2H5-2in2

サジェストおよびトピックモデルを用いた多様な話題のウェブページの選択的収集

Collecting Web Pages of Diverse Contents with Search Engine Suggests and a Topic Model

聶 添[*1*2] Tian Nie　徐 凌寒[*1] Linghan Xu　趙 辰[*1] Chen Zhao　宇津呂 武仁[*3] Takehito Utsuro　河田容英[*4] 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

[*3](株) ログワークス Logworks Co., Ltd.

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. Evaluation result shows that the proposed clustering approach proves to be quite useful for efficiently overviewing Web search information needs of a given query keyword. 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. Finally, we show the effectiveness of the interface in terms of the aggregation of Web search results.

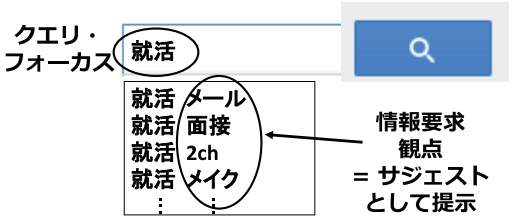

図 1: 検索エンジン・サジェストにおける情報要求観点の例

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
マンション	951	14,695
花粉症	872	11,144
3D プリンタ	763	7,586

1. はじめに

現代の情報社会においては，インターネットの普及により，ウェブ上に膨大な量の情報が溢れている．このような膨大な量の情報の中から，ユーザが求める情報を見つけ出すための手段としては，Google 等の検索エンジンの利用が一般的である．検索エンジン会社はユーザの検索行動支援のため，検索エンジン・サジェストというサービスを提供している．このサービスにおいては，検索者が入力した検索語のログを蓄積し，それらを用いて強い関連を持つ語が検索エンジン・サジェストとして提供されている．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「**クエリ・フォーカス**」と呼ぶ．そして，それに対してより詳細な情報を得るために，どのような側面に着目するのかを表す部分，すなわち，クエリ・フォーカスと 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)．検索エンジン・サジェストは検索者のログに基づいて作られているため，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・サジェストをウェブ検索者の関心事項であると見なし，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．

本論文の枠組み [井上 16] においては，一つのクエリ・フォーカスに対して，最大約 1,000 語のサジェストを収集する．そして，クエリ・フォーカスに加えて一つの検索エンジン・サジェストを指定した AND 検索によってウェブページを収集する．最大約 1,000 個の検索エンジン・サジェストに対してこの方法を用いることにより，あるクエリ・フォーカスに関する大規模なウェブページ集合を収集することが出来る．しかし，収集されるサジェスト，および，それらを用いて収集されるウェブページ集合では，多くは話題が重複しており冗長である．そこで本論文では，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検索者の情報要求観点を集約・俯瞰することを目的とする．

特に，本論文では，トピックモデルの一種である潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) [Blei 03b] を用いた話題集約の手法を提案する．本論文で提案する手法においては，まず，一つのクエリ・フォーカスあたり最大約 1,000 語のサジェストを収集し，それらサジェストを用いてウェブページの収集を行う．そして収集されたウェブページ集合に対して，LDA を適用しトピックと呼ばれる話題のまとまりごとにウェブページのクラスタリングを行う．各ウェブページはサジェストを用いて収集されたものであるため，各ウェブペー

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

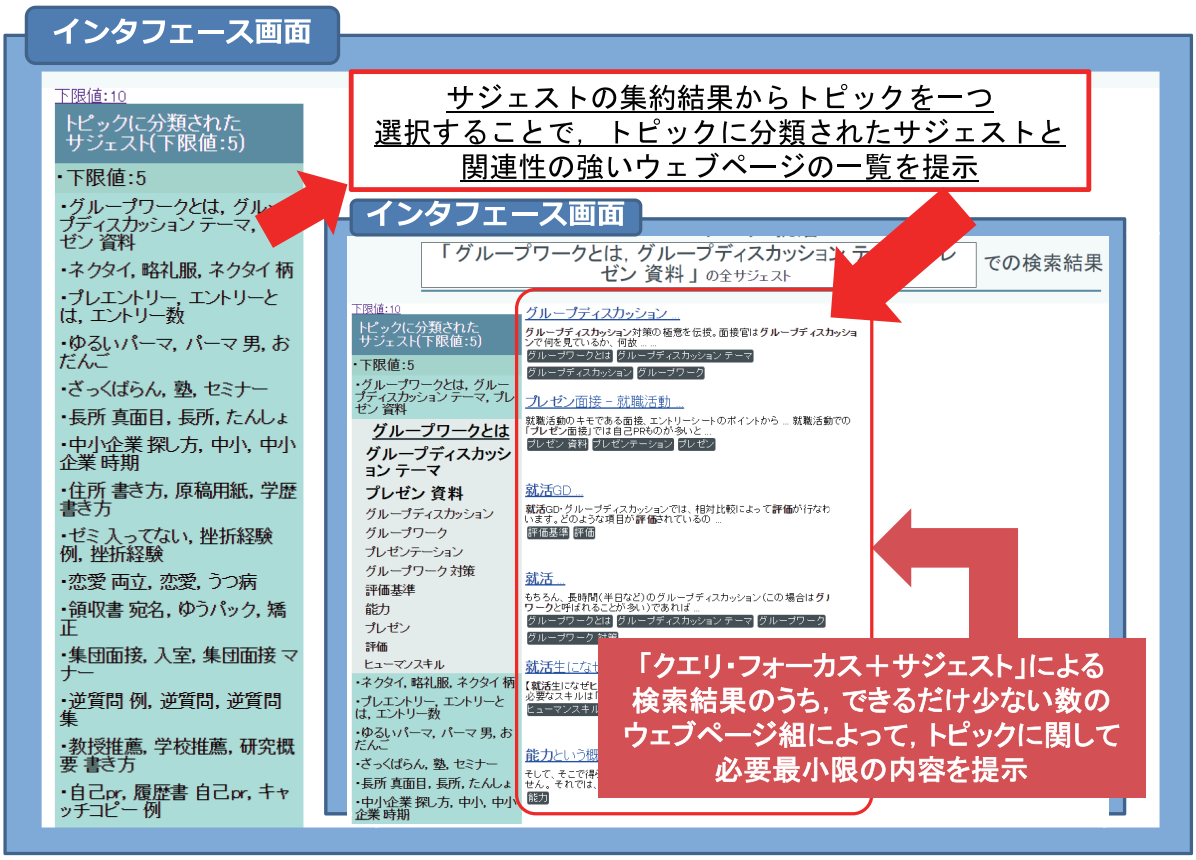

図 2: ウェブ検索結果の俯瞰インタフェース画面 (クエリ・フォーカス: 「就活」)

ジには最低一つ以上のサジェストを対応付けることが出来る．この対応付けによりサジェストの集約を行う．これにより，約 1,000 語あったサジェストを数十個程のまとまりへと集約することが出来る．

ここで，各トピックにおいてサジェストを集約した結果においては，互いに類似するサジェストを用いてウェブページが収集されているため，相互に類似する冗長なウェブページが多数収集されているのが現状である．これらのウェブページ集合を効率よく俯瞰するためには，冗長性を無くしてできるだけ多様な話題を示すウェブページ集合へと集約した上で閲覧する必要がある．そこで，本論文では，各トピック中のサジェストを用いて，できるだけ多様なサジェストを含むウェブページを選択的に提示する手法を提案する．また，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，ユーザ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そのトピックにおける選定されたウェブページの一覧を提示するインタフェース (図 2 参照) の作成を行う．以上の手順のうち，本論文では特に，ウェブ検索結果の集約におけるウェブページの選定に関して評価を行い，その有効性を示す．

2. 多様な話題のウェブページの選択的収集

2.1 概要

収集したサジェスト全てをそのまま一覧で提示した場合，全体でいくつの話題の情報要求観点が提示されているかを俯瞰することは困難である．また，サジェストを用いて検索を行う際には，話題が重複する冗長なサジェストを指定した検索を繰り返し行なうなどの非効率的な検索を余儀なくされることが予測され，できるだけ多様な話題の情報を効率よく収集する場合には大きな障害となる．この問題を解決するために，本論文のインタフェースにおいては，各サジェストをクラスタに集約し，各クラスタ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，この機能によって情報要求観点の俯瞰を実現した．また，図 2 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，クラスタ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．これにより，話題が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．

2.2 手順

本節では，トピックに属するサジェストを用いて収集されるウェブページ集合において，冗長性を集約しつつも出来るだけ多様な話題を表すようなウェブページ集合の選定方法について述べる．

クエリ・フォーカスに加えてサジェスト s を指定した AND 検索によって上位 N 件以内に検索されるウェブページ集合 $D(s, N)$ において，ウェブページ d の検索順位を $rank(d, s)$ とする．ここで，本論文の提案手法におけるウェブページ選定のある段階において，既に選定済みのウェブページ集合を D_r，未選定のウェブページ集合を D_{nr} とする．

$$D_{nr} = \left(\bigcup_{s \in S} D(s, N) \right) - D_r$$

また，選定済みのウェブページ集合 D_r の各ウェブページ d に対応付けられているサジェスト s の集合 $S(d)$ の和集合を S_r

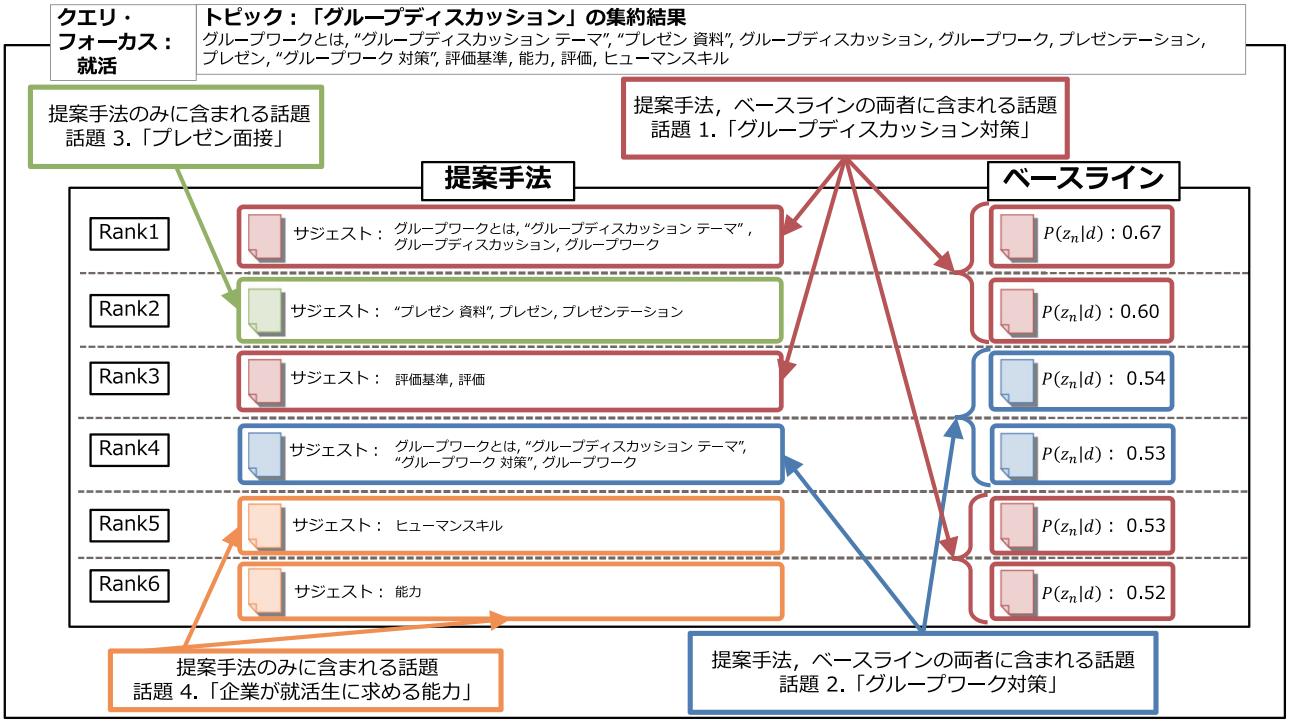

図 3: ウェブ検索結果の集約の例 (クエリ・フォーカス: 「就活」, トピック: 「グループディスカッション」)

として，それら以外のサジェストの集合を S_{nr} とする．

$$S_r = \bigcup_{d \in D_r} S(d)$$
$$S_{nr} = S - S_r$$

冗長性を集約しつつも出来るだけ多様な話題を表すようなウェブページ集合の選定するために，各ウェブページ d に対して，S_{nr} に中のサジェストのうち，出来るだけ多くのものに対応付けられ，検索された際の順位が高いほど，小さくなるようなコストを次式により定義し，ウェブページ選定の各段階においてこのコストが最小となるウェブページを順に選定する貪欲法によって，ウェブページの選定を行う．

$$cost(d, D_r) = \sum_{s \in S} r(d, D_r)$$

$$r(d, D_r) = \begin{cases} rank(d, s) & (s \notin S_r \text{ かつ } d \in D(s, N) \text{ の場合}) \\ N + 1 & (\text{それ以外の場合}) \end{cases}$$

D_r の初期値を ϕ とし，$S_{nr} = \phi$ となるまで以下の手順を行う．

(1) $cost(d, D_r)$ が最小のウェブページ $\hat{d}$ を選択する．

$$\hat{d} = \underset{d \in D_{nr}}{\operatorname{argmin}}\, cost(d, D_r)$$

(2) 集合 D_r を以下の式によって更新する．

$$D_r \leftarrow D_r \bigcup \{\hat{d}\}$$

作成したインタフェース画面の例を図 2 に示す．作成したインタフェースにおいては，以上の方法により選定されたウェブページの一覧をリスト形式で表示する．また，選定されたウェブページ d に対し，対応するサジェスト $s \in S(d)$ をタグ情報として付与し，ウェブページの情報とともに表示する．提案手法により，話題が重複する冗長なサジェストは単一のウェブページに付与されるため，ユーザはその単一のウェブページを見ることで，冗長なサジェストを把握できる．次節にて，以上の方法により選定されたウェブページに対する評価を行う．

3. 評価

ウェブ検索結果の集約に関する評価においては，表 1 に示す 5 つのクエリ・フォーカスの各々において，トピックを無作為に 5 つ選択し，合計 25 トピックを評価の対象とした．集約されたウェブページに対して，各ウェブページが示す話題を人手で分析することにより，集約されたウェブページ集合に含まれる話題数を，提案手法とベースライン手法との間で比較した．ここで，各トピックにおける話題分析の際には，提案手法によって選定されるウェブページの数を $|D_r|$ とすると，ベースライン手法においても，確率値 $P(z_n|d)$ の降順でウェブページ $d \in D(z_n)$ を順位付けし，順位付けの上位より $|D_r|$ と同数のウェブページを選定し分析対象とした．

3.1 例

提案手法による集約結果とベースライン手法による集約結果の比較を行った際の例の一部を図 3 に示す．図 3 では，クエリ・フォーカス「就活」のトピック「グループディスカッション」におけるウェブ検索結果の集約結果の比較を示している．図の左半分では，提案手法による集約結果を示す．この例においては，提案手法により選定されたウェブページの数は 6 件であり，それらには合計 4 個の話題が含まれていた．選定された 6 件のウェブページのうち，2 件は同一の話題「グループディスカッション対策」のページであり，また，他の 2 件も同一の話題「企業が就活生に求める能力」のページであった．残りの 2 件のウェブページはそれぞれ「プレゼン面接」,「グルー

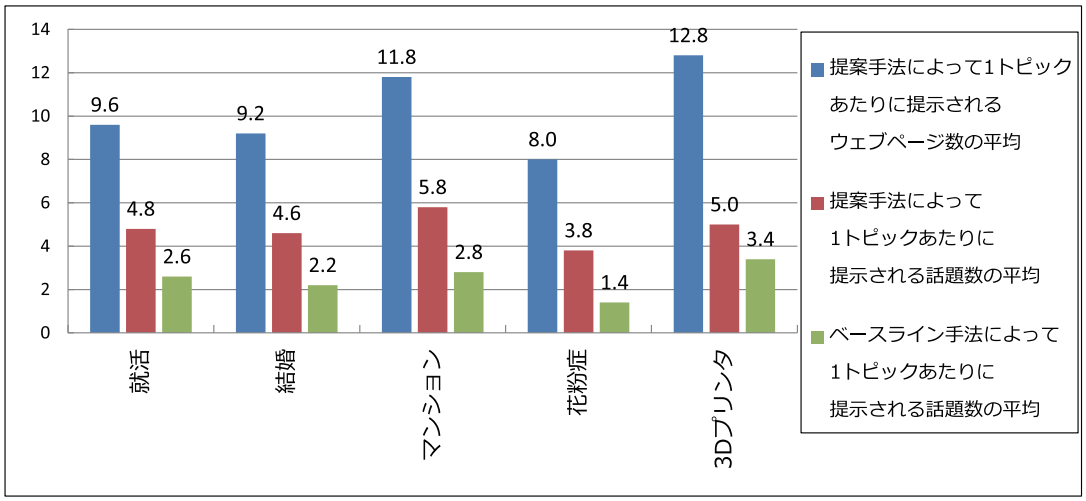

図 4: ウェブ検索結果の集約の評価

プワーク対策」という異なる話題のページであった．

一方，図の右半分では，ベースライン手法による集約結果を示す．ベースライン手法では，トピック z_n におけるウェブページ記事集合 $D(z_n)$ において，確率値 $P(z_n|d)$ の降順でウェブページ $d \in D(z_n)$ のランキングを行った．また，そのランキングのうち，上位 N(N は提案手法により選定されたウェブページの件数を表す．この例においては $N = 6$ となる) 件をベースライン手法における集約結果とした．ベースライン手法では，「グループディスカッション対策」，「グループワーク対策」の 2 個の話題のみが含まれていた．このように，提案手法によるウェブページの集約では，ベースライン手法に比べ，より少ない数のウェブページで多様な話題を得ることができた．

3.2 評価結果

次に，表 1 に示す 5 つのクエリ・フォーカスを対象として，提案手法によりウェブ検索結果を集約の評価を行った結果を図 4 に示す．この結果においては，

- 提案手法によって 1 トピックあたりに提示されるウェブページ数および話題数の平均
- ベースライン手法によって 1 トピックあたりに提示される話題数の平均

を比較した結果を示す．この結果から，ベースライン手法における集約結果と比較すると，提案手法による集約によって約 2 倍の数の話題が提示されることがわかる．

4. 関連研究

本論文に関連して，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．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[*1] が知られている．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．一方，本論文の提案手法においては，閲覧対象の文書集合からラベルを抽出するのではなく，その文書集合に対して検索を行った検索者が情報要求観点として指定した語をラベルとして用いており，この点において関連研究の手法とは大きく異なっている．

*1 http://yippy.com/

5. おわりに

本論文では，ウェブ検索者の関心事項に着目し，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し，集約を行った．特に，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集合に対してトピックモデルを適用し，ウェブページのクラスタリングを行うことによって，ウェブページに対応付けられたサジェストの集約を行った後 [井上 16]，各トピックに対応して収集されたウェブ検索結果に対して，多様なサジェストを含むウェブページを選択的に提示することによって，ウェブ検索結果を集約し，多様な話題のウェブページを選択的に提示できることを示した．

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* (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)

[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)

[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)