

ノウハウサイト群のトピック閲覧インタフェース

川畑 修人[†] 大川 遥平[†] 牛 文彬[†] 趙 辰[†] 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 理工学群 工学システム学類 〒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 一般に、質問回答サイトやその他の多様なウェブサイトには、多種多様な内容および質の情報が分散しているため、ウェブページのノウハウ知識を集約してユーザに提示することが必要となる。この要求を満たすため、先行研究では、ある決まったクエリ・フォーカスにおいて、ノウハウ知識の候補を集約・俯瞰している。さらに、後続研究では、収集したウェブページ集合にトピックモデルを適用した結果に対し、複数のトピックにまたがってウェブページを保持するドメインはノウハウサイトのドメインである可能性が高いと仮定し、いくつかのクエリ・フォーカスにおいて、収集された候補サイトの半数以上がノウハウサイトであるという結果を得た。これに対し本論文では、ノウハウサイト群の各ノウハウサイトにどのような話題のウェブページが存在するかの一覧を表示するトップページ、および、それらノウハウサイトのウェブページ群を効率良く閲覧するためのインタフェースを作成する。

キーワード ノウハウ, インタフェース, 検索エンジン, トピックモデル, word2vec, 分類器学習

An Interface for Browsing Topics of Know-How Sites

Shuto KAWABATA[†], Ohkawa YOUHEI[†], Wenbin NIU[†], Chen ZHAO[†], 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

現在、私たちは、インターネットやソーシャルメディアを活用することによって、日常生活に役立つ知識を容易に得ることができる。また、様々な知識をまとめたウェブサイトも多数存在し、Yahoo!知恵袋、OKWAVE等の質問回答サイトでは、「就活のエントリーシート書き方」、「結婚式にかかる具体的な費用」などのノウハウ知識が多数掲載されている。しかしながら、このような質問回答サイトやその他の多様なウェブサイトには、多種多様な内容および質の情報が分散して存在しているため、ウェブページのノウハウ知識を集約して利用者に提示することが不可欠である。

この要求を満たすため、文献 [8] では、特定のクエリ・フォーカス (検索エンジンにおいて検索クエリとして用いるキーワードに相当) において、ノウハウ知識の候補を集約・俯瞰している。さらに、文献 [9] では、分類器学習手法として SVM (Sup-

port Vector Machine) を用いることにより、いくつかのクエリ・フォーカスを対象として、自動でノウハウサイトを同定する手法を提案している。これに対し本論文では、文献 [9] において自動同定されたノウハウサイト群を効果的に閲覧可能なインタフェースを作成することを目的とする。本閲覧インタフェースは、各ノウハウサイトにどのような話題のノウハウ記事が存在するかの一覧を表示するトップページ、および、各ノウハウサイトの記事を効率良く閲覧するためのサブページによって構成される。本論文では特に、ノウハウに関する質問・応答組を作成するタスクにおいてこの閲覧インタフェースの有効性を評価し、検索エンジンを利用した場合と比較して、より効率的に質問・応答組が作成できることを示す。

2. 検索エンジン・サジェストを用いたウェブページ収集

検索エンジンには、過去の検索者が検索したときの検索ロ

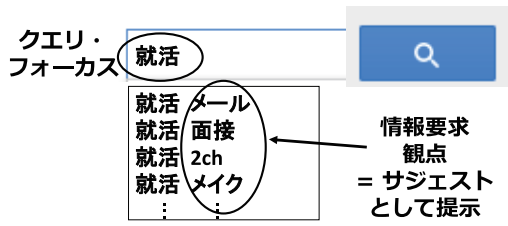


図 1 検索エンジン・サジェスト(クエリ・フォーカス「就活」の例)

グを用いることで、次の検索者が入力するクエリ・フォーカスに関係のある語群を、検索ワードの隣にサジェストとして提示する機能がある(図 1)。クエリ・フォーカスについて検索者が関心を持つ事柄を収集するため、本論文では検索エンジン・サジェストを用いる。本論文では、クエリ・フォーカスとして、「就活」、「結婚」、「虫歯」、「花粉症」の 4 種類を対象とし、検索エンジンには Google 社の検索エンジンを用いる。サジェスト収集時、一つのクエリ・フォーカスあたり 100 通りの文字列(五十音、濁音、半濁音、および、開やく音を含む)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可能なため、理論上、最大 1,000 語のサジェストを収集することが可能である。また、ウェブページの収集においては、検索語として、「クエリ・フォーカス AND サジェスト」の AND 条件を設定し、Google Custom Search API^(注1)を用いて、各検索結果の上位 20 ページを収集する。以上の手順により収集されたサジェスト数、および、ウェブページ数を表 1 に示す。

以下、本論文では、上述のサジェスト収集から始まり、図 2 に示すウェブページ収集、トピック単位でのウェブページの話題集約、サブトピック単位でのウェブページの話題集約、の流れに沿って収集されたウェブページ集合の話題集約を行う。

3. 話 題 集 約

本節では、図 2 に示す流れのうち、トピック単位でのウェブページの話題集約、および、サブトピック単位でのウェブページの話題集約、の詳細について述べる。

3.1 トピックモデルの適用およびサジェストの集約

検索エンジン・サジェストを利用することにより、大規模なウェブページ集合を収集することができるが、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ため、収集されたサジェスト、および、ウェブページの集約・俯瞰が必要となる。そこで、収集したウェブページ集合に対して、潜在的ディリクレ配分法(LDA; Latent Dirichlet Allocation) [1] を適用する。LDA 適用時には、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。LDA の適用により、各トピック $z_n (n = 1, \dots, K)$ 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 $P(z_n|d) (n = 1, \dots, K)$ が得られる。各文書 d におけるトピック z_n の確率分布 $P(z_n|d) (n = 1, \dots, K)$ 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。そして、トピック z_n に対して、次式を用いてウェブペー

ジ集合 $D(z_n)$ を割り当てる。

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

ここで、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。そのため、各ウェブページ d に対して一つ以上のサジェストから構成されるサジェスト集合 $S(d) (d \in D(z_n))$ が対応付けられる。そこで、各トピック z_n に対してサジェスト集合 $S(z_n)$ を割り当てるため、各トピック z_n におけるウェブページ集合 $D(z_n)$ 中のウェブページ $d (d \in D(z_n))$ に対応付けられたサジェストを収集し、それらのサジェスト集合の和集合を $S(z_n)$ とする。

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

一方、各トピック z_n のサジェスト集合 $S(z_n)$ 中のサジェスト $s (s \in S(z_n))$ に対しては、 s が付与されたウェブページ $d (d \in D(z_n))$ の集合 $D(s, z_n)$ の要素数を求めることにより、トピック z_n での頻度に相当する値 $f(s, z_n)$ を求めることができる。

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

$$f(s, z_n) = |D(s, z_n)|$$

そこで、本論文の閲覧インタフェースでは、各トピック z_n の話題を表す方法の一つとして、サジェスト頻度 $f(s, z_n)$ の上位 3 位までのサジェストを連結して提示する方式を採用する。

3.2 サブトピックへの話題集約

前節までの手順による、トピックモデルを用いたウェブページ集合の話題集約の粒度においては、その粒度が粗いという問題がある。そこで本論文では、文献 [2] に従い、トピック単位の粒度の粗い話題集約結果を、さらに粒度の細かいサブトピック単位に分割する。具体的には、各クエリ・フォーカスごとに、日本語版 Wikipedia のテキスト、および、収集されたウェブページ集合の混合文書集合に対して word2vec [7] を適用することにより、頻度下限値^(注2)以上のサジェストに対して分散表現を求める^(注3)。ここで、各ウェブページ d に対し、サジェスト集合 $S(d)$ の要素のうち、トピック z_n でのサジェスト頻度 $f(s, z_n)$ が最大となるサジェスト $s(d)$ を d に割り当てる。

$$s(d) = \underset{s \in S(d)}{\operatorname{argmax}} f(s, z_n)$$

$$(ただし, z_n = \underset{z_u (u=1, \dots, K)}{\operatorname{argmax}} P(z_u|d))$$

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

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

を用いて、ウェブページ d と d' の類似度尺度を定義する。

(注2)：頻度下限値を 5 とする。

(注3)：各クエリ・フォーカスごとの分散表現を持つ検索エンジン・サジェストの数を表 1 の「分散表現を持つサジェスト」欄に示す。

(注1)：<https://developers.google.com/custom-search/>

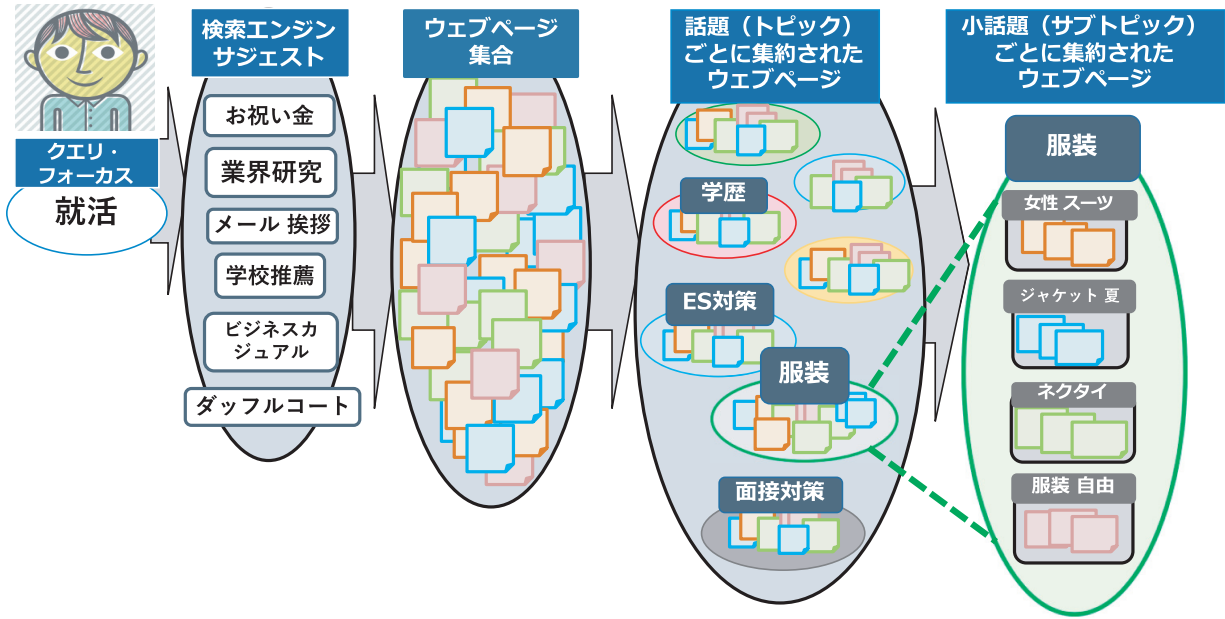


図 2 サジェスト収集からサブトピック集約までの流れ (クエリ・フォーカス「就活」の例)

表 1 データセット

クエリ・フォーカス	サジェスト数		ウェブページ集合	トピック数	ノウハウサイト候補数	参照用ノウハウサイト数
	ウェブページ収集時	分散表現を持つサジェスト				
就活	923	627	11,840	50	81	50
結婚	978	841	12,880	50	55	39
虫歯	855	694	6,947	50	53	29
花粉症	849	707	6,997	50	44	25

ここで、次式により下限値 θ_{lbd} 以上の類似度を持つウェブページ組を集めることにより、各サブトピックに対応するウェブページ集合

$$D_i(z_n) = \left\{ d \in D(z_n) \mid \exists d' \in D_i(z_n), \right. \\ \left. sim(v(d), v(d')) \geq \theta_{lbd} \right\}$$

を構成する。これにより、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 (サブトピック集合^(注4)) へと分割される [2]。

4. 閲覧対象のノウハウサイト候補群

本節では、文献 [5] に従い、閲覧対象のノウハウサイトを選定する。文献 [5] では、ウェブページ収集後の文書集合にトピックモデルを適用した結果の各トピック確率上位 30 位以内の文書集合に対して、

複数のトピックにまたがって出現するドメインはノウハウサイトである可能性が高い

という仮説 (図 3) を立て、この仮説を検証した結果、

複数のトピックにまたがってウェブページを持つドメインの半数以上がノウハウサイト^(注5)である

との結果となった。

文献 [5] におけるノウハウサイト候補選定手順の詳細を以下に述べる。まず、あるドメインを dm_x 、異なるトピック z_i および z_j ($i \neq j$) に割り当てられた文書 d 、および d' 、それらの URL を $u(d)$ 、および $u(d')$ とする。このとき、 $dm(u(d))$ 、および $dm(u(d'))$ がともにあるドメイン dm_x と等しい場合 (次式)、

$$\exists i, \exists j, i \neq j, \exists d \in D(z_i), \exists d' \in D(z_j)$$

$$dm(u(d)) = dm(u(d')) = dm_x$$

ドメイン dm_x を、「当該ドメインのウェブページが複数のトピックにまたがる」ドメインとする^(注6)。ただし、i) 質問回答サイト、ii) ニュースサイト、iii) 販売サイト、iv) ブログホスト以下のブログ、のサイトのウェブページにおいては、特定ジャンル

(注5)：本論文では、ノウハウサイト候補の集合 S の各ドメインのうち、人手によって、表 2 の基準のうちの A 群、B 群、C 群のいずれかであると判定されたドメインを指す。以下では、このドメインの集合を「参照用ノウハウサイト集合 R 」と呼ぶ。

(注6)：例えば、 $u(d) = \text{http://xxx.com/yyy}$ 、 $u(d') = \text{http://xxx.com/zzz}$ の場合に、 $dm(u(d)) = dm(u(d')) = \text{http://xxx.com/}$ となる。

(注4)：本論文では、類似度下限値 θ_{lbd} を 0.7 とし、要素数 3 以上となるサブトピックのみを有効とした。

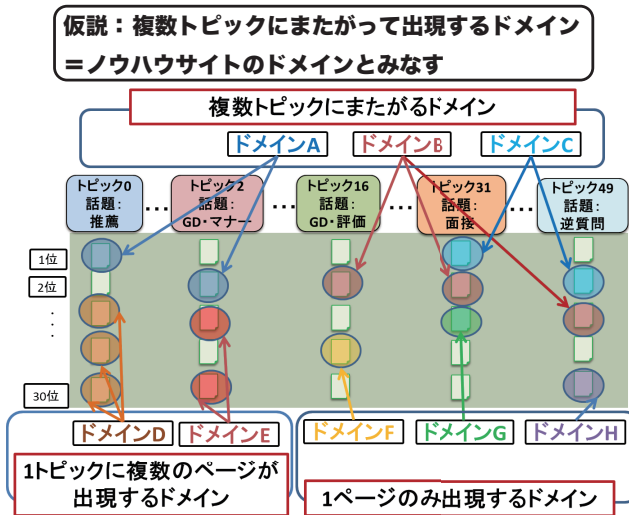


図3 ノウハウサイト候補の選定

履歴書Do

応募準備

履歴書の書き方や職務経歴書やエントリーシートの書き方、志望動機、自己PR、履歴書の写真、用紙の選択、学歴欄・職歴欄、郵送方法などなど。

ノウハウ記事へのリンク

このカテゴリーの人気記事

1. [履歴書ダウンロード] 書き方・Word面対応テンプレートと記入例

2. 履歴書を送る封筒の色・サイズ・宛名などの正しい書き方

3. [職務経歴書ダウンロード] すぐ使えるフォーマットと記入例

4. 履歴書の扶養家族・扶養義務・配偶者の書き方

5. 履歴書の職歴・学歴で使う「現在に至る」の書き方

6. 履歴書の特技欄の書き方おすすめ記入例7選

7. 履歴書の書き方はどっち!? 手書きvsパソコン作成

8. 在職中の転職は履歴書の書き方が重要! 本人希望欄を活用する方法

9. 履歴書の職歴が書ききれないときの対応

HOME

応募準備

就職試験

自分磨き

会社選び

就活

特集一覧

応募準備に役立つ新着記事

更新された記事

幼稚園教諭の志望動機の書き方では実は「企業研究」が最も大切

出版社の志望動機の書き方は損得勘定を抜きにした強い熱意を示すのがコツ

エステへの志望動機がグッとよくなる書き方ポイント

塾講師の志望動機の書き方は企業研究を大事にするのがポイント

栄養士の志望動機は本気度を見せる書き方がポイント

自己PRでクレーム対応経験をアピールするときのコツ

自己PRでゼミの活動を企業に響くように書くコツ

保育士の自己PRでアピールしたい鉄板ポイントと書き方

自己PRの無遅刻無欠席は皆勤を能力や人格に変換してアピールすると効果的

図4 ノウハウサイト例(クエリ・フォーカス「就活」の例)

「履歴書 Do」, (<https://www.rirekisyodo.com/>)

のノウハウの一覧が参照できる可能性が低いと仮定し、複数のトピックにまたがるドメインが上記 i)~iv) のいずれかに該当する場合には、そのドメインを除外する。そして、以上の一連の条件を満たすドメインの集合をノウハウサイト候補の集合 S とする。

クエリ・フォーカス「就活」および「結婚」における参照用ノウハウサイトの例を、図4、および、図5にそれぞれ示す。

5. ノウハウサイト候補群閲覧インターフェース

5.1 概要

前節の手順によって選定されたノウハウサイト候補群を閲覧するためのインターフェースを作成する。インターフェースの画面例を図6に示す。

トップページでは、各ノウハウサイト候補におこえるノウハウの話題の分布を俯瞰するため、各サイトにおけるトピックの有無を表形式で表示する(図6におけるマス目の色の有無が、各トピックにおけるウェブページの有無を示す)。各ドメイン



図5 ノウハウサイト例(クエリ・フォーカス「結婚」の例)
「結婚レシピ」, (<http://www.wedding-recipe.com/>)

トップページ

クエリ

ドメイン (ノウハウサイト候補群)

就活

ドメイン絞り込みボタン

ノウハウ (トピック)

操作の主な流れ

1 ドメインを選択, デモページへ遷移
2 ノウハウ (トピック) を閲覧・選択
3 サブトピックを閲覧・選択
4 興味ある記事にジャンプ

選択中のノウハウを表示

サブトピックを表示するためのバー

デモページ

グループ直接 マナー集団面接・グループ面接(top)

就活 手順

就活 集団面接 マナー

グループ面接でよく聞かれる質問と対策 - 履歴書Do

就活 説明会 マナー

就活の座談会の主な流れ・どんな質問が好印象につながる? - 履歴書Do

図6 ノウハウサイト候補群閲覧インタフェース(クエリ・フォーカス「就活」の例)

dm_x のウェブページ集合を $D(dm_x)$ とすると, 各ドメインに含まれるトピックの集合 $Z(dm_x)$ は次式で定義される.

$$Z(dm_x) = \{z_n \mid d \in D(dm_x),$$

$$z_n = \underset{z_u (u=1, \dots, K)}{\operatorname{argmax}} P(z_u | d)\}$$

本インタフェースでは, 文献 [9] において SVM(Support Vec-

表2 ノウハウサイト候補のドメインに対する評価基準

ドメインそのものがノウハウ知識を提示する個別ページへのリンクを一覧するページである			A 群
ドメインそのものがノウハウ知識を提示する個別ページへのリンクを一覧するページではない	ノウハウ知識を提示する個別ページへのリンクを一覧するページが存在する	ドメインのトップからノウハウ知識を提示する個別ページへのリンクを一覧するページに容易に辿り着ける	B 群
		ドメインのトップからノウハウ知識を提示する個別ページへのリンクを一覧するページには容易に辿り着けない	C 群
	ノウハウ知識を提示する個別ページへのリンクを一覧するページが存在しない	ノウハウ知識を提示する個別ページが存在する	D 群
		ノウハウ知識を提示する個別ページが存在しない	E 群

tor Machine) モデルによってノウハウサイトか否かの自動判定を行う際の信頼度を参照して、利用者が指定した下限値以上の信頼度を持つドメインのみを表示させる機能を持ち、SVM モデルによる自動判定結果のうちの高信頼度ノウハウサイト候補群のみを閲覧可能としている。また、各ウェブページに対し、ページタイトル、概要を示すスニペット^(注7)、各ウェブページに付与されたサジェスト提示機能も持ち、利用者によるウェブページ閲覧過程の支援を可能としている。

5.2 利用例

図6に示すように、各ドメインのデモリンク部分(ボタン1)をクリックすると、そのドメインのウェブページを効果的に閲覧するためのページに遷移する。このページでは、そのドメインの持つトピックがオレンジ色で示され、各トピック(ボタン2)をクリックすると、そのトピックに含まれるサブトピックが表示される。サブトピック(ボタン3)をクリックすると、そのサブトピック中のウェブページへのリンクがリストされる。ウェブページへのリンク(ボタン4)をクリックすると、実際にノウハウが記述されているページに遷移する。

図6の例では、ノウハウサイトの一つである「履歴書 Do」のドメイン(ボタン1)をクリックした後、当該サイトのトピックの一つであるグループ面接に関するトピック「グループ面接 マナー・集団面接・グループ面接」^(注8)(ボタン2)をクリックし、次に、「集団面接 マナー」、「説明会 マナー」、「グループワーク」、「グループディスカッション」等、選択したトピックに関するサブトピックがリストで表示される。その他、サブトピックとは別に、サブトピックに分類されなかった話題をリストで表示する機能も持つ。

本インタフェースの以上の機能を用いることによって、クエリ・フォーカスにおける俯瞰的なノウハウ知識を得ることができ、ノウハウサイトに絞ったウェブページの効果的な閲覧が可能となる。

6. ノウハウサイト候補の順位付け方式の比較

ノウハウサイト候補の集合 S の順位付け方式として、「トピック

数降順」や「ウェブページ数降順」、「SVM 信頼度降順」などいくつかの方法が考えられる。順位付け方式 r において上位 θ 位以内のノウハウサイト候補の集合 $S_r^\theta (\in S)$ 中のドメインのうち、参照用ノウハウサイト集合 R に含まれるドメインの数 $|S_r^\theta \cap R|$ のプロット(クエリ・フォーカス「就活」の場合)の比較結果を図7に示す。図7に示す通り、ノウハウサイト候補の数 $|S|$ は81、参照用ノウハウサイトの数 $|R|$ は50となった。順位付け上位はどの方式でもほとんど違いがないのに対し、下位はSVM 信頼度降順で順位付けした場合が参照用ノウハウサイト集合 R のドメインを多く含む結果となった。

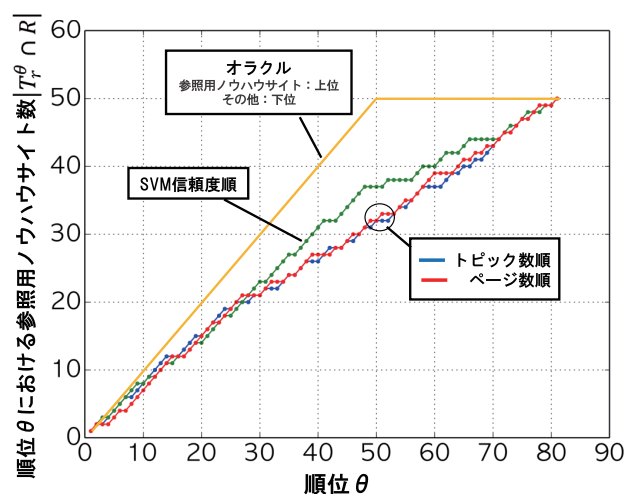


図7 ノウハウサイト候補の順位付け方式の比較(クエリ・フォーカス「就活」の例)

7. トピックモデル訓練時に未知のウェブページのトピック分類

7.1 分類手順

前節までに述べた、検索エンジン・サジェストを用いて収集したウェブページ群とは別に、任意に指定したドメインの全ウェブページを収集し、「検索エンジン・サジェストを用いて収集したウェブページ群によって訓練したトピックモデル」のトピック群へ分類する方式[1]を適用した。これにより、検索エンジン・サジェスト、および、検索エンジンAPIの組み合わせでは収集しきれなかったウェブページ群を閲覧インタフェース

(注7): ページタイトル、および、スニペットについては、Google Custom Search API によって利用できるものを提示する。

(注8): 頻度上位となる3サジェスト「グループ面接 マナー」、「集団面接」、「グループ面接」によって示される。

表 3 「トピックモデル訓練時に未知のウェブページのトピック分類」によるウェブページ数・トピック数の変化 (クエリ・フォーカス「就活」の例)

「トピックモデル 訓練時に未知の ウェブページ」の トピック分類前後	ノウハウサイト									
	Jobweb		Jobrass 新卒		履歴書 Do		キャリアタツ新卒		ハタラクティブ	
	ページ数	トピック数	ページ数	トピック数	ページ数	トピック数	ページ数	トピック数	ページ数	トピック数
前	17	11	55	18	40	16	193	29	34	14
後	4,401	38	1,218	35	1,378	34	279	33	6,002	31

に反映させることが可能となる。

具体的には、特定のドメインのウェブページ d のうち、前節までに述べた手順でのトピックモデルの訓練においては未知であったウェブページに対して、訓練済みトピックモデルにおけるトピックの確率分布を求める。そして、確率値 $P(z_n|d)$ が最大となるトピック z_n に対して、確率値 $P(z_n|d)$ が下限値 (本論文では 0.3 とした) 以上となる場合に、本論文の閲覧インタフェースに格納する。

7.2 閲覧インタフェースへの導入例

クエリ・フォーカス「就活」において、ノウハウサイトとなる 5 種類のドメインを対象として、前節で述べた手順により、ウェブページの収集、および、訓練済みトピックへのウェブページ分類を行った。これらの 5 種類のドメインにおける、訓練済みトピックへのトピック分類前後のウェブページ数、および、分類結果のトピック数を表 3 に示す。本閲覧インタフェース利用者は、この機能を利用することにより、任意に選択したドメインに含まれる全てのウェブページに対して、的確にトピック分類された形での集約・俯瞰・閲覧が可能となる。

8. 評価実験

8.1 評価手順

本論文の閲覧インタフェースを評価するため、既存の検索エンジンを比較対象として比較実験を行う。被験者延べ 11 名に対し、あらかじめ指定したクエリ・フォーカス「就活」、「結婚」、「花粉症」に関するノウハウについての質問応答事例 [6,10] を作成するタスクを実施させる。このタスク実施時において、検索エンジンを用いた場合、および、本論文の閲覧インタフェースを用いた場合、それぞれにおいて、ノウハウに関する質問応答事例 10 例を作成してもらい^(注9)、その際の所要時間、クリック数、話題数^(注10)、および、1 時間あたりに作成された質問応答事例の話題数を評価する。その際、参考にしたページの URL、ノウハウに関する質問内容、質問に対する回答、および、その回答が含まれる段落をエクセル形式で保存するものとする。

(注9)：各被験者は、検索エンジンを用いて質問応答事例 10 例を作成した後、検索エンジンと本論文の閲覧インタフェースの間で質問応答事例が重複しないという条件のもと、本論文の閲覧インタフェースを用いて、新たな質問応答事例 10 例の作成を行った。

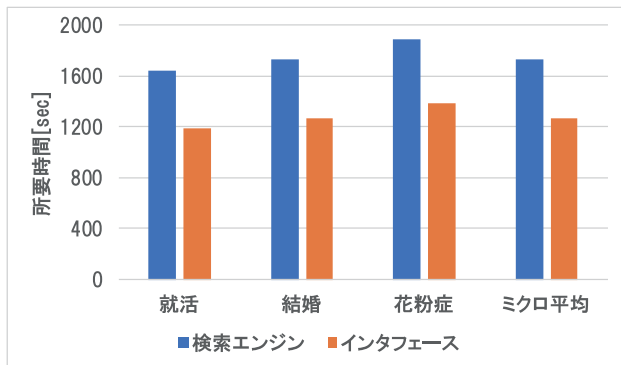
(注10)：作成された質問応答事例における話題の広がりを手で測定した。その際の話題の粒度は、トピックモデルにおけるトピックの粒度とほぼ同等となった。ここで、各クエリ・フォーカスにおけるトピック数 K はいずれも 50 であったが、そのうちノウハウとは無関係なトピックが除外された結果、作成された質問応答事例における話題の広がり数は、おおそ 20~30 程度となった。

8.2 評価結果

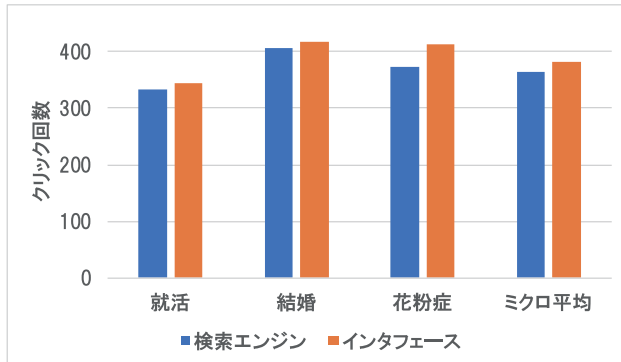
評価結果を図 8 に示す。クエリ・フォーカスごとの平均では、検索エンジンを用いた場合の所要時間は、それぞれ、1,640 秒、1,731 秒、1,891 秒、クリック回数は、333 回、407 回、374 回となった。一方、閲覧インタフェースを用いた場合の所要時間は、それぞれ、1,189 秒、1,272 秒、1,387 秒、クリック回数は、344 回、417 回、413 回となった。平均所要時間については、全てのクエリ・フォーカスにおいて検索エンジンを用いた場合を上回ったのに対して、クリック回数においてはほとんど差が見られなかった。また、被験者延べ 11 人各々の所要時間については、検索エンジンを用いた場合よりも閲覧インタフェースを用いた場合の方が圧倒的に短い結果となった。さらに、全クエリ・フォーカスでのミクロ平均の比較においては、所要時間においては、検索エンジンが 1,733 秒、閲覧インタフェースが 1,266 秒となり、一方、クリック数においては、検索エンジンが 364 回、閲覧インタフェースが 382 回となった。ここで延べ 11 名を母集団として、検索エンジンと閲覧インタフェースの間で所要時間の差の t 検定を行ったところ、その差は有意 (有意水準 1%) となった。この理由として、検索エンジンを用いた場合は、被験者自らが検索語を考え出して検索窓に入力する必要があるのに対し、閲覧インタフェースを用いた場合には、あらかじめクエリ・フォーカスに関する俯瞰的な話題、および、各話題に対するウェブページ集合が用意されているため、被験者自らが検索語を考えて入力する必要がなく、相対的に所要時間が短縮されたと考えられる。さらに、話題数においては、両者はほぼ同等であったが、1 時間あたりの話題数に換算すると、閲覧インタフェースが検索エンジンを上回り、その差は有意 (有意水準 1%) となった。ここで、本論文の比較実験では、被験者が作成するノウハウ質問応答事例数を 10 としたが、これを数十～百程度とした場合には、検索エンジンと閲覧インタフェースの間の所要時間の差がさらに大きくなることが予想される。以上の結果から、本閲覧インタフェースの有用性を示すことができると考えられる。

9. 関連研究

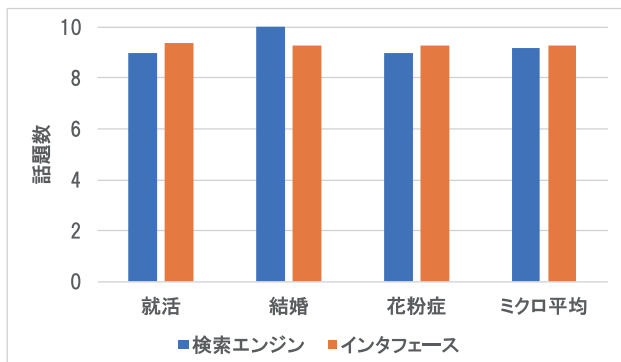
関連研究として、文献 [3] では、クエリ・フォーカスに対して、サジェストおよびウェブページを収集し、トピックモデルを適用することにより話題を集約した。そして、その話題集約結果を俯瞰・閲覧するためのインタフェースを作成した。文献 [8] では、文献 [3] に対し、Yahoo 知恵袋などの質問回答サイトおよびウェブページから収集した混合文書集合に対して LDA を



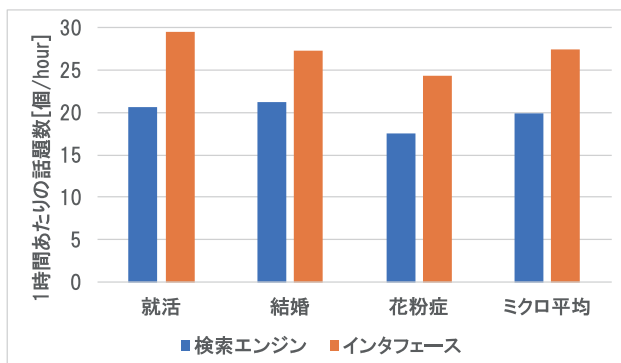
(a) 所要時間



(b) クリック回数



(c) 話題数



(d) 1時間あたりの話題数

図8 評価結果

適用し、より有用なノウハウ知識が得られることを示した。文献 [9] では、トピックモデルおよび分類器学習を適用することによって、ノウハウサイトの自動同定を行なった。さらに、文

献 [4] では、人手で収集したノウハウサイト群を効果的に閲覧するインタフェースを作成し、各ノウハウサイトにどのようなウェブページが含まれるのかを一覧する機能を実現した。

一方、本論文は、文献 [3,8,9] におけるトピックモデルの適用に加え、粒度の粗い話題をサブトピック化する機能を備えた閲覧インタフェースを実現した点が異なっている。また、文献 [4] と比べると、人手で同定したノウハウサイトの閲覧にとどまらず、分類器学習によって自動同定されたノウハウサイト候補群の閲覧インタフェースを実現した点が異なっている。

10. おわりに

本論文では、4種類のクエリ・フォーカスについて、ウェブページを自動収集し、そのウェブページ集合をトピックおよびサブトピックへ分類後、分類器によってノウハウサイトを自動同定した結果に対して、利用者がノウハウサイト候補群を効果的に閲覧するための閲覧インタフェースを作成し、評価実験を行った。評価実験の結果より、検索エンジンとの比較において、本閲覧インタフェースの有用性を示すことができた。今後の課題として、トピック分類の精度の改善、および、分類されたトピックの内容がノウハウ知識に相当するか否かの自動同定が挙げられる。その他、特定のドメインから網羅的に収集したウェブページに対して、当該ページのウェブ検索において用いられることが推定される検索クエリを自動同定することが挙げられる。

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] 丁易, 趙辰, 川畑修人, 宇津呂武仁, 河田容英. 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. 第10回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2018.
- [3] 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.
- [4] 川畑修人, 牛文彬, 宇津呂武仁, 河田容英. ノウハウサイト群およびトピックモデルによる話題集約結果の閲覧インタフェース. 第32回人工知能学会全国大会論文集, 2018.
- [5] 李佳奇, 趙辰, 林友超, 馬場瑞穂, 宇津呂武仁, 河田容英, 神門典子. トピックモデルにおける話題分布特性を手がかりとするノウハウサイトの収集. 第9回 DEIM フォーラム論文集, 2017.
- [6] 前田竜治, 陳騰揚, 大川遥平, 宇津呂武仁, 河田容英. ウェブ上のコラムページからのノウハウ質問応答事例の収集. 第33回人工知能学会全国大会論文集, 2019.
- [7] 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.
- [8] 守谷一朗, 井上祐輔, 今田貴和, 轟添, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例および検索エンジン・サジェストを用いたノウハウ知識の相補的収集. 第7回 DEIM フォーラム論文集, 2015.
- [9] Y. Ohkawa, S. Kawabata, C. Zhao, W. Niu, Y. Lin, T. Utsuro, and Y. Kawada. Identifying tips Web sites of a specific query based on search engine suggests and the topic distribution. In *Proc. 3rd ABCSS*, pp. 4347–4353, 2018.
- [10] 山本航平, 前田竜治, 陳騰揚, 川畑修人, 大川遥平, 宇津呂武仁, 河田容英. ノウハウ質問応答におけるニューラル読解モデルの評価. 言語処理学会第25回年次大会論文集, 2019.