2A1-5

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

A study on Auto-Tagging Method based on Latent Dirichlet Allocation

加藤　亮　　吉川　大弘　　古橋　武
Ryo Kato　Tomohiro Yoshikawa　Takeshi Furuhashi

名古屋大学大学院工学研究科
Graduated School of Engineering Nagoya University

Recently, huge amounts of documents are posted on review sites. These documents contain valuable information for consumers and companies. However, it is difficult to read all of documents in point of time and efforts so that the system which organizes knowledge in each document is needed. For this purpose, this study focuses on "tags" which can represent the overview of contents. This paper tries to construct the auto-tagging system and proposes a tagging method based on topic information which are available from Latent Dirichlet Allocation. To investigate the performance of the proposed method, an experiment to assign tags is carried out.

1. はじめに

近年，インターネットの普及により，様々な人々が自由に情報の収集や発信を行うことが可能となっている．特に，価格.com[*1]，Amazon.co.jp[*2] などの通販サイトでは，様々な商品に対する消費者の感想や評価が大量に投稿・公開されている．企業や消費者は，それら投稿された文書を読むことで，有益な情報を得ることができる．例えば，企業にとっては，消費者の声を直接分析することで，商品のマーケティングに活用することができる．また，消費者にとっては，ある商品の購入を考える際に，既に購入した数多くの人の意見を参考にすることができる．しかしインターネット上では，膨大な量の文書が，無秩序に投稿されており，有益な情報を含む文書を探し出すことは容易ではない．そのため，所望する文書を効率的に探し出すための，文書の内容に基づいた知識整理を行うシステムが必要であると考えられる．

知識整理を行う方法として，タグと呼ばれる短い言葉を整理対象に付与する方法が報告されている．写真共有サイトの Flickr[Marlow 06] や，ソーシャルブックマークの Delicious[Golder 06] 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であるフォークソノミー (folksonomy = folks, taxonomy : 民衆による分類法) を採用し，大きな成功を収めている．また，文書に自動でタグを付与し，知識整理を行う手法も数多く研究され，成果を挙げている [Nishida 10][Xiance 09][Brooks 06][Ohkura 06][Fujimura 07]. これらの手法では，あらかじめタグが付与された文書を用いて学習を行う必要がある．

一方，文書の内容を捉える手法として，トピックモデルが注目されている．トピックとは，話題や意味のまとまりのことであり，トピックモデルとは，単語の出現の背景にトピックを仮定した言語モデルである．トピックモデルでは，各文書に出現した単語の種類と，その出現回数の情報を基に，辞書などを用いることなく，トピックの推定を行うことができる．推定されたトピックは，明示的にトピックの名前は得られないものの，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．

本稿では，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の検討を行う．システムでは，事前にタグ付き文書を用意することなく，各文書の内容に適したタグの付与を行うことを目指す．そこで，各文書から最も出現するトピックと，そのトピックに特徴的な単語の情報を用いて文書中の単語を抽出し，タグとして付与するシステムを提案する．

また本稿では，実際のレビュー文書を用いて，提案システムの評価実験を行う．初めに，提案システムにおける単語の特徴量について，抽出された単語の例を示し，評価を行う．さらに，タグ付与の性能の評価実験において，提案システムで付与されるタグの精度，タグの種類数について検討する．

連絡先: 加藤亮，名古屋大学大学院工学研究科，名古屋市千種区不老町，052-789-2793，052-789-3166，katou@cmplx.cse.nagoya-u.ac.jp

∗1 http://kakaku.com/
∗2 http://www.amazon.co.jp/

2. 従来研究

2.1 自動タグ付与手法

文書への自動タグ付与の研究として，brooks ら [Brooks 06] は，文書中の単語から TF-IDF 値が大きい語をタグとして付与する手法を提案している．この手法では，文書に特徴的な単語をタグとして付与することはできるものの，文書間のタグの共通性については考慮しておらず，タグの種類数が膨大となってしまうという問題がある．また，大量のタグ付与済み文書を用いてあらかじめ学習を行い，入力文書に対しタグを付与する手法も提案されている [Xiance 09][Ohkura 06][Fujimura 07]. しかし，多くの文書に対し，事前に人手でタグ付与を行うことは労力の面で困難である．また，これらは既存のタグからの選択によるタグ付与手法であり，新たなタグを生成することはできない．そのため，学習時に存在しない内容を含む文書に対し，適切なタグを付与できない．

文書に含まれるトピックを自動で推定する手法には，確率的潜在意味解析 (PLSA: Probabilistic Latent Semantic Analysis)[Hofmann 99] や潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation)[Blei 03] がある．しかし，これらの手法をそのまま自動タグ付与に用いることはできない．LDA をタグ付与手法に発展させた Tag-LDA[Xiance 09] も提案されているものの，事前のタグ付き文書での学習が必要であり，本稿の目的と異なる．

2.2 TF-IDF

TF-IDF は，文書における単語の特徴量であり，TF-IDF は TF(Term Frequency) 値と IDF(Inverse Document Frequency) 値の積で計算される．文書 d における語彙 v の TF-IDF 値は以下のように求められる．

$$TF-IDF(v_d) = TF(v_d) \cdot IDF(v) \tag{1}$$

TF 値は，文書内で出現する回数が多い単語ほど，値が大きくなる指標である．$n_w(v_d)$ を文書 d 中の語彙 v の出現回数，N_d を文書 d の総単語数とすると，TF 値は式 (2) で表される．

$$TF(v_d) = \frac{n_w(v_d)}{N_d} \tag{2}$$

また，IDF 値は，ある特定の文書にのみ出現する単語ほど値が大きくなる指標である．$n_d(v)$ を語彙 v の出現する文書数，D を全文書数とすると，IDF 値は式 (3) で表される．

$$IDF(v) = \log \frac{D}{n_d(v)} \tag{3}$$

つまり TF-IDF は，特定の文書に偏って出現する単語が，文書中で多く用いられる場合に，大きな値を取る指標である．

2.3 潜在的ディリクレ配分法（LDA）

LDA は，文書が複数の潜在的なトピックを持ち，それらのトピックを媒介して単語が生成されることを仮定した代表的なトピックモデルである．LDA では，文書におけるトピックの出現と，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，トピックの推定を可能にしている．LDA における文書の生成を以下に示す．また，グラフィカルモデルを図 1 に示す．

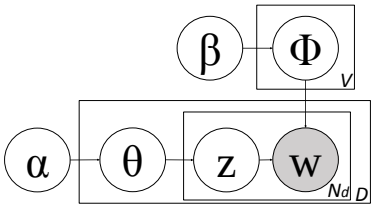

図 1: LDA のグラフィカルモデル

(1)For each choose $\Theta \sim \mathrm{Dir}(\alpha)$.
(2)Choose $\Phi \sim \mathrm{Dir}(\beta)$.
(3)For each of N words w_n :
　(a)Choose a topic $z_n \sim \mathrm{Multinomial}(\Theta)$.
　(b)Choose a word w_n with $p(w_n|z_n, \Phi)$.

ここで，Dir(·) はディリクレ分布，Multinomial(·) は多項分布を表し，α，β はそれぞれのディリクレ分布のハイパーパラメータである．また，N は単語の総出現回数である．

3. 提案システム

本稿では，文書のトピック情報を用いた文書への自動タグ付与システムを提案する．提案システムのフローチャートを図 2 に示す．また，各部分の処理について 3.1 から 3.5 で順に述べる．

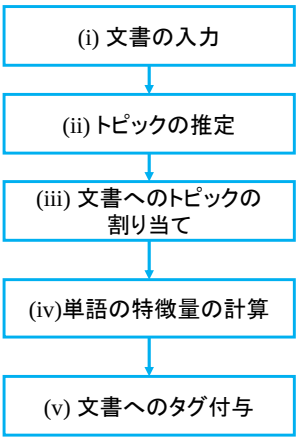

図 2: 提案システムのフローチャート

3.1 文書の入力

システムにタグ付けを行う文書を入力する．文書中のテキストに対して形態素解析を行い，各文書に出現する単語と，その出現回数を求める．本稿では，名詞のみを単語情報として用いる．なお，複合語のように，名詞が連続した場合は，それらを一つの名詞として扱う．また，事前に *stopword* が与えられている場合は，これを取り除く．

3.2 トピックの推定

3.1 で得られた単語情報を基に，トピックを推定する．本稿では，2.3 で述べた LDA を用いてトピック推定を行う．推定されたトピックから，各文書のトピック分布と，各トピックの単語分布を算出する．

3.3 文書へのトピック割り当てによる分類

3.2 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決定する．本稿では，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．

3.4 単語の特徴量の計算

提案システムでは，意味のまとまりであるトピックに着目した単語の特徴量として，term-score[Blei 09] を用いる．term-score は，TF-IDF の考え方をトピックに応用した指標であり，TF-IDF における文書をトピックに置き換えたものであるといえる．トピックの総数を K，あるトピック k における語彙 v の出現確率を $p(v_k)$ とすると，term-score は式 (4) で表される．

$$term-score(v_k) = p(v_k) \cdot \log\left(\frac{p(v_k)}{\prod_{k=1}^{K} p(v_k)}\right) \tag{4}$$

式 (4) において，$p(v_k)$ が TF 値に，log(·) が IDF 値にそれぞれ対応する．term-score は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，大きな値をとる指標である．

3.5 文書へのタグ付与

各文書に対し，n_t 個のタグを付与する．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．

4. 実験

実際のレビュー文書を用いて，提案システムの評価実験を行う．初めに，単語の特徴量として用いる term-score に対する評価実験を行った．実験では，トピックにおける出現確率上位の単語と，term-score 上位の単語を比較し，特徴量としての妥当性を定性的に評価した．次に，提案システムによるタグ付与性能の評価を行った．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と比較し，付与されるタグの精度，タグの種類数について検討した．

4.1 実験条件

本実験では，楽天市場 *3 に公開されるレビュー文書のうち，カテゴリ「家電」,「AV 機器」,「カメラ」に属するレビューを用いた．その中から名詞が 70 回以上出現する 2000 文書を抽出し，実験用データセットを構築した．

また，LDA におけるトピックの推定方法はギブスサンプリングを用い，推論の反復回数は 300 回とした．LDA におけるディリクレ分布のパラメータは α=0.01，β=0.01 とし，トピック数は 30 とした．

4.2 term-score の評価実験

4.2.1 実験方法

単語の特徴量として，term-score に対する評価実験を行った．トピックにおける出現確率上位の単語 10 語と，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．

4.2.2 結果

結果の例を表 1 に示す．トピック 3 に注目すると，term-score では,「香り」「匂い」「空気清浄機」「部屋」「空気」といった，関連性が高い特徴的な単語が上位となっていることがわかる．また,「空気清浄機」が現れていることで,「空気清浄機」による「部屋」の「空気」に対する「香り」や「匂い」についてのレビューであることが予測できる．一方，出現確率では,「香り」「匂い」「部屋」が上位ではあるものの，何の「香り」や「匂い」についてのレビューなのかは想像できない．また,「気」「購入」「私」といった，特徴があまりないと思われる単語も上位となる結果となっていた．

トピック 8，トピック 14 も同様に，term-score では対象であると思われる「イヤホン」,「扇風機」が，出現確率よりも上位に現れていることが確認できる．さらに特徴的とはいえない単語の順位が出現確率よりも下がっていることが確認できる．特に,「私」「使用」など，どの文書でも用いられる一般語は，複数のトピックで確率上位になっており，出現確率をそのまま用いた場合，適切ではない語がタグとして付与されてしまう可能性があると考えられる．

4.3 提案システムの性能評価実験

4.3.1 実験方法

文書へのタグ付与性能について，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と，term-score を用いる提案システムとの比較を行った．実験ではまず，データセットの全レビュー文書に対し，各手法を用いて 3 個（$n_t = 3$）のタグを付与した．続いて，全文書からランダムに 25 文書を抽出し，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．一致と判定された回数を，抽出された文書数で割ることで精度を計算した．また，各手法により付与されたタグの種類数についても検討した．

*3 http://www.rakuten.co.jp/

4.3.2 結果

各手法のタグ付与の精度（3 名の被験者における精度の平均）を表 2 に示す．表から，文書単位の特徴量である TF-IDF を用いる方法に対し，トピック情報に基づいてタグ付与を行った 2 手法の精度が上回っていることがわかる．また，term-score を用いた提案システムが，出現確率に基づくタグ付与の方法よりも，精度が高いことがわかる．実際に付与されたタグを確認すると，TF-IDF による方法では，単に他の文書に出現しないだけの単語が抽出され，内容を表しているとはいえないタグが多く付与されていた．また，出現確率を用いる方法では，4.2 で示した通り，一般語がタグとして付与されやすい傾向があることが確認できた．それらに対し提案システムでは，まだ十分に内容を捉えたタグ付けができているとはいえないものの，トピックに特徴的と思われる単語も多くタグとして付与できていた．これらの結果から，term-score を用いた提案システムが，トピック情報に基づく自動タグ付与手法において有効な手法であることが確認できた．

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

また，各手法が全 2000 文書に付与したタグの種類数を表 3 に示す．表から，TF-IDF によるタグ付けでは，タグの種類数が著しく多くなっていることがわかる．理論上のタグの最大種類数が $2000 \times 3 = 6000$ であるため，7 割近いタグが文書固有のタグとなる結果であった．種類数が多い場合，共通のタグを手掛かりに文書の絞り込みを行うことが困難となるため，TF-IDF は文書の分類や検索の面でも適切ではないと考えられる．一方，トピックに基づく 2 手法では，平均 12 文書程度に同一のタグが付与されている結果となった．これは，トピックに特徴的な単語をタグとして付与することで，文書間の内容の共通性をタグ付与に反映させやすいためであると考えられる．ただし，付与されたタグには少なからず一般語も含まれており，一般語などが共通のタグとして付与されることは検索の上で有用とはいえないため，今後，一般語を除去する工夫などが必要である．

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

5. おわりに

本稿では，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により，提案システムで用いる term-score の単語抽出性能評価を行った．実験結果から，term-score によって，トピックに特徴的な単語を捉え，一般語の順位を下げることができることを確認した．また，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．

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

参考文献

[Marlow 06] C.Marlow, M.Naaman, D. Boyd, and M. Davis: “Ht06, tagging paper, taxonomy, flickr, academic article, to read”, Proc. 17rd Conference on Hypertext and Hypermedia, pp.31-40, 2006.

[Golder 06] S.A. Golder, and B.A. Huberman, ”Usage patterns of collaborative tagging systems”ACM SIGKDD Explor. Newsletter, vol.32, no.2, pp198-208, 2006.

[Nishida 10] 西田京介, 藤村考: “階層的オートタギングによる Q & A コミュニティの知識整理”, DEIM Forum 2010, D3-4, 2010.

[Xiance 09] Xiance Si, Maosong Sun: “Tag-LDA for Scalable Real-time Tag Recommendation”, Journal of Information & Computational Science 6, 1, 23-31, 2009.

[Brooks 06] C.H. Brooks,and N. Montanez,“Improved annotation of the blogosphere via autotagging and hierarchical clustering”Proc. 15th International Conference on World Wide Web, pp.625-632, 2006.

[Ohkura 06] T. Ohkura, Y. Kiyota, and H. Nakagawa, “Browsing system for weblog articles based on automated folksonomy”Proc. WWW 2006 Works. Weblogging Ecosystem: Aggregation,Analysis, and Dynamics, 2006.

[Fujimura 07] S. Fujimura, K. Fujimura, and H. Okuda, “Blogosonomy: Autotagging any text using blogger’s knowledge” Proc. 2007 IEEE/WIC/ACM International Conference on Web Intelligence, pp.205-212, 2007.

[Hofmann 99] Hofmann, T. Probabilistic Latent Semantic Analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 289-296, 1999.

[Blei 03] D.M.Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”Journal of Machine Learning Research, 3:993-1022, 2003.

[Blei 09] D.M.Blei, and J. D. Lafferty,“TOPIC MODELS” In A.Srivastava and M. Sahami, editors, Text Mining: Theoryand Applications. Taylor and Francis, 2009.