DEIM Forum 2017 D3-4

トビ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

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††††

† 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
†††† 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 This paper presents techniques of retrieving know-how sites from the collection of Web pages. The proposed techniques are designed to discover the maximum possible amount of know-how knowledge from such collections of Web pages, where know-how knowledge is defined as text contents qualified as information source regarding specific domain of questions. Techniques in this paper primarily collect sites, spanning multiple topics as know-how sites by classifying Web pages aggregated by a topic model. Collected sites are manually verified to counter check whether the sites are truly supplemental know-how knowledge.

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，話題分布，収集・集約

Collecting Know-How Sites based on the Topic Distribution per Site

Jiaqi LI†, Chen ZHAO†, Youchao LIN†, Mizuho BABA†, 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 151-0053, Japan
†††† National Institute of Informatics, Tokyo 101-8430, Japan

1. はじめに

現在，インターネットの普及やソーシャルメディアの隆盛により，多くのユーザはウェブページから日常生活に役立つ知識を得ている．様々な知識をまとめたウェブサイトが多数存在する．Yahoo!知恵袋等の質問回答サイトでは,「就活の準備方法」,「結婚式スケジュールガイド」など，ユーザにとって，日常生活の手助けになるノウハウ知識が掲載されている．しかし，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おいては，多種多様な内容および質の情報が散在しているため，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必要とされている．このような要求を満たすことを目的として，[4] では，指定されたクエリ・フォーカスにおいて，ノウハウ知識の候補を多く収集し集約・俯瞰している．

これに対して，本論文では，[4] において，ノウハウ知識を含むサイトの候補として収集されたサイト群に対して，実際にノウハウ知識を多く掲載したノウハウサイト，および，それ以外の一般サイトを識別する手法について提案する．具体的には，[4] においてウェブページ (および質問回答事例) を収集した結果の文書集合に対してトピックモデルを適用した結果において，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のドメインである可能性が高いという仮説を提案し，その有用性を評価する．本論文では，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．そして，トピックモデルの適用結果から，各トピック上位 30 件のウェブページ文書を抽出する．得られた文書集合から複数トピックにまたがって出現するドメインを収集し，ノウハウサイトのドメインであるか否かを人手によって評価する．

2. 質問回答事例の収集

本論文では，Yahoo!知恵袋から提供された 5 年間 (2004 年 4 月 1 日 ~2009 年 4 月 7 日) の質問回答事例のデータを用いた．クエリ・フォーカス q において，カテゴリ名，質問タイトル，質問本文の中に含まれている質問を抽出して，その質問に基づく回答本文全てを結合し，一つの質問回答事例の文書集合 d_q を作成する．各クエリ・フォーカス q に対応する質問回答事例の文書集合を D_q として，D_q を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

本論文では，分析の対象であるクエリ・フォーカス「就活」，「結婚」，および,「花粉症」について，検索エンジン・サジェストを利用してウェブページを収集する．そのためにまず，Google 検索エンジンを用いて，一クエリ・フォーカスにおいて，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．クエリ・フォーカス「就活」,「結婚」，および,「花粉症」において，収集されたサジェストの数を表 1 に示す．ここで，クエリ・フォーカスに対して収集されたサジェストの集合 $\mathbb{S}$ として，あるサジェスト $s \in \mathbb{S}$ に対して，クエリ・フォーカスとの 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 p の集合を $\mathbb{P}(s, N)$ とする．本論文では，$N = 20$ とする．各クエリ・フォーカスに対して収集されるウェブページの集合 D_w を次式で定義する．

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

ウェブページの収集においては，Yahoo! Search BOSS API(注1) を用いた．各ウェブページ p を検索する際に指定されたサジェスト s を集めた集合 $\mathbb{S}(p)$ を次式で定義する．

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

4. トピックモデルを用いた文書集合の集約

2. 節で収集した質問回答事例の文書集合 D_q，および，3. 節で収集したウェブページの文書集合 D_w の混合文書集合

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

を作成する．各クエリ・フォーカスに対する混合文書集合の文書数を表 1 に示す．

(注1)：http://developer.yahoo.com/search/boss

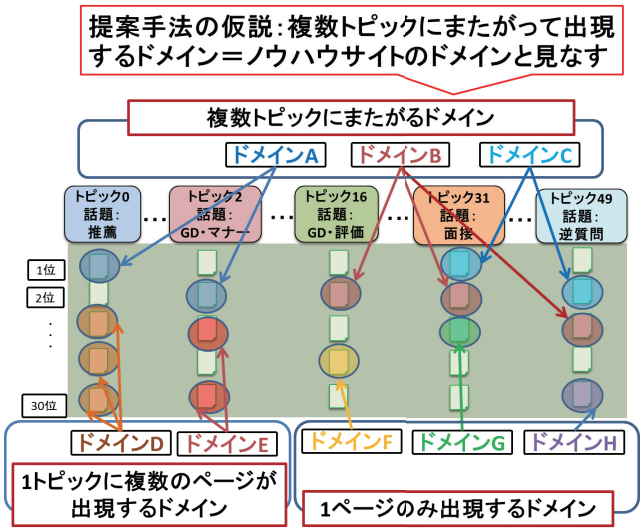


図 1 トピックモデルにおけるサイトの分布に基づくノウハウサイト発見手法の考え方

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 を用いてトピックモデルを推定する際には，語 w の集合を V として，語 $w (w \in V)$ の列により表現された文書の集合およびトピック数 K を指定して，各トピック $z_n \quad (n = 1, \ldots, K)$ 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 $P(z_n|d)$ $(n = 1, \ldots, K)$ を推定する．

本論文では，文書 d におけるトピックの確率分布において確率が最大となるトピックを文書 d に割り当てる．その結果，全文書集合を D として，トピック $z_n (n = 1, \ldots, K)$ が割り当てられた文書の集合 $D(z_n)$ は次式で表される．

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

5. ノウハウサイトの収集

本論文では，前節において，混合文書集合に対してトピックモデルを適用し，トピックごとに話題集約を行なった結果に対して，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のドメインである可能性が高いという仮説を提案しその有用性を評価する．

まず，あるドメインを dm_x とし，異なるトピック z_i および z_j $(i \neq j)$ を割り当てられた文書 d，および，d' が存在し，それらの URL を $u(d)$，および，$u(d')$ とする．ここで，$dm(u(d))$，および，$dm(u(d'))$ がともにあるドメイン dm_x と等しいとき(注2)，dm_x を複数トピックにまたがるドメインとする．

$$\begin{aligned} &\exists i, \exists j, i \neq j, \exists d \in D(z_i), \exists d' \in D(z_j) \\ &dm(u(d)) = dm(u(d')) = dm_x \end{aligned} \tag{1}$$

(注2)：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．

表 1　各クエリ・フォーカスにおけるサジェスト数，および，混合文書集合の記事数

クエリ・フォーカス	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	
就活	11,366	934	13,211	24,587
結婚	35,426	956	14,409	49,835
花粉症	14,059	872	11,144	25,203

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			A 群
ドメインそのものはノウハウ知識を提示しない	ドメインの中の下位ページにノウハウ知識を提示するページが存在する	ノウハウ知識を提示するページに容易に辿り着ける	B 群
		ノウハウ知識を提示するページに容易には辿り着けない	C 群
	ドメインの中の下位ページにもノウハウ知識を提示するページが存在しない	末端ページ中にノウハウ知識が提示されているページが存在する	D 群
		ノウハウ知識を提示するページが存在しない	E 群

本論文では，

> 上記の式 (1) を満たし，トピックモデル推定結果において複数のトピックにまたがるドメイン dm_x に対して，ノウハウサイトのドメインである可能性が高い．

という仮説を立ててその有用性を評価する．

6. 評　　価

6.1 評価対象のウェブページ

表 1 の各クエリ・フォーカスに対して，質問回答事例 D_q およびウェブページ集合 D_w の混合文書集合 D_{qw} を対象とした場合，および，ウェブページ集合 D_w のみを対象とした場合の二通りについて，評価実験を行う．トピック数 $K = 50$ としてトピックモデルを適用した結果から，各トピックの確率上位 30 件のウェブページ集合のみを抽出して，ノウハウサイトか否かの評価を行う．ここで，収集された各ウェブページに対して，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各ブログ，の各サイトのページにおいては，

> 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，

ことを期待できる可能性が低いとみなし，それらのサイトから収集されたウェブページを機械的に排除し，それ以外のウェブページのドメインを評価対象とする

6.2 評 価 手 順

本論文では，図 1 に示すように，評価対象のウェブページのドメインを以下の三種類に分類する．

(a)　複数トピックにまたって出現するドメイン (5. 節の式 (1))

(b)　1 トピックに複数のページが出現するドメイン — トピック z_i において次式を満たすウェブページ d および d' が存在するが，ドメイン dm_x を持つウェブページはいずれもトピック z_i を割り当てられており，z_i 以外のトピックを割り当てられたウェブページは収集されていない．

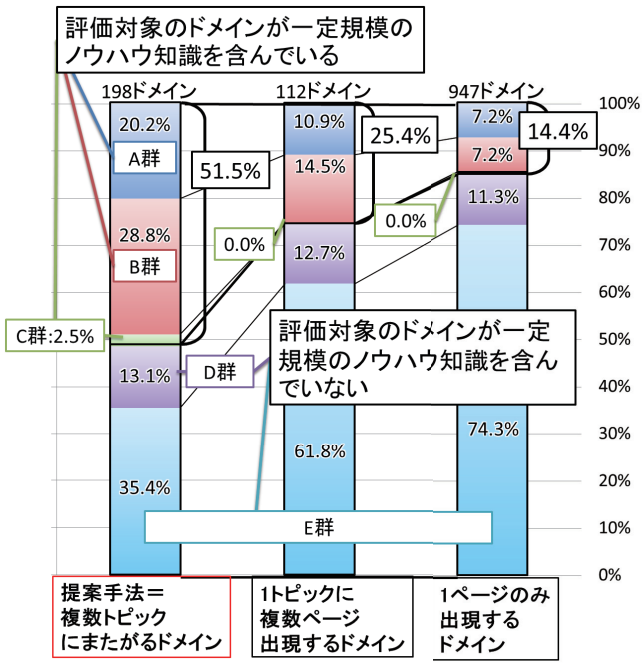
$$\exists i, \exists d, d' \in D(z_i), d \neq d' \\ dm(u(d)) = dm(u(d')) = dm_x \tag{2}$$

(c)　1 ページのみ出現するドメイン — (a) および (b) 以外．すなわち，ドメイン dm_x を持つウェブページは 1 ページしか収集されていない．

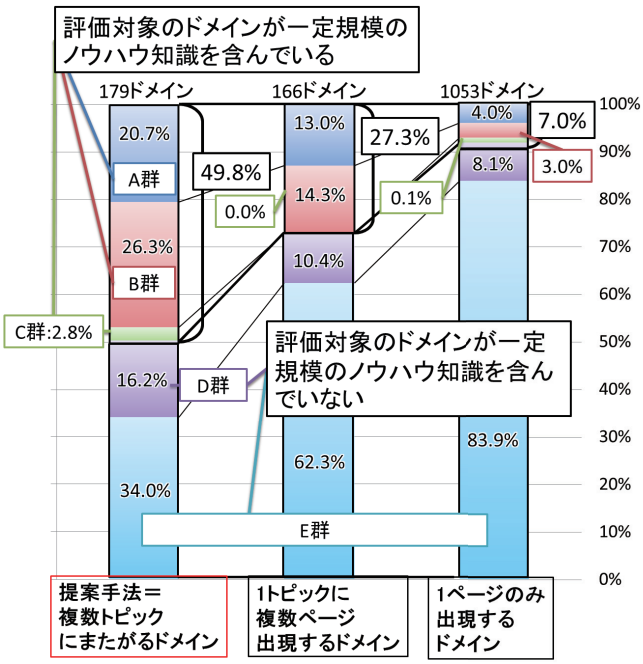
そして，収集された各ウェブページのドメインを対象として，表 2 の評価基準を適用して人手による評価を行う．実際の評価においては，(a) 複数トピックにまたって出現するドメインの全て，(b) 1 トピックに複数のページが出現するドメインのうち，無作為に選んだ半数，(c) 1 ページのみ出現するドメインのうち，無作為に選んだ 10%，をそれぞれ評価対象とする．

6.3 評 価 結 果

クエリ・フォーカスとして「就活」,「結婚」，および,「花粉症」を対象とした場合の評価結果を図 2 に示す．混合文書集合 D_{qw} を対象とした場合，一定規模のノウハウ知識を含んでいると判定されたドメインの割合は,「複数トピックにまたがるドメイン」の場合が最も多く，51.5%となった．また，ウェブページ集合 D_{qw} を対象とした場合も，一定規模のノウハウ知識を含んでいると判定されたドメインの割合は,「複数トピックにまたがるドメイン」の場合が最も多く，49.8%となった．同様に，クエリ・フォーカスとして「就活」,「結婚」，および,「花粉症」の各々を対象とした場合の評価結果を，それぞれ，図 3，図 4，および，図 5 に示す．　また，表 2 の評価基準において,「A 群: ドメインそのものがノウハウ知識を提示する」と判定されたドメインの例を図 6，図 7，および，図 8 に示す．例えば，図 6 のドメインでは，就活を行う際のノウハウを網羅的に掲載している．

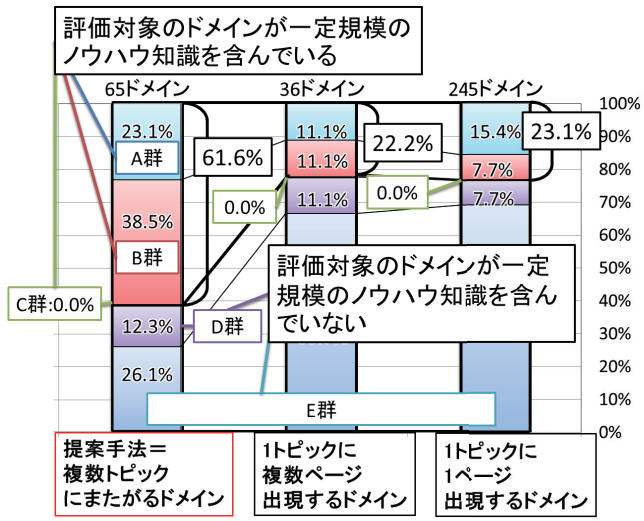


(1) 質問回答事例・ウェブページの混合文書集合

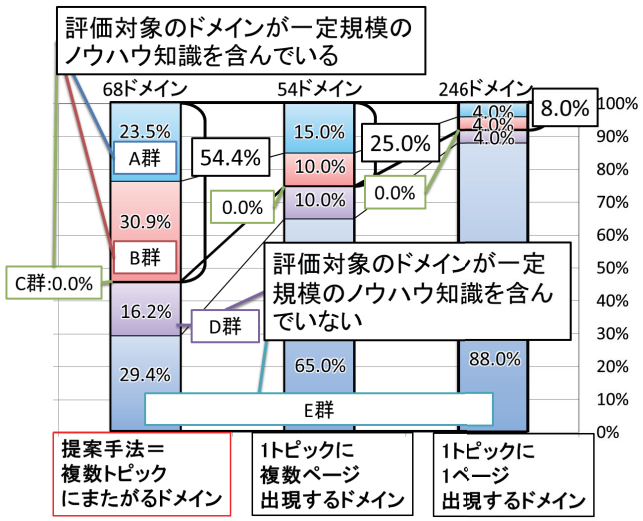


(2) ウェブページ集合

図 2　ノウハウサイト候補のドメインの評価結果 (クエリ・フォーカス: 「就活」,「結婚」, および,「花粉症」)



(1) 質問回答事例・ウェブページの混合文書集合



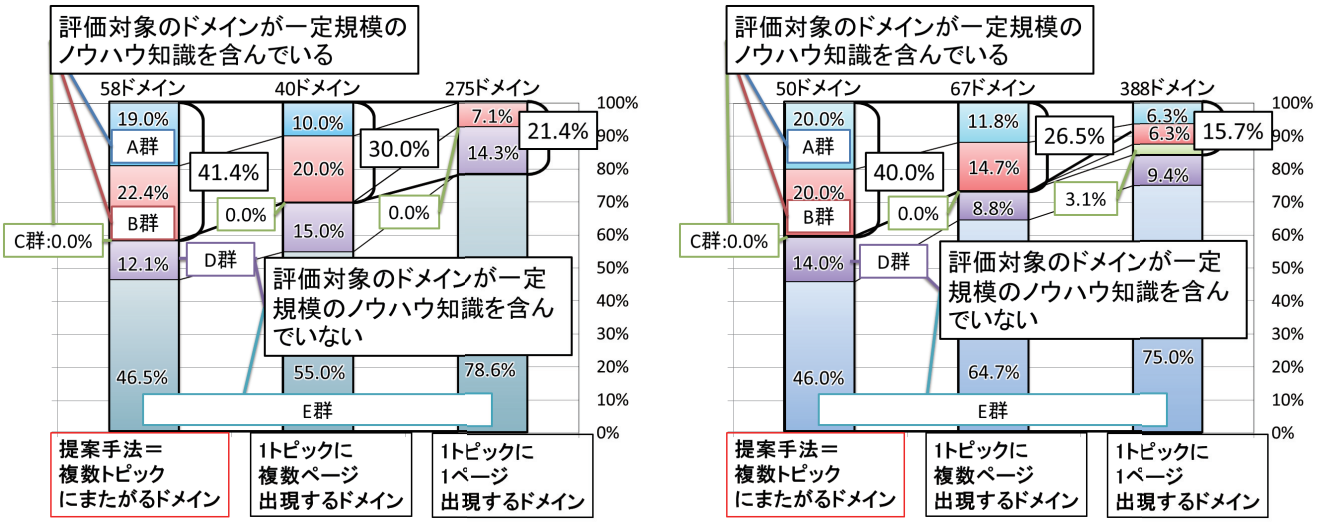
(2) ウェブページ集合

図 3　ノウハウサイト候補のドメインの評価結果 (クエリ・フォーカス: 「就活」)

7. 関 連 研 究

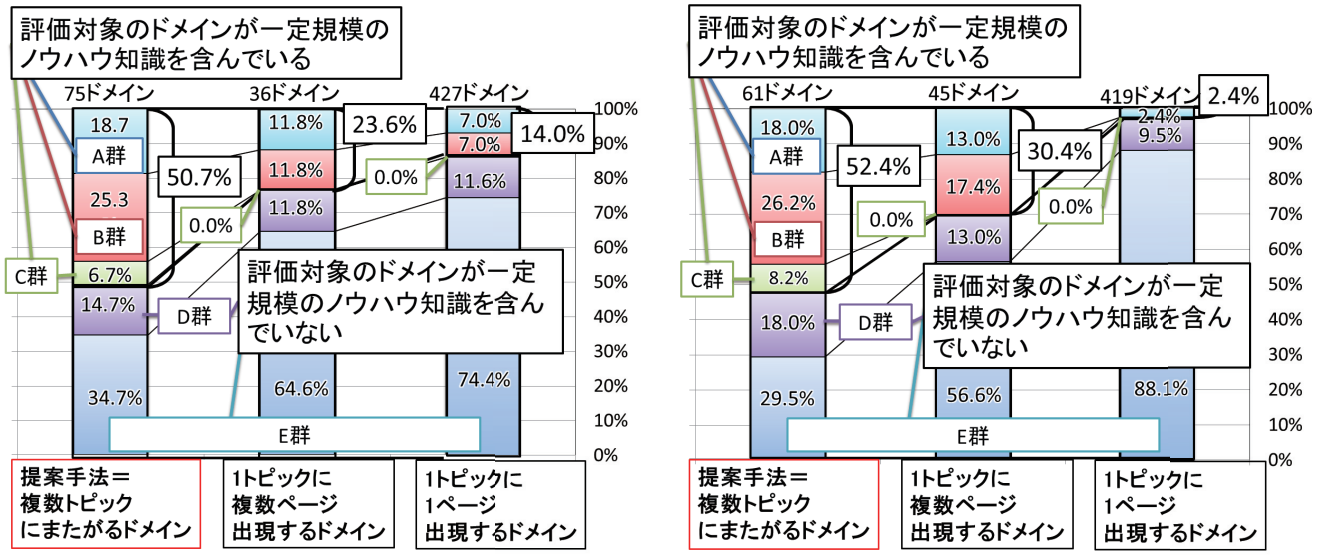
関連研究として，[2] では，クエリを実現するためのサブタスクの収集方式において，行為を表す動詞表現の形式を用いる方式を提案している．2014 年 12 月に開催された NTCIR-11(注3) においては，著者らの主催により，ほぼ同様の仕様で Task Mining Task も実施された．本研究においても，Task Mining Task で用いられたクエリリストおよび評価手順 [3] のもとで，本論文の提案手法を適用する方式の可能性を検討する必要があると考えられる．ただし，Task Mining Task のタスク設定では，クエリを実現するためのサブタスク群を動詞表現の形式で出力することにとどまっており，ノウハウ知識そのものは収集の対象

(注3)：http://research.nii.ac.jp/ntcir/ntcir-11/index-ja.html



(1) 質問回答事例・ウェブページの混合文書集合　　(2) ウェブページ集合

図 4　ノウハウサイト候補のドメインの評価結果 (クエリ・フォーカス:「結婚」)



(1) 質問回答事例・ウェブページの混合文書集合　　(2) ウェブページ集合

図 5　ノウハウサイト候補のドメインの評価結果 (クエリ・フォーカス:「花粉症」)

として扱われていない．これに対して，本研究において収集の対象となるのは，ウェブページ群の形式で表現されたノウハウサイトであり，この点において上記の関連研究とは大きく異なっている．

8. おわりに

本論文では，あるクエリ・フォーカスについて，ウェブからノウハウサイトを収集する手法として，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集合に対して，トピックモデルを適用し，ドメインの単位を対象としてノウハウサイトを収集する手法を提案した．そして，各トピックの確率上位 30 件のウェブページを対象とした場合において，提案方式によって収集されたドメインのうちの半数が有用なノウハウサイトであるという結果が得られた．

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] 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.

[3] 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.

[4] 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,

図 6 A 群「ドメインそのものがノウハウ知識を提示する」のサイトの例
（「賢者の就活」，http://kenjasyukatsu.com/）

図 7 A 群「ドメインそのものがノウハウ知識を提示する」のサイトの例
（「結婚レシピ」，http://www.wedding-recipe.com/）

2015.

図 8 A 群「ドメインそのものがノウハウ知識を提示する」のサイトの例
(「花粉症対策と症状対策」, http://www.kafuntaisaku1.com/)