2H3-2in

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

Aggregating Search Engine Suggests based on a Topic Model

土井 俊弥[*1] 井上 祐輔[*1] 今田 貴和[*1] 宇津呂 武仁[*2] 河田容英[*3] 神門 典子[*4]
Syunya Doi Yusuke Inoue Takakazu Imada Takehito Utsuro Yasuhide Kawada Noriko Kando

[*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]筑波大学システム情報系 Fclty. Eng, Inf. & Sys, Univ. of Tsukuba
[*3](株) ログワークス Logworks Co., Ltd.
[*4]国立情報学研究所 National Institute of Informatics

In this paper, we address the issue of how to overview the knowledge of a given query keyword. We especially focus on concerns of those who search for Web pages with a given query keyword, and study how to efficiently overview the whole list of Web search information needs of a given query keyword. First, we collect Web search information needs of a given query keyword through search engine suggests. Although we collect up to around 1,000 suggests given a query keyword, some of them are redundant in that they originate from almost the same Web search information needs. In order to aggregate such redundant search engine suggests, we take an approach of clustering search engine suggests based on a topic model. We also develop an interface system for overviewing those aggregated search engine suggests of a given query keyword as well as links to top ranked Web pages that are closely related to those aggregated search engine suggests.

1. はじめに

近年のインターネットの普及に伴い，多くの人がウェブページ上から情報を得ている．情報を収集する手段としては，Google 等の検索エンジンを用いてウェブ検索を行うのが一般的である．各検索エンジン会社においては，検索者が入力した検索語のログが蓄積されており，多数の検索者が検索した検索語に対して，強い関連を持つ語を検索エンジン・サジェストとして提示するシステムを提供している．ここで，本論文では，検索者が詳細な情報を検索したい対象を「**検索対象**」と呼ぶ．また，検索対象に対して，より詳細な情報を得るために，AND 検索の形で二つ目以降に入力する語を「**情報要求観点**」と呼ぶ (図 1).

ここで，検索エンジン・サジェストの形で表現された情報要求観点においては，ウェブ検索者の関心事項そのものが反映されていると考えられる．そこで，本論文では，検索エンジン・サジェストに着目し，それらはウェブ検索者の関心事項であるとみなして，検索エンジン・サジェストを収集したものを集約・俯瞰することを目的とする．

本論文においては，まず，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を収集する．具体的には，一つの検索対象に対して，最大約 1,000 語のサジェストを収集する．ただし，収集されるサジェストの多くは話題が重複し冗長である．これを改善するために，冗長性を考慮してサジェストの集約を行う．特に，本論文では，トピックモデルの一つである潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) [Blei 03b] を用いて話題集約を行う手法を提案する．具体的には，まず一つの検索対象あたり最大約 1,000 語のサジェストを収集する．その後，各サジェストを用いて収集されるウェブページ集合に対して，LDA を適用してウェブページ集合をトピックと呼ばれる話題ごとのまとまりに集約するとともに，各ウェブページを収集する際に用いられたサジェストの集約を行う．その結果，約 1,000 語あったサジェストを数十個程度のまとまりに集約することができる．

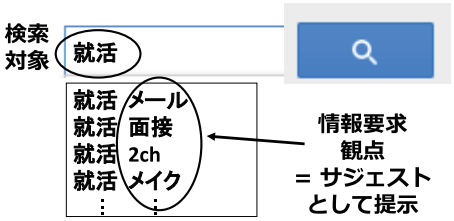

図 1: 検索エンジン・サジェストにおける情報要求観点の例

提案手法を用いることにより，サジェストが示す話題を考慮し，類似する話題ごとに集約してサジェストを提示することが可能となる．閲覧者が検索対象に関する前提知識をほとんど持たない場合には，より詳細な情報を得るための情報要求観点を自身で思いつくことが難しい．しかし，本論文の手法によって提示されるサジェストの集約結果を参照することにより，検索対象に関して収集された膨大な数の情報要求観点を容易に俯瞰することができ，情報を効率よく収集することができる．本論文では，以上の考え方に基づき，集約したサジェストをトピックごとに一覧で提示し，閲覧者があるトピックを選択すると，そのトピックに分類されたサジェストと関連性の強いウェブページの一覧を提示するインタフェース (図 4 参照) を作成し，その有効性を示す．

2. 検索エンジン・サジェストの収集

評価用検索対象 (本論文では,「就活」,「結婚」を検索対象とする) に対して，Google[*1] 検索エンジンを用いて，一検索対象当たり約 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「就活 あ」と入力すると,「あいさつ」や「あなたの強み」等がサジェストとして提示されるので，それらの

連絡先: 土井 俊弥 , 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: 各検索対象のサジェスト数，および，ウェブページ数

検索対象	サジェスト数	ウェブページ数
就活	934	13,221
結婚	989	14,413

収集を行う．検索対象毎に得られたサジェストの数を表 1 に示す．

3. 検索エンジン・サジェストの集約

3.1 概要

本節では，トピックモデルを適用することにより，前節において収集したサジェストを自動的に集約し，トピックと呼ばれる話題毎にまとめる．

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 20 件のウェブページを収集する．ここでの検索クエリは，各検索対象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．各検索対象ごとに得られたウェブページ数を表 1 に示す．収集されたウェブページの集合を D として，D を対象としてトピックモデルを適用することによってトピックを推定する．そして，推定されたトピックを用いることによって，サジェストの集約を行う．

3.2 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) [Blei 03b] 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，$K = 50$ を採用した．

3.3 文書に対するトピックの割り当て

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．

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

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．

3.4 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，各検索対象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/

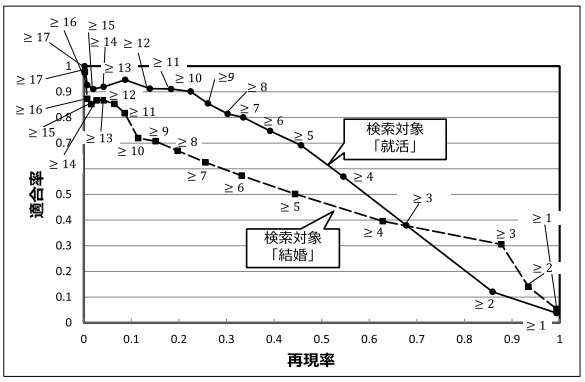

図 3: 検索エンジン・サジェストの集約： 評価結果 (検索対象「就活」,「結婚」におけるミクロ平均．サジェストの頻度の下限値を変化させた場合)

実際に，検索対象「就活」の場合，934 個のサジェストが 50 個のいずれかに割り当てられた (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．以上の手順より出力された集約結果の例の一部を表 2 に示す．

3.5 評価

特定のサジェストがトピックにおいて割り当てられている文書の数を，トピックにおけるサジェストの頻度として定義する．そして，こ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サジェストがそのトピックに属しているとみなして評価を行う．参照用に作成したトピックを用いて，頻度の下限値ごとに，次式の再現率，適合率を算出し，プロットした結果を図 3 に示す．

$$\text{再現率} \quad = \quad \frac{\begin{array}{c}\text{出力された各トピックに}\\ \text{含まれるサジェスト組のうち，}\\ \text{参照用トピックに含まれるサジェスト組数の和}\end{array}}{\text{参照用トピックに含まれるサジェスト組数の和}}$$

$$\text{適合率} \quad = \quad \frac{\begin{array}{c}\text{出力されたトピックに}\\ \text{含まれるサジェスト組のうち，}\\ \text{参照用トピックに含まれるサジェスト組数の和}\end{array}}{\begin{array}{c}\text{出力された各トピックに}\\ \text{含まれるサジェスト組数の和}\end{array}}$$

4. 検索エンジン・サジェストおよびウェブ検索結果の集約・俯瞰インタフェース

本論文のインタフェースにおいては，各サジェストをトピックに集約し，各トピック内のサジェストをリスト形式で閲覧する仕様とした．これにより，閲覧者は，話題が類似するサジェストをまとめて俯瞰することができるようになり，この機能によってサジェストの俯瞰を実現した．また，図 4 に示すように，収集されたウェブページについても，話題が重複するウェブページを集約した上で，トピックに分類されたサジェストとの関連性の強いウェブページを一覧で提示した重複する冗長なウェブページをスキップするとともに，話題が関連するウェブページを集約的にまとめて提示することによって，ウェブ検索結果の俯瞰を実現した．

表 2: 提案手法による検索エンジン・サジェストの集約結果の例

検索対象	人手によりトピックに付与したラベル	トピックに割り当てられたサジェスト (各トピック 10 サジェストを抜粋)
就活	髪型	“ヘアスタイル 女”，“くせ毛 女”，“写真 髪型”，ロングヘア，まとめ髪，おだんご，ゆるいパーマ，襟足，美容院，シュシュ
	身に着けるもの	ネクタイ，シューズ，“ベルト 色”，かばん，ピーコート，シャツ，“パンプス おすすめ”，“グレー スーツ”，“ジャケット ボタン”，防寒
	グループディスカッション	グループワークとは，“グループディスカッション テーマ”，グループディスカッション，グループワーク対策，評価基準，評価，“プレゼン 資料，グループワーク，能力，プレゼン
	自己分析	“長所 真面目”，長所，座右の銘，どうなりたいか，あなたの夢，将来の夢，どんな人，こだわり，なりたい自分，軸
	恋愛との両立	“恋愛 両立”，ふられた，恋愛，寂しい，脈あり，結婚，“うまくいかない 彼氏”，“プレゼント 彼女”，わがまま，プレッシャー
	メイク	ノーメイク，ビューラー，チーク，化粧，つけま，まつエク，ネイル、まゆげ，“証明写真 メイク “，ペディキュア
結婚	お祝い，メッセージ	“友人 スピーチ”，“お祝い メッセージ”，ぬいぐるみ電報，祝辞，“電報 バルーン”，一言メッセージ，“祝電 文例”，“ぬいぐるみ メッセージ”，ビデオメッセージ，“文例 電報 友人”
	条件，決めて	“妥協 顔”，ルックス，見極め，美人，理想，価値観，“男性 条件”，“決め手 女性”，容姿，相手
	求める収入	高望み，条件，“条件 年収”，収入，平均年収，高望み，ランキング職業，“条件 ランキング”，年収，求めるもの
	結婚祝い	プレゼント，“ぬいぐるみ 手作り”，“ぬいぐるみ うさぎ”，印鑑，プチギフト，祝い，贈り物，ペアウォッチ，サプライズ，寄せ書き
	手続き	“入籍 手続き”，住所変更，“苗字 変更”，必要書類，パスポート，住民票，“会社 手続き”，外国人，名義変更，グリーンカード
	写真	“写真 東京”，前撮り，写真，ポーズ，“写真 大阪”，“写真 札幌”，写真だけ，ビデオ，和装，ビデオメッセージ

5. 関連研究

本論文において提案した手法に関連して，ウェブページの検索結果を分類し，各分類に対して適切な要約文を付与する手法 [原島 10]，検索された個々の Web ページに対してラベルの付与を行い，付与されたラベルに基づいて分類を行う手法 [戸田 05,de Winter 07,馬場 09]，階層的なトピックの体系を推定する手法 [Blei 03a] 等の手法が提案されている．これらの手法においては，いずれも，閲覧対象の文書集合のみを用いて，ファセット体系およびファセットラベルに相当する情報を抽出している．また，メタ検索エンジンにおいてウェブページ検索結果の上位 200 記事程度を対象にして，検索結果のクラスタリングおよびラベル付けをした結果を提示するサービスとして，Yippy*4 が知られている．これらの先行研究においては，いずれも，与えられた文書集合における話題の広がりを俯瞰することに焦点が当てられている．

その他，[小池 14] においては，本論文の枠組みにおいて，トピックモデルを用いて検索エンジン・サジェストの集約を行うのではなく，各サジェストを用いた検索によって収集されるウェブページのスニペットをサジェストに付与し，これをクラスタリングすることにより，冗長なサジェストを集約する方式を提案している．

6. おわりに

本論文では，ウェブ検索者の関心事項の収集手段として検索エンジン・サジェストを用いた．ここで，多数のウェブ検索者が共通の事項について調べるためにウェブ検索を行う場合でも，それぞれ異なる表記のキーワードで検索を行う場合も多いため，冗長なサジェストが多数存在する．そこで，本論文では，冗長なサジェストを話題ごとに集約し，ウェブ検索者の関心事項の全体像を俯瞰する枠組みを提案した．特に，本論文では，トピックモデルの一つである潜在的ディリクレ配分法を用いて話題集約をする手法を提案し，その有効性を示した．

*4 http://yippy.com/

参考文献

[馬場 09] 馬場 康夫, 黒橋 禎夫：キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰, 情報処理学会論文誌, Vol. 50, No. 4, pp. 1399–1409 (2009)

[Blei 03a] Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B.: Hierarchical Topic Models and the Nested Chinese Restaurant Process, in *NIPS'03* (2003)

[Blei 03b] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[de Winter 07] de Winter, W. and de Rijke, M.: Identifying Facets in Query-Biased Sets of Blog Posts, in *Proc. ICWSM*, pp. 251–254 (2007)

[原島 10] 原島 純, 黒橋 禎夫：PLSI を用いたウェブ検索結果の要約, 言語処理学会第 16 回年次大会論文集, pp. 118–121 (2010)

[小池 14] 小池 大地, 鄭 立儀, 今田 貴和, 守谷 一朗, 井上 祐輔, 宇津呂 武仁, 河田 容英, 神門 典子：ウェブ検索者の情報要求観点の集約, 言語処理学会第 20 回年次大会論文集, pp. 328–331 (2014)

[戸田 05] 戸田 浩之, 中渡瀬 秀一, 片岡 良治：特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案, 情報処理学会論文誌：データベース, Vol. 46, No. SIG 13(TOD 27), pp. 40–52 (2005)

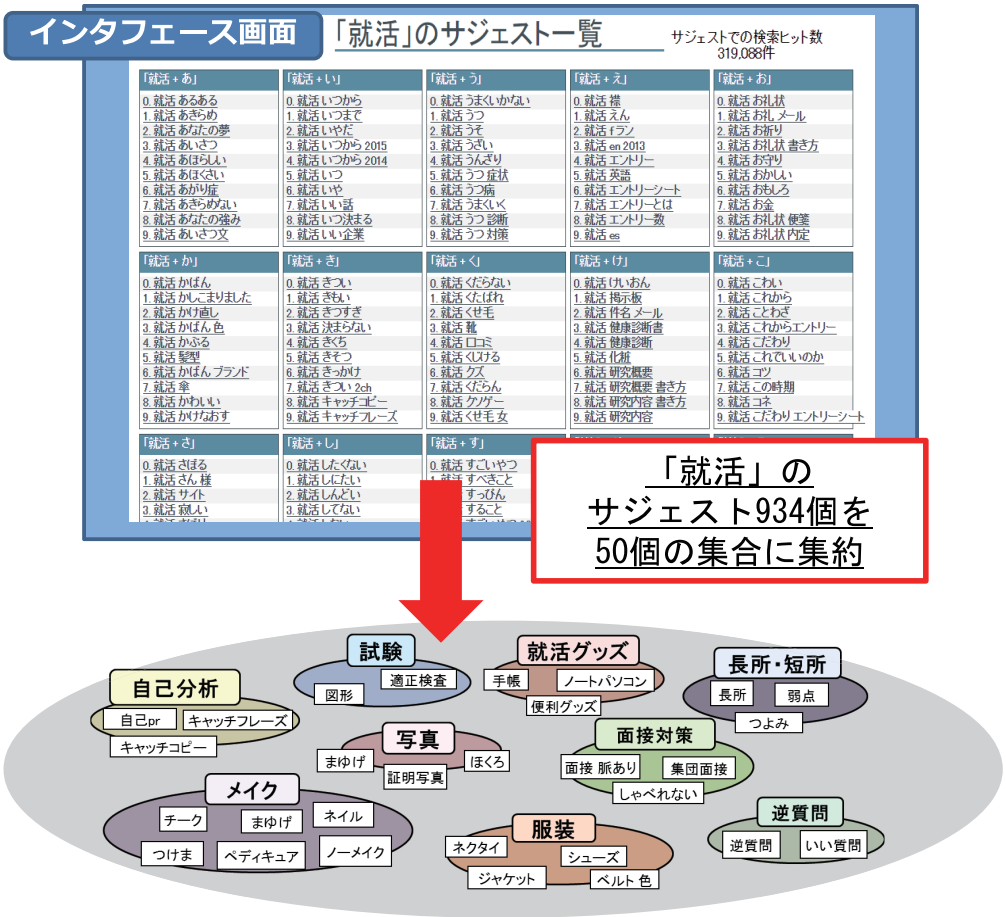

図 2: 検索エンジン・サジェストの集約 (検索対象: 「就活」)

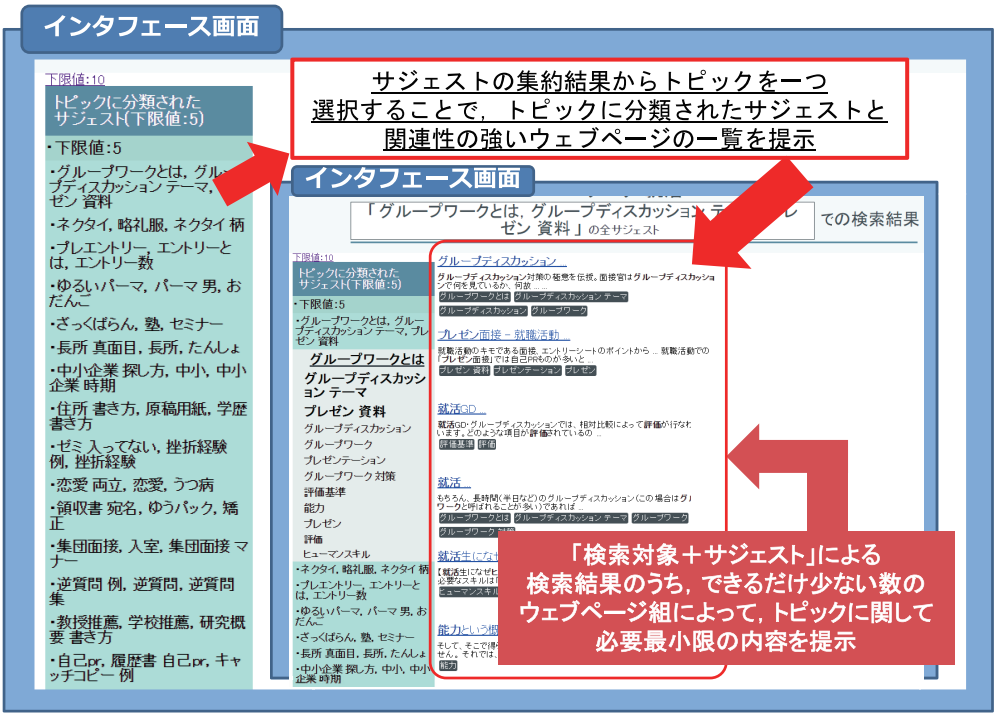

図 4: インタフェース画面: ウェブ検索結果の俯瞰 (検索対象: 「就活」)