1L3-2in1

トピックモデルおよび分散表現の併用による検索エンジン・サジェストの集約

Aggregating Search Engine Suggests by Integrating a Topic Model and Word Embeddings

聶 添[*1*2] 丁 易[*1] 趙 辰[*1] 李 佳奇[*1] 宇津呂 武仁[*3] 河田容英[*4]
Tian Nie Yi Ding Chen Zhao Jiaqi Li Takehito Utsuro Yasuhide Kawada

[*1]筑波大学大学院システム情報工学研究科
Grad. Sch. Sys. & Inf. Eng, Univ. of Tsukuba

[*2]パイオニア 商品統括部
Product Management Division, Pioneer Corporation

[*3]筑波大学システム情報系
Fclty. Eng, Inf. & Sys, Univ. of Tsukuba

[*4](株) ログワークス
Logworks Co., Ltd.

The background of this paper is the issue of how to overview the knowledge of a given query keyword. Especially, we focus on concerns of those who search for Web pages with a given query keyword. The Web search information needs of a given query keyword is collected through search engine suggests. Given a query keyword, we collect up to around 1,000 suggests, while many of them are redundant. We cluster redundant search engine suggests based on a topic model. However, one limitation of the topic model based clustering of search engine suggests is that the granularity of the topics, i.e., the clusters of search engine suggests, is too coarse. In order to overcome the problem of the coarse-grained clusters of search engine suggests, this paper further applies the word embedding technique to the Web pages used during the training of the topic model, in addition to the text data of the whole Japanese version of Wikipedia. Then, we examine the word embedding based similarity between search engines suggests and further classify search engine suggests within a single topic into finer-grained sub-topics based on their word embedding based similarity.

1. はじめに

現代社会では，インターネット上に様々な情報が溢れている．ウェブ検索者が欲しい情報を手に入れる手段の一つとしては，Google 等の検索会社が提供している検索エンジンを利用するのが一般的である．ここで，検索会社は，検索行動支援の一環として，検索エンジン・サジェスト・サービスを提供している．このサービスの特徴として，入力された検索語に対して，関連が強い語を検索エンジン・サジェストとして提示する．ここで，本研究では，ウェブ検索者が詳細な情報を得たい対象を**「クエリ・フォーカス」**と呼ぶ．そして，AND 検索の形でクエリ・フォーカスの次に入力され，より詳細な情報を得たい観点を示す語を**「情報要求観点」**と呼ぶ (図 1)． 検索エンジン・サジェストは，ウェブ検索者の検索履歴の収集結果に基づいて作られている．このことから，検索エンジン・サジェストには，ウェブ検索者の関心事項が反映されていると言える．そこで，本研究では，検索エンジン・サジェストを情報収集の情報源として，ウェブ検索者の情報要求観点を収集する．

本研究の枠組みにおいて，一つのクエリ・フォーカスに対して，最大約 1,000 個のサジェストを収集する．そして，クエリ・フォーカスにサジェストを加えて，AND 検索によってウェブページの収集を行う．最大約 1,000 個のサジェストに対してこの方法を使うことによってウェブページの収集を行う．ここで，収集されるサジェスト，および，ウェブページ集合の双方において，話題の内容が重複し，かつ，冗長である点が問題である．そこで，まず，トピックモデル (具体的には，潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) [Blei03]) を適用して話題集約を行う．この手法では，収集されたウェブページ集合に対して LDA を適用することによって，話題ごとにウェブページのクラスタリングを行う．各ウェブページを検索する際には，サジェストが指定されているため，ウェブページごとに少なくとも一つのサジェストが対応付けられる．この対応関係を用いることにより，約 1,000 個のサジェストを数十個程のまとまりに集約する [井上 16].

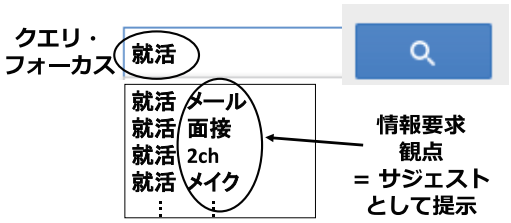


図 1: 検索エンジン・サジェストにおける情報要求観点の例

ここで，トピックモデルを用いた集約の問題点として，話題集約の粒度が粗い点が挙げられる．一方，近年，深層学習技術により，単語の意味表現を分散表現によって表す方式が提案され，その有効性が報告されている [Mikolov13]．この方式では，大規模なコーパスから語の意味のベクトル表現を学習し，これを用いて各語の周囲の語 (文脈) の予測を行う．そこで，本論文では，分散表現によって表現される話題の粒度が相対的に細かいところに着目し，分散表現とトピックモデルを併用することによって，検索エンジン・サジェストの類似度をより詳細に測定し，これを用いることによって，検索エンジン・サジェストの話題をより精密に集約する手法を提案する．

2. 検索エンジン・サジェストの収集

本研究では，評価対象となるクエリ・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・フォーカスにつき約 100 通りの文字列を指定する．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．そ

連絡先: 聶 添，筑波大学大学院システム情報工学研究科，
〒 305-8573 茨城県つくば市天王台 1-1-1, 029-853-5427

*1 https://www.google.com/

表 1: クエリ・フォーカス，サジェスト数，ウェブページ数，および，トピック数

クエリ・フォーカス	サジェスト数			ウェブページ数	総トピック数	評価対象トピック数	小分類数 / 評価対象トピック数
	ウェブページ収集時	Wikipediaのみで頻度下限以上	Wikipedia+ウェブページにおいて頻度下限以上				
就活	925	575	671	13,222 (30.8 MB)	50	28	2.4
結婚	947	637	694	14,413 (31.0 MB)	50	27	4.0

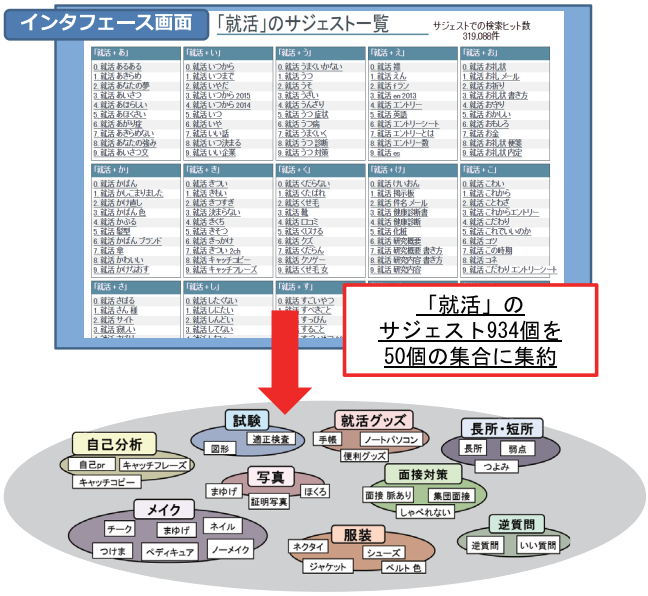


図 2: トピックモデルを用いた検索エンジン・サジェストの集約 (クエリ・フォーカス: 「就活」)

して，1 通りの文字列当り約 10 個のサジェストを収集することにより，一クエリ・フォーカスあたり最大約 1,000 個のサジェストを収集し，これを集合 S とする．例えば検索窓に「就活あ」と入力すると,「あいさつ」や「あなたの強み」等のサジェストが表示されるので，それらの収集を行う．クエリ・フォーカス「就活」および「結婚」に対して収集されたサジェストの数を表 1 「サジェスト数，ウェブページ収集時」欄に示す．

3. トピックモデルを用いた検索エンジン・サジェストの集約

3.1 サジェストを用いたウェブページの収集

収集したサジェストを用いてウェブページの収集を行う． クエリ・フォーカスにサジェスト s を加えて AND 検索することによって上位 N 件以内に検索されるウェブページ d の集合を $D(s,N)$(ただし，本研究では，$N=20$ とする) とする． ウェブページの収集では，Yahoo! Search BOSS API[*2]を用いる．また，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 を集めた集合を次式 $S(d)$ とする．

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

クエリ・フォーカス「就活」および「結婚」に対して収集したウェブページの数およびテキストサイズを表 1 に示す．集めたウェブページの集合を D とする．そして，この D にトピックモデルを適用し，トピックの推定を行い，推定されたトピック分布に従ってサジェストの集約を行う．

*2 http://developer.yahoo.com/search/boss

3.2 トピックモデル

本論文で適用するトピックモデルは，潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [Blei03] である．トピックモデルの推定においては，入力として，語 w の列によって表す文書集合，および，トピック数 K を指定する．出力として，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 $P(z_n|d)$ $(n=1,\ldots,K)$ を得る．LDA 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の α，β については，GibbsLDA++のデフォルト値である $\alpha = 50/K$，$\beta = 0.1$ を使い，Gibbs サンプリングの反復回数を 2,000 に設定した． 語 w の集合 V としては，日本語 Wikipedia 中のタイトルおよびそのリダイレクトの集合[*4]を用いた． また，トピック数 K としては，K を 10 から 80 まで変化させてトピック推定を行った上で，トピック推定による話題のまとまりが最もよいトピック数 K を人手で選定した．クエリ・フォーカス「就活」および「結婚」においては，$K=50$ を採用した．

3.3 文書に対するトピックの割り当て

各ウェブページに対して最も確率が高いトピック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集約する．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 $z_n(n=1,\ldots,K)$ のウェブページ集合 $D(z_n)$ は次式で表される．

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

3.4 トピックに対するサジェスト割り当てによるサジェストの集約

各ウェブページは，クエリ・フォーカスに一つのサジェストを加えて AND 検索することによって収集される．そのため，各ウェブページに対して一つ以上のサジェストが対応する．ここで，ウェブページ d にはサジェスト集合 $S(d)$ 中のサジェストが対応する．また，ウェブページ d には，トピック z_n が割り当てられている．すると，トピック z_n に対して割り当てられたウェブページ $d \in D(z_n)$ に対応するサジェスト s を集めることによって，トピック z_n に対してサジェスト s が割り当てられる．トピック z_n に割り当てられたサジェストの集合 $S(z_n)$ は次の式で表される．

$$S(z_n) \quad = \quad \bigcup_{d \in\ D(z_n)} S(d)$$

さらに，トピック z_n におけるサジェスト s の頻度 $f(s,z_n)$ は

*3 http://gibbslda.sourceforge.net/

*4 トピックモデルにおける語の集合として用いる日本語 Wikipedia としては，2014 年 3 月にダウンロードしたエントリ数約 140 万 7,000 のものを用いた．

下記の式で表される．

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

例えば，クエリ・フォーカス「就活」の場合，934 個のサジェストが 50 個のトピックに割り当てられた (図 2)．以上のように，検索エンジン・サジェストを用いて収集されたウェブページ集合に対してトピックモデルを適用することによって，検索エンジン・サジェストが集約される．

4. 分散表現を用いた検索エンジン・サジェストの類似度測定

本節では，前節において収集されたサジェストに対して word2vec [Mikolov13] を適用し，サジェストの分散表現を計算する．そして，得られた分散表現を用いて，同一トピックに集約されているサジェスト同士の類似度を測定することによって，検索エンジン・サジェストの話題をより精密に集約する．

4.1 分散表現

分散表現 [Mikolov13][*5] においては，訓練コーパスを単語列 $w_1, w_2, \ldots, w_T$ で表し，位置 t の単語 w_t の前後 δ 語の単語列を文脈窓 $C_{w_t} = (w_{t-\delta}, ..., w_{t-1}, w_{t+1}, ..., w_{t+\delta})$ とする．そして，文脈窓 C_{w_t} から単語 w_t を予測する条件付き確率分布関数 $p(w_t|C_{w_t})$ を定義し，次式の対数尤度 L を目的関数として，次式を最大化する分散表現を学習する．

$$L \quad = \quad \sum_{t=1}^{T} \log p(w_t|C_{w_t})$$

本論文では，word2vec における分散表現の実装の中でも，特に，skip-gram による実装を用いた場合について述べる．skip-gram では，条件付き確率分布関数 $p(w_t|C_{w_t})$ を，各文脈語 $c \in C_{w_t}$ から単語 w_t を予測する条件付き確率 $p(w_t|c)$ の積に分解する．

$$L^{SG} \quad = \quad \sum_{t=1}^{T} \sum_{c \in C_{w_t}} \log p(w_t|c)$$

そして，コーパス中の全語彙集合を V として，単語 c のベクトル v_c，および，単語 w を予測するベクトル $\tilde{v}_w$ の二種類のベクトルを導入し，さらに，ソフトマックス関数を用いることによって，単語 c から単語 w_t を予測する条件付き確率 $p(w_t|c)$ を次式で定式化する．

$$p(w_t|c) \quad = \quad \frac{\exp(v_c \cdot \tilde{v}_{w_t})}{\sum_{w' \in V} \exp(v_c \cdot \tilde{v}_{w'})}$$

ここで，上式の分母の計算においては，コーパス中の全語彙 V を用いるのではなく，無作為に選んだ K 個の疑似負例単語 $\tilde{w}'$ を用いる．さらに，シグモイド関数 σ を用いた近似により，次式を最大化する分散表現を求め，最終的に単語 c のベクトル v_c を得る．

$$L^{SG} \quad = \quad \sum_{t=1}^{T} \sum_{c \in C_{w_t}} (\log \sigma(v_c \cdot \tilde{v}_{w_t}) + \sum_{k=1}^{K} \log \sigma(-v_c \cdot \tilde{v}_{\tilde{w}'}))$$

*5 本節における分散表現の定式化は，[岡崎 16] にしたがう．

4.2 分散表現の類似度

前節の方式によって，語 w_a および w_b の分散表現 v_{w_a} および v_{w_b} を求めた後，分散表現であるベクトル間の余弦類似度を求め，これを分散表現間の類似度と定義する．

$$Sim(w_a, w_b) = \frac{v_{w_a} \cdot v_{w_b}}{| v_{w_a} || v_{w_b} |}$$

4.3 検索エンジン・サジェストの分散表現の作成手順

検索エンジン・サジェストの分散表現を求めるために，分散表現訓練用コーパスを用意する．検索エンジン・サジェストを含む標準的な日本語単語の典型的な用例コーパスとして，日本語 Wikipedia の全ページテキストを用いる[*6]．さらに，各クエリ・フォーカスに対して収集された検索エンジン・サジェスト特有の特性を反映した分散表現を得るために，各クエリ・フォーカスに対して収集されたウェブページを加え，これら二種類のテキストデータの混合集合を分散表現訓練用コーパスとして，検索エンジン・サジェストの分散表現を求める．また，ベースラインとして，日本語 Wikipedia の全ページテキストのみを用いて分散表現を訓練した場合との比較を行う．ここで，分散表現訓練時には，各サジェストが頻度下限値 (本論文では，頻度下限値を 5 とする) を満たす必要がある．分散表現訓練用コーパスとして，(i) Wikipeida のみを用いた場合，および，(ii) Wikipedia および各クエリ・フォーカスに対して収集されたウェブページの混合集合を用いた場合，の両方について，この頻度下限値を満たす検索エンジン・サジェストの数を表 1 に示す．なお，word2vec を適用する際には，分散表現の次元数を 300 とし，前後の文脈幅 $\delta = 15$ とする．

4.4 サジェスト間の類似度測定例

一例として，クエリ・フォーカス「就活」の場合について，トピックモデル推定結果におけるトピックのうち,「グループディスカッション，グループワーク，プレゼンテーション，プレゼン」等の検索エンジン・サジェストが割り当てられたトピックを対象として，一部の検索エンジン・サジェスト間で分散表現の類似度を測定した結果を図 3 に示す．ここで，分散表現間の類似度の下限値を 0.6 に設定した場合を想定すると,「グループディスカッション，グループワーク，プレゼンテーション，プレゼン」の四つのサジェストのうちの任意の二組のうち，類似度下限値の条件を満たすサジェスト組は,「グループディスカッション，グループワーク」および「プレゼンテーション，プレゼン」の二組のみとなることが分かる．主観評価においても，この二組のみが類似関係にあると言えるため，分散表現の類似度を用いることによって，主観評価の結果と一致する類似度が測定できていることが分かる．

4.5 評価

分散表現を用いたサジェスト間類似度の評価を行うために，「就活」および「結婚」の各クエリ・フォーカスを対象として，各トピック中の検索エンジン・サジェストに対して，人手で類似サジェストをまとめることにより，「検索エンジン・サジェストの参照用小分類」を作成した．その際，(i) サジェストの意味的まとまりがないトピック，および，(ii) 逆に全てのサジェストが均一的にまとまっており，より詳細な小分類を作ることが困難なトピック，は評価対象から除外した．その結果，表 1 の「評価対象トピック数」欄に示す数のトピックが評価対象となった．また，評価対象のトピック一つ当たりの小分類数の平均は，表 1 の「小分類数/評価対

*6 2016 年 2 月時点の日本語 Wikipedia 全 100 万ページ，約 4.8 億語，2.91GB を用いる．

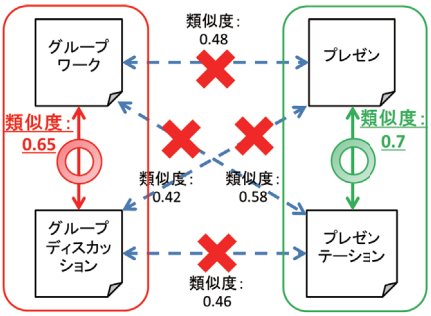


図 3: 分散表現を用いたサジェスト間の類似度の測定例 (クエリ・フォーカス: 「就活」, 類似度の下限値: 0.6 の場合)

象トピック数」欄に示す値となった.

次に, 二つのサジェストの組に対して,

1. サジェスト間の分散表現類似度の下限値,
2. あるサジェストに対して, 分散表現類似度の降順に他のサジェストを順位付けした場合の
   (a) 全サジェスト中の順位の上限値,
   (b) トピック内のサジェスト中の順位の上限値.

の三つの素性を設定する. そして, 三つの素性のあらゆる可能な組み合わせに対して,「検索エンジン・サジェストの参照用小分類」中のサジェスト組の集合を R, 三つの素性の組み合わせを満たすサジェスト組の集合を S_f として, 次式で算出される再現率・適合率をプロットする.

$$\text{再現率} = \frac{|R \cap S_f|}{|R|}, \quad \text{適合率} = \frac{|R \cap S_f|}{|S_f|}$$

ここで, 分散表現訓練用コーパスとして,

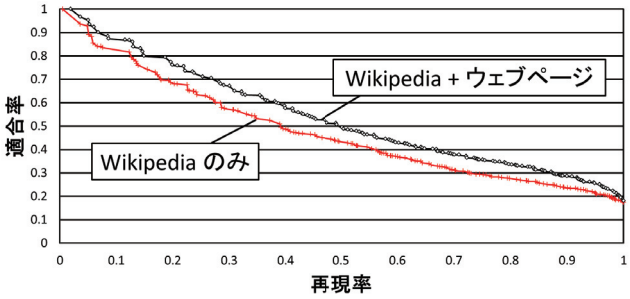
(i) Wikipeida のみを用いた場合,
(ii) Wikipedia および各クエリ・フォーカスに対して収集されたウェブページの混合集合を用いた場合,

の間でこのプロットを比較した結果を図 4 に示す. この結果から, クエリ・フォーカス「就活」および「結婚」の両方において, 分散表現訓練用コーパスとして (ii) を用いた場合の効果を示すことができた.

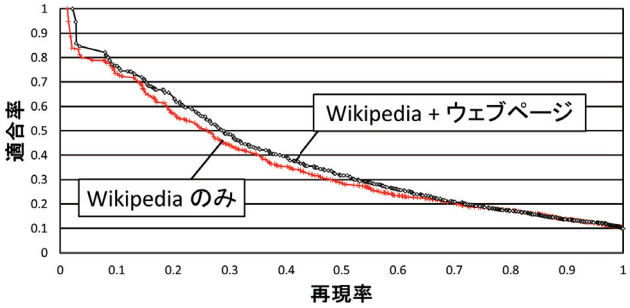
5. 関連研究

本論文に関連して, 従来型のトピックモデルにおいては, 語そのものに対する確率分布としてトピックモデルを推定するのに対して, [Li16] においては, 語に対して求めた分散表現に対する確率分布としてトピックモデルを推定する方式を提案している. 一方, 本論文では, 従来型のトピックモデルにおいても, 粒度の粗いトピックとしては十分良質なトピックが得られているという立場に立ち, その上で, 各トピックに対してより詳細かつ精密な話題集約を実現するために, 検索エンジン・サジェストの分散表現を利用する方式を提案した.

その他の関連研究としては, クリックスルーデータを用いて検索クエリのクラスタリングを行う手法 [Ma08,Guo10,Leung08] が挙げられる. [Ma08,Guo10] では, 大量のクリックスルーデータを用いて検索クエリのクラスタリングを行った後, 出力されたクラスタをベースにして検索クエリを推薦する手法を提案している. [Leung08] では, 検索クエリのクラスタリングをユーザ毎に行う手法を提案している. この研究では, ユーザの個人特徴に着目し, 各ユーザの趣味を考慮した検索クエリのクラスタリングを行っている. 評価実験では, 30 人程度の検索ユーザを対象として, 約 150 個の検索クエリについて, クラスタリング結果を評価している.



(a) クエリ・フォーカス: 「就活」



(b) クエリ・フォーカス: 「結婚」

図 4: 検索エンジン・サジェストの類似度の評価結果

6. おわりに

本論文では、検索エンジン・サジェストとウェブページの間の関連性に着目し, 文書集合にトピックモデルを適用することによって, サジェストの自動集約を行なった. また, トピックモデルによる話題集約の粒度は粗い, という問題点に対して, 分散表現によって表現された語の意味表現を併用することによって, より詳細かつ精密な話題集約手法を提案した.

参考文献

[Blei03] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[Guo10] Guo, J., Cheng, X., Xu, G. and Shen, H.-W.: A Structured Approach to Query Recommendation with Social Annotation Data, *Proc. 19th CIKM*, pp. 619–628 (2010).

[井上 16] 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).

[Leung08] Leung, K. W.-T., Ng, W. and Lee, D. L.: Personalized Concept-Based Clustering of Search Engine Queries, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 11, pp. 1505–1518 (2008).

[Li16] Li, X., Chi, J., Li, C., Ouyang, J. and Fu, B.: Integrating Topic Modeling with Word Embeddings by Mixtures of vMFs, *Proc. the 26th COLING*, pp. 151–160 (2016).

[Ma08] Ma, H., Yang, H., King, I. and Lyu, M. R.: Learning Latent Semantic Relations from Clickthrough Data for Query Suggestion, *Proc. 18th CIKM*, pp. 709–718 (2008).

[Mikolov13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Proc. NIPS*, pp. 3111–3119 (2013).

[岡崎 16] 岡崎直観：言語処理における分散表現学習のフロンティア, 人工知能学会誌, Vol. 31, No. 2, pp. 189–201 (2016).