviet	afghanistan	israel	foreign
4	aleen	macmin	art	entertain
5	polic	bush	counti	car
6	stock	bank	price	market
7	game	team	player	coach

1. 独立話題分析によって k 個の話題を得る．
2. Separate Link 制約を与える話題 z とその話題を構成する任意の単語 p と単語 q をユーザが選択する．
3. 単語 p が重要単語となる話題 x を生成する．
4. ユーザが選択しなかった話題 $k-1$ 個を独立話題分析によって求める．
5. 単語 q が重要単語となる話題 y を生成する．
6. Separate Link 制約を満たす新たな $k+1$ 個の独立な話題を得る．

以上のアルゴリズムで Separate Link 制約付き独立話題分析のイメージを図 1 に図示する．

しかし，ステップ 2. において，ユーザは分離したい話題だけでなく，その話題を構成する重要単語の中から 2 個の単語を選択しなければならない．これはユーザへの負担が非常に大きい．. 特に話題数が増えると，それにともなって重要単語も増えていくため，ユーザが実際に単語にまで制約を与えるのは難しい．そこで，本稿では Separate Link 制約においてユーザへの負担を軽減し，かつより独立性の高い話題が得られる制約を提案する．

Simplification of user-constraints in Independent Topic Analysis

Takahiro NISHIGAKI†, Kenta YAMAMOTO†and Takashi ONODA†
†College of Science and Engineering, Aoyama Gakuin University
252-5258, Sagamihara, Japan

{nishigaki, onoda}@ise.aoyama.ac.jp

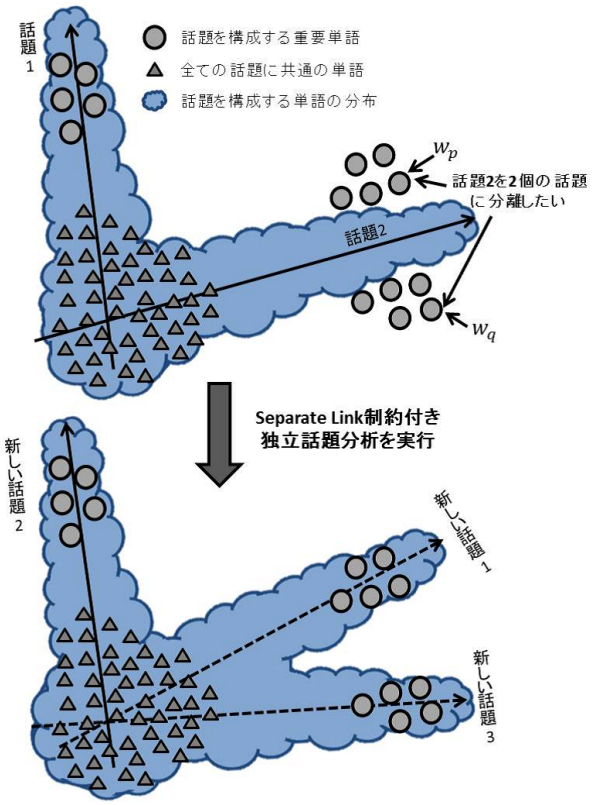

図 1: Separate Link 制約付き独立話題分析のイメージ

3 ユーザ負担を軽減した Separete Link 制約付き独立話題分析の提案

Separate Link 制約においてユーザの負担を軽減し，かつより独立性の高い話題が得られる制約を提案する．提案する新しい Separate Link 制約のアルゴリズムではユーザは分離したい話題 z だけを選択する．そして，その選択した話題をより独立性が高い 2 個の話題に分離するということを行う．より独立背板高い 2 個の話題への分離方法は，選択された話題 z を構成する単語の中から，最も独立性の高い単語を 2 個抽出し，その単語が重要単語となるように話題を生成する．これの方法によって，ユーザは分離したい話題 1 個を選択するだけで，自動的に独立性が高くなるように分離された話題を得ることができる．

4 評価実験

提案した新しい Separate Link 制約付き独立話題分析の評価実験を行った．評価方法として，1) Separate Link 制約の充足性，2) 得られた話題の独立性，の二点で評価を行った．表 1 の話題 6 に対して Separate Link 制約を提案手法を適用した．

4.1 Separate Link 制約の充足性の評価方法

Separate Link 制約の充足性の評価には，各話題における単語の重要度 **V** と各話題における文書の重要度 **U** を使用する．制約として選択された単語が制約を満たす話題の時に最も大きければ制約を満たしていると言える．話題数と分離する話題の組み合わせを適当に変更して行ったが，いずれの場合においても新たに得られた話題は制約は満たしていることが確認できた．

4.2 Separate Link 制約の独立性の評価方法

Separate Link 制約によって得られた話題の独立性の評価には，相互情報量を用いて比較を行う．話題間の相互情報量の値が 0 の場合，その話題間は完全に独立していることを意味し，相互情報量の値が小さい方が，話題間の独立性が高い．ただし，話題数が増えれば増えるほど，話題間が完全な独立でない限り，話題間の相互情報量は増加していく．得られた話題の単語の重要度の行列 **V** を用いて，[4] の方法で求める．話題数 7 の話題 6 に制約を与える場合，ランダムに単語を選択して，通常の Separate Link 制約付き独立話題分析を行う場合との比較を行った結果を表 2 に示す．

表 2: 表 1 の話題 6 に制約を与えた時の相互情報量

	提案手法	ランダム選択
相互情報量	0.00383	0.00680

5 おわりに

本稿では Separate Link 制約付き独立話題分析における，ユーザへの負荷を減らしてかつより独立性の高い話題を抽出する方法を提案した．また，提案した方法をベンチマークデータに適用し，提案手法によって得られる話題は，制約を満たしていることおよび，従来手法よりも独立性が高い話題であることを示した．

今後の課題として，複数個の制約が同時に，あるいは逐次的に与えられる場合における制約の充足方法を考慮する必要があると考えている．

参考文献

[1] 篠原 靖志: 独立話題分析－独立性最大化による特徴的話題の抽出, 信学技法, OFS99-14, 1999.

[2] 西垣 貴央 et al.: 制約付き独立話題分析, 人工知能学会誌, Vol.31，No.4，pp.D-FB1_1-13, 2016.

[3] George Karypis: CLUTO - A Clustering Toolkit, http://glaros.dtc.umn.edu/gkhome/views/cluto/, 2002.

[4] Gavin Brown et al.: Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature, JMLR, Vol. 13, pp.27-66, 2012.