

潜在的ディリクレ配分法を利用した 文書への自動タグ付与に関する検討

加藤 亮 吉川 大弘 古橋 武

文書へのタグ付与の目的

インターネット上の文書は膨大かつ無秩序で、有益な情報を含む文書を探し出すことは容易ではない

文書の内容に基づいて知識整理を行うシステムが必要



整理対象に**タグ**と呼ばれる短い言葉を付与する方法が研究されている

既存手法の問題点

- FlickrやDeliciousなどは、多数のユーザが複数のタグを自由に付け加えていく手法を採用しているが、これはいわゆる人力
- 自動でタグ付けする手法も、あらかじめタグ付けされた文書を用いて学習する必要があった

提案手法の目標

事前のタグ付き文書なしに、各文書の内容に適したタグの付与を行うことを目指して、

文書に含まれる**トピック**に着目した自動タグ付与システムを検討する

(**トピック**：話題や分野、意味のまとまりのこと)

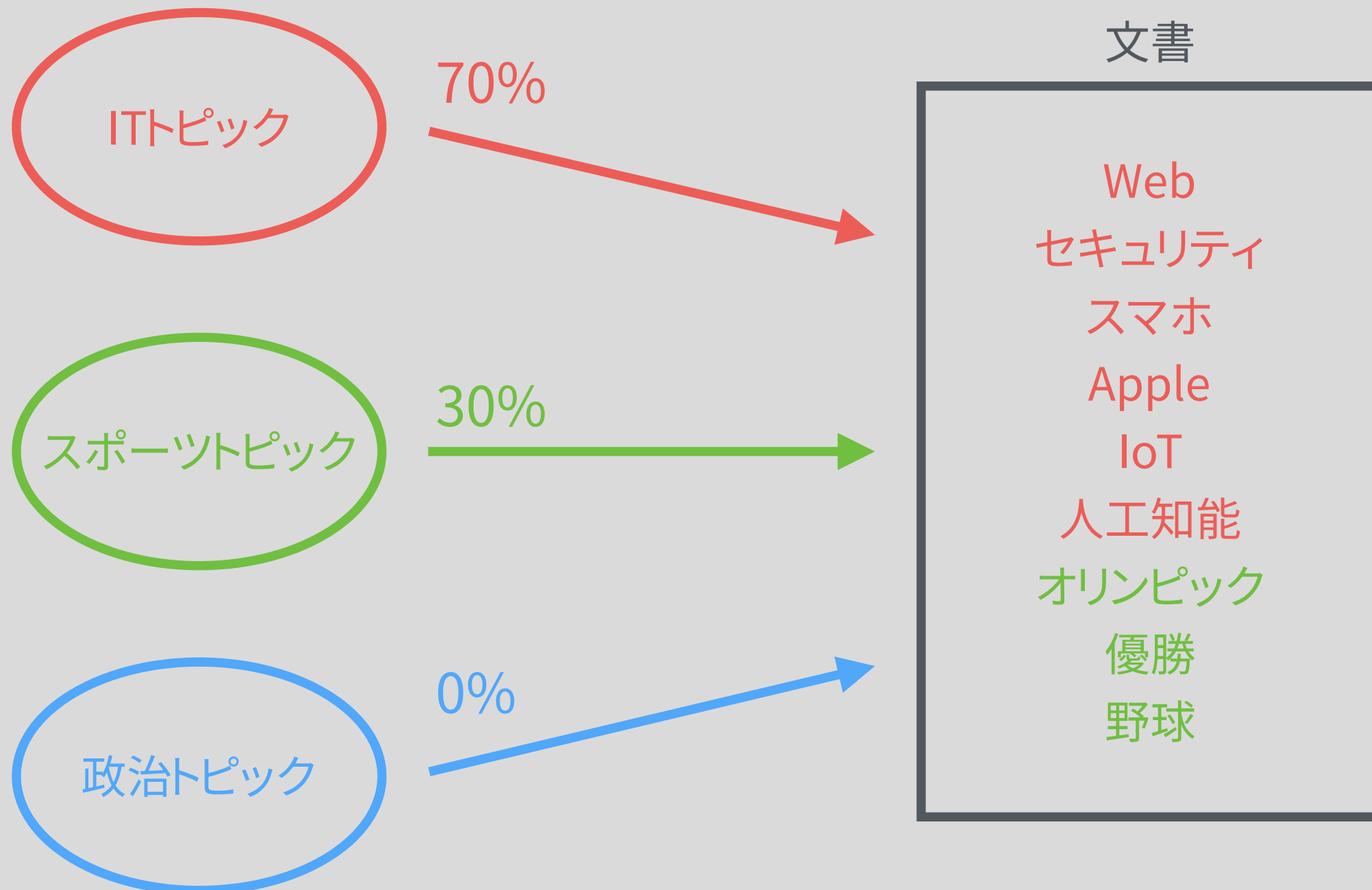
トピックモデル

データ(文書)が持つ潜在的なトピックを推定する確率モデル

- この文書は、大体どういう内容？
- この文書群の中には、どんな「話題」がある？
- この単語は、大体どんな意味？

トピックモデルの仮定

文書において、単語は特定のトピックを基に出現すると仮定する



トピックの推定

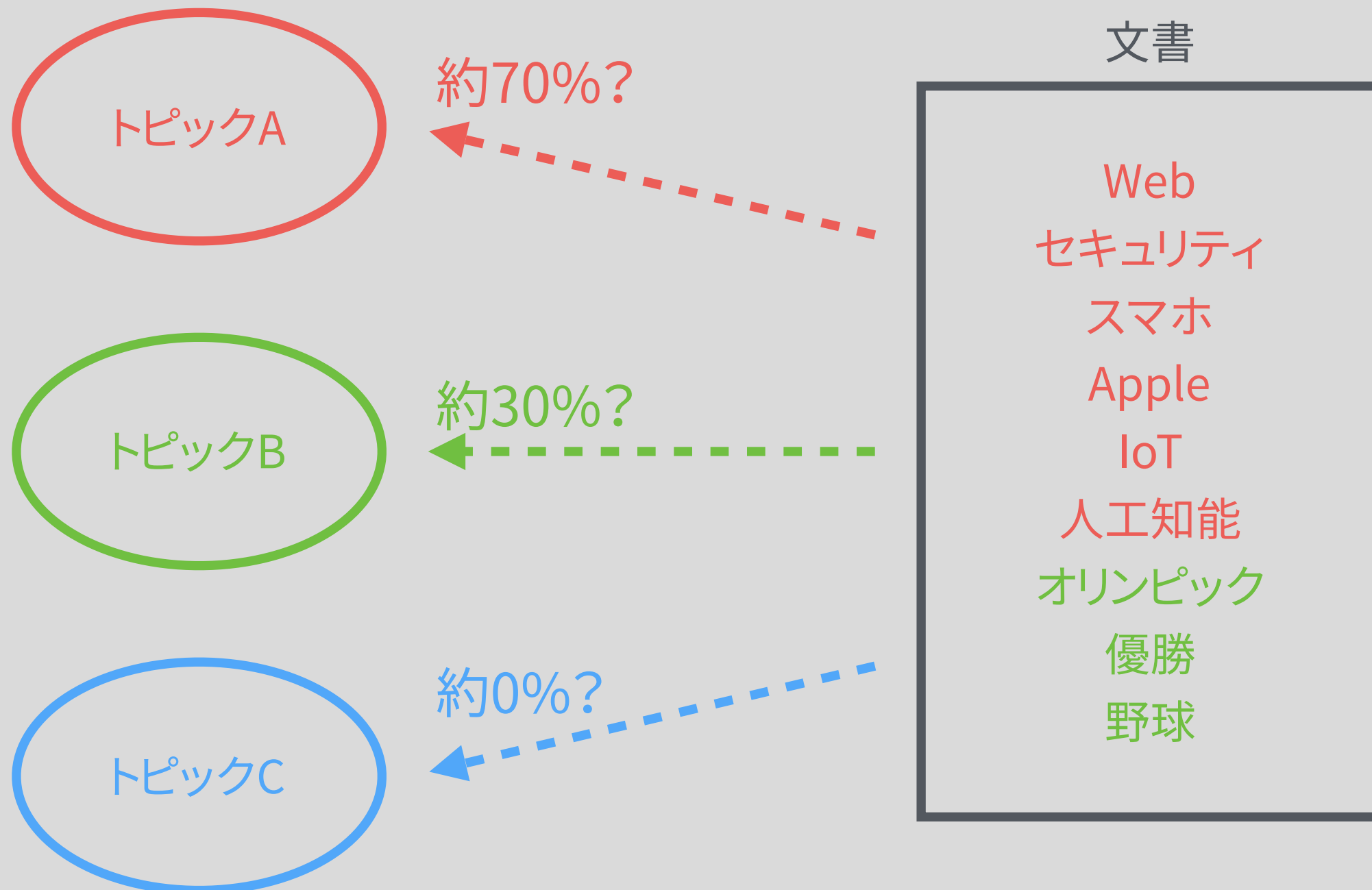
文書内の単語の出現回数の情報から、その文書が持つトピックを推定する

文書

Web
セキュリティ
スマホ
Apple
IoT
人工知能
オリンピック
優勝
野球

トピックの推定

文書内の単語の出現回数の情報から、その文書が持つトピックを推定する



トピックモデルで文書にタグを

推定されたトピックの名前は明示的には得られないが、
文書から出現するトピックの確率やトピックから出現する単語の確率を用
いることで、

文書間や単語間の意味の関係性を得ることができる

↓

文書への自動タグ付与に利用できるのでは？

提案手法の手順

1. 文書の入力
2. トピックの推定
3. 文書へのトピック割り当てによる分類
4. 単語の特徴量の計算
5. 文書へのタグ付与

1. 文書の入力

- タグ付けしたい文書をシステムに入力する過程
- 形態素解析を行い、各文書に出現する単語とその出現回数を求める

2. トピックの推定

ステップ1で得られた単語情報を基に、

潜在的ディリクレ配分法 (Latent Dirichlet Allocation, **LDA**)

を用いてトピックを推定する

潜在的ディリクレ配分法(LDA)

- 代表的なトピックモデル
- 文書が複数の潜在的なトピックを持ち、それらを媒介して単語が生成されることを仮定したトピックモデル
- トピックの出現と、各トピックにおける単語の出現を多項分布で仮定し、各事前分布にディリクレ分布を導入することでトピックを推定する

LDAの生成過程

(1) 各トピック $k = 1, \dots, K$ について：

(a) デイリクレ分布に従って単語分布 ϕ_k を生成

$$\phi_k \sim \text{Dir}(\beta)$$

(2) 各文書 $d = 1, \dots, D$ について：

(a) デイリクレ分布に従ってトピック分布 θ_d を生成

$$\theta_d \sim \text{Dir}(\alpha)$$

(b) 文書 d における各単語 $n = 1, \dots, N_d$ について：

i. 多項分布に従ってトピックを生成

$$z_{dn} \sim \text{Multi}(\theta_d)$$

ii. 多項分布に従って単語を生成

$$w_{dn} \sim \text{Multi}(\phi_{z_{dn}})$$

ϕ_k : トピック k の単語分布

θ_d : 文書 d のトピック分布

z_{dn} : 文書 d の n 番目の単語の潜在的トピック

w_{dn} : 文書 d の n 番目の単語

$\text{Dir}(\cdot)$: デイリクレ分布

$\text{Multi}(\cdot)$: 多項分布

LDAを用いたトピック推定

トピックの推定手法には以下のようなものがある

1. 変分ベイズ法
2. Collapsed変分ベイズ法
3. ギブスサンプリング

中でもギブスサンプリングは、十分な反復回数を得られるならば変分ベイズ法よりも高い精度でモデル推定を行えることが分かっている

本研究でも**ギブスサンプリング**による推論を行う

提案手法の手順(再掲)

1. 文書の入力
2. トピックの推定
3. 文書へのトピック割り当てによる分類
4. 単語の特徴量の計算
5. 文書へのタグ付与

3. 文書へのトピック割り当てによる分類

- ステップ2で得られた各文書のトピック分布から、各文書に割り当てるトピックを1つ決定する
- 今回は、各文書において出現確率が最大となるトピックを割り当てる
- ここで選択されたトピックには、タグに使うべき単語が含まれていると考えられる

4. 単語の特徴量の計算

トピックに着目し、タグに使うべき単語を見つける過程

今回はterm-scoreを用いる

- TF-IDFの考え方をトピックに応用した指標
- 特定のトピックに偏って出現する単語が、トピックにおいて高確率に出現する際に、大きな値をとる

term-score

$$term - score(v_k) = p(v_k) \cdot \log\left(\frac{p(v_k)}{\prod_{k=1}^K p(v_k)}\right)$$

$p(v_k)$ がTF値に、 $\log(\cdot)$ がIDF値にそれぞれ対応する

5. 文書へのタグ付与

文書に割り当てられたトピックにおけるterm-score上位の単語から、
文書中に出現する単語を順に n_t 個指定し、
タグとして付与する

評価実験

- term-scoreの有用性の評価実験
- 提案システムの性能評価実験

term-scoreの評価実験

- トピックにおける出現確率上位の単語10語
- term-score上位の単語10語

を比較し、どちらの手法がよりトピックの特徴的な単語を捉えられているかを確認した

term-scoreの評価実験結果

表 1: トピックにおける出現確率上位の単語と term-score 上位の単語の比較

トピック 3		トピック 8		トピック 14	
出現確率	term-score	出現確率	term-score	出現確率	term-score
気	香り	音	音質	気	扇風機
購入	匂い	使用	音	デザイン	デザイン
香り	空気清浄機	音質	イヤホン	方	音
私	部屋	私	曲	音	風
匂い	空気	イヤホン	耳	扇風機	サイズ
今	気	方	充電	使用	気
使用	効果	耳	使用	感じ	方
部屋	ソリューション	曲	音楽	サイズ	普通
方	コーヒー	場合	カバー	風	ファン
値段	私	時	場合	普通	感じ

提案システムの性能評価実験

- TF-IDFを用いる方法
- トピックの単語出現確率を用いる方法
- term-scoreを用いる提案システム

の比較を行った

1. 各手法を用いて各文書に3個ずつタグを付与し、
2. ランダムに抽出した25文書に付与された3個のタグに対し、被験者3名が文書の内容との一致/不一致を判定した

提案システムの性能評価実験(1/2)

表 2: 各手法によるタグ付与の精度

	TF-IDF	出現確率	term-score
精度	0.0533	0.36	0.56

term-scoreを用いた提案システムが有効

提案システムの性能評価実験(2/2)

表 3: 全文書に付与されたタグの種類数

	TF-IDF	出現確率	term-score
種類数	4013	514	649

TF-IDFは文書の分類や検索の面で不適切

トピックに基づく2手法のタグも、少なからず一般語も含んでいた

まとめ

- トピックモデルにより得られたトピック情報に基づく文書への自動タグ付与システムを提案した
- 実際のレビュー文書を用いた実験の結果から、term-scoreが有用であることを確認した
- 提案システムのタグ付与性能について、タグの精度と種類数を基に検討を行い、提案システムの有効性を示した
- 今後の課題として、一般語の除去や、トピックへのタグ付与についての検討が挙げられる