2A1-1

論文著者の貢献度推定

Contribution Estimation of Authors

市瀬 龍太郎[*1] 渡辺 曜大[*2]
Ryutaro Ichise Yodai Watanabe

[*1]国立情報学研究所 [*2]会津大学コンピュータ理工学部
National Institute of Informatics The University of Aizu

We propose a method to estimate the contributions to a paper by authors. The contribution estimation method is based on a probabilistic model for writing papers. We conducted experiments with two data sets and discuss about the model for estimating the contributions.

1. はじめに

近年になり，学術研究の細分化が進んでいる．それに伴い，学術研究においては，多くの知見を統合しながら研究を進めなければならなくなっている．Jones らの研究 [Jones 08] によると，科学と工学の分野において，1975 年には，単著の論文が 3 割程度であったのに対して，2005 年には，1 割程度にまで低下している．その一方で，複数の大学にまたがる多様な研究者で，研究チームを構成して書かれた論文が 3 割以上に増えてきている．このような現象は，科学と工学の分野のみならず，社会科学の分野でも見られ，多様な知の統合が，新たな知を生み出すのに欠かせなくなって来ていると言えるであろう．

たくさんの研究者がお互いに知を出し合い，研究をすれば，常に新しく画期的な研究ができるわけではない．どのような組み合わせで研究を行えば，新たな知が生み出されるのかを予測するために，リンク予測を用いた研究 [Wohlfarth 08] などが行われてきている．しかし，そのような研究では，研究者の論文に対する貢献度は，等価なものとして取り扱われている．しかし，実際には，論文に対して，研究者が貢献する部分は異なっているであろう．

そこで，本研究では，どのような共同研究がうまくいくかを予測するために，論文に対する著者の貢献度を推定する手法を提案する．提案手法は，著者貢献モデルと呼ばれる確率モデルに基づいており，変分ベイズ法を用いることで，論文著者の貢献度を推定する．そして，人工データと実際の論文執筆データを用いて，このモデルの検証を行った結果について報告する．

2. 関連研究

文書から，確率モデルを用いて著者と文書の関係を推定する研究は，これまでにもいくつかの研究が行われている．著者-話題モデル [Steyvers 04] では，文書の情報から，著者がどの話題に関連しているかを示す分布を計算するための確率モデルを提案している．また，著者-ペルソナ-話題モデル [Mimno 07] では，著者-話題モデルを拡張し，著者が複数の話題分布を持つことができる．また，潜在興味話題モデル [川前 09] では，著者の話題分布間に類似性を導入した確率モデルを提案している．しかし，このようなモデルには，本研究で提案するような，著者の貢献度を推定するモデルは入っていない．

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，Fax:03-3556-1916，E-mail:ichise@nii.ac.jp

3. 著者貢献モデル

3.1 論文の生成モデル

本論文で提案する著者貢献モデルにおいて，論文の生成モデルは次のようになっている．

- 文書 d の著者集合 A_d に対して，著者貢献度の多項分布 ψ_d から，著者 a を選択．
- 選択した著者 a に応じて，トピック選択の多項分布 θ_a から，トピック z を選択
- 選択したトピック z に応じて，単語選択の多項分布 ϕ_z から，単語 w を選択

この時，観測可能なのは，文書 d の著者集合と，文書中の単語 w である．

3.2 変分ベイズによる確率推定

本論文では，変分ベイズ法 [Attias 99] によりパラメータを推定する．変分ベイズ法では，潜在変数 Z およびパラメータ θ のテスト分布 $q(Z)$, $q(\theta)$ を導入し，適当な初期分布を仮定して以下の 2 つのステップを収束するまで繰り返すことによって真の事後分布 $p(Z, \theta|D)$ を近似する．

- VB-E ステップ

$$q(Z) \leftarrow C \exp\langle \log p(D, Z|\theta) \rangle_{q(\theta)} \tag{1}$$

- VB-M ステップ

$$q(\theta_i) \leftarrow C p(\theta_i) \exp\langle \log p(D, Z|\theta) \rangle_{q(Z)q(\theta_{-i})} \tag{2}$$

ただし，C は規格化定数であり，θ_{-i} は θ_i 以外の θ の成分をあらわしている．

前節の確率モデルでは，$D = (\boldsymbol{d}, \boldsymbol{w})$, $Z = (\boldsymbol{a}, \boldsymbol{z})$ であり，

$$p(D, Z|\theta) = \prod_i p(d_i) p(a_i|d_i) p(z_i|a_i) p(w_i|z_i)$$

とあらわされる．多項分布の自然共役事前分布は Dirichlet 分布だから，事前分布を

$$p(a|d) \sim \mathrm{Dir}(\theta^0_{da}),\ p(z|a) \sim \mathrm{Dir}(\theta^0_{az}),\ p(w|z) \sim \mathrm{Dir}(\theta^0_{zw})$$

とおけば，事後分布もパラメータ $\theta_{da}, \theta_{az}, \theta_{zw}$ をもちいて

$$q(a|d) \sim \mathrm{Dir}(\theta_{da}),\ q(z|a) \sim \mathrm{Dir}(\theta_{az}),\ q(w|z) \sim \mathrm{Dir}(\theta_{zw})$$

表 1: 人工データを用いた実験の設定値とその結果

n_d	n_a	n_z	n_w	t_w	誤差
1000	100	10	100	1000	0.0352347
1000	100	10	100	100	0.16419
100	100	10	100	1000	0.253495
1000	100	10	1000	1000	0.0332936
1000	100	50	100	1000	0.040309

とあらわすことができる．これらを (1) 式に代入して計算すると，潜在変数 $\boldsymbol{a}, \boldsymbol{z}$ のテスト分布 $q(\boldsymbol{a}, \boldsymbol{z})$ は

$$q(\boldsymbol{a}, \boldsymbol{z}) = \prod_i q_i(a_i, z_i)$$

と因数分解され，分布 q_i は d_i, w_i のみに依存することがわかる．これを q_{dw} とあらわせば，q_{dw} の更新則は

$$q_{dw}(a, z) \leftarrow C \exp(\psi(\theta_{da}) - \psi(\theta_d) + \psi(\theta_{az}) - \psi(\theta_a) + \psi(\theta_{zw}) - \psi(\theta_z)) \quad (3)$$

であたえられる．ただし，C は正規化定数，ψ は Digamma 関数をあらわし，

$$\theta_d = \sum_a \theta_{da}, \quad \theta_a = \sum_z \theta_{az}, \quad \theta_z = \sum_w \theta_{zw}$$

とおいた．一方 (2) 式より，パラメータ $\theta_{da}, \theta_{az}, \theta_{zw}$ の更新則は，論文 d における単語 w の頻度を F_{dw} とするとき，

$$\theta_{da} \leftarrow \theta^0_{da} + \sum_{z,w} F_{dw} q_{dw}(a, z) \quad (4)$$

$$\theta_{az} \leftarrow \theta^0_{az} + \sum_{d,w} F_{dw} q_{dw}(a, z) \quad (5)$$

$$\theta_{zw} \leftarrow \theta^0_{zw} + \sum_{d,a} F_{dw} q_{dw}(a, z) \quad (6)$$

とあらわされる．これらを用いてパラメータの推定を行う．

4. 実験

提案モデルの能力を調べるために，2 つの実験を行った．まず，最初に，提案モデルの推定能力を調べるために，人工的にデータを作成し，推定の誤差を評価した．人工データは，それぞれ，論文数 n_d，著者数 n_a，トピック数 n_z，単語数 n_w，一論文中の延べ単語数 t_w として，以下の要領で生成した．

- 論文数 n_d，著者数 n_a，トピック数 n_z，単語数 n_w を定める．
- 各論文の著者数，各著者のトピック数，各トピックの単語数を適当な分布を仮定してランダムに定める．
- 各論文の著者分布，各著者のトピック分布，各トピックの単語分布を適当な分布を仮定してランダムに定める．
- 一論文中の延べ単語数 t_w を定め，前節の確率モデルに従って各論文中の単語を生成する．

実験の設定値とその結果を表 1 に示す．ここで，真の貢献度分布と推定貢献度分布の間の変動距離をすべての論文について平均したものを誤差としている．

この実験結果より，以下のことがわかった．

- 著者数 n_a を固定した際に，論文数 n_d，単語数 n_w，一論文中の延べ単語数 t_w が大きい方がエラーが小さくなる．
- 論文数 n_d と一論文中の延べ単語数 t_w では，論文数 n_d を大きくした方がエラーを小さくするのに有効である．

次に，実際の論文データを用いて実験した．実験では，DBLP のデータから，論文概要が取得可能な 28650 本の論文を選択し，その情報を用いた．そのデータの著者数は 43,715，単語数は 45,607，論文全ての延べ単語数は 930,481 であった．また，トピック数は 50 を用いた．実験の結果，各著者の貢献度は，人工データを用いた実験で誤差が大きい時とよく似た出力が得られた．この原因の一つとして，用いたデータの性質が挙げられる．本実験では，国際会議で発表された論文の概要を用いた．そのため，一論文中の延べ単語数が非常に小さくなったと考えられる．また，著者が書いた論文の本数が限られるため，精度の低下につながったと考えられる．そのため，実データに適用する際には，フィルタリング処理によってデータの質を上げるなどの修正が必要になると考えられる．また，確率モデルやアルゴリズムを改良することで，このようなデータにも対応可能なように改良することも今後の課題となる．

5. おわりに

本研究では，著者が論文に対してどの程度貢献しているのかを推定するモデルを提案し，その計算アルゴリズムを示した．また，実験により，提案手法の特性を明かにした．今後は，実データにおける実験を複数のデータに対して行っていくと同時に，定性的な評価も行っていく予定である．

参考文献

[Attias 99] Attias, H.: Inferring Parameters and Structure of Latent Variable Models by Variational Bayes, in Laskey, K. B. and Prade, H. eds., *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 21–30, (1999)

[Jones 08] Jones, B. F., Wuchty, S., and Uzzi, B.: Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science, *Science*, Vol. 322, pp. 1259–1262 (2008)

[Mimno 07] Mimno, D. and McCallum, A.: Expertise Modeling for Matching Papers with Reviewers, in Berkhin, P., Caruana, R., and Wu, X. eds., *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*, pp. 500–509, ACM (2007)

[Steyvers 04] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T.: Probabilistic author-topic models for information discovery, in *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315 (2004)

[Wohlfarth 08] Wohlfarth, T. and Ichise, R.: Semantic and Event-Based Approach for Link Prediction, in Yamaguchi, T. ed., *Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management*, Vol. 5345 of *Lecture Notes in Computer Science*, pp. 50–61, Springer (2008)

[川前 09] 川前 徳章, 山田 武士：著者の興味と文書の内容の依存関係に着目した潜在変数モデル, 信学技報, Vol. 109, No. 51, pp. 19–24 (2009)