DEIM Forum 2018 J7-4

トピックモデル・分散表現の併用による ウェブ検索結果話題集約におけるサブトピック化

丁　易† 趙　辰† 川畑　修人† 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　〒 305-8573 茨城県つくば市天王台 1-1-1
†† 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
††† (株) ログワークス　〒 151-0053 東京都渋谷区代々木 1-3-15 天翔代々木ビル 6F

あらまし　本論文では，トピックモデルによってウェブ検索結果の話題集約を行った結果に対して，各トピックによる話題集約の粒度が粗いことに着目する．そして，より粒度の細かい話題集約の単位として，各トピックをサブトピックに分割するアプローチをとる．具体的には，分散表現の尺度を用い，各ウェブページの類似度を計算することによって，各トピック内のウェブページ集合の話題集約の粒度をサブトピック単位へと詳細化する．さらに，本論文では，先行研究で設計したインタフェースの仕様をふまえて，サブトピック単位での話題集約結果を提示するインタフェースを設計・実装する．

キーワード　検索エンジン・サジェスト，トピックモデル，話題集約，分散表現

Subtopic Labeling in Web Search Results Aggregation by Integrating a Topic Model and Word Embeddings

Yi DING†, Chen ZHAO†, Shuto KAWABATA†, Takehito UTSURO††, and Yasuhide KAWADA†††

† Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba 305-8573 Japan
†† Faculty of Engineering, Information and Systems, University of Tsukuba,　Tsukuba 305-8573 Japan
††† Logworks Co., Ltd.　Tokyo 151-0053, Japan

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．そして，クエリ・フォーカスに対して，より詳細な情報を得るために指定する語を**「情報要求観点」**と呼ぶ (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ウェブ検索者の情報要求観点の収集を行う．[5] においては,「情報要求観点」に着目して，ウェブ検索結果に対する集約を行った．[5] においては,「クエリ・フォーカス」に関する情報要求観点を用いてウェブページの収集を行い，収集されたウェブページに対してトピックモデルを適用することによって話題の集約を行った．これに対して，本論文では，トピックモデルによってウェブ検索結果の話題集約を行った結果に対して，各トピックによる話題集約の粒度が粗いことに着目する．そして，より粒度の細かい話題集約の単位として，各トピックをサブトピックに分割するアプローチをとる．

具体的には，本論文では，まず，[5] と同様に，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．次に，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り，ウェブページ集合の話題集約を行う．さらに，[13] の手法に基づき，word2vec [10]，および doc2vec [8] の二つの分散表現の尺度を用い，各ウェブページの類似度を計算することによって，各トピック内のウェブページ集合の話題集約の粒度をサブトピック単位へと詳細化する．ここで，word2vec のモデル訓練時においては，訓練用テキストとして，日本語 Wikipedia エントリの本文テキストに加えて，クエリ・フォーカスごとに，収集結果のウェブページのテキストを追加することによって，各クエリ・フォーカスに特化した分散表現を訓練した．そして，人手で作成した参照用サブトピックを用いた評価実験によって，分散表現を用いたサブトピック単位への話題集約の性能評価を行った．また，各サブトピックごとに，収集されたウェブページ集合のテキストに対して複数文書要約手法である LexRank [3] を適用することによっ

て，各サブトピックの内容を要約する．

さらに，本論文では，[5] におけるウェブ検索結果の話題集約インタフェースの仕様をふまえて，サブトピック単位での話題集約結果を提示するインタフェースを設計・実装する．具体的には，[5] におけるウェブ検索結果の話題集約インタフェースにおいては，各トピックごとに，検索結果のウェブページの URL およびスニペットを表示する仕様となっている．一方，本論文におけるサブトピック単位での話題集約結果提示インタフェースにおいては，トピック粒度の提示だけでなく，各サブトピックことのウェブページ情報，および，各サブトピックにおける複数文書要約結果を提示する機能を実装した．これらの機能によって，ウェブ検索者の関心事項を効率的に俯瞰することを実現した．

2. 検索エンジン・サジェストを用いたウェブページの収集

2.1 検索エンジン・サジェスト

検索エンジン会社は，過去の検索者による検索ログを用いて，利用者が入力した検索語と強い関連を持つ語をサジェストとして提示する．過去の検索者の検索ログに基づいて提供されることから，検索エンジン・サジェストには，検索者の関心そのものが反応されていると考られる．そこで，本論文では，検索者の関心事項を収集するために検索エンジン・サジェストを利用する．特に，本論文では，四種類のクエリ・フォーカス「就活」，「結婚」,「花粉症」，および,「マンション」を対象として，検索エンジン・サジェストを収集・分析する．

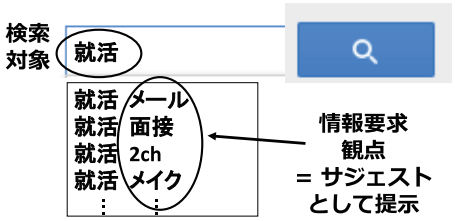


図 1　検索エンジン・サジェストにおける情報要求観点の例

2.2 サジェストの収集

本論文では，Google 社の検索エンジン・サジェストを用いて，クエリ・フォーカスに対する情報要求観点を収集する．一つのクエリ・フォーカスあたり 100 通りの文字列を指定する．一文字列あたり最大 10 個のサジェストを収集できるため，理論上，最大 1，000 語のサジェストを収集することができる．100 通りの文字列としては，五十音，濁音，半濁音および開拗音を含む．この手順により収集したサジェストの数を表 1 に示す．

2.3 ウェブページの収集

ウェブページの収集においては，Google Custom Search API [(注1)] を用いて，検索語として,「クエリ・フォーカス AND サジェスト」の AND 条件での検索語を指定して，ウェブページを収集する．収集されたウェブページの数を表 1 に示す．

(注1)：https://cse.google.com/cse/

3. トピックモデルを用いたウェブ検索結果の集約

3.1 トピックモデル

検索エンジン・サジェストを用いることにより大規模なウェブページ集合を収集することができるが，サジェストおよびそれを用いて収集されるウェブページには重複する話題が多く含まれる．そのため，収集されたサジェストおよびウェブページ集合の集約・俯瞰が不可欠となる．そこで，本論文では，トピックモデルとして，潜在的ディリクレ配分法 [1](LDA; Latent Dirichlet Allocation) を用いる．LDA を適用することにより，語 w の列で表現された文書の集合，および，トピック数 K を入力することによって，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_n)$ $(w \in V)$，および，各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ $(n = 1, \ldots, K)$ を得ることができる．これらを推定するためのツールとしては，GibbsLDA++[(注2)] 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，Gibbs サンプリングの反復回数は 2,000 とした．また，本論文においては，語 w の集合 V として日本語 Wikipedia 中のタイトル，および，そのリダイレクトの集合 [(注3)] を用いた．

3.2 ウェブページに対するトピックの割り当て

ウェブページ集合を D，人手で指定したトピック数を K として，次式によって，トピック z_n に対してウェブページ集合 $D(z_n)$ を割り当てる．

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

各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ $(n = 1, \ldots, K)$ に基づき，確率 $P(z_n|d)$ が最大となるトピックにウェブページ d を割り当てる．また，人手でトピック数 K を指定する際には，各トピックにおける話題のまとまりが最もよいトピック数 K を決める．具体的には，クエリ・フォーカス「結婚」におけるトピック数 K を 60 とし，その他のクエリ・フォーカスにおけるトピック数 K を 50 とした．

3.3 トピックに対するサジェストの割り当て

ウェブページは，クエリ・フォーカスおよびサジェストの AND 検索によって収集されるため，各ウェブページに対して，一つ以上のサジェストが対応付けられている．そこで，各トピック z_n に対応付けられたウェブページ集合 $D(z_n)$ 中の各ウェブページに対応付けられたサジェストを収集することにより，各トピックに対してサジェスト集合を割り当てることができる．

4. ウェブ検索結果の集約粒度の詳細化：サブトピックへの話題集約

前節までの手順によって, 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，ウェブページ収集結果の話題の集約を行った．ここで，各トピッ

(注2)：http://gibbslda.sourceforge.net/

(注3)：日本語 Wikipedia には，2014 年 3 月にダウンロードを行った総記事数約 89 万 7,000 ページのものを用いた．

表 1　評価対象のサジェストおよびウェブページの数

クエリ・フォーカス	結婚	就活	花粉症	マンション
サジェスト数	959	926	850	958
日本語 Wikipedia のみを用いて分散表現を訓練した場合に，分散表現を持つサジェストの数	709	559	632	684
日本語 Wikipedia およびクエリ・フォーカスに関するウェブページを用いて分散表現を訓練した場合に，分散表現を持つサジェストの数	771	671	695	758
収集したウェブページの数	13,256	12,078	9,745	13,742

クにおける話題集約結果の問題点として，集約された話題の粒度が粗いことが挙げられる．そこで，本論文では，[5] におけるウェブ検索結果の話題集約インタフェースにおいて機能を高度化するアプローチとして，トピック単位の粒度の粗い話題集約結果を，さらにサブトピックに分割する方式を提案する．また，各サブトピックの単位においても，依然として数十から数百のウェブページが収集された状態であり，各サブトピックの単位での情報集約という点においては不十分である．そこで，本論文では，各サブトピック内のウェブページ集合に対して，複数文書要約手法を適用することによって，情報の俯瞰効率を大幅に改善するアプローチを提案する．

以上をふまえて，本節では，ウェブページ集合に対してトピックモデルを適用した結果を対象として，より粒度の細かいサブトピックへの話題分割を行う手順を定式化する．この手順においては，トピック z_n に割り当てられたウェブページ集合 $D(z_n)$ は，互いに素な部分集合へと分割され，これらの各部分集合が，各サブトピックに対して分割されたウェブページ集合に対応する．

$$D(z_n) = D_1(z_n) \cup D_2(z_n) \cup \ldots \cup D_m(z_n)$$
$$D_i(z_n) \cap D_j(z_n) = \emptyset \quad (i \neq j)$$

ここで，各サブトピックに対応するウェブページ集合は，下限値 θ_{lbd} 以上の類似度を持つウェブページ組を集めて構成される．

$$D_i(z_n) = \left\{ d \in D(z_n) \,\middle|\, \exists d' \in D_i(z_n), sim(v(d), v(d')) \geqq \theta_{lbd} \right\}$$

5. 文書間類似度尺度

ウェブページ集合において，話題を高精度に集約する過程を実現するためには，まず，文書間の類似度を導入することが不可欠である．この課題に対して，本論文では，[13] にしたがい，分散表現に基づくアプローチとして，word2vec [10] および doc2vec [8] を採用する．

5.1 サジェストの分散表現に基づく類似度尺度

ウェブページ収集結果の文書集合に対して word2vec [10] を適用した結果において，出現頻度が下限値以上となり分散表現を持つ検索エンジン・サジェストの数は，表 1 に示す通りである．このうち，各ウェブページ d に対して，以下の条件を満たすサジェスト $s(d)$ を一つ設定する．

(a)　$s(d)$ は，ウェブページ d を検索する際に指定されたサジェストの一つである．

(b)　(a) を満たすサジェストのうち，ウェブページ d が割り当てられたトピックにおける頻度が最大となるサジェストが $s(d)$ である．

なお，サジェストのトピックにおける頻度は次式によって定義する．

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

そして，ウェブページ d および d' に対して，サジェスト $s(d)$ および $s(d')$ の分散表現の間の類似度

$$sim(v(d), v(d')) = \frac{v(s(d)) \cdot v(s(d'))}{\| v(s(d)) \| \| v(s(d')) \|}$$

によって，ウェブページ d と d' の間の類似度尺度を定義する．

5.1.1 Wikipedia のみで訓練したモデル

word2vec のモデル訓練において用いる日本語テキストとして，日本語 Wikipedia の全エントリ本文テキストを用いた(注4)．このテキストに対して，品詞体系・形態素辞書として IPAdic(注5) を用いて，形態素解析ツールとして MeCab(注6) を用いて形態素単位への分割を行った後，word2vec の訓練を行った．ただし，表 1 に示す各サジェストについては，形態素解析結果における形態素の単位を超えて，複数の形態素を結合した複合語の単位で一つの分散表現を訓練するという調整を行った．結果的に，分散表現を持つサジェストの数を表 1 に示す．

5.1.2 Wikipedia および収集されたウェブページを用いて訓練したモデル

日本語 Wikipedia の全文テキストデータにおいては，日本語の典型的な用例は含まれているが，各クエリ・フォーカスに密接に関連する用語が高頻度で出現することは難しい．そこで，各クエリ・フォーカスごとに，2.3 節で収集したウェブページ集合を訓練データに追加したうえで，word2vec のモデルの訓練を行う．表 1 に示すように，各クエリ・フォーカスごとのウェブページ数は約 1 万前後で，およそ 30MB のサイズとなる．本節の手順で訓練した word2vec のモデルにおいては，前節のモデルと比べて，より多くのサジェストに対して分散表現が生成された．

(注4)：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．
(注5)：http://sourceforge.jp/projects/ipadic/
(注6)：http://taku910.github.io/mecab/

5.2 doc2vec に基づく類似度尺度

前節の word2vec に基づく尺度はウェブページと対応しているサジェスト間の word2vec 類似度のみを用いる手法である．しかし，ウェブページとサジェストの間の対応付けが不正確な場合も起こり得る．そこで，ウェブページに対応付けられたサジェストは用いず，文書自体をベクトル化して，文書間の類似度を測定する方式として doc2vec [8] を用いる．本論文では特に，Distributed Memory Model of Paragraph Vectors (PV-DM) のオプションを用いてウェブページのベクトルを生成する．そして，次式によりウェブページのベクトル間の類似度を測定する．

$$sim(v(d), v(d')) = \frac{p(d) \cdot p(d')}{\| p(d) \| \| p(d') \|}$$

6. 評　　価

6.1 評 価 方 法

表 1 に示す四種類のクエリ・フォーカスについて，各 10 個のトピック，合計 40 個のトピックを評価対象とする．各トピックにおいて，ウェブページ d におけるトピック z_n の確率 $P(z_n|d)$ の降順に上位 30 個のウェブページを評価対象とし，各トピックにおける 30 個のウェブページに対して，参照用サブトピックを人手で作成し評価を行った．評価においては，-1 から 1 まで 0.02 刻みで類似度下限値 θ_{lbd} を変化させて，次式の再現率，適合率のマクロ平均，マイクロ平均を測定した結果をプロットした．

$$\text{再現率} = \frac{\begin{array}{c}\text{出力された各サブトピックに含まれる} \\ \text{ウェブページ組のうち，参照用サブトピック} \\ \text{に含まれるウェブページ組数の和}\end{array}}{\begin{array}{c}\text{参照用サブトピックに含まれる} \\ \text{ウェブページ組数の和}\end{array}}$$

$$\text{適合率} = \frac{\begin{array}{c}\text{出力されたサブトピックに含まれる} \\ \text{ウェブページ組のうち，参照用サブトピック} \\ \text{に含まれるウェブページ組数の和}\end{array}}{\begin{array}{c}\text{出力された各サブトピックに} \\ \text{含まれるウェブページ組数の和}\end{array}}$$

6.2 評 価 結 果

評価対象の四種類のクエリ・フォーカスにおける評価結果をそれぞれ，図 2，図 3，図 4，および，図 5 に示す．全体的な傾向として，word2vec の方が doc2vec よりも高い性能を示しており，word2vec の中では，日本語 Wikipedia に加えて，クエリ・フォーカスごとに収集したウェブページを追加して訓練された word2vec モデルの方が高い性能を示した．これらの結果から，ウェブページのテキスト情報を直接用いて分散表現を訓練する doc2vec 方式よりも，各ウェブページに割り当てられたサジェストの分散表現を用いる word2vec 方式の方が，ウェブページ集合におけるノイズの影響を受けにくいことが推測される．また，word2vec 方式においては，日本語 Wikipedia に加えて，クエリ・フォーカスごとに収集したウェブページを追加して訓練した word2vec モデルの方が，日本語 Wikipedia のみで訓練した word2vec モデルよりも，分散表現を持つサジェストの数が多いため，より高い性能を示したと考えられる．

7. サブトピックにおける複数文書要約

前節までの手順によって，各サブトピックに分割されたウェブページ集合に対して，複数文書要約技術を適用することによって，各サブトピックに対応する要約を生成し，ウェブページ集約結果インタフェースにおいて提示する．本論文では，特に，複数文書要約手法として，文間の関連ネットワークに対して PageRank アルゴリズム [4] を適用する文書要約手法 [3]LexRank を用いて，各サブトピックごとの複数文書要約を生成し，インタフェースにおいて提示する．ツールとしては，Summpy(注7) を用いた．

8. ウェブ検索結果の集約提示インタフェース

本論文は，[5] におけるウェブ検索結果の集約粒度が粗いという問題に着目し，ウェブ検索結果をサブトピック単位に分割して提示するインタフェースを設計し実装した．

8.1 トピックレベルの粒度の提示

本論文のウェブ検索結果集約提示インタフェースの機能の模式図を図 6 に示す．図 6 の左半分は，[5] におけるトピックレベルでのウェブ検索結果集約提示インタフェースの機能を流用した仕様となっている．この仕様においては，トピックごとのサジェスト，各サジェストの頻度，および，各トピックのウェブページ，つまり，トピックレベルの粒度のウェブ検索結果の集約結果を提示している．図 6 の例では，「就活」における「研究概要」や「学校推薦」等に関するトピックおよびそのサジェストを表示している．このインタフェースにおいては，各トピックをクリックすると，そのトピックに関するサジェストが表示される．各サジェストの部分をクリックすると．そのトピックにおいて，各サジェストによって検索されたウェブページが表示される．

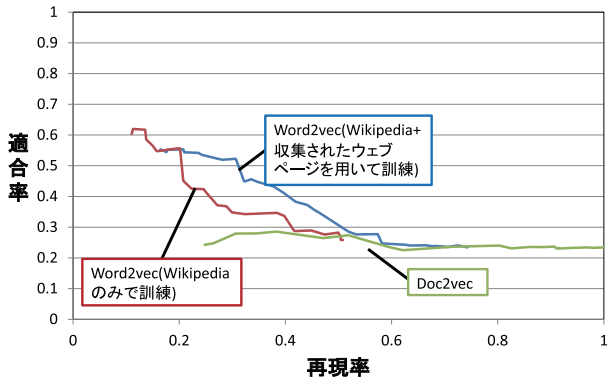
8.2 サブトピックレベルの粒度の提示

次に，図 6 の右半分において，本論文のウェブ検索結果集約提示インタフェースにおけるサブトピックレベルでのウェブ検索結果集約提示インタフェースの機能を示す．

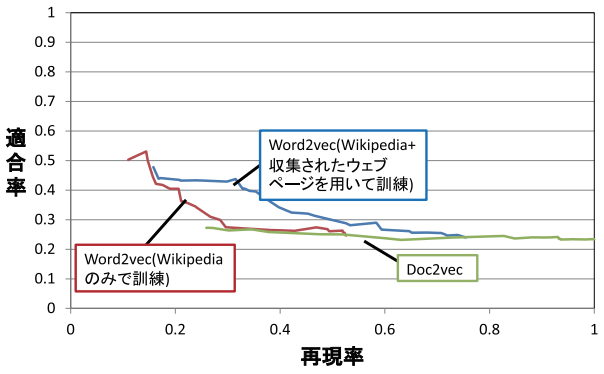
一般に，各トピックに含まれるウェブページの数は数百ページとなり，膨大であるため，[5] のインタフェースでは，各トピック内のウェブページを部分的に選定し，表示している．これに対して，図 6 の右半分に示す本論文のインタフェースにおいては，サブトピックごとにウェブページを表示する．具体的には，左側の各トピックをクリックすると，図 6 の右半分に示すように，当トピックに所属するサブトピックおよび各サブトピックのウェブページが表示される．

図 6 の例では，クエリ・フォーカス「就活」における，「研究概要」，「学校推薦」等に関するトピック，および，そのサブトピックの内容を示している．具体的には，左側のインタフェースのトピック部分をクリックすると，「学校推薦」，「研究内容」等のサブトピックごとに，右側の各サブトピック内のウェブページが表示される．

(注7)：https://github.com/recruit-tech/summpy

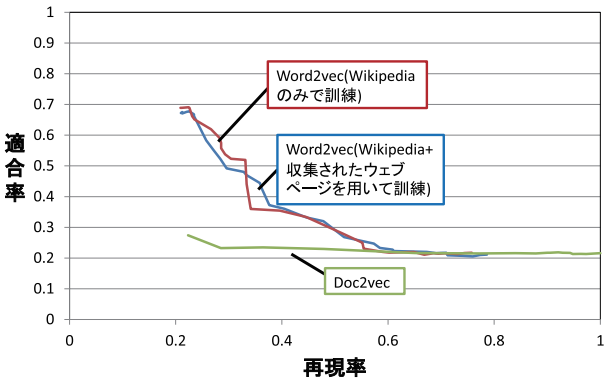


(a) マクロ平均

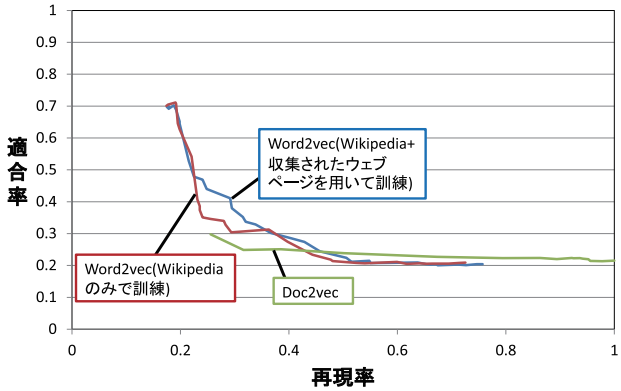


(b) マイクロ平均

図 2　クエリ・フォーカス「就活」の評価結果

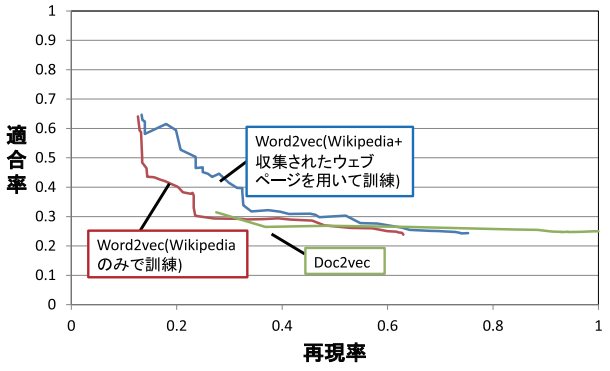


(a) マクロ平均

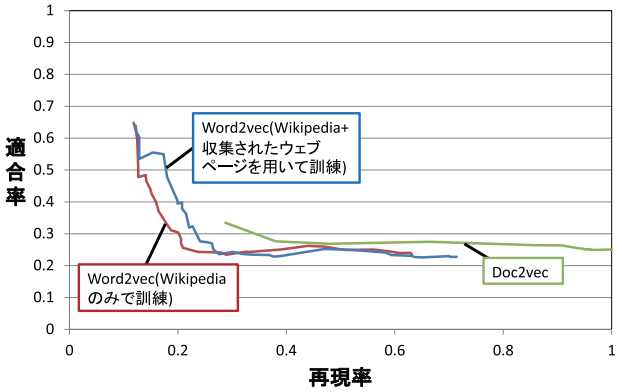


(b) マイクロ平均

図 3　クエリ・フォーカス「結婚」の評価結果



(a) マクロ平均



(b) マイクロ平均

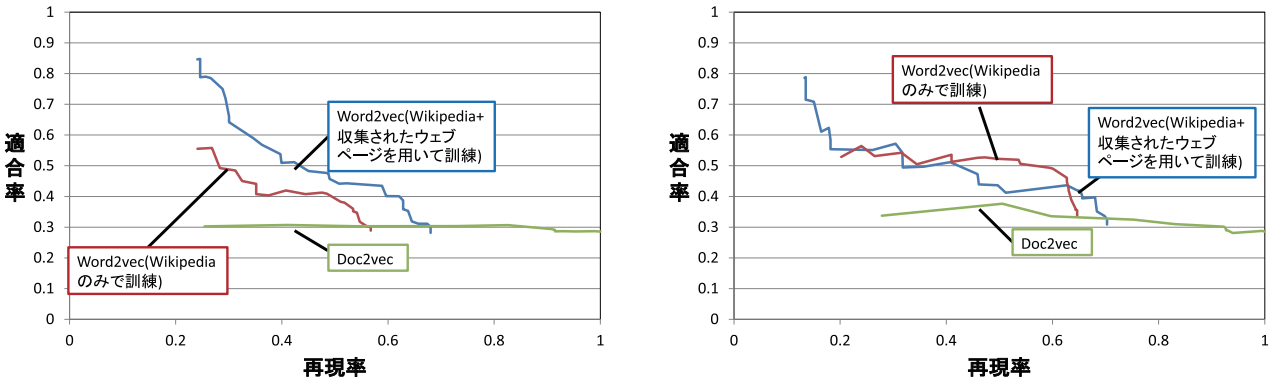
図 4　クエリ・フォーカス「花粉症」の評価結果

8.3　複数文書要約の提示

本論文のウェブ検索結果集約提示インタフェースにおける複数文書要約機能の模式図を図 7 の右半分に示す．要約機能においては，図 7 の左半分のトピックに対して，そのトピックにおける「学校推薦」,「研究概要」等のサブトピックごとに別々に，図 7 の右半分に示す要約を提示する．

9.　関 連 研 究

本論文の基盤となる研究として，[5] においては，収集したウェブページに対して LDA を適用し，各トピック中のウェブページのうちで代表的なものを選択して提示する方式によってウェブ検索結果の集約を実現している．一方，[5] の後段の研究の一つである [11, 12] においては，トピックモデルおよび分散表現を併用し，本論文において対象とした検索結果のウェブ



(a) マクロ平均　　(b) マイクロ平均

図 5　クエリ・フォーカス「マンション」の評価結果

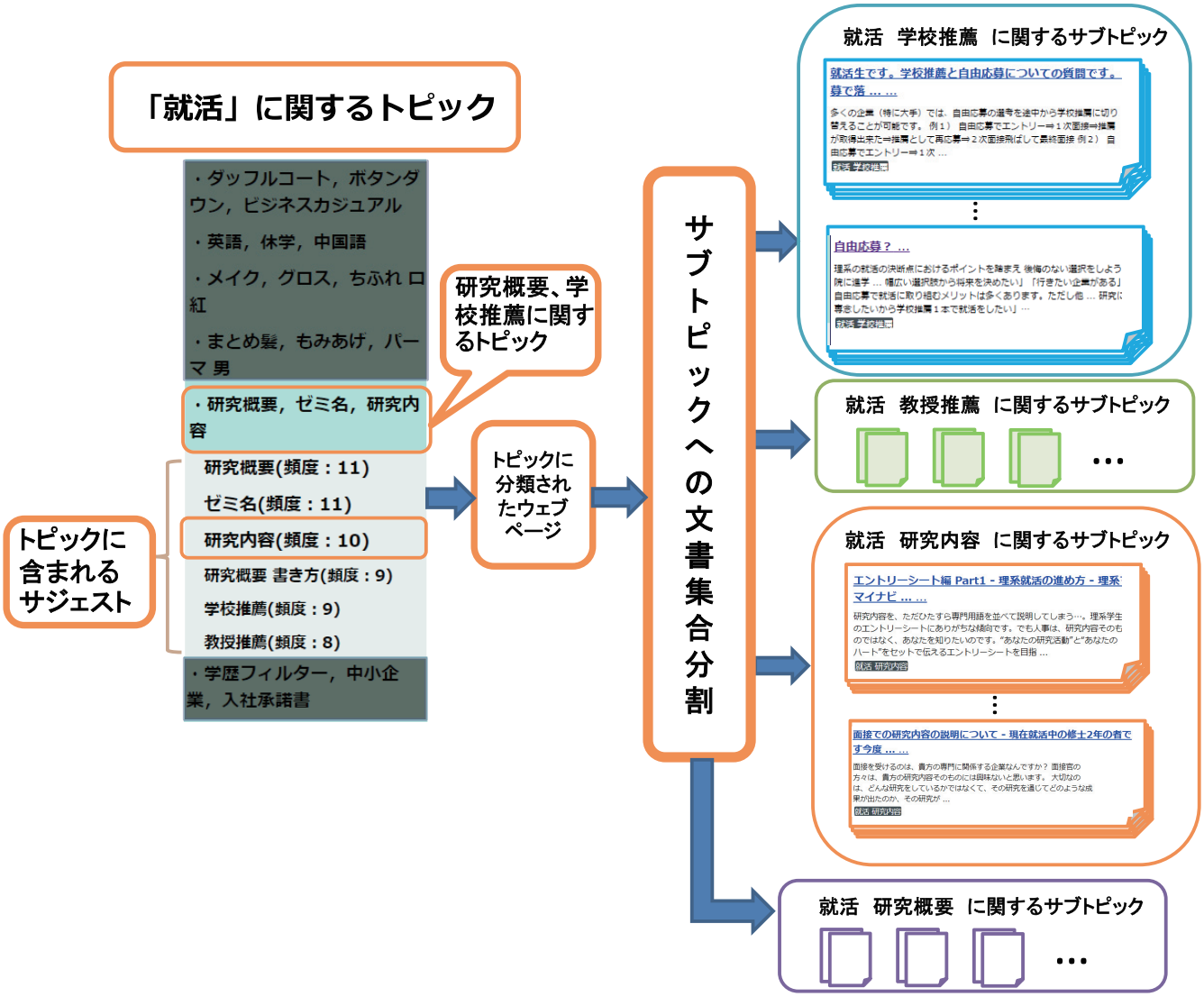


図 6　ウェブ検索結果の集約提示インタフェースにおけるサブトピックへの文書集合分割機能

ページではなく，検索エンジン・サジェストの集約における粒度を詳細化する手法を提案している．一方，本論文の前身となる [13] においては，LDA を適用した集約結果に対し，本論文で行ったサブトピックラベル付けの結果をふまえて，同一サブトピックとなるウェブページ数が 3 以上となるサブトピックを主要なサブトピックとみなして，トピック中のウェブページに対して，主要サブトピックか否かに二分するタスクを対象として，本論文と同様の分散表現を用いる手法を提案している．[13] において対象としているタスクは，本論文のタスクの粒度を粗くしたものに相当する．

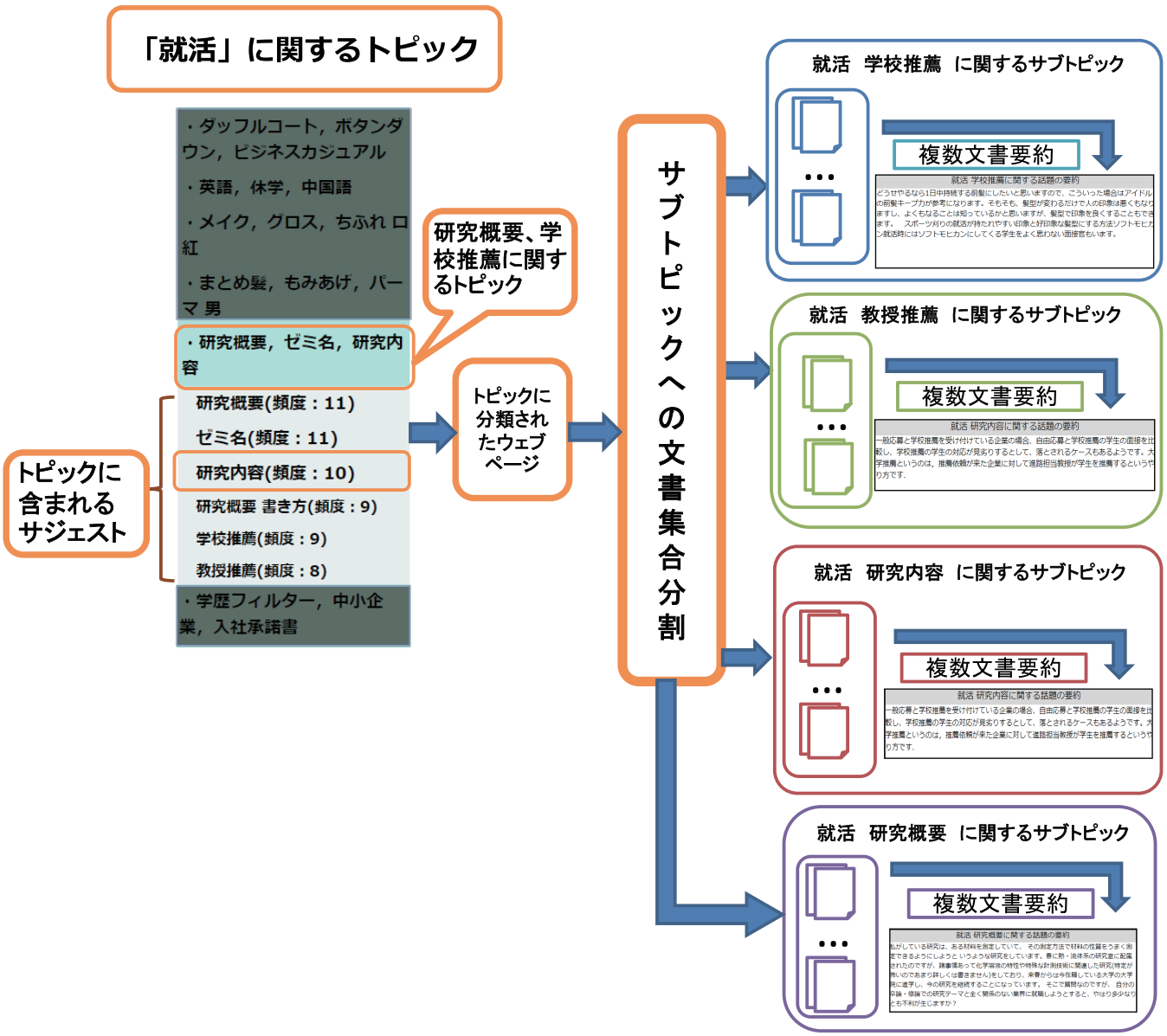


図 7　ウェブ検索結果の集約提示インタフェースにおける複数文書要約機能

また，[9] においては，語に対して求めた分散表現の上の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，語の上の確率分布として推定される従来型のトピックモデルとの比較を行なっている．一方，本論文では，従来型のトピックモデルに対してより粒度の細かいサブトピックを得るためのアプローチとして，検索エンジン・サジェストの分散表現の間の類似度を用いてウェブページ間の類似度を測定し，粒度の粗いトピックをサブトピックに分割する方式を提案した．

さらに，本論文では，各サブトピックの粒度に対して，LexRank [3] の複数文書要約手法を適用することによって，各サブトピック単位での要約結果を提示するインタフェースを設計・実装した．ここで，本論文における要約対象はウェブページ集合であるが，本論文で適用した要約手法は，ウェブページの HTML 文書ではなく，テキストで書かれた文書を対象とする要約手法である．したがって，今後は，ウェブページの HTML タグ等を考慮したウェブページ要約手法 (例えば，[6,7]) の適用を検討する必要があると考えられる．その他，ウェブページに対してリンク元のウェブページの内容を利用した要約手法 [2] 等も提案されているので，これらの適用についても検討が必要であると考えられる．

10. おわりに

本論文では，トピックモデルによってウェブ検索結果の話題集約を行った結果に対して，各トピックによる話題集約の粒度が粗いことに着目した．そして，より粒度の細かい話題集約の単位として，各トピックをサブトピックに分割するアプローチをとった．具体的には，分散表現の尺度を用い，各ウェブページの類似度を計算することによって，各トピック内のウェブページ集合の話題集約の粒度をサブトピック単位へと詳細化した．さらに，本論文では，先行研究で設計したインタフェースの仕様をふまえて，サブトピック単位での話題集約結果を提示するインタフェースを設計・実装した．

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] J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi. Enhanced Web document summarization using hyperlinks. In *Proc. 14th HYPERTEXT*, pp. 208–215, 2003.

[3] G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, Vol. 22, pp. 457–479, 2004.

[4] S. Grin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 107–117, 1998.

[5] 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.

[6] 井山晃洋, 砂山渡, 谷内田正彦. HTML テキストの文章化による Web ページ要約. 第 16 回人工知能学会全国大会論文集, 2002.

[7] 清田陽司, 黒橋禎夫. WWW テキストの自動要約と KWIC インデックスの作成. 情報処理学会研究報告, Vol. 2000–NL–137, pp. 31–38, 2000.

[8] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proc. 31st ICML*, pp. 1188–1196, 2014.

[9] X. Li, J. Chi, C. Li, J. Ouyang, and B. Fu. Integrating topic modeling with word embeddings by mixtures of vMFs. In *Proc. the 26th COLING*, pp. 151–160, 2016.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. 26th NIPS*, pp. 3111–3119, 2013.

[11] T. Nie, Y. Ding, C. Zhao, Y. Lin, T. Utsuro, and Y. Kawada. Clustering search engine suggests by integrating a topic model and word embeddings. In *Proc. 18th SNPD*, pp. 581–586, 2017.

[12] 聶添, 丁易, 趙辰, 李佳奇, 宇津呂武仁, 河田容英. 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. 第 31 回人工知能学会全国大会論文集, 2017.

[13] C. Zhao, T. Nie, Y. Ding, J. Li, T. Utsuro, and Y. Kawada. Identifying major documents with search engine suggests by unsupervised subtopic labeling. 第 31 回人工知能学会全国大会論文集, 2017.