DEIM Forum 2015 A1-4

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

守谷　一朗† 　今田　貴和† 　井上　祐輔† 　聶　　添† 　宇津呂武仁††

河田　容英††† 　神門　典子††††

† 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　〒 151-0051 東京都渋谷区千駄ヶ谷 5-13-18
†††† 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2

あらまし　本論文では，検索エンジン・サジェストを索引として収集される情報に加えて質問回答サイトから得られる情報を相補的に利用し，それらを混合して集約する手法を提案する．さらに，収集対象とする知識を，特に，特定の目的のもとでのノウハウに関する知識に制限することにより，有用性の高い知識を選択的に収集する枠組みを実現する． 本方式においては，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，それらの話題のまとまりからノウハウ知識を選定する．

キーワード　ノウハウ知識, 質問回答サイト, 検索エンジン・サジェスト, トピックモデル, 収集・集約

A Complementary Framework for Collecting Know-How Knowledge based on Question-Answer Examples and Search Engine Suggests

Ichiro MORIYA†, Takakazu IMADA†, Yusuke INOUE†, Tian NIE†, Takehito UTSURO††,

Yasuhide KAWADA†††, and Noriko KANDO††††

† 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 151-0051, Japan
†††† National Institute of Informatics, Tokyo 101-8430, Japan

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(注1) をはじめとする百科事典サイトや Yahoo!知恵袋(注2) 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．そこで，本研究では，ある検索対象についてのノウハウ知識を網羅的に収集し，集約・俯瞰する手法を確立する．ここで，質問回答サイトには多くのノウハウ知識が含まれているが，質問回答サイトだけでは十分でないことが想定される．例えば「結婚式の電報の文例」を考えると，質問回答サイトから得られる限定的な文例情報だけでなく，電報の文例を専門的に扱っているサイトが紹介している網羅的な文例情報を合わせて参照した方がより有益である．そこで，本論文では，質問回答サイトおよび一般のウェブページという二種類の情報源を併用することにより，ノウハウ知識を相補的に収集し，集約・俯瞰する手法を提案する．

本研究の全体の流れを図 1 に示す．本研究では，まず，質問

(注1)：http://www.wikipedia.org/
(注2)：http://chiebukuro.yahoo.co.jp/

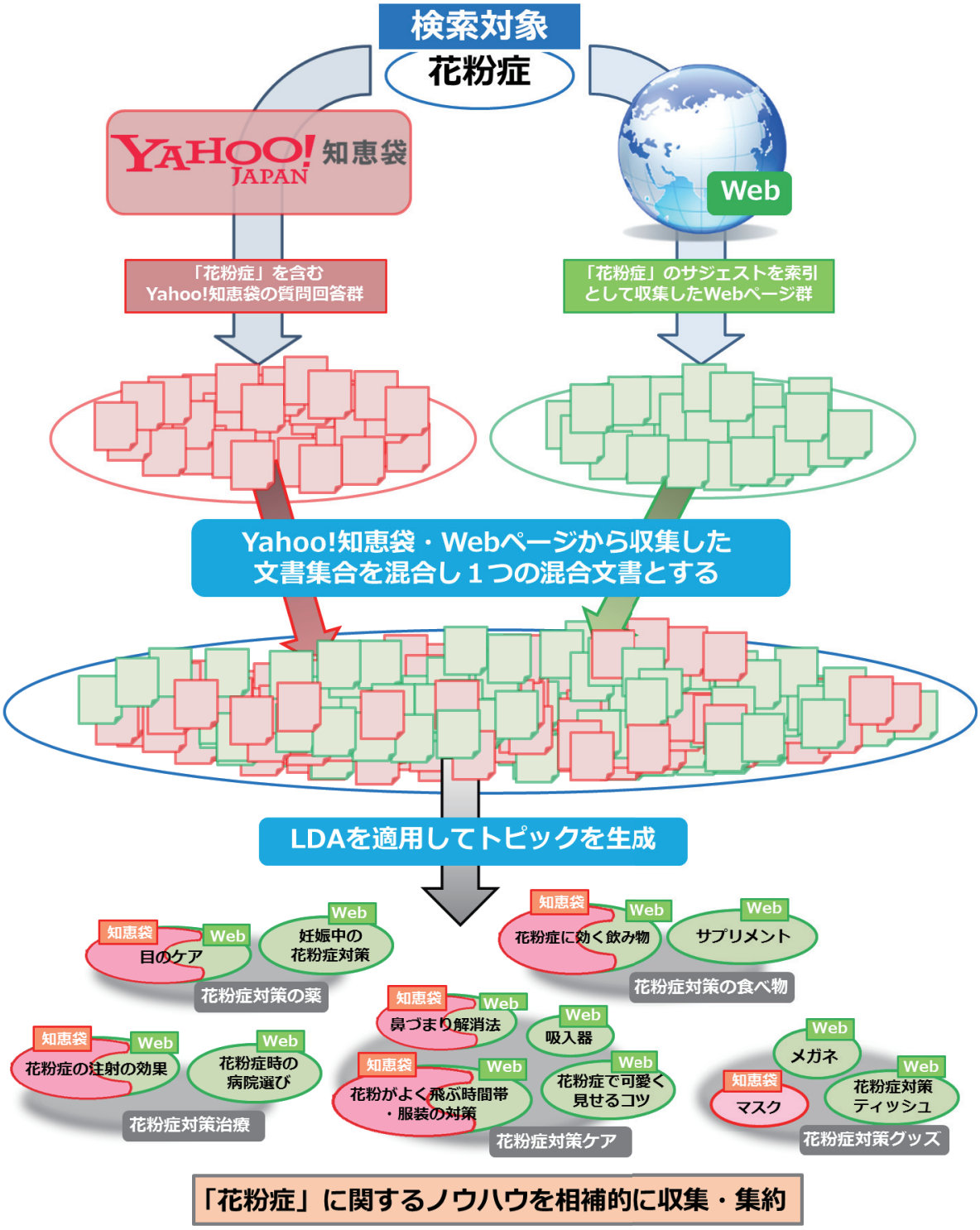

図 1　質問回答サイトのノウハウ収集・集約およびウェブからの新ノウハウ補充の流れ

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，話題のまとまりごとに二種類の情報源の記事を確率の高い順に 10 件ずつ目視にて分析し，3 件以上同一とされる内容の話題を情報源ごとに抽出する．次に，抽出した話題を，「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，ノウハウ知識を人手で選定する．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．また，検索対象「結婚」に関するノウハウ知識を収集した結果においては，合計 35 個の話題が収集された．収集された話題の中には，「結婚生活での夫婦円満の秘訣」のように，ウェブページからのみ得られるノウハウ知識が合計で 7 個であり，全話題の 20%と

表 1　各検索対象における LDA のトピック数 K

検索対象	質問回答サイト	ウェブページ集合	混合文書集合
花粉症	40	30	50
結婚	30	40	50

表 2　ノウハウ知識の話題数: 質問回答サイト単独での文書集合から収集／ウェブページ単独での文書集合から収集

検索対象	質問回答サイト単独			ウェブ単独		
	大分類の数	トピック数	話題数	大分類の数	トピック数	話題数
花粉症	10	28	31	10	19	24
結婚	3	16	18	3	15	18

なった．このように，本研究において，質問回答サイトのノウハウ知識を集約し，さらに，質問回答サイトには含まれないノウハウ知識をウェブページから補えることを示した．

2. トピックモデルを用いた文書集合中の話題の集約およびノウハウ知識の収集

2.1 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) [1] 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．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 には，GibbsLDA++ の基本設定値である $\alpha = 50/K$，$\beta = 0.1$ を用いた．LDA ではトピック数 K を人手で与える必要があるが，

今回は，トピック数を 10 から 100 程度まで変化させてトピック推定を行い，得られたトピックを人手で見比べ，トピックの推定結果の性能がより高くなったトピック数を採用するという手順を採った．なお，このツールは推定の際に Gibbs サンプリングを用いているが，その反復回数は 2,000 とした．

本論文で分析の対象とする検索対象「花粉症」および「結婚」について，文書収集対象として質問回答サイトおよびウェブページ集合を用いた場合について，それぞれ，まとまりが良いと判定し，分析に用いたトピック数 K の値を表 1 に示す．

また，本論文において，語 w の集合 V としては，日本語 Wikipedia 中のタイトルの集合[(注4)]を用いる．

また，GibbsLDA++では，各トピック z_n において確率 $P(w|z_n)$ の高い順に語 w を N 件出力することができる．本研究においては，$N = 20$ として，トピックの話題分析の際に参考情報として用いている．

2.2 文書に対するトピックの割り当て

本研究では，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．

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

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．

2.3 トピックモデル適用結果における話題分析の手順

本研究では生成された各トピックについて，割り当てられた文書 d の確率の高い順に 20 件を分析し，6 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．また，話題分析の際には，各トピック z_n における確率 $P(w|z_n)$ の高い語 w を参照した．これにより，一トピックあたり最大 3 つの話題が含まれることになる．

例として，5. 節において，検索対象「花粉症」について，質問回答サイトおよびウェブの混合文書集合を用いて行った話題分析の場合について述べる．この例においては，トピック数を 50 として LDA を適用したところ，意味的にまとまったトピックは 42 個となった．これらのトピックに対して話題分析を行い，59 個の話題が選定された．これらのトピックおよび話題の例として，例えば,「花粉症対策のメガネやマスク」のトピックにおいては,「オススメのマスクやメガネが曇りにくいマスク」の話題，および、「花粉症対策のメガネ」の話題が含まれていた．

2.4 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[(注5)]．

（1） ノウハウ知識
（2） ノウハウ以外の知識
（3） 意見
（4） その他

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．本研究では，ユーザの行動につながる要素があればノウハウ知識であると考え，判定を行った．例えば，検索対象「花粉症」について，前節の手順の例において同定された 59 個の話題のうち，ノウハウ知識であると判定された話題は 55 個であった．

「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．

(注3)：http://gibbslda.sourceforge.net/

(注4)：日本語 Wikipedia としては，2014 年 3 月にダウンロードした，エントリ数約 140 万 7,000 のものを用いた．

(注5)：ここでの分類作業においては，文献 [6] で提案した作業インタフェースを用いる．

また，ノウハウ知識であると判定された話題については，さらに，意味的なまとまりである大分類への分類を行う．例えば，検索対象「花粉症」について，前節の手順の例において同定された 55 個のノウハウ知識に関する話題の場合は，「花粉症対策治療」，「花粉症対策の薬」，「花粉症対策ケア」などの 10 個の大分類にまとめられた．

3. 質問回答サイトからのノウハウ知識の収集・集約

3.1 質問回答事例の収集

本研究では，質問回答サイトのデータとして，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．質問には，カテゴリ情報が付与されており，最下位層の分類として 453 種のカテゴリが存在している．453 種のカテゴリは，それぞれ親カテゴリ，さらにその親カテゴリを持つ三層構造になっており，各カテゴリに数万 ~ 数十万の質問が含まれている．

本研究では，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を作成した．この一つの質問回答事例を d_q とする．各検索対象あたりの質問回答事例の文書集合を D_q とし，次のように定義する．

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

なお，「結婚」については，知恵袋の記事数が 357,760 件と多かったために，ランダムで 50,000 件を抽出し，知恵袋のみで LDA を適用したあと，話題分析を行い，ノイズと判定したトピックの記事を取り除くという手順を用いた．

3.2 トピックモデルの適用およびノウハウ知識の人手選定

2. 節の手順に従い，質問回答事例の文書集合 D_q に LDA を適用し，文書に対するトピックの割り当て，話題分析，ノウハウ知識の人手選定を行う．検索対象毎の記事数については表 4 の知恵袋記事数に等しい．

検索対象「花粉症」および「結婚」について収集されたノウハウ知識の話題数を表 2 に示す．

検索対象「花粉症」については，合計 31 個のノウハウ知識の話題が収集された．収集された話題の例として，「妊娠中の花粉症対策」についての話題では，花粉症の薬の胎児への影響についてのノウハウ知識を得ることができる．一方，検索対象「結婚」については，合計 18 個のノウハウ知識の話題が収集された．収集された話題の例として，「結婚後の家具について」についての話題では，家具にかかる費用や婚礼家具選びのノウハウ知識を得ることができる．

4. 検索エンジン・サジェストを用いたウェブからのノウハウ知識の収集・集約

4.1 概　　要

本節では，検索エンジン・サジェストを用いて得られるウェブページ集合からノウハウ知識を収集する方法について述べる．

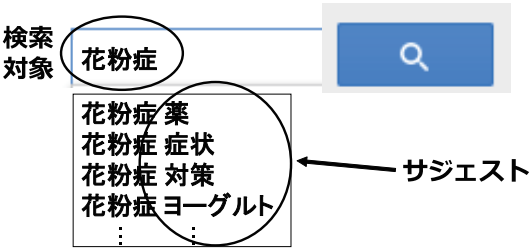

図 2　検索エンジン・サジェストの例

表 3　検索対象，および，サジェスト数

検索対象	サジェスト数
花粉症	872
結婚	956

なお，検索エンジン・サジェストの収集，およびウェブページの収集は 2014 年 6 月から 7 月にかけて行った．

4.2 検索エンジン・サジェスト

各検索エンジン会社においては，ウェブ検索者の検索ログが蓄積されており，多数のウェブ検索者が検索したキーワードに対して，検索者が強い関心を持つ語を抽出し，検索エンジン・サジェストとして提示するサービスを提供している．ここで，本論文では，詳細な情報を検索したい対象を**「検索対象」**と呼ぶ．また，検索対象に対して，検索者が AND 検索の形で二つ目以降のキーワードとして指定し，検索対象に対して詳細な情報を得るために用いる観点を「情報要求観点」と呼ぶ．すると，検索エンジン・サジェストとして提示される言葉は，「検索対象」に対して，多数のウェブ検索者が「情報要求観点」として指定した語に相当しており，ウェブ検索者の関心事項そのものを反映していることが分かる(注6)．そこで，本論文では，検索エンジン・サジェストに着目することによって，ウェブ検索者に焦点を当て，ウェブ検索者の関心事項の収集を行う．

4.3 検索エンジン・サジェストの収集

選定した評価用検索対象に対して，Google(注7) 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．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」などの開拗音である．例えば検索窓に「花粉症 た」と入力すると，「対策」や「食べ物」など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．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．本論文で分析の対象とする検索対象「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 3 に示す．

4.4 検索エンジン・サジェストを用いたウェブページの収集

$s \in \mathbb{S}$ 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

(注6)：図 2 の例では，検索窓に「花粉症 」を入力すると，「薬」，「症状」，「対策」などが検索エンジン・サジェストとして提示される．この例では，「花粉症」が検索対象であり，「薬」，「症状」，「対策」等がサジェストである．また，実際の検索ログにおいては，「花粉症 AND 薬」のように，検索対象とサジェストの AND 検索の形式で表現された検索要求が蓄積されている．

(注7)：https://www.google.com/

表 5　ノウハウ知識の話題数: 質問回答サイト・ウェブページの混合文書集合から収集
(a) 混合文書集合から生成されたトピック全体に含まれるノウハウ知識

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

(b) 混合文書集合から生成されたトピックのみに含まれるノウハウ知識

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	9	14	2	9	6	17
結婚	3	11	5	4	3	12

表 4　各検索対象における混合文書集合の記事数

検索対象	知恵袋記事数	ウェブ記事数	合計
花粉症	14,059	11,144	25,203
結婚	35,426	14,409	49,835

$\mathbb{P}(s,N)$ (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．

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

なお，ウェブページの収集には Yahoo! Search BOSS API (注8) を用いた．

4.5　ウェブページに対するサジェストの割り当て

各ウェブページは，検索対象および各サジェストの AND 検索によって検索されたものである．したがって，あるウェブページには，一つ以上のサジェストが対応することになる．

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．

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

4.6　トピックモデルの適用およびノウハウ知識の人手選定

2. 節の手順に従い，ウェブページの文書集合 D_w に LDA を適用し，文書に対するトピックの割り当て，話題分析，ノウハウ知識の人手選定を行う．検索対象毎の記事数については表 4 のウェブ記事数に等しい．

また，各ウェブページには，トピックが対応付けられている．一つのトピックに対して割り当てられた一つ以上のウェブページに対応するサジェストを収集することにより，一つのトピックに一つ以上のサジェストが割り当てられていることになる．あるトピック z_n^w に割り当てられたウェブページ集合を $D(z_n^w)$ とすると，トピックに割り当てられたサジェスト集合 $\mathbb{S}(z_n^w)$ は以下のようになる．

$$\mathbb{S}(z_n^w) = \bigcup_{p \in D(z_n^w)} \mathbb{S}(p)$$

話題分析を行う際には，$\mathbb{S}(z_n^w)$ 中のサジェストのうち頻度上位 20 個を参照することによって話題を分析する．

検索対象「花粉症」および「結婚」について，収集されたノウハウ知識の話題数を表 2 に示す．検索対象「花粉症」については，合計 24 個のノウハウ知識の話題が収集された．収集された話題の例として,「花粉症を悪化させる食べ物」についての話題では，揚げ物など花粉症の際に注意すべき食べ物についてのノウハウ知識を得ることができる．一方，検索対象「結婚」については，合計 18 個のノウハウ知識の話題が収集された．収集された話題の例として,「招待状のマナー」についての話題では，結婚式で招待状を返信する際のマナーに関するノウハウ知識を得ることができる．

5.　質問回答サイトおよびウェブからのノウハウ知識の相補的収集

5.1　二種類の情報源からの混合文書集合の作成

3.1 節および 4.4 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，

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

である．各検索対象における混合文書集合の記事数を表 4 に示している．

5.2　トピックモデルの適用およびノウハウ知識の人手選定

2. 節の手順に従い，混合文書集合 D_{qw} に LDA を適用し，文書に対するトピックの割り当て，話題分析，ノウハウ知識の人手選定を行う．

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(注9)．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．収集されたノウハウ知識の話題数を表 5(a) に示す．

以下に，ノウハウ知識以外に分類した話題の例を挙げる．検索対象「花粉症」においては,「環境問題と花粉症」,「花粉症の

(注8)：http://developer.yahoo.com/search/boss

(注9)：ここでの作業においては，文献 [6] で提案した作業インタフェースを用いる．

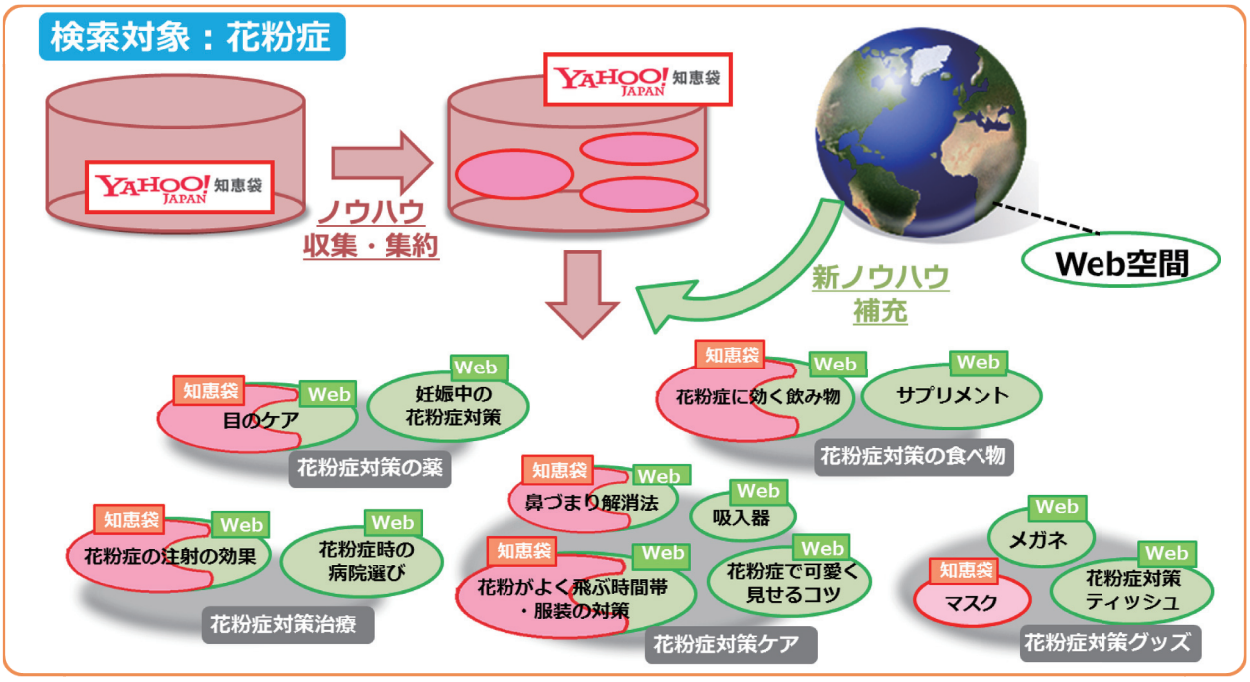

図 3　質問回答サイトのノウハウ収集・集約およびウェブからの新ノウハウ補足の例 (検索対象:「花粉症」)

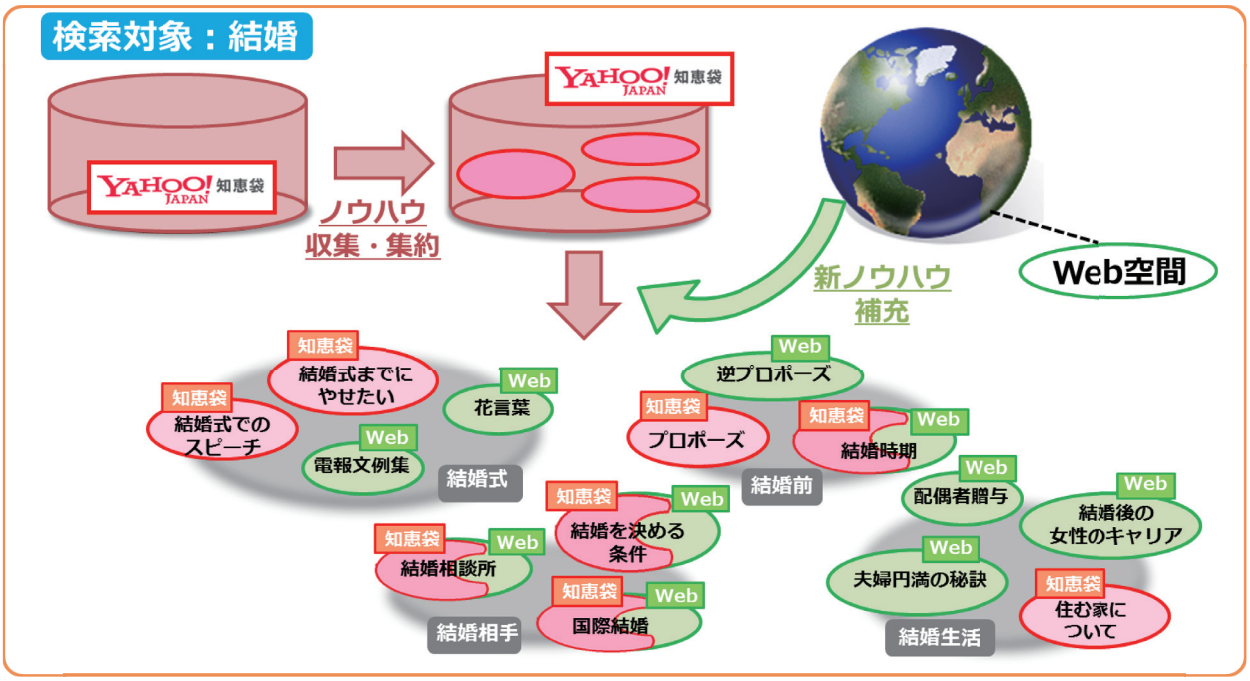

図 4　質問回答サイトのノウハウ収集・集約およびウェブからの新ノウハウ補足の例 (検索対象:「結婚」)

研究」をノウハウ以外の知識,「病院の診察時のトラブル」を意見,「花粉症の広告」をその他に分類した．検索対象「結婚」においては,「芸能人の結婚」等をノウハウ以外の知識,「結婚相手の外見についての相談」等を意見,「結婚占い」をその他に分類した．

5.3　ノウハウ知識収集結果の分析

5.3.1　情報源ごとのノウハウ知識の分析

表 5(a) に示すように，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．一方で質問回答サイトのみから得られるノウハウ知識は合計で 6 個あり，全話題の約 11%となった．一方,「結婚」に関するノウハウ知識を収集した結果においては，合計 35 個の話題が収集された．収集された話題の中には,「結婚生活での夫婦円満の秘訣」のように,ウェブページからのみ得られるノウハウ知識が合計で 7 個であり，全話題の 20% となった．一方で質問回答サイトのみから

表 6　各検索対象におけるウェブページ集合に含まれる質問回答サイトの割合 (%)

検索対象	確率上位 10 ページ		ウェブページ全体	
	Yahoo!知恵袋	質問回答サイト全体	Yahoo!知恵袋	質問回答サイト全体
花粉症	5.0 (25/500)	8.4 (42/500)	8.5 (946/11,144)	16.6 (1,847/11,144)
結婚	5.6 (28/500)	16.8 (84/500)	7.0 (1,007/14,409)	22.1 (3,179/14,409)

得られるノウハウ知識は合計で 12 個あり，全話題の約 34%となった．

質問回答サイトおよびウェブを情報源として相補的にノウハウ知識を収集した結果の抜粋を図 3 および図 4 に示す．

収集された話題の例として,「おすすめのマスクやメガネが曇りにくいマスク」についての話題では,「メガネを曇りにくくする方法」として「快適ガードプロ」や「ノーズマスク・ピット」というマスクを着けると良いといったノウハウ知識を得ることができる．一方，検索対象「結婚」については,「新婚旅行について」の話題等が質問回答サイトのみから収集された．

収集された話題の例として,「花粉症の温熱治療のための吸入器」についての話題では,「花粉症の温熱治療のための吸入器」として「ホットシャワー 3」という超音波吸入器等についてのノウハウ知識が得られる．その他にも,「花粉症対策のメガネ」の話題では,「スポーツをする際の花粉症対策としてオススメのメガネ」や「花粉症対策専用のメガネ」に関するノウハウ知識が得られる．また,「花粉症で可愛く見せるコツ」という話題では,「花粉症を逆手に取ったドライブデート中の必殺テクニック」といった独特のノウハウ知識が得られた．検索対象「結婚」についても,「結婚祝い電報の例文集」や「夫婦円満の秘訣」等のノウハウ知識の話題がウェブのみから得られた．

収集された話題の例として,「鼻づまり解消法」についての話題では,「鼻づまり解消法」として「鼻うがい」,「ブリーズライトを使う」等のノウハウ知識を得ることができる．ここで，質問回答サイトおよびウェブの双方において同一の内容であると判定した話題においても，部分的に内容に異なりがある場合がある．例えば,「鼻づまり解消法」の話題においては,「ブリーズライトを使う」というノウハウ知識はウェブからしか収集されていない．このように，詳細なノウハウ知識の中には，片方の情報源からしか収集できないノウハウ知識が存在する．

また，今回収集したウェブページ集合においては，Yahoo!知恵袋やその他の質問回答サイトも含まれている．検索対象「花粉症」および「結婚」について，ウェブページ集合 D_w 中に含まれる質問回答サイトの割合を表 6 に示す(注10)．ただし，質問回答サイトとしては，Yahoo!知恵袋およびその他の質問回答サイト(注11) を区別して割合を算出した．表 6 からわかるように，ウェブページ集合における Yahoo!知恵袋の影響は最大でも 8.5%であり，その影響は小さいと考えられる．また，その他の質問回答サイトを含めた場合の割合においても，その影響は最大で 2 割程度であり，残りの 8 割は質問回答サイト以外のウェブページである．ただし，トピックモデルを適用した結果においては，トピックごとに質問回答サイトの割合に偏りが生じると考えられるので，より詳細な分析を行う必要がある．

5.3.2　質問回答サイトおよびウェブページの文書混合方式の有効性の分析

表 5(b) に示すように，検索対象「花粉症」および「結婚」において，質問回答サイトおよびウェブページの混合文書集合から収集されたノウハウ知識のうち，約 3 割は質問回答サイトまたはウェブページ単独の文書集合からは収集できなかった話題であった．このことから，質問回答サイトおよびウェブページを混合することによって，有用なノウハウ知識が新たに収集可能であることが示された．具体的には，検索対象「花粉症」においては，合計 55 個の話題のうち，17 個が混合文書集合から生成されたトピックにのみ含まれる話題であった．例えば,「花粉症時の病院選び」や「花粉症対策の服装・帽子や外出・帰宅時のケア」等の話題がこれらの話題に該当する．一方，検索対象「結婚」においては，合計 35 個の話題のうち，12 個が混合文書集合から生成されたトピックにのみ含まれる話題であった．例えば,「プロポーズのタイミングや結果について」,「逆プロポーズについて」,「結婚式でのスピーチでの話し方」,「配偶者贈与について」等の話題がこれらの話題に該当する．

以上の結果から，二種類の情報源から収集された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，有用なノウハウ知識を新たに発見することができることがわかった．

6. 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，文献 [7] 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(注12) 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．NTCIR-11 では，Task Mining Task の研究として，ウェブページを用いた手法 [15] や質問回答サイトを用いた手法 [11] が採用されており，一定の成果を挙げている．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 [10] 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

(注10)：表 6 における「確率上位 10 ページ」とは，LDA の各トピックに割り当てられた確率上位 10 件のウェブページを指す．

(注11)：本論文では，チエノワ (chienowa-qa.com)，Yahoo!知恵袋 (chiebukuro.yahoo)，発言小町 (komachi.yomiuri)，OKWave(okwave)，@nifty 教えて広場 (oshiete1.nifty)，教えて!goo(oshiete.goo)，人力検索はてな (q.hatena)，エキサイトみんなの相談広場 (qa.excite)，楽天みんなで解決!Q&A(qanda.rakuten)，Sooda!(sooda.jp)，BIGLOBE なんでも相談室 (soudan1.biglobe) のいずれかを URL に含むものを抽出した．

(注12)：http://research.nii.ac.jp/ntcir/ntcir-11/index-ja.html

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，特に，質問回答サイトおよびウェブからの相補的な知識収集の部分に関連して，文献 [14] がある．この研究では，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．

その他，ウェブからノウハウを発見することを目的とした研究として，文献 [3] においては，ノウハウに関連する単語を抽出したものを手がかりとして，ウェブ上のノウハウ情報を効率的に収集する手法を提案している．これに対して，本研究においては，質問回答サイトの情報およびトピックモデルを利用しており，これらの点が大きな違いである．一方，文献 [9] においては，モノとその使われ方に着目してノウハウを収集する手法を提案している．具体的には，手がかり情報や品詞情報等の言語表現のパターンを用いて，ノウハウか否かの判定を行っている．本論文においても，今後，これらの手法を導入することによって，ノウハウ知識の自動判定を実現する必要がある．

また，本研究の前段の研究 [2,4,8,12] においては，検索エンジン・サジェストを情報源とすることにより，ウェブ検索者の関心の高い知識を優先的・選択的に収集する方式を提案している．しかし，知識を収集・集約した結果においては，多種多様な有用性の高い知識だけにとどまらず，有用性の低い知識や瑣末的な興味に基づく関心事項に関する知識が混在するという問題も散見された．この問題に対して，本論文の方式は，有用性の低い知識や瑣末的な興味に基づく関心事項に関する知識の混在による悪影響を軽減し，有用性の高い知識を選択的に収集・集約することを目的とした手法として位置付けることができる．一方，文献 [5] では，本論文の手法によって収集したノウハウ知識を閲覧するインタフェースを提案している．また，文献 [13] では，本論文の手法によって日中二言語において収集したノウハウ知識を二言語間で比較対照分析した結果を紹介している．

7. おわりに

本論文では，質問回答サイトおよびウェブからノウハウ知識を相補的に収集する手法を提案した．特に，Yahoo!知恵袋から得た質問回答事例と，検索エンジン・サジェストを索引として収集したウェブページ文書の混合文書に対して，トピックモデルの LDA を適用し，各トピックの確率上位の文書の内容を分析することで，検索対象に対するノウハウ知識を幅広く収集した．実際の分析例においては，「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．「花粉症の温熱治療のための吸入器」の話題など，そのうち 19 個の話題がウェブページからのみ得られ，全話題の約 35%となった．「結婚」に関するノウハウ知識を収集した結果においては，合計 35 個の話題が収集された．「結婚生活での夫婦円満の秘訣」の話題など，そのうち 7 個の話題がウェブページからのみ得られ，全話題の 20%となった．このように，本研究において，質問回答サイトのノウハウ知識を集約し，さらに，質問回答サイトには含まれないノウハウ知識をウェブページから補えることを示した．今後の課題として，人手によって選定したノウハウ知識を正例として分類器学習手法を適用することにより，ノウハウ知識を自動判定する方式を実現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] 土井俊弥, 井上祐輔, 今田貴和, 宇津呂武仁, 河田容英, 神門典子. トピックモデルを用いた検索エンジン・サジェストの集約. 第 29 回人工知能学会全国大会論文集, 2015.

[3] 服部元, 武吉朋也, 小野智弘, 滝嶋康弘. Web からのノウハウ検索手法の提案. 電子情報通信学会技術研究報告, NLC2009-35, pp. 13–18, 2010.

[4] 井上祐輔, 今田貴和, 守谷一朗, 陳磊, 宇津呂武仁, 河田容英, 神門典子. 冗長な情報要求観点の集約によるウェブ検索結果の集約. 第 28 回人工知能学会全国大会論文集, 2014.

[5] 井上祐輔, 今田貴和, 宇津呂武仁, 河田容英, 神門典子. 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. 第 29 回人工知能学会全国大会論文集, 2015.

[6] 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.

[7] 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.

[8] 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.

[9] 小澤俊介, 内元清貴, 松原茂樹. モノの使われ方の情報がノウハウ獲得に与える影響. 電子情報通信学会論文誌, Vol. J95-D, No. 3, pp. 506–517, 2012.

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

[11] S. Mine, T. Matsumoto, T. Yoshida, T. Shinohara, and D. Kitayama. InteractiveMediaMINE at the NTCIR-11 IMine search task. In *Proc. 11th NTCIR Workshop Meeting*, pp. 84–87, 2014.

[12] 守谷一朗, 小池大地, 今田貴和, 宇津呂武仁, 河田容英, 神門典子. Wikipedia 掲載事項との間の差分に着目したウェブ検索者の情報要求観点の分析. 第 6 回 DEIM フォーラム論文集, 2014.

[13] 聶添, 守谷一朗, 井上祐輔, 今田貴和, 李雪山, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析. 言語処理学会第 21 回年次大会論文集, pp. 948–951, 2015.

[14] 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.

[15] T. Yumoto. University of Hyogo at NTCIR-11 TaskMine by dependency parsing. In *Proc. 11th NTCIR Workshop Meeting*, pp. 24–27, 2014.