DEIM Forum 2016 C6-3

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

井上 祐輔† 今田 貴和† 陳 磊† 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 〒 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，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，集約を行う．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う．さらに，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す．

キーワード 検索エンジン・サジェスト, トピックモデル, 収集・集約, 情報要求観点, クラスタリング

Web Search Results Aggregation based on Search Engine Suggests and a Topic Model

Yusuke INOUE†, Takakazu IMADA†, Lei CHEN†, Linghan XU†, 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

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

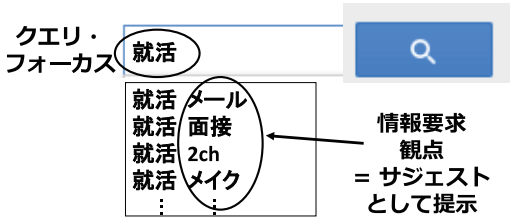


図 1　検索エンジン・サジェストにおける情報要求観点の例

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

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) [3] 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 4 参照) の作成を行う．本論文では，以上のサジェストの集約手法，および，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.　検索エンジン・サジェストの収集

本論文において，評価対象とするクエリ・フォーカスに対して，Google(注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」などの開拗音であり，一文字列あたり最大 10 個のサジェストを収集可能することが出来る．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，それらの収集を行う．クエリ・フォーカス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

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トピックと呼ばれる話題の単位へと自動的に集約する．そして，自動集約の結果に対する評価を行う．

3.1　検索エンジン・サジェストを用いたウェブページの収集

まず，前節において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 d の集合を $D(s,N)$(ただし，本論文においては，$N=20$ とする) とする．ウェブページの収集においては，Yahoo! Search BOSS API(注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．

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

ここで，各クエリ・フォーカスごとに収集し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，以下の各節においては，この 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の推定を行う．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 (LDA; Latent Dirichlet Allocation) [3] 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++(注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，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，

(注1)：https://www.google.com/
(注2)：http://developer.yahoo.com/search/boss
(注3)：http://gibbslda.sourceforge.net/

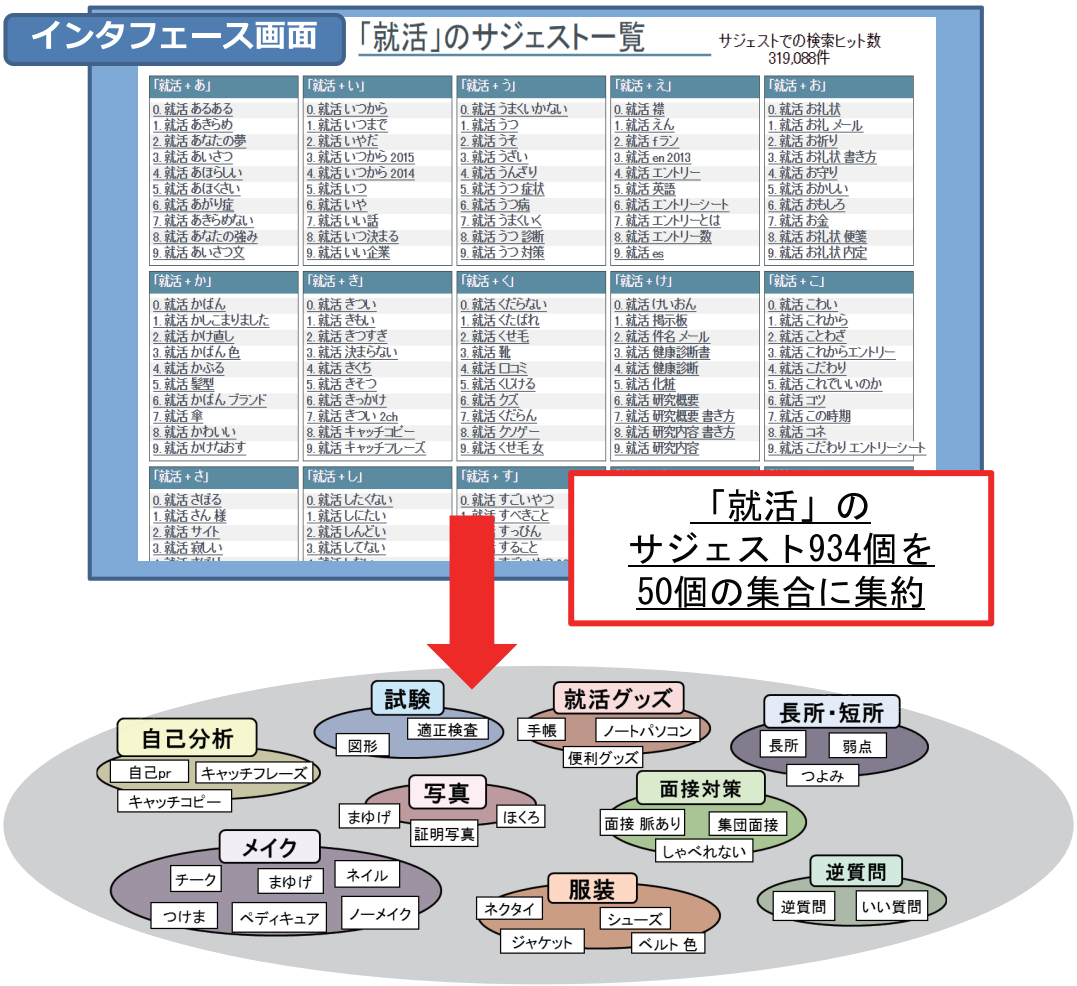


図 2　検索エンジン・サジェストの集約 (クエリ・フォーカス:「就活」)

そのリダイレクトの集合(注4)を用いた.

LDA を用いたトピック推定においては，LDA 適用時にトピック数 K を人手で指定する必要がある．そのため，本論文では，トピック数 K を 10 から 80 まで変化させてトピック推定を行い，その結果を人手で見比べ，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 による推定結果を採用した．その結果，クエリ・フォーカス「就活」,「結婚」,「花粉症」においては $K = 50$ を,「マンション」においては $K = 40$ を,「3D プリンタ」においては $K = 25$ を，それぞれ採用した.

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.

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

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，クエリ・フォーカスに加えて一つのサジェストを指定した AND 検索によって収集されたものである．そのため，各ウェブページには，最低一つ以上のサジェストが対応することになる．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付けられている．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．以上のことから，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 を集めることにより，トピック z_n に対するサジェスト s の割り当てを行うことが出来る．この時，トピック z_n に割り当てられたサジェスト集合 $S(z_n)$ は次式のように表される.

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

また，トピック z_n におけるサジェスト s の頻度 $f(s, z_n)$ は以下の式で表される.

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

実際に，クエリ・フォーカス「就活」の場合，934 個のサジェ

(注4)：日本語 Wikipedia には，2014 年 3 月にダウンロードを行ったエントリ数約 140 万 7,000 のものを用いた.

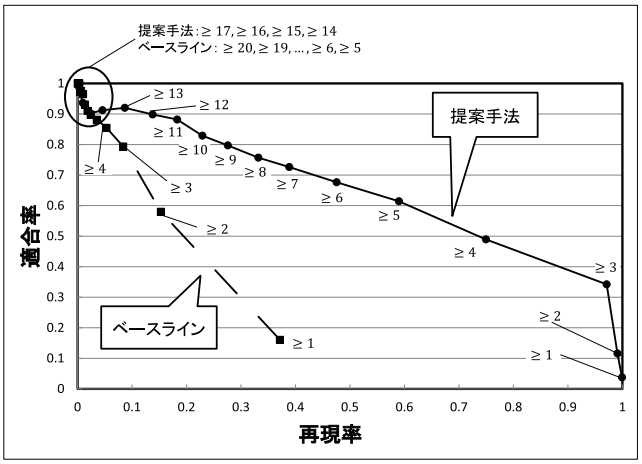


図 3　検索エンジン・サジェストの集約の評価結果 (サジェストの頻度の下限値を変化させた場合)

ストが 50 個のいずれかに割り当てられた (図 2)．このように，検索エンジン・サジェストを用いて収集されたウェブ検索結果に対してトピックモデルを用いることにより検索エンジン・サジェストの集約を行う．

本論文において，提案手法によるクラスタリング結果の評価を行う際には，トピック z_n におけるサジェスト s の頻度 $f(s, z_n)$ に対する下限値を導入し，下限値 f_{lbd} 以上の頻度を持つサジェスト s のみがクラスタ $C(z_n, f_{lbd})$ に含まれるとみなして評価を行う．

$$C(z_n, f_{lbd}) = \left\{ s \in S(z_n) \;\middle|\; f(s, z_n) \geqq f_{lbd} \right\}$$

上式を用いると，サジェストに対する頻度下限値 f_{lbd} の条件のもとで，提案手法によるクラスタリングによって生成されるクラスタの集合 $\mathbb{C}(f_{lbd})$ は次式となる．

$$\mathbb{C}(f_{lbd}) = \left\{ C(z_n, f_{lbd}) \;\middle|\; z_n = 1, \ldots, K \right\}$$

3.5 評　　価

サジェストの集約に関する評価においては，表 1 に示すクエリ・フォーカスのうち,「就活」および「結婚」の二つのクエリ・フォーカスを対象とした．また，提案手法の有効性の検証のためベースライン手法との比較を行った．

3.5.1 ベースライン手法

ベースライン手法におけるサジェストのクラスタリングにおいては，サジェスト s および s' によって収集されたウェブページ集合 $D(s, N)$ および $D(s', N)$ に対して，重複するウェブページの数 $\left|D(s, N) \cap D(s', N)\right|$ が下限値 n_{lbd} 以上となるサジェスト組 $\langle s, s' \rangle$ を同一クラスタに含めるという制約のもとで多重クラスタリングを行う．ウェブページ集合間の重複ウェブページ数の下限値 n_{lbd} の条件のもとで，このベースライン手法により作成されるクラスタの集合 $\mathbb{C}_b(n_{lbd})$ は，

$$\mathbb{C}_b(n_{lbd}) = \left\{ C_b \;\middle|\; \forall s, \forall s' \in C_b,\ \left|D(s, N) \cap D(s', N)\right| \geq n_{lbd} \right\}$$

となる．

3.5.2 評 価 結 果

評価対象のクエリ・フォーカス「就活」,「結婚」についてそれぞれ参照用クラスタ集合 $\mathbb{C}_r$ を作成し評価を行った．参照用クラスタ集合 $\mathbb{C}_r$ を作成する際には，提案手法による出力クラスタ集合 $\mathbb{C}$，および，ベースライン手法による出力クラスタ集合 $\mathbb{C}_b$ の和集合を初期集合として，1) クラスタからのサジェスト削除，2) 意味的まとまりのないクラスタそのものの削除，3) 提案手法によるクラスタとベースライン手法によるクラスタの併合，の三種類の操作のみを許容して参照用クラスタ集合 $\mathbb{C}_r$ を作成した．

次に，いずれかの参照用クラスタ $C_r (\in \mathbb{C}_r)$ に含まれる任意のサジェスト組 $\langle s, s' \rangle$ を集めた参照用サジェスト組集合 R，および，頻度下限値 f_{lbd} の条件のもとで，提案手法によって出力されるクラスタ $C(\in \mathbb{C}(f_{lbd}))$ のうちのいずれかに含まれる任意のサジェスト組 $\langle s, s' \rangle$ を集めたサジェスト組集合 $S(f_{lbd})$ を，それぞれ次式によって作成する．

$$R = \bigcup_{C_r \in \mathbb{C}_r} \left\{ \langle s, s' \rangle \;\middle|\; \exists C_r,\ s, s' \in C_r \right\}$$

$$S(f_{lbd}) = \bigcup_{C \in \mathbb{C}} \left\{ \langle s, s' \rangle \;\middle|\; \exists C,\ s, s' \in C \right\}$$

そして，参照用サジェスト組集合 R と提案手法によるサジェスト組集合 $S(f_{lbd})$ との間の重複を用いて，次式の再現率 $\mathrm{recall}(f_{lbd})$ および適合率 $\mathrm{precision}(f_{lbd})$ によって評価を行う．

$$\mathrm{recall}(f_{lbd}) = \frac{\left|R \cap S(f_{lbd})\right|}{\left|R\right|}$$

$$\mathrm{precision}(f_{lbd}) = \frac{\left|R \cap S(f_{lbd})\right|}{\left|S(f_{lbd})\right|}$$

一方，ベースライン手法に対しても，同様に，下限値 n_{lbd} の条件のもとで，ベースライン手法によって出力されるクラスタ $C_b (\in \mathbb{C}_b(n_{lbd}))$ のうちのいずれかに含まれる任意のサジェスト組 $\langle s, s' \rangle$ を集めたサジェスト組集合 $S_b(n_{lbd})$ を次式によって作成する．

$$S_b(n_{lbd}) = \bigcup_{C_b \in \mathbb{C}_b} \left\{ \langle s, s' \rangle \;\middle|\; \exists C_b,\ s, s' \in C_b \right\}$$

そして，参照用サジェスト組集合 R とベースライン手法によるサジェスト組集合 $S(n_{lbd})$ との間の重複を用いて，次式の再現率 $\mathrm{recall}_b(n_{lbd})$ および適合率 $\mathrm{precision}_b(n_{lbd})$ によって評価を行う．

$$\mathrm{recall}_b(n_{lbd}) = \frac{\left|R \cap S_b(n_{lbd})\right|}{\left|R\right|}$$

$$\mathrm{precision}_b(n_{lbd}) = \frac{\left|R \cap S_b(n_{lbd})\right|}{\left|S_b(n_{lbd})\right|}$$

提案手法，ベースライン手法における評価結果をそれぞれプロットした結果を図 3 に示す(注5)．図 3 に示す通り，提案手法

(注5)：図 3 においては，2 つのクエリ・フォーカス「就活」および「結婚」に対する評価結果のマクロ平均をプロットした．

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 (クエリ・フォーカス: 就活)

クエリ・フォーカス	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”，まとめ髪，おだんご，襟足，ロングヘア，ゆるいパーマ，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，“グループディスカッション　テーマ”，評価，グループワーク対策，評価基準，プレゼン，“プレゼン　資料，グループワーク，能力
	自己分析	“長所　真面目”，長所，座右の銘，軸，どうなりたいか，あなたの夢，こだわり，将来の夢，どんな人，なりたい自分
	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

によって，ベースライン手法よりも高い適合率および再現率が達成できた．提案手法により得られたサジェストの集約結果の例の一部を表 2 に示す．表 2 に示すように，各トピックにおいて，クエリ・フォーカスとの関連において相互に類似するサジェスト群が同一のトピックに割り当てられていることが分かる．

4. ウェブ検索結果の集約

4.1 概　　要

図 2 に示すように，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 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，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．

4.2 多様な話題のウェブページの選択的収集

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．

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

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 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 4 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

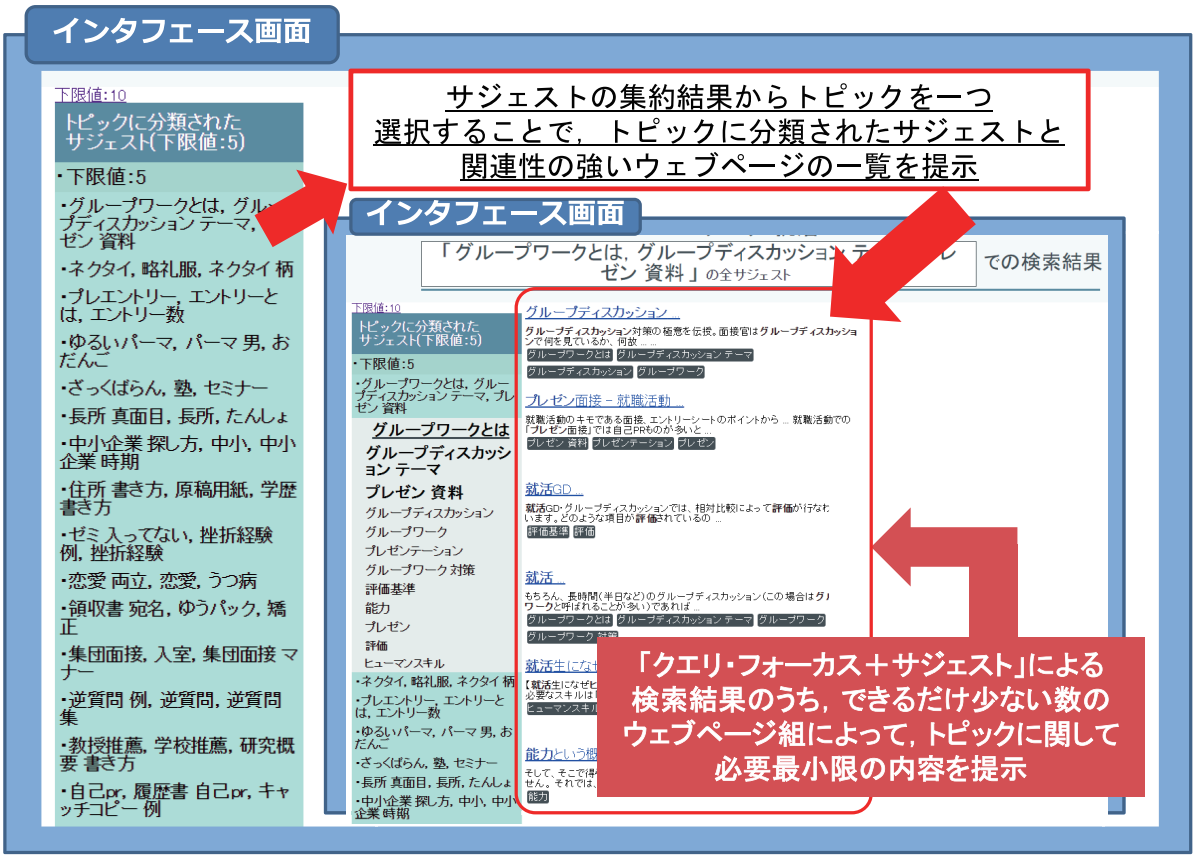


図 4　ウェブ検索結果の俯瞰インタフェース画面 (クエリ・フォーカス: 「就活」)

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．

4.3　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(注6)．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．

4.3.1　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 5 に示す．図 5 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．

4.3.2　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 6 に示す．まず，図 6(a) 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

(注6)：話題分析の際，クエリ・フォーカスとの関連が無いウェブページに関しては，話題「関連無し」として，話題数の数え上げの際には対象外とした．

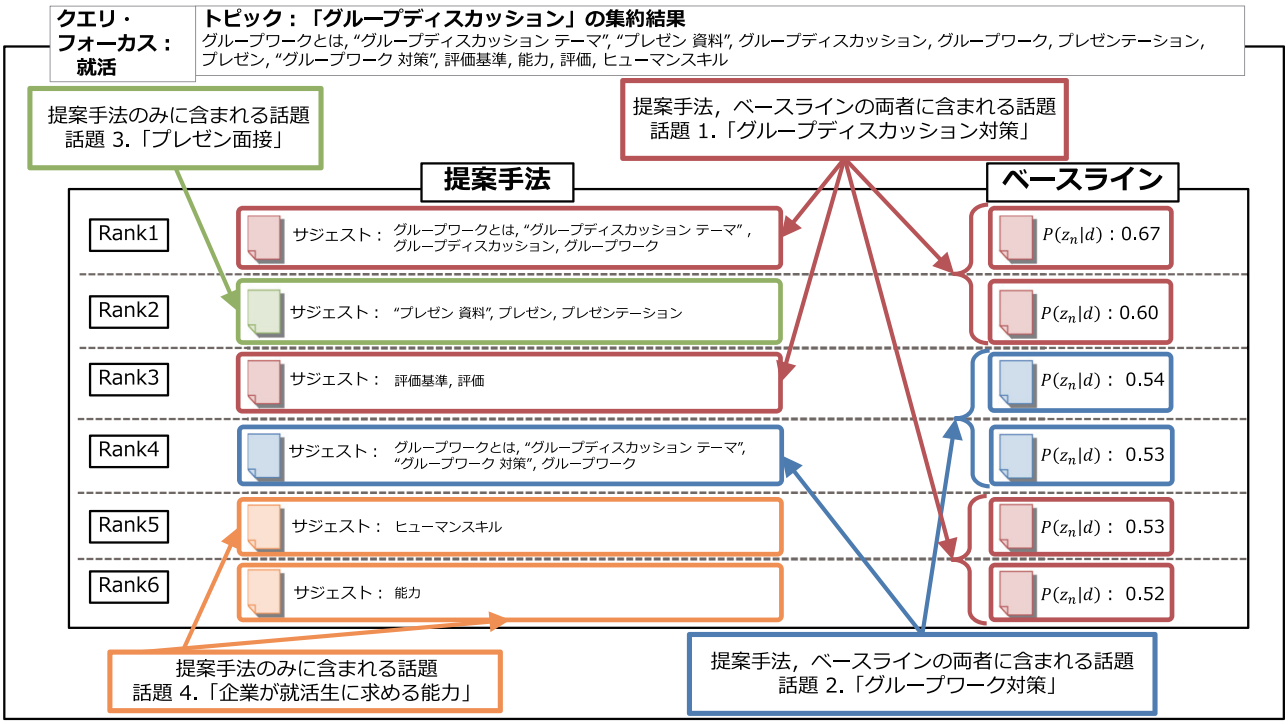


図 5　ウェブ検索結果の集約の例 (クエリ・フォーカス: 「就活」, トピック: 「グループディスカッション」)

- 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．一方，図 6(b) においては，

$$\frac{\text{1 トピックあたりに提示される話題数}}{\text{1 トピックあたりに提示されるウェブページ数}}$$

をトピック間で平均した結果を提案手法とベースライン手法との間で比較した結果を示す．これらの結果から，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．

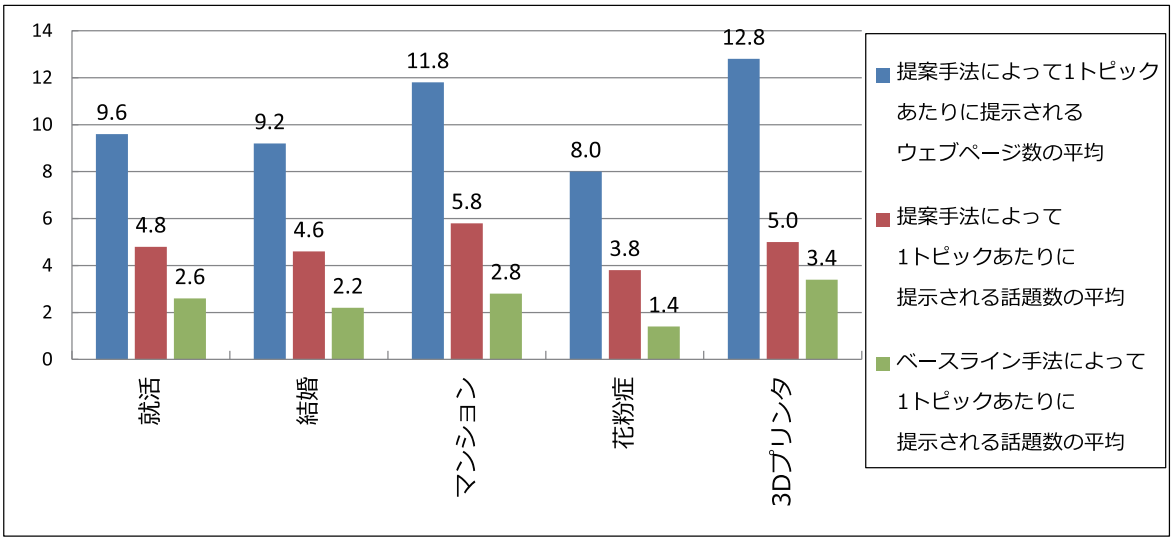
5. 関 連 研 究

本論文において提案した手法に関連する手法として，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 [5,8,9] が挙げられる．文献 [5,9] においては，数ヶ月分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い，作成されたクラスタに基づいて検索クエリを推薦する手法を提案している．一方，文献 [8] においては，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，ユーザプロファイリングの観点に基づいて，各検索ユーザの嗜好を考慮した検索クエリのクラスタリングを行っている．評価実験においては，30 人程度の検索ユーザを対象として，最大 150 個の検索クエリを評価対象として，検索クエリのクラスタリングを評価している．これらの関連研究のうち，特に，文献 [5,9] において用いられているクリックスルーデータにおいては，ユーザの ID，入力された検索クエリ，クリックされた URL，その URL の検索順位，検索クエリが入力された日時等の情報が含まれており，約数百万の検索クエリに対して約 1,000 万のクリックスルーデータを収集した研究資源となっている．それに対して，本論文の手法においては，最大約 1,000 語の検索エンジン・サジェスト，および，それらを用いて収集される約 10,000 件程度のウェブページ集合を対象としている．したがって，クリックスルーデータを対象とした関連研究と，本論文の手法とでは，対象とする研究資源，および，その規模が大きく異なっている．

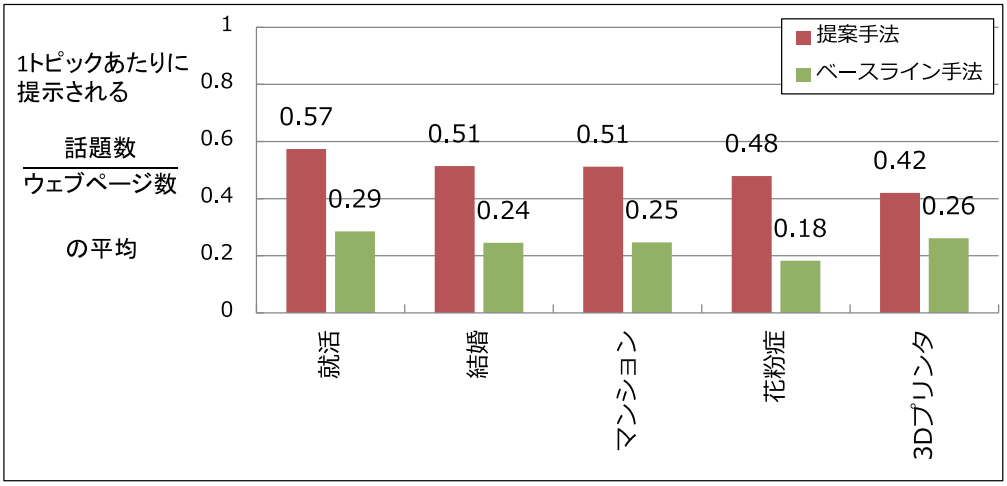
また，他の先行研究として，ウェブページの検索結果を分類し，各分類に対して適切な要約文を付与する手法 [6]，検索された個々の Web ページに対してラベルの付与を行い，付与されたラベルに基づいて分類を行う手法 [1,4,10]，階層的なトピックの体系を推定する手法 [2] 等の手法が提案されている．また，メタ検索エンジンにおいてウェブページ検索結果の上位 200 記事程度を対象にして，検索結果のクラスタリングおよびラベル付けをした結果を提示するサービスとして，Yippy(注7) が知られている．これらの手法においては，いずれも，閲覧対象の文書集合のみを用いて，ファセット体系およびファセットラベ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．

その他，文献 [7] 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

(注7)：http://yippy.com/



(a) 1 トピックあたりに提示されるウェブページ数/話題数



(b) (1 トピックあたりに提示される話題数/1 トピックあたりに提示されるウェブページ数) の平均

図 6　ウェブ検索結果の集約の評価

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．さらに，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．

文　献

[1] 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.

[2] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS*, 2003.

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

[4] W. de Winter and M. de Rijke. Identifying facets in query-biased sets of blog posts. In *Proc. ICWSM*, pp. 251–254, 2007.

[5] J. Guo, X. Cheng, G. Xu, and H.-W. Shen. A structured approach to query recommendation with social annotation data. In *Proc. 19th CIKM*, pp. 619–628, 2010.

[6] 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.

[7] 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.

[8] K. W.-T. Leung, W. Ng, and D. L. Lee. Personalized concept-based clustering of search engine queries. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 11, pp. 1505–1518, 2008.

[9] H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proc. 18th CIKM*, pp. 709–718, 2008.

[10] 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.