2E1-5in

質問回答事例およびウェブから収集したノウハウ知識の閲覧インタフェース

Interface for Browsing Know-How Knowledge collected from Question Answer Examples and Web

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]
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

This paper proposes an interface for browsing know-how knowledge collected from question answer examples and Web. In this framework, any single knowledge source of the question answer site and the Web is not sufficient to collect know-how knowledge, but the two types of knowledge source function in a complementary fashion. Given a certain query whose know-how knowledge is to be collected, the procedure starts from collecting question answer examples as well as Web pages that are relevant to the query. To the document set that are mixture of those collected from the two types of knowledge source, we apply a topic model and generate a set of topics. In this framework, certain portion of the generated topics can be regarded as know-how knowledge of the query. Among those collected know-how knowledge, those that exist both in a question answer site and on the Web are limited, while certain portion of them exist only in the question answer site or on the Web.

1. はじめに

インターネット上には様々な情報があり，多くのユーザはウェブページから日常の行動に役立つ知識を得ている．知識を得るための代表的なウェブサイトとして，Wikipedia をはじめとする百科事典サイトや Yahoo!知恵袋[*1] をはじめとする質問回答サイトが挙げられる．特に，質問回答サイトでは，「花粉症の対策方法」や「結婚式でのスピーチの仕方」といったユーザの日常の行動に役立つノウハウ知識が多く掲載されている．一方で，質問回答サイトやウェブ上に含まれる情報は膨大であり，ユーザにとって役立つノウハウ知識を集約して提示することが求められる．このような要求に対する研究として，[守谷 15] では，質問回答サイトから収集した質問回答事例，および，検索エンジン・サジェストを索引として収集されたウェブページの混合文書集合に対してトピックモデルを適用することにより，話題のまとまりを生成した．この手法を用いることにより，検索対象に対するノウハウ知識を幅広く収集することが可能となる．そこで，本論文では，ある検索対象についてのノウハウ知識の候補を網羅的に収集し，集約・俯瞰するとともに，ノウハウ知識とノウハウ知識以外の話題を選別して，効率的にノウハウ知識を収集する．この過程全体の流れを図 1 に示す．この過程においては，まず，質問回答サイトから収集した質問回答事例，および，検索エンジン・サジェストを索引として収集されたウェブページの混合文書集合に対してトピックモデルを適用することにより，話題のまとまりを生成する．次に，提案するインタフェースを用いて，話題を，「ノウハウ知識」，「ノウハウ以外の知識」，「意見」，「その他」の 4 つに分類することで，ノウハウ知識を選定する．最後に，得られたノウハウ知識を内容ごとに人手で大分類にまとめる．一例として，検索対象「花粉症」に関するノウハウ知識を収集した結果においては，合計 55 個の話題が収集された．収集された話題の中には，「花粉症の温熱治療のための吸入器」のように，ウェブページのみから得られるノウハウ知識が合計で 19 個あり，全話題の約 35%となった．本論文では，以上の手順によって収集したノウハウ知識を効率的に閲覧するインタフェースを提案する．

連絡先: 井上 祐輔，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 http://chiebukuro.yahoo.co.jp/

2. 質問回答事例の収集

本論文では，質問回答事例のデータとして，Yahoo!知恵袋から提供されている 2004 年 4 月 1 日 ~2009 年 4 月 7 日の 5 年間の質問回答事例のデータ (質問: 16,257,413 件，回答: 50,053,894 件) を用いた．本論文では，カテゴリ名，質問タイトル，質問本文のいずれかに検索対象 q が含まれている質問を抽出し，その質問に対する回答本文全てを結合し，一つの質問回答事例 d_q を作成した．各検索対象 q あたりの質問回答事例の文書集合を D_q とし，以下のように定義する．

$$D_q = \{d_q^1, \ldots, d_q^k\}$$

3. 検索エンジン・サジェストを用いたウェブページの収集

本研究では，検索エンジン・サジェストに着目し，ウェブ検索者の関心事項を収集する．本論文で分析の対象とする「花粉症」および「結婚」の各々について，Google 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．この際，ある検索対象に対して収集されたサジェストの集合を $\mathbb{S}$ とする．「花粉症」および「結婚」の各々について，それぞれ収集したサジェストの数を表 1 に示す．ここで，$s \in \mathbb{S}$ となるサジェスト s に対して，検索対象との AND 検索により上位 N 件以内に検索されるウェブページ p の集合を $\mathbb{P}(s, N)$ (ただし，本論文においては，$N = 20$ とする) とし，各検索対象あたりのウェブページの文書集合 D_w を以下のように定義する．

$$D_w = \bigcup_{s \in \mathbb{S}} \mathbb{P}(s, N)$$

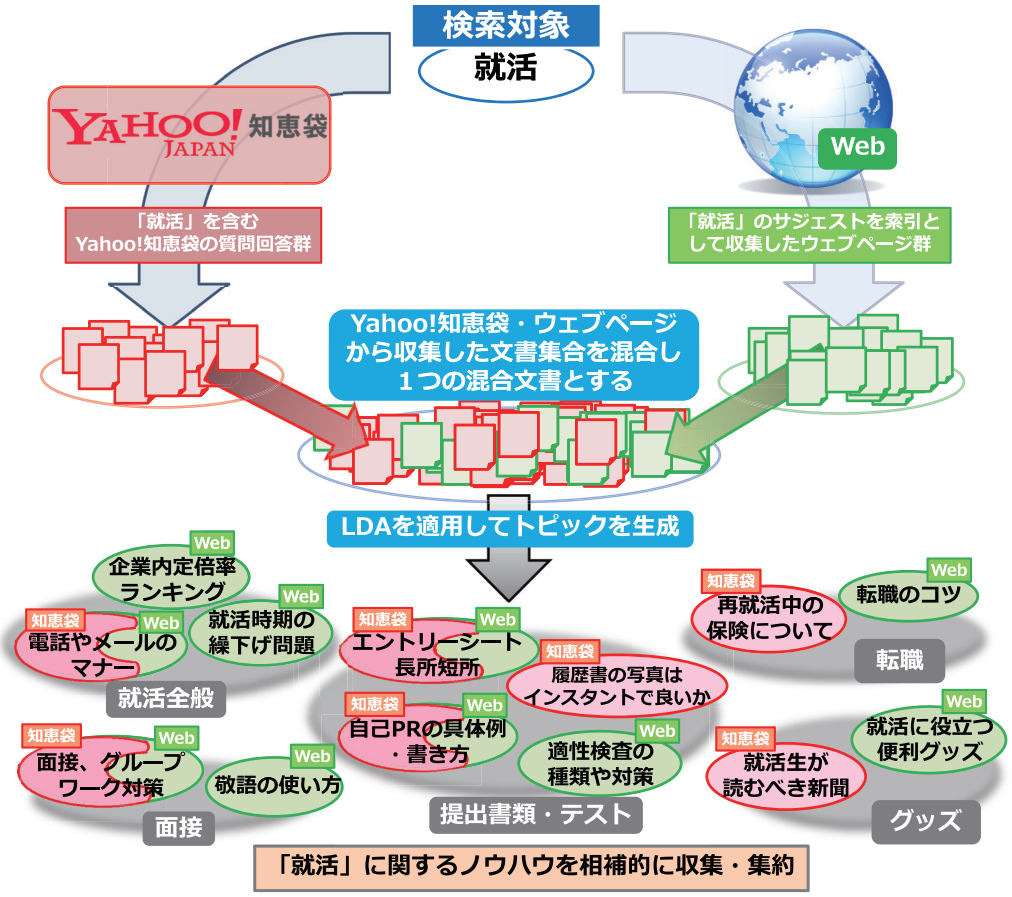

図 1: 質問回答サイトのノウハウ収集・集約およびウェブからの新ノウハウ補足の流れ

表 1: 各検索対象におけるサジェスト数，および，混合文書集合の記事数

検索対象	知恵袋記事数	ウェブページ		知恵袋記事数＋ウェブページ数
		サジェスト数	ページ数	
花粉症	14,059	872	11,144	25,203
結婚	35,426	956	14,409	49,835
就活	11,366	934	13,211	24,587

なお，ウェブページの収集には Yahoo! Search BOSS API [*2] を用いた．各ウェブページ p に対して，$p \in \mathbb{P}(s, N)$ となるサジェスト s を集めた集合を $\mathbb{S}(p)$ とし，以下のように定義する．

$$\mathbb{S}(p) \quad = \quad \left\{ s \in \mathbb{S} \;\middle|\; p \in \mathbb{P}(s, N) \right\}$$

4. トピックモデルの適用

2. 節および 3. 節で収集した質問回答事例の文書集合 D_q とウェブページの文書集合 D_w の混合文書集合 D_{qw} を作成する．すなわち，

$$D_{qw} \quad = \quad D_q \bigcup D_w$$

である．各検索対象における混合文書集合の記事数を表 1 に示している．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) [Blei 03] 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 の集合を V として，語 $w (w \in V)$ 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 b におけるトピック z_n の確率分布 $P(z_n|b)$ $(n = 1, \ldots, K)$ を推定する[*3]．本論文では，各文書に対してトピックを一意に割り当てることで，各文書を分類することとした．記事集合を D，トピック数を K，1 つの文書を $d\,(d \in D)$ とすると，トピック z_n $(n = 1, \ldots, K)$ の記事集合 $D(z_n)$ 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，文書 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，文書 d を割り当てていることになる．

5. ノウハウ知識の収集

以下の 5.1 節，5.2 節の手順に従ってノウハウ知識の収集を行う．

5.1 トピックモデル適用結果における話題分析の手順

4. 節の手順に従い，各トピックに割り当てられた確率上位 20 件の記事を分析したところ，トピックによっては，いずれかの

*2 http://developer.yahoo.com/search/boss

*3 推定のためのツールは GibbsLDA++を用いた．LDA のハイパーパラメータである α，β は，$\alpha = 50/K$，$\beta = 0.1$，Gibbs サンプリングの反復回数は 2,000，トピック数は $K = 50$ を用いた．

表 2: ノウハウ知識の話題数

検索対象	大分類の数	トピック数	話題数			
			質問回答サイト	ウェブ	質問回答サイト＋ウェブ	合計
花粉症	10	40	6	19	30	55
結婚	4	26	12	7	16	35
就活	6	33	15	17	17	49

情報源に偏るものがあった．そこで，今回の分析では，情報源ごとに確率上位 10 件の記事を分析し，そのうち 3 件以上同一とされる話題があった場合に，そのトピックの話題として抽出した[*4] ．これにより各トピックの情報源毎に最大 3 つの話題を抽出した．なお，話題分析の際には，各トピックにおける確率 $P(w|z_n)$ の高い語 w とトピック及びウェブページに割り当てられたサジェストを参照して分析を行う．

5.2 ノウハウ知識の人手選定

各トピックから得られた各話題を以下の 4 つに分類する．

1. ノウハウ知識
2. ノウハウ以外の知識
3. 意見
4. その他

以下，各分類について詳しく説明する．「ノウハウ知識」はやり方についての情報など閲覧した人の行動につながるものである．具体的にはレシピサイト，方法や手順が書かれているもの，対策やマナー，コツなどがノウハウ知識にあたる．本論文では，ユーザの行動につながる知識は全てノウハウ知識であるとみなした．収集されたノウハウ知識の話題数を表 2 に示す．「ノウハウ以外の知識」は，それを見てもユーザの行動に影響を与えない情報である．例えば，「花粉症が増えた背景」や「芸能人の結婚」がこれにあたる．「意見」は，多くの人の意見を求める相談や，自分の意見を主張しているものである．例えば，「花粉症で病院に行った際のトラブル」や「結婚後の嫁姑の問題」がこれにあたる．「その他」は，上記 3 つのいずれにも分類できないものである．例えば，「花粉症の広告」や「結婚占い」がこれにあたる．

6. ノウハウ知識の閲覧インタフェース

本論文では，5. 節の手順により収集したノウハウ知識を効率的に閲覧するインタフェースを作成した．作成したインタフェースの例を図 2 に示す．図 2 に示すように，収集したノウハウ知識をインタフェース画面左部分にてリスト形式に表示した．表示されているノウハウ知識を選択することにより，選択したノウハウ知識に割り当てられた質問回答事例，および，ウェブページの一覧をインタフェース画面右部分において閲覧する．ユーザは，一覧に表示されている質問回答事例，および，ウェブページを選択することにより，実際の記事内容を閲覧する．また，選択したノウハウ知識の情報源としてウェブページが含まれる場合には，そのウェブページに対応付けられているサジェストをインタフェース画面左部分のノウハウ知識の下にリスト形式で表示する．

*4 ここでの作業においては，[井上 15] で提案した作業インタフェースを用いる．また，異なるトピックから同一の話題が収集される場合においても，本論文の分析の範囲においては，別の話題として数えた．

7. 関連研究

先行研究として，特に，ノウハウ知識収集部分に関連して，[加藤 14] 等がある．この研究では，「部屋を掃除する」，「花粉症対策をする」といったクエリを実現するためのサブタスクを，行為を表す動詞表現の形式で収集する方式を提案している．また，2014 年 12 月開催の NTCIR-11[*5] においては，この論文の著者らによる主催で，この論文の課題とほぼ同様の仕様のもとでの Task Mining Task も実施されている．今後，本研究においても，本論文の手法を Task Mining Task で用いられたクエリリストおよび評価手順 [Liu 14] に適用し，有効性を検証する必要がある．ただし，Task Mining Task のタスク設定においては，クエリを実現するためのサブタスク群を動詞表現の形式で出力するだけにとどまっており，実際にそれらのサブタスクをどのようにして実現すればよいのかについてのノウハウ知識そのものを収集の対象とはしていない．一方，本研究において収集・集約の対象とするのは，質問回答事例あるいはウェブページ群の形式で表現されたノウハウ知識そのものであり，この点において上記の先行研究とは大きく異なっている．また，他の先行研究として，[高田 10] では，質問回答サイトに対する検索結果において，検索者の検索要求を満たす回答を数個選択した後，それらの回答に対する別解をウェブから収集する方式を提案している．一方，本研究においては，数個の質問回答事例における質問事項および回答といった小さい粒度のノウハウ知識を対象とするのではなく，質問回答事例およびウェブ検索結果を数万文書程度収集した結果に対して，多種多様なノウハウ知識を網羅的に収集するとともに，質問回答事例由来のノウハウ知識を補足する新ノウハウ知識を，一般のウェブページを情報源として収集・集約する方式を研究対象としている点が大きく異なっている．

8. おわりに

本論文では，ある検索対象について収集・集約されたノウハウ知識を俯瞰し，閲覧するインタフェースを提案した．

参考文献

[Blei 03] 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 15] 井上 祐輔, 守谷 一朗, 今田 貴和, 聶 添, 宇津呂 武仁, 神門 典子：質問回答事例および検索エンジン・サジェストを情報源とするノウハウ知識の収集インタフェース, 言語処理学会第 21 回年次大会論文集, pp. 700–703 (2015)

[加藤 14] 加藤 龍, 大島 裕明, 山本 岳洋, 加藤 誠, 田中 克己：タスクの汎化と特化に着目した Web からのタスク検索, 第 6 回 DEIM フォーラム論文集 (2014)

[Liu 14] Liu, Y., Song, R., Zhang, M., Dou, Z., Yamamoto, T., Kato, M., Ohshima, H., and Zhou, K.: Overview of the NTCIR-11 IMine Task, in *Proc. 11th NTCIR Workshop Meeting*, pp. 8–23 (2014)

[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)

[高田 10] 高田 夏希, 大島 裕明, 田中 克己：Web と QA コンテンツの相互補完に基づくソーシャルサーチ, WebDB Forum 2010 論文集 (2010)

*5 http://research.nii.ac.jp/ntcir/ntcir-11/index-ja.html

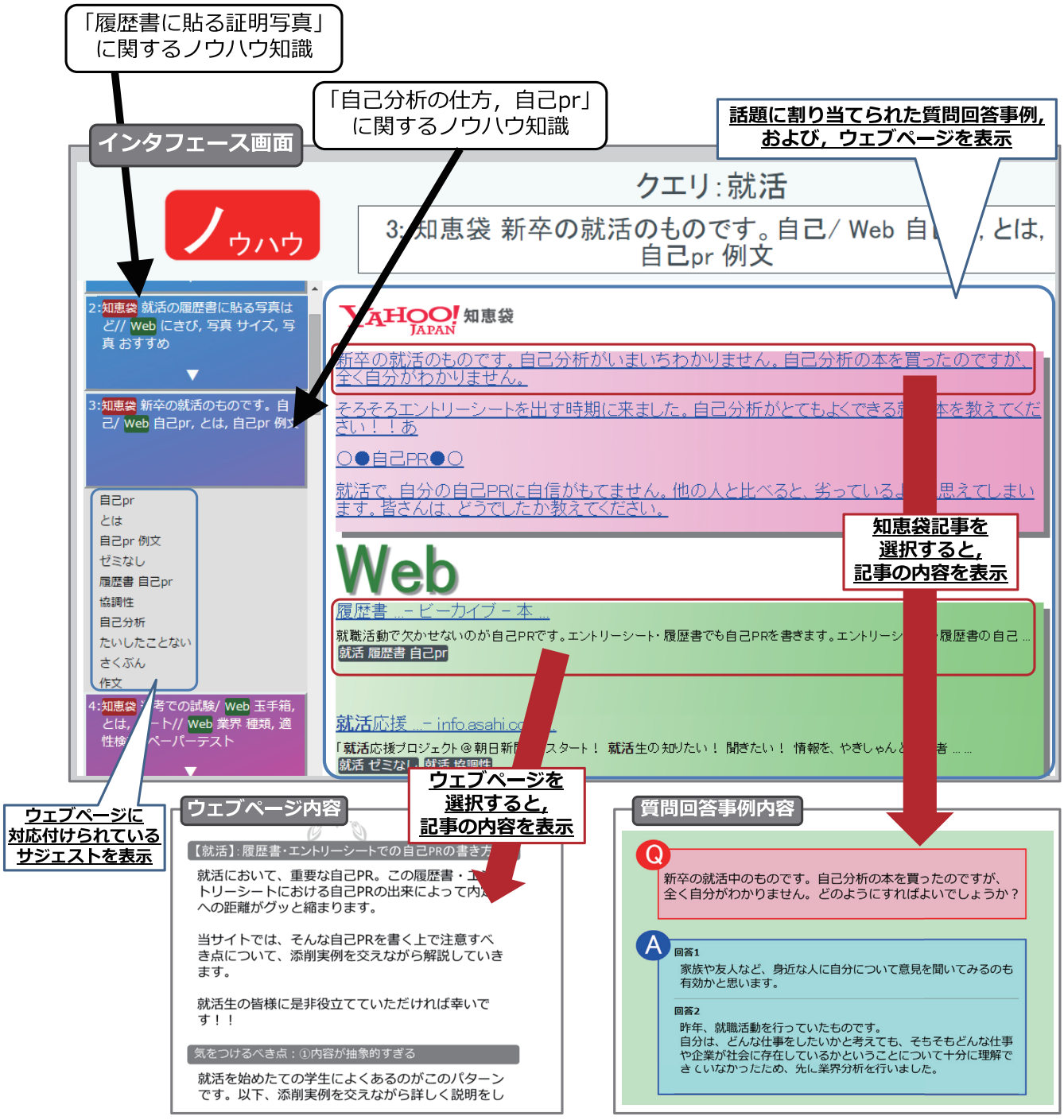

図 2: ノウハウ知識の閲覧インタフェース画面