サジェストを用いて収集したウェブ検索結果のトピックモデリングにおける文書フィルタリング*

陳 磊[†] 井上 祐輔[†] 今田 貴和[†] 徐 凌寒[†] 宇津呂 武仁[‡] 河田 容英[§]
筑波大学大学院 システム情報工学研究科[†] 筑波大学システム情報系[‡] (株) ログワークス[§]

1 はじめに

本論文では，検索者が詳細な情報を検索したい対象を「クエリ・フォーカス」と呼ぶ．また，検索エンジン・サジェストとして提示される語は，クエリ・フォーカスに対して，多数のウェブ検索者が AND 検索の形で 2 つ目以降に入力した語を情報源として抽出されたものである (図 1)．本論文では，検索エンジン・サジェストを情報源として収集されたウェブページの集合に対して，トピックモデルを適用し，生成されたトピック中に対応付けられたサジェストの頻度を用いて主要話題を含むウェブページを選定することにより，トピックモデルにおける話題同定を高精度に行う手法を提案する．

本論文では，まず，検索エンジンを用いて，1 つのクエリ・フォーカスに対して，最大約 1,000 語の検索エンジン・サジェストを収集する．次に，クエリ・フォーカスと収集されたサジェストを用いた AND 検索によって，上位 20 件のウェブページを収集する．収集されたウェブページを集約するために，トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [1] を用いる．収集されたウェブページの集合に対して，トピックモデルを適用することにより，ウェブページを数十個のトピックに集約する．ウェブページを収集する際，検索において用いたサジェストをウェブページのラベルとみなす．これによって，全 1,000 個のサジェストも，同様に数十個のトピックに集約される．次に，1 つのトピックに対して，トピック中の各サジェストの頻度を求める．1 つのトピックにおいて，1 つのウェブページのラベルであるサジェストの頻度の最大値を，そのトピックにおいて当該ウェブページに付与されたサジェストの頻度とみなす．そして，そのトピックにおいて，各ウェブページの頻度に対して下限値を設けて，下限値以上の頻度を持つウェブページを，そのトピックの主要話題を含むウェブページとみなして選定する．下限値未満の頻度を持つウェブページを，そのトピックの副次的な話題を含むウェブページとみなして除外する．各トピックに対して，このような処理を行い，各トピック中の話題を厳選し，各トピックの主要話題を含むウェブページの集合を収集することによって，トピックモデルにおける話題同定の高精度化を実現する．

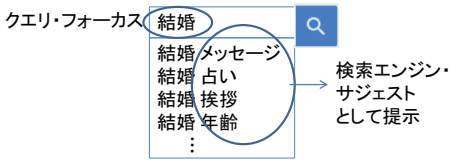


図 1: 検索エンジン・サジェストの例

表 1: 各クエリ・フォーカスのサジェスト数，および，ウェブページ数

クエリ・フォーカス	サジェスト数	ウェブページ数
就活	934	13,221
結婚	989	14,413

2 検索エンジン・サジェストの収集

評価用クエリ・フォーカスに対して，Google[1] 検索エンジンを用いて，一クエリ・フォーカス当たり約 100 通りの文字列を指定し，最大約 1,000 語のサジェストを収集する．100 通りの文字列とは具体的には，五十音，濁音，半濁音および「きゃ」や「ぴゃ」などの開拗音である．例えば検索窓に「結婚 あ」と入力すると,「挨拶」や「相性」等がサジェストとして提示されるので，それらの収集を行う．本論文において収集したサジェストの数を表 1 に示す．

*Document Filtering in Topic Modeling of Web Pages collected with Search Engine Suggests

†Lei Chen, Yusuke Inoue, Takakazu Imada, Linghan Xu, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba
§Yasuhide Kawada, Logworks Co., Ltd.

[1] https://www.google.com/

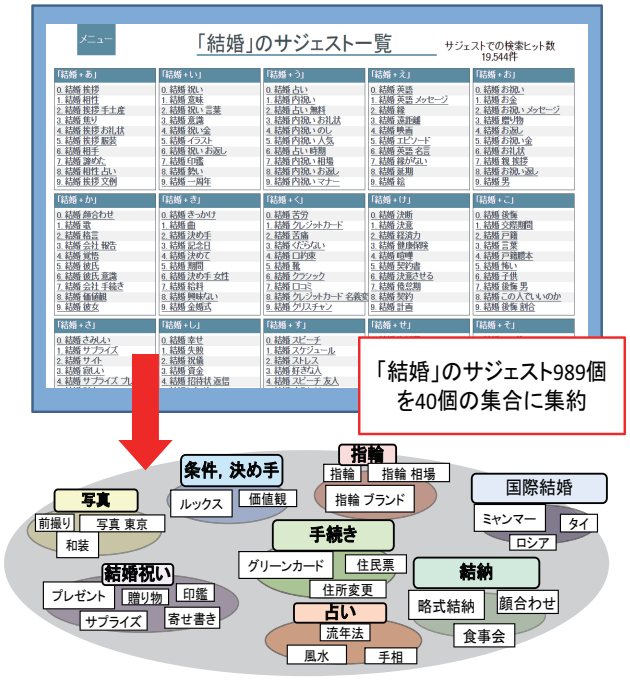


図 2: 検索エンジン・サジェストの集約 (クエリ・フォーカス: 「結婚」)

3 トピックモデルを用いた文書集合およびサジェストの集約

本節では，検索エンジン・サジェストを用いてウェブページを収集する．そして，収集されたウェブページ集合を対象としてトピックモデルを適用することにより，ウェブページをトピックに集約する．さらに，ウェブページに付与されたサジェストについても，同様に集約を行う．

3.1 ウェブページの収集

まず，Yahoo! Search BOSS API[2] に対して検索クエリを指定することにより，上位 N 件のウェブページを収集する (本論文においては，$N = 20$ とする)．ここでの検索クエリは，各クエリ・フォーカスおよび前節において収集した各サジェストの AND 検索の形で作成する．各クエリ・フォーカスごとに得られたウェブページ数を表 1 に示す．

3.2 ウェブページへの検索エンジン・サジェストの割り当て

各ウェブページは，クエリ・フォーカスおよび各サジェストの AND 検索によって検索されたものである．異なるサジェストによって，同一のウェブページが検索される場合も存在する．したがって，複数のサジェストが対応しているウェブページも多数存在する．そこで，各ウェブページ d に対して，$d \in D(s, N)$ となるサジェスト s を集めた集合 $S(d)$ を次式で定義する．

$$S(d) = \left\{ s \in S \middle| d \in D(s, N) \right\}$$

3.3 トピックモデル

本研究では，トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [1] を用いる．LDA を用いたトピックモデルの推定においては，語 w の列によって表現された文書の集合と，トピック数 K を入力として，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_n)$ $(w \in V)$，及び，各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ $(n = 1, \ldots, K)$ を推定する．これらを推定するためのツールとしては，GibbsLDA++[3]を用いた．LDA のハイパーパラメータである α，β としては，GibbsLDA++の基本設定値である $\alpha = 50/K$，$\beta = 0.1$ を用いた．LDA を用いたトピック推定においては，トピック数 K を人手で与える必要があるが，今回の評価においては，各トピックにおける記事のまとまりが最もよくなる場合のトピック数として，クエリ・フォーカス「結婚」の場合は $K = 40$ を採用した，クエリ・フォーカス「就活」の場合は $K = 50$ を採用した．

3.4 文書に対するトピックの割り当てによる文書集合の集約

本論文では，各ウェブページに対してトピックを一意に割り当てることによって，ウェブページ集合をトピックに分類する．ウェブページ集合を D，トピック数を K，1 つのウェブページを $d (d \in D)$ とすると，トピック $z_n (n = 1, \ldots, K)$ のウェブページ記事集合 $D(z_n)$ は以下の式で表される．

$$D(z_n) = \left\{ d \in D \middle| z_n = \underset{z_u \ (u=1,\ldots,K)}{\operatorname{argmax}} P(z_u|d) \right\}$$

これはつまり，ウェブページ d におけるトピックの分布において，確率が最大のトピックに，ウェブページ d を割り当てていることになる．

3.5 トピックに対するサジェストの割り当て

各ウェブページは，各クエリ・フォーカスおよび各サジェストの AND 検索によって検索されたものである．したがって，あるウェブページには，一つ以上のサジェストが対応することになる．また，各ウェブページには，トピックが対応付けられている．以上のことから，一つのトピックに対して割り当てられた一つ以上のウェブページに対応するサジェストを収集することにより，一つのトピックに一つ以上のサジェストが割り当てられていることになる．

[2] http://developer.yahoo.com/search/boss

[3] http://gibbslda.sourceforge.net/

実際に，クエリ・フォーカス「結婚」の場合，989個のサジェストが 40 個のトピックに割り当てられた(図 2)．このことから，一般には，各トピックに対して複数のサジェストが対応しており，これによって，複数のサジェストが各トピックに集約されたとみなす [3]．

4 トピックにおける文書フィルタリングによる話題同定の高精度化

4.1 概要

前節では，ウェブページ集合を対象としてトピックモデルを適用することによって，ウェブページ集合をトピックの集合に集約した．しかし，通常，一つのトピックに対しても，膨大な数のウェブページが割り当てられており，クエリ・フォーカスとの関連性が低いウェブページやトピックにおける副次的な話題を含むウェブページ等も数多く存在する．そのため，必要な情報を厳選して効率よく情報の収集を行う目的においては，大きな障害となる．この問題を解決するために，本節では，トピックに対応付けられたサジェストの頻度を用いて，トピックから主要話題を含むウェブページだけを選定する手法について述べる．

4.2 トピックにおけるサジェストの頻度を用いた文書フィルタリング

本節では，トピックに割り当てられているウェブページの集合から主要話題を含むウェブページをフィルタリングする手法を述べる．トピック z_n においてサジェスト s が割り当てられているウェブページ d の数を，トピック z_n におけるサジェスト s の頻度 $f(s, z_n)$ として次式により定義する．

$$f(s, z_n) \quad = \quad \left| \left\{ d \in D(z_n) \middle| \ s \in S(d) \right\} \right|$$

また，トピック z_n において，ウェブページ d に割り当てられたサジェストの頻度の最大値 $f_{max}(d)$ を次式で定義する．

$$f_{max}(d) \quad = \quad \left| \underset{s \in S(d)}{\operatorname{argmax}} f(s, z_n) \right|$$

そして，ウェブページの頻度に対して下限値 f_{lbd} を設けて，下限値以上の頻度を持つウェブページだけを選定する．各トピック z_n において，下限値 f_{lbd} 以上の頻度を持つウェブページの集合 $D(z_n, f_{lbd})$ を次式で表す．

$$D(z_n, f_{lbd}) \quad = \quad \left\{ d \in D(z_n) \middle| \ f_{max}(d) \geq f_{lbd} \right\}$$

また，ウェブページの頻度の下限値 f_{lbd} の条件のもとで，一つ以上のウェブページを持つトピックの集合 $T(f_{lbd})$ は次式で表される．

$$T(f_{lbd}) \quad = \quad \left\{ z_n \middle| \ D(z_n, f_{lbd}) \neq \phi \right\}$$

4.3 評価

本節では，トピックに対応付けられたサジェストの頻度を用いて主要話題を含むウェブページをフィルタリングする手法を評価する．

まず，各トピック z_n において，確率値 $P(z_n|d)$ の降順の上位 r 件のウェブページを評価対象とし[4]，評価対象のウェブページの集合を次式 $D_{rank}(z_n, r)$ とする．

$$D_{rank}(z_n, r) \ = \ \left\{ d \in D(z_n) \middle| \ D(z_n) \text{ 中での } P(z_n|d) \text{ の降順の } d \text{ の順位} \leq r \right\}$$

次に，集合 $D_{rank}(z_n, r)$ 中の各ウェブページに対して，人手で話題分析を行う．特定のウェブページ $d(\in D_{rank}(z_n, r))$ に対して，d と同じ話題を持つウェブページが集合 $D_{rank}(z_n, r)$ 中に (d を含めて) 合計 d_f 件以上含まれている場合に[5]，集合 $D_{rank}(z_n, r)$ は話題としてまとまっているとみなし，これらのウェブページを，集合 $D_{rank}(z_n, r)$ 中の参照用ウェブページとみなす．以上の手順によって，トピック z_n において収集された参照用ウェブページの集合 $D_{ref}(z_n, r, d_f)$ を次式で表す．

$$D_{ref}(z_n, r, d_f) \ = \ \left\{ d \in D(z_n) \middle| \ D_{rank}(z_n, r) \text{ 中に, } d \text{ と同じ話題のウェブページが } d_f \text{ 件以上含まれる} \right\}$$

また，各トピック z_n における確率値 $P(z_n|d)$ の降順の上位 r 件のウェブページ集合 $D_{rank}(z_n, r)$ から，提案手法により，頻度の下限値 f_{lbd} を満たすウェブページを選定して収集した集合 $D'(z_n, f_{lbd}, r)$ は次式で定義される．

$$D'(z_n, f_{lbd}, r) \quad = \quad D(z_n, f_{lbd}) \cap D_{rank}(z_n, r)$$

本節では，この集合 $D'(z_n, f_{lbd}, r)$ と参照用ウェブページ集合 $D_{ref}(z_n, r, d_f)$ の間の重複を測定することによって提案手法の評価を行う．

本論文では，ウェブページの頻度の下限値 f_{lbd} の条件のもとでの，トピック間での再現率と適合率のマクロ平均およびマイクロ平均として，次式で示す値を

[4] 本論文においては，$r = 30$ とする．
[5] 本論文においては，$d_f = 3$ とする．

求め，これを評価尺度として用いて提案手法の評価を行う．

$$\text{再現率 (マクロ平均)}(f_{lbd}) = \frac{\sum_{z_n \in T(f_{lbd})} \frac{\left|D_{ref}(z_n, r, d_f) \cap D'(z_n, f_{lbd}, r)\right|}{\left|D_{ref}(z_n, r, d_f)\right|}}{\left|\left\{z_n (n = 1, \ldots, K) \middle| D_{ref}(z_n, r, d_f) \neq \phi\right\}\right|}$$

$$\text{適合率 (マクロ平均)}(f_{lbd}) = \frac{\sum_{z_n \in T(f_{lbd})} \frac{\left|D_{ref}(z_n, r, d_f) \cap D'(z_n, f_{lbd}, r)\right|}{\left|D'(z_n, f_{lbd}, r)\right|}}{\left|T(f_{lbd})\right|}$$

$$\text{再現率 (マイクロ平均)}(f_{lbd}) = \frac{\sum_{z_n \in T(f_{lbd})} \left|D_{ref}(z_n, r, d_f) \cap D'(z_n, f_{lbd}, r)\right|}{\sum_{z_n (n=1,\ldots,K)} \left|D_{ref}(z_n, r, d_f)\right|}$$

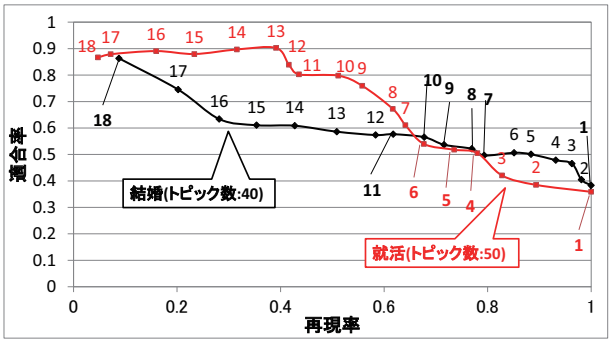
$$\text{適合率 (マイクロ平均)}(f_{lbd}) = \frac{\sum_{z_n \in T(f_{lbd})} \left|D_{ref}(z_n, r, d_f) \cap D'(z_n, f_{lbd}, r)\right|}{\sum_{z_n \in T(f_{lbd})} \left|D'(z_n, f_{lbd}, r)\right|}$$

評価対象として,「結婚」および「就活」をクエリ・フォーカスとして用いた場合について，$r = 30$，$d_f = 3$ として，ウェブページの頻度の下限値 f_{lbd} を変化させて再現率および適合率をプロットした結果を図 3 に示す．

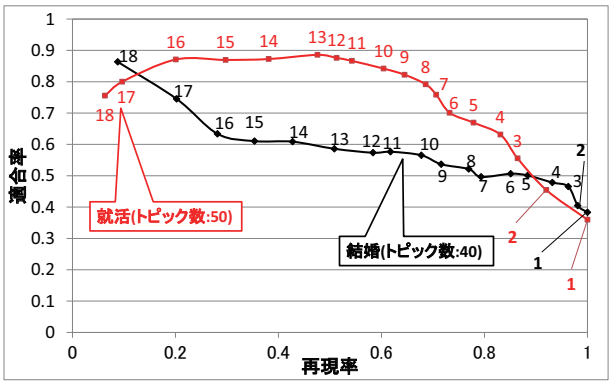
比較対象として，ウェブページの頻度の下限値 $f_{lbd} = 1$ として，提案手法を用いず，トピックモデルをそのまま適用した結果をベースラインとすると，図 3 の結果より，提案手法によってベースラインよりも高い適合率が達成できていることが分かる．以上の結果によって，本論文の提案手法の有効性を示すことができた．

5 関連研究

本論文に関連して，文献 [2] においては，SNS データに対してトピックモデルを適用する際に，人気度の高い語が引き起こす弊害を解消するためのモデル化を施したトピックモデルを提案している．また，文献 [4] においては，トピックモデルの適用結果において，話題の多様性をなるべく大きくするとともに冗長な話題を集約するために，異なる粒度でのトピックモデルの適用結果における関連するトピックを対応付けする手法を提案している．一方，本論文で用いている検索エンジン・サジェスト集約の方式については，文献 [3] においてその詳細な評価結果を報告している．



(a) マクロ平均



(b) マイクロ平均

図 3: 評価結果 (ウェブページの頻度の下限値 f_{lbd} を変化させた場合の再現率・適合率のプロット)

6 おわりに

本論文では，検索エンジン・サジェストを情報源として収集されたウェブページの集合に対して，トピックモデルを適用し，生成されたトピック中に対応付けられたサジェストの頻度を用いて主要話題を含むウェブページを選定することにより，トピックモデルにおける話題同定を高精度に行う手法を提案した．

参考文献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] Y. Cha, B. Bi, C.-C. Hsieh, and J. Cho. Incorporating popularity in topic models for social network analysis. In *Proc. 36th SIGIR*, pp. 223–232, 2013.

[3] 井上祐輔, 今田貴和, 陳磊, 徐凌寒, 宇津呂武仁, 河田容英. 検索エンジン・サジェストおよびトピックモデルを用いたウェブ検索結果の集約. 第 8 回 DEIM フォーラム論文集, 2016.

[4] 井上祐輔, 小池大地, 宇津呂武仁, 神門典子. 複数の粒度での LDA 適用結果におけるトピック集約. 言語処理学会第 20 回年次大会論文集, pp. 924–927, 2014.