The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

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

An Interface of the Collection of Know-How Sites and the Topics aggregated by a Topic Model

川畑 修人*1 Shuto Kawabata　牛 文彬*1 Wenbin Niu　宇津呂 武仁*2 Takehito Utsuro　河田容英*3 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筑波大学システム情報系 Fclty. Eng, Inf. & Sys, Univ. of Tsukuba　*4(株) ログワークス Logworks Co., Ltd.

Since information of various quality are scattered in pieces on the Web, it is necessary to efficiently present useful information to users who are not familiar with the search query. In order to address this request, we propose an interface to present Web pages with aggregated topics extracted through a collection of know-how sites. In the proposed method, we collect the Web pages and aggregate them to topics and subtopics by using a topic model and word2vec. Collected know-how sites are integrated in the top page of the interface which presents the distribution of topics throughout the collection of know-how sites. For each know-how site, the interface has a function of efficiently browsing through aggregated topics and subtopics.

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で，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の書き方」,「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，[守谷 15] では，ある決まったクエリ・フォーカス (検索エンジンを利用する際に入力する文字列) 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，[李 17] では，[守谷 15] において，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．このインタフェースを用いることで，ユーザは，クエリ・フォーカスについての網羅的な話題と有益なノウハウ知識を効率的に閲覧することが可能となる．

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サジェストとして検索ワードの隣に誇示するという便利な機能がある．検索者が関心を持つ事柄を収集するため，本論文では，検索者のクエリ・フォーカスに関する関心そのものが反映される，検索エンジン・サジェストを利用する．また，本論文では，クエリ・フォーカス「就活」を対象に，検索エンジン・サジェストを収集する．また，検索エンジンには，Google 社の検索エンジン・サジェストを用いる．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，半濁音及び開やく音を含む) 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．以上の手順で収集したサジェストの数は表 1 に示すとおり，925 となった．ウェブページの収集は，検索語として,「クエリ・フォーカス AND サジェスト」の AND 条件の検索設定の元，Google Custom Search API を用いて行う．収集されたウェブページの数は表 1 に示すとおり，13,222 となった．

連絡先: 川畑 修人，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

3. トピックモデルの適用とサブトピックへの話題集約

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・サジェストを用いることで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，潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [Blei03] を，収集したウェブページ集合に対して適用する．LDA を適用することで，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．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．本論文では，トピック z_n に対してウェブページ集合 $D(z_n)$ を割り当てる際，次式を用いる．

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

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，各ウェブページに対して一つ以上のサジェストが対応づけられている．そこで，各トピックに対してサジェスト集合を割り当てるため，各ト

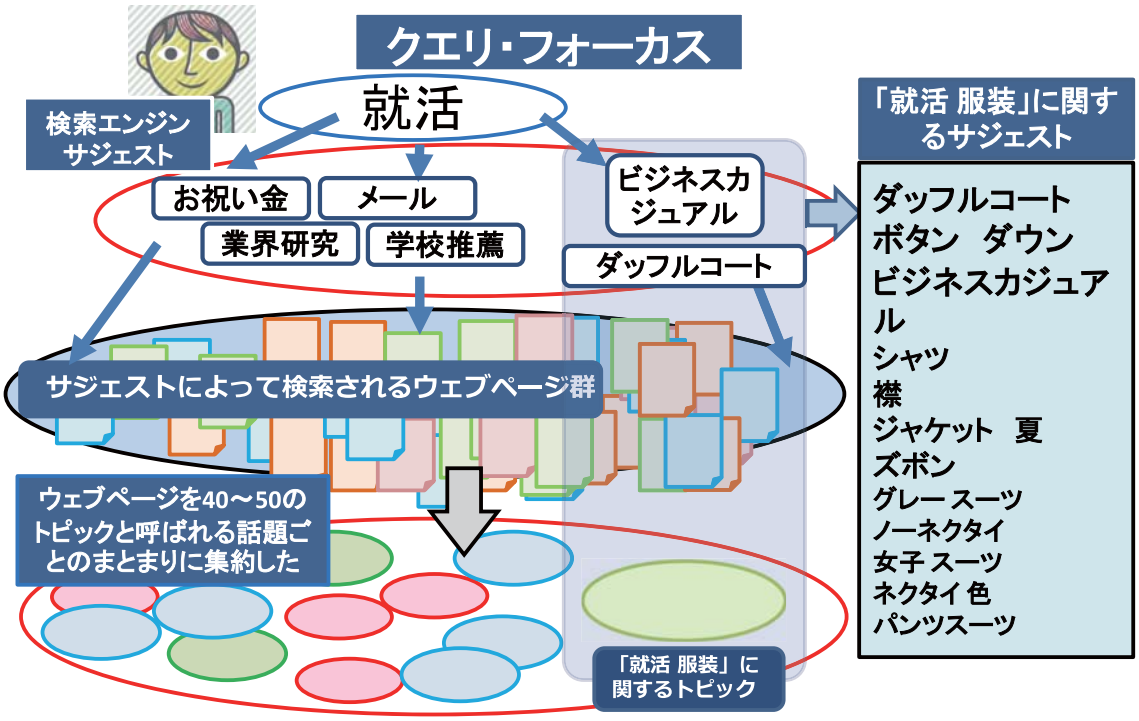

図 1: サジェストおよびトピックモデルを用いたウェブ検索結果の集約 (クエリ・フォーカス「就活」の例)

表 1: クエリ・フォーカス，サジェスト数，ウェブページ数，および，トピック数

クエリ・フォーカス	サジェスト数		ウェブページ数	総トピック数	複数トピックにまたがるドメイン数	A, B, C 群に属するドメイン数
	ウェブページ収集時	頻度下限以上，分散表現を持つサジェスト				
就活	925	671	13,222 (30.8 MB)	50	91	51

ピック Z_n に対応づけられたウェブページ集合 $D(Zn)$ 中の各ウェブページに対応づけられたサジェストを収集する．

3.2　サブトピックへの話題集約

図 1 に示すように，前節までの手順によりウェブページ収集およびその話題集約を行なった．ここで，集約された話題の粒度が粗いという問題があるため，本論分では，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に分割する手法を提案する．これは，[井上 16] 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ける機能を向上させることに等しい．各トピックにおける話題を高精度に集約するには，文書間の類似度が必要不可欠となる．これに対し，本論文では [Zhao17] に従い，word2vec [Mikolov13] を分散表現に基づくアプローチとして用いる．　まず，収集した文書集合に対して word2vec を適用する．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 に示すとおり，671 となった．　このうち，各ウェブページ d に対し，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．

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

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

から，ウェブページ d と d' の類似度尺度を定義する．

この，ウェブページ類似度尺度から，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る．これらの各部分集合が，各サブトピックに対応する．

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

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組とする．

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

表 2: ノウハウサイト候補のドメインに対する評価基準

<table>
<tr><td colspan="3">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</td><td>A 群</td></tr>
<tr><td rowspan="4">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</td><td rowspan="2">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</td><td>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</td><td>B 群</td></tr>
<tr><td>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</td><td>C 群</td></tr>
<tr><td rowspan="2">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</td><td>ノウハウ知識を提示する個別ページが存在する</td><td>D 群</td></tr>
<tr><td>ノウハウ知識を提示する個別ページが存在しない</td><td>E 群</td></tr>
</table>

4. 閲覧対象のノウハウサイト

本節では，[李 17] に従い，閲覧対象のノウハウサイトを選定する．[李 17] では，ウェブページ収集した後の文書集合にトピックモデルを適用した結果に対して，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を立て，実際に収集された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有用なノウハウサイトであるという結果を得ている．この，複数のトピックにウェブページを持つドメインについて，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，dm_x を複数のトピックにウェブページがまたがるドメインとする[*1]．また，複数のトピックにまたがるドメイン dm_x は，以下の式を満たすものとする．

$$\exists i, \exists j, i \neq j, \exists d \in D(z_i), \exists d' \in D(z_j)$$
$$dm(u(d)) = dm(u(d')) = dm_x$$

ノウハウサイトであるかどうかの評価は表 2 の基準を複数のトピックにウェブページがまたがるドメイン集合に対して人手によって適用し，A，B，C 群に属するドメインをノウハウサイトとする．また，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は，トップページのドメイン，あるいは，当該サイト中のいずれかのページにおいて，特定ジャンルのノウハウの一覧が参照できる，可能性が低いと仮定し，ドメインが上記 i) iv) のいずれかのサイトタイプにあたる場合にはそのドメインをドメイン集合から排除する．本研究では，この [李 17] においてノウハウサイトと判断されたドメインを閲覧対象のノウハウサイトとして使用する．表 1 に示すとおり，実際に得られた複数トピックにまたがるドメインの数は 91，そのうち，表 2 の A，B，C 群に属するドメインの数は 51 となった．

5. ノウハウサイト閲覧インタフェース

前節の手順により収集したノウハウサイト群を閲覧するためのインタフェースを作成する．作成したインタフェースの例を図 2 に示す．インタフェースのトップページでは，ノウハウサイト群中の各ノウハウサイトの持つトピックが一目でわかるように，トピックとノウハウサイトの表形式 (図 2 奥) で表示する．ノウハウサイトの表示順としては,「そのサイトに含まれるトピック数の多い順」や「ウェブページ数の多い順」,「ユーザが任意で指定したトピックを持つサイト優先」など様々な方法があるが，本論文では，ユーザに対し多くのトピックを含むサイトを優先的に示すため，[李 18] に従い，トピックの数の多いノウハウサイト順に表示する．また，今回，ノウハウサイトのタイトルは人手で取得する．各ドメインのトピック集合 $Z(dm_x)$ は以下の式で定義する．

$$Z(dm_x) \quad = \quad \left\{ z_i \middle| \exists d \in D(z_i), dm(u(d)) = dm_x \right\}$$

各ノウハウサイトのデモページへのリンク部分をクリックすると，図 2 手前のような，各ノウハウサイトごとのインタフェースが表示される．各ノウハウサイトにおいて，ウェブページを持つトピックは，オレンジ色で示され，クリックすると，トピック中のサブトピックとサブトピック以外に分類されたウェブページを表示するためのバーが表示される．例えば，就活のネクタイに関するウェブページがサブトピックに分類されると，

「就活 ネクタイ に関するサブトピック」

サブトピック以外のクラスタに分類されると，

「就活 ネクタイ に関するウェブページ」

のように表示を行う．ウェブページはサブトピックまたはサブトピック以外またはその両方に存在するので，ユーザはウェブページの存在するオレンジ色で示されたバーをクリックすることで，ウェブページのタイトル，リンクおよびそのスニペット (ウェブページの概略) を閲覧することができる．

6. 関連研究

関連研究として，[井上 16] では，検索エンジン，クエリ・フォーカスを用いて，サジェストおよびウェブページを収集，トピックモデルを用いることで話題を集約，ユーザの入力したクエリ・フォーカスに対し網羅的に話題とそのウェブページを示すためのインタフェースを作成した． [守谷 15] では，[井上 16] 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文書集合を一つの文書集合とし，この混

*1 例えば，$u(d)$ = `http://xxx.com/yyy`，$u(d')$ = `http://xxx.com/zzz` の場合に，$dm(u(d)) = dm(u(d'))$ = `http://xxx.com/` となる．

図 2: ノウハウサイト群閲覧インタフェース

合文書集合に LDA トピックモデルを適用することで，より有用なノウハウ知識を得られることを示した．また，[李 17] では，収集したウェブページにトピックモデルを適用した結果に対し，複数のトピックにウェブページがまたがるドメインは，ノウハウサイトである可能生が高いと仮定し，いくつかのクエリ・フォーカスにおいて実際に評価を行い，候補サイトの半数以上がノウハウサイトであることを確認した．本研究は，[井上 16]，[守谷 15] におけるトピックモデルの適用に加え，さらに，粒度の粗い話題のサブトピック化を行い，ノウハウサイトに限定したウェブページページを示すという点で大きく異なっている．また，[李 17] に対しては，収集したノウハウサイトのウェブページを，話題と小話題に分類し，クエリ・フォーカスに関する網羅的な知識を提供するためのインタフェースを作成したという点で大きく異なっている．

7. おわりに

本論文では，ひとつのクエリ・フォーカスについて，ウェブページを自動収集し，トピックとサブトピックに分類，ノウハウサイトに限定してユーザに示すことで，網羅的な話題と共に，ウェブページ群を効果的に閲覧するためのインタフェースを作成した．

参考文献

[Blei03] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[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).

[李 17] 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).

[李 18] 李佳奇, 趙辰, 林友超, 丁易, 川畑修人, 宇津呂武仁, 河田容英：トピックモデルおよび分類器学習を用いたノウハウサイトの同定, 第 10 回 DEIM フォーラム論文集 (2018).

[Mikolov13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Proc. 26th NIPS*, pp. 3111–3119 (2013).

[守谷 15] 守谷一朗, 井上祐輔, 今田貴和, 聶添, 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).

[Zhao17] Zhao, C., Nie, T., Ding, Y., Li, J., Utsuro, T. and Kawada, Y.: 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).