DEIM Forum 2014 C1-2

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

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†††††

† 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
††† 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　本論文では，検索エンジン・サジェストを情報源としてウェブ検索者の情報要求観点を収集し，それらの観点の中でも，Wikipedia には掲載されていない観点に焦点を当てて，Wikipedia とは異なる観点についての情報を収集して集約し，提示することを目的とする．特に，Wikipedia においては，物事を解決するための実用的な知識や経験談，些細な雑談の類いや最新の話題等が掲載されることはあまり多くない．その一方で，検索エンジン・サジェストを分析することによって，ウェブ検索者がそれらの話題についても高い関心を持っていることが容易に分かる．そこで，Wikipedia に未掲載のそれらの情報に焦点を当て，検索エンジン・サジェストを通してウェブページ集合を収集し，その要点を要約・集約することにより，Wikipedia とは相補的な情報を掲載した百科事典を作成することを目的とする．

キーワード　検索エンジン・サジェスト, 観点, Wikipedia, 集約, 俯瞰

Analyzing Viewpoints of Web Search Information Needs considering Differences with Wikipedia

Ichiro MORIYA†, Daichi KOIKE†, Takakazu IMADA††, Takehito UTSURO†††, Yasuhide KAWADA††††, and Noriko KANDO†††††

† Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba 305-8573 Japan
†† College of Eng. Sys., School of Science and 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

インターネットの普及により，日頃からウェブサイトを閲覧する機会が増えている．そうしたウェブ閲覧者の多くは，自らの関心事項について Google や Yahoo!，Bing といった検索エンジンを用いてウェブ検索を行っている．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(注1)．すると，検索エンジン・サ

(注1)：図 1 の例では，検索窓に「コーラ 」を入力すると,「工場見学」,「音楽」,「依存性」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,「コーラ」が検索対象であり,「工場見学」,「音楽」,「依存性」等が情報要求観点である．また，実際の検索ログにおいては,「コーラ AND 工場見学」のように，検索対

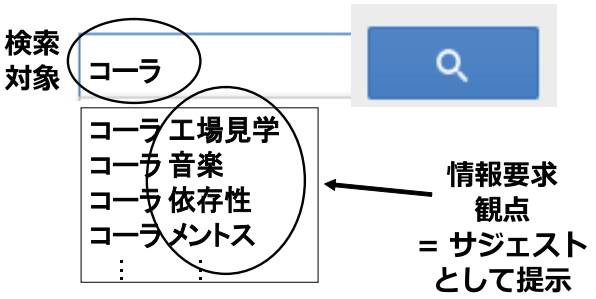

図 1　検索エンジン・サジェストにおける情報要求観点の例

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．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，ウェブ閲覧者の中でも，特にウェブ検索者に焦点を当てて，その関心事項の収集を行う．

ウェブ検索者が，ある検索対象について情報収集を行う場合，検索エンジンを用いて直接的にウェブページを検索する以外にも，Wikipedia に代表される何らかの知識アーカイブサイトを閲覧することにより，情報を収集する方法が考えられる．ここで，検索エンジンを利用する場合には，ウェブ上の大量のページ集合を情報源として検索結果が提示されるため，情報の量の点においては，Wikipedia 等の知識アーカイブを上回る場合もあるが，情報の質の点においては，Wikipedia 等の統制のとれた知識アーカイブよりも劣っている．一方，Wikipedia 等の統制のとれた知識アーカイブの場合は，人手によって記事を執筆することによって知識の蓄積がなされるため，説明のための観点が不十分である，説明のための観点が執筆者の視点に偏る，あるいは，未登録の用語がある，等の問題がある．

ここで，上述の検索エンジン・サジェストと，Wikipedia 記事中で段落タイトルとして掲載されている事項を比較してみると，一般の検索者が，Wikipedia において段落タイトルとして掲載されている事項について高い関心を持っている場合もあれば，Wikipedia において段落タイトルとしては全く掲載されていない事項に高い関心を持っている場合も多く観測される．具体的には，Wikipedia においては，物事を解決するための実用的な知識や経験談，ささいな雑談の類いや最新の話題等が掲載されることはあまり多くない．その一方で，検索エンジン・サジェストを分析することによって，検索者がそれらの話題についても高い関心を持っていることが容易に分かる．そこで，本論文では，検索エンジン・サジェストを情報源として検索者の情報要求観点を収集し，それらの観点の中でも，Wikipedia には掲載されていない観点に焦点を当てて，Wikipedia とは異なる観点についての情報を収集した百科事典を作成することを目的とする．

2.　Wikipedia 非掲載事項の事典化の枠組み

本節では，Wikipedia に非掲載の事項を収集し事典化する枠組みを以下の 6 つの手順に分けて説明する．この枠組のうち，検索エンジン・サジェストから情報要求観点を収集する手順，および，情報要求観点と検索結果のウェブページ集合中の話題に対して Wikipedia 掲載の有無を分析する手順の概要を図 2 に示す．

2.1　評価用検索対象の選定

評価用検索対象の選定においては，日本最大のブロガー・コミュニティ・サービスである「にほんブログ村」(注2) (登録ブロガー数約 68 万人，カテゴリ数 121，サブカテゴリ数約 5,500) のカテゴリ名及びサブカテゴリ名を参照し，検索対象の候補とした(注3)．

2.2　サジェスト (情報要求観点) の収集

選定した評価用検索対象に対して，Google(注4) 検索エンジンを用いて，一検索対象当り約 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 (3. 節)．

2.3　サジェストと Wikipedia 中の段落タイトルの比較およびサジェストの選定

検索対象に対して収集したサジェスト (情報要求観点) と，検索対象をタイトルとする Wikipedia 記事中の段落タイトルの比較を行う．次に，Wikipedia 記事中において段落タイトルとして取り上げられていないサジェストを 10~30 個程度選定する (4. 節)．

2.4　ウェブページの収集

選定したサジェスト (情報要求観点) と検索対象の AND 検索によって，ウェブページを収集する (5. 節)．

2.5　収集したウェブページの話題分析・集約・要約

収集したウェブページ中の話題の分析・集約，および，要約を行う．本論文では，収集したウェブページ集合に対してトピックモデルを適用することにより，話題の集約を行うというアプローチを採用する．本論文では，このアプローチに対して，検索結果上位のウェブページに対するスニペットを用いて話題の集約・要約を行う方式との比較を行う (6. 節)．

2.6　収集・集約された話題に対する Wikipedia 掲載の有無の分析

トピックモデルを用いて収集・集約された話題，および，検索結果上位のスニペットから収集・集約された話題に対して，Wikipedia 掲載の有無の分析を行う (7. 節)．

3.　検索エンジン・サジェストからの情報要求観点の収集

選定した評価用検索対象に対して，Google 検索エンジンを用いて，一検索対象当り約 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，半濁音及び「きゃ」や「びゃ」などの

象と情報要求観点の 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2)：http://www.blogmura.com
(注3)：予備調査においては，日常的な文化・慣習に関連する語を検索対象として検索エンジン・サジェストを収集したところ，意外性に富んだサジェストを多く収集できることが判明したため，日常的な文化・慣習に関連する語を中心に評価用の検索対象を選定した．
(注4)：http://www.google.com/

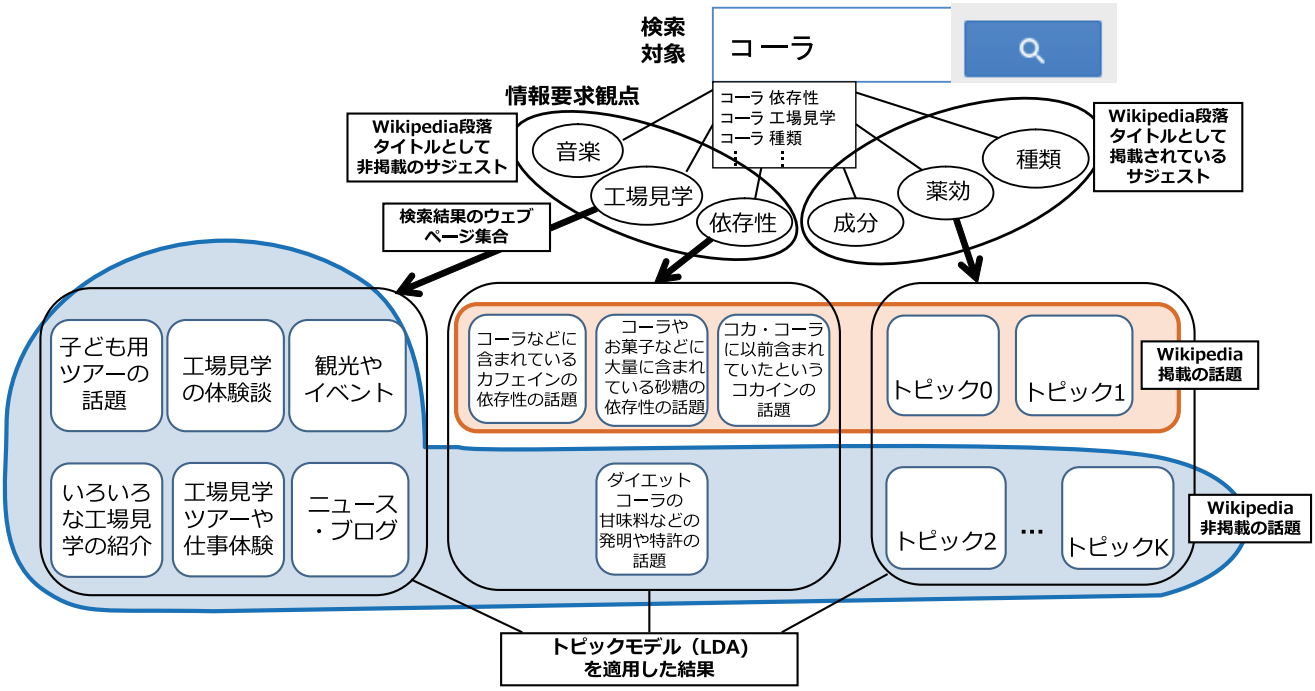

図 2　検索エンジン・サジェストからの情報要求観点の収集及び Wikipedia 掲載の有無の分析

表 1　収集したサジェストと Wikipedia 中の段落タイトルの比較

検索対象	比較対象の Wikipedia エントリ	サジェスト数（Wikipedia 非掲載/全サジェスト）
コーラ	コーラ (飲料)	779/791
貯金	預金	721/729
禁煙	禁煙	803/818
共働き	共働き	615/619
結婚式	結婚式	896/917
	結婚披露宴	
節電	節電	736/783

開拗音である．例えば検索窓に「コーラ こ」と入力すると，「工場見学」や「凍らせる」などがサジェストとして掲示されるので，それらの収集を行う．

4. サジェストと Wikipedia 中の段落タイトルの比較およびサジェストの選定

Google，Yahoo!，Bing の三つの検索エンジンのうち，収集されたサジェストの数が最も多い Google を対象として，Wikipedia 中の段落タイトルとサジェストの比較を行った．検索対象に対して収集された各サジェストについて，検索対象をタイトルとする Wikipedia 記事中の段落タイトルとして掲載されているか否かを集計した結果を表 1 に示す．この結果から分かるように，各検索対象に対して収集されたサジェスト (情報要求観点) の大部分は，Wikipedia 記事中の段落タイトルとしては掲載されていない．ただし，この比較作業においては，検索対象をタイトルとする Wikipedia 記事のみを比較対象としており，関連語や情報要求観点そのものの Wikipedia 記事との比較は行っていない．

5. ウェブページ収集の手順

ウェブページの収集においては，Yahoo! Search BOSS API(注5) を用い，検索エンジン API に対して，サジェスト (情報要求観点) と検索対象の AND 検索クエリを指定することにより，日本語のサイトを対象として収集を行った．Yahoo! Search BOSS API では 1 クエリ当り最大 1,000 件のウェブページを取得することが可能である．そこで，本論文では，AND 検索クエリごとに最大 1,000 件のウェブページを取得し，それらを分析対象とした．

6. トピックモデルを用いたウェブページ集合の話題同定

6.1 手　　順

本研究では，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6)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，本論文では，トピック数を 10 から 50 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

(注5)：http://developer.yahoo.com/search/boss/
(注6)：http://gibbslda.sourceforge.net/

表 2　トピックモデルを用いたウェブページ集合の話題同定結果

ID	検索対象	情報要求観点	収集されたウェブページ数	LDA のトピック数	話題数 (Wikipedia 非掲載数+掲載数)		
					LDA のみにより同定	検索順位上位のスニペット中のみで観測	LDA・スニペット共通
1	コーラ	工場見学	572	15	2+0	1+1	4+0
2	コーラ	依存性	375	15	1+0	4+0	0+3
3	貯金	コツ	540	25	5+0	2+1	4+0
4	貯金	平均	411	20	3+0	2+0	3+0
5	禁煙	アプリ	460	15	2+0	2+0	1+1
6	禁煙	うつ病	369	15	1+2	1+2	1+1
7	共働き	保険	509	20	2+0	2+0	4+1
8	共働き	病気	402	20	2+0	2+0	2+0
9	共働き	教育	534	20	4+1	2+0	3+0
10	共働き	習い事	261	15	3+1	1+1	2+0
11	結婚式	髪型	396	15	5+1	2+0	3+0
12	結婚式	ゲスト	557	25	5+6	2+2	2+2
13	節電	アプリ	493	20	5+0	4+0	1+0
14	節電	アイデア	715	20	4+6	0+2	2+1

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．

6.2　検索順位上位のスニペットにおける話題分布との比較

ウェブページ集合にトピックモデルを適用して得られた話題と，検索エンジンとして Google を用いた場合の検索順位上位 20 位以内のスニペットから得られる話題との比較を行った結果を表 2 に示す．ただし，トピックモデルの適用結果から得られた話題については，各トピックに対する確率値の上位 5 記事程度を人手で分析し，記事数が最大となる話題を採用した．この結果から分かるように，検索順位の上位には現れないが，同一話題の記事が複数存在し，トピックモデル特有の話題として同定された話題，および，同一話題の記事は少数しか存在しないものの，検索順位の上位で観測されスニペットのみから同定された話題の双方が観測された．このことから，両方の手法は相補的な位置付けにあると言える．

一例として，AND 検索クエリ「コーラ AND 工場見学」によって収集されたウェブページ集合から得られた話題の抜粋を図 3 に示す．

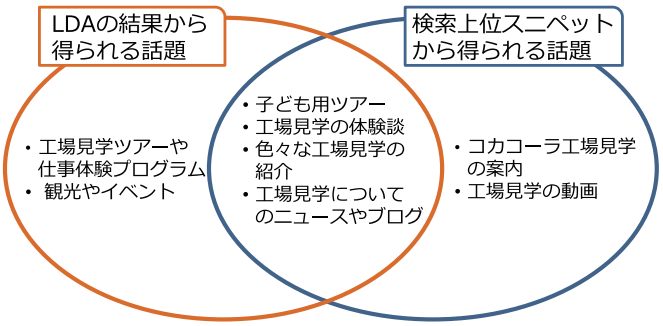

図 3　AND 検索クエリ「コーラ AND 工場見学」によって収集されたウェブページ集合における話題分布

7.　Wikipedia 掲載の有無の分析

表 2 においては，各検索対象について，サジェスト (情報要求観点) と検索対象の AND 検索クエリによって収集されたウェブページ集合に対してトピックモデルを用いて収集・集約された話題，および，検索エンジンとして Google を用いた場合の検索順位上位 20 位以内のスニペットから得られる話題の双方について，検索対象をタイトルとする Wikipedia 記事，および，関連する Wikipedia 記事における掲載の有無の分析結果を示す．この結果から，トピックモデルを用いて収集・集約された話題，および，スニペットから得られる話題の双方において，Wikipedia に未掲載の話題が多数含まれることが分かった．

8.　検索エンジン・サジェストによって収集される情報の信頼性に関する予備調査

検索エンジン・サジェストを情報源として収集される情報の中には，信頼性の低いものが散見される場合も少なくない．そこで，本節では，収集される情報の信頼性に関して，信頼性が低いと考えられる場合を対象として，以下の三種類の項目に該当する情報の有無の割合を調査した．

(1)　事実と異なる，あるいは，真偽が判定できない，
(2)　有用性が低い，
(3)　検索対象との関連性が低い，

また，信頼性が高いと考えられる場合についても，以下の三種類の項目に分類し，

(4)　関心を持つ検索者が少ない，
(5)　有用性が中程度，
(6)　有用性が高い，

このうち，特に，「(6) 有用性が高い」に該当する場合に着目して，その割合を求めた．

具体的には，比較的信頼性の低そうな情報が多く収集されると予想される「コーラ」，および，その逆に，比較的信頼性の高そうな情報が多く収集されると予想される「就活」および「禁

表 3　検索エンジン・サジェストによって収集される情報の信頼性に関する予備調査結果 (抜粋)

(a) 検索対象「コーラ」 (調査対象サジェスト数: 52)

分類		割合 (%)	サジェストの例
信頼性が低い	(1) 事実と異なる，あるいは，真偽が判定できない	11.5	造影剤，ゴキブリ，緑色，虫，生肉，避妊
	(2) 有用性が低い	9.6	ヘリウム，“ヘリウム ゲップ”，蛇口，大好き，ローラ
	(3) 検索対象との関連性が低い	3.9	硫酸，氷点
信頼性が高い	(4) 関心を持つ検索者が少ない	0	—
	(5) 有用性が中程度	46.2	都市伝説 (コーラに関する都市伝説)，マジック (コーラを使ったマジック)
	(6) 有用性が高い	28.8	風邪 (コーラを用いた風邪の民間療法)，炭酸 (炭酸ガス抜けを防ぐ方法)，まずい (世界一まずいコーラの種類)，毒入り (1977 年青酸コーラ事件)
合計		100	—

(b) 検索対象「禁煙」 (調査対象サジェスト数: 62)

分類		割合 (%)	サジェストの例
信頼性が低い	(1) 事実と異なる，あるいは，真偽が判定できない	6.5	屁が臭い，“尿 臭い”，前頭葉，鼻炎
	(2) 有用性が低い	3.2	バカ，ワロタ
	(3) 検索対象との関連性が低い	3.2	ミュージックオルグ，恋愛運
信頼性が高い	(4) 関心を持つ検索者が少ない	4.8	茗荷谷 (茗荷谷周辺の禁煙施設)，龍谷大学 (龍谷大学および周辺の禁煙事情)，ニューカレドニア (公的場所での禁煙法)
	(5) 有用性が中程度	24.2	待ち受け (禁煙の待ち受け画像)，ロードバイク (ロードバイク乗車と喫煙の関係)
	(6) 有用性が高い	58.1	腰痛 (喫煙者はなりやすい)，業績 (飲食店の業績と禁煙の関係)，記憶力 (喫煙と記憶力の関係)，ビタミン (ビタミン C 破壊)
合計		100	—

(c) 検索対象「就活」 (調査対象サジェスト数: 76)

分類		割合 (%)	サジェストの例
信頼性が低い	(1) 事実と異なる，あるいは，真偽が判定できない	6.6	“中国 臓器”，2ch ビューア，スレ，“ランキング 2ch”，なん j
	(2) 有用性が低い	6.6	けいおん，ざまあ，ざまぁ，メシウマ，ヌルゲー
	(3) 検索対象との関連性が低い	1.3	“world class”
信頼性が高い	(4) 関心を持つ検索者が少ない	7.9	ヒューマントラスト (会社名)，ヒューリック (会社名)
	(5) 有用性が中程度	13.2	オワタ (就活終了)，きくち (個人の逆就活サイト)
	(6) 有用性が高い	64.4	北海道 (就活支援のための飛行機運賃割引)，ご利益 (就活にご利益のある神社・お守り)，うそ (就活でつく「うそ」についての議論)，youtube(就活に関する動画)，なぜ銀行 (銀行の志望動機)，
合計		100	—

煙」を対象として，上記の (1)〜(6) に該当する検索結果が含まれる可能性が高そうなサジェストを重点的に調査した．

以上の調査結果，および，各分類におけるサジェストの具体例を表 3 に示す．この結果から，「(6) 有用性が高い」の割合は，「コーラ」において 28.8%とやや低い値となったのに対して，「就活」においては 64.4%，「禁煙」においては 58.1%と過半数を越える割合となり，実際に有用性が高いことが確認できた．

9. 関 連 研 究

Wikipedia 記事の自動生成，あるいは，記事中の項目の補完に関する関連研究として，文献 [2–4] がある．このうち，文献 [3, 4] においては，Wikipedia カテゴリごとに，段落タイトル一覧を観点集合として設定し，Wikipedia 記事ごとに，欠落した観点の内容をウェブから収集・補完する方式を提案している．この方式は，動物カテゴリや病気カテゴリのように，典型的観点が Wikipedia カテゴリ中に網羅されている場合には効果的だが，本論文において百科事典化の対象とする文化・慣習に関する情報要求観点のように，非常に多様でカテゴリ化が困難な情報を収集する目的には適さない．また，文献 [2] においては，検索対象をタイトルとする Wikipedia 記事本文，および，検索対象をクエリとして収集したウェブページ集合に LDA を適用し，Wikipedia に未掲載の新規観点を生成する，あるいは，Wikipedia に未掲載の新規の内容を補完する手法を提案している．

これらの関連研究と比較すると，本論文のアプローチにおいては，ウェブ検索者の関心事項に着目し，検索エンジン・サジェストを情報源として検索者の情報要求観点を収集し，Wikipedia

に未掲載の情報要求観点を網羅的に収集する点に新規性があり，収集される情報の種類において，これらの関連研究とは大きく異なっている．

一方，[7] においては，特定のタスクを達成するために必要なタスク集合をウェブ検索結果から抽出する手法を提案している．この関連研究と比較すると，本論文のアプローチにおいて収集されるウェブ検索者の関心事項は，一部，タスク情報を含んだものとなっていると言える．今後は，ウェブ検索者の関心事項を類型化することにより，タスク情報以外の有用性の高い情報がどの程度収集されているかの評価・分析を行う必要がある．

10. おわりに

本論文では，ウェブ検索者の関心事項に着目し，検索エンジン・サジェストを情報源として，ウェブ検索者の情報要求観点を収集し，それらの観点のもとで収集される情報を百科事典化する枠組みを提案した．また，実際に，ウェブ検索者の情報要求観点のもとで収集された話題と Wikipedia 中の掲載事項を比較することにより，Wikipedia に未掲載の事項が多数収集可能であり，Wikipedia との間で相補的な関係を持つ百科事典の作成が可能であることを示した．

本研究およびその周辺の研究課題の展開の一つとして，一検索対象当り最大約 1,000 語あるサジェストに対して，一語を一つの情報要求観点とみなすのではなく，冗長なサジェストを集約して得られるサジェスト集合を一つの情報要求観点ととらえ，サジェスト集合単位でウェブページの収集及び話題の集約を行う方式 [6, 8] が有効であると考えている．その他に，検索対象とサジェストとの間の関係を提示する方式 [5]，日中検索エンジン・サジェストを情報源として，ウェブ検索者の情報要求観点を収集し，他国と自国との間の文化・関心・意見の違いを発見する過程を支援する方式 [9] 等についても研究を進めている．

また，百科事典化技術を実現するための重要な要素技術の一つとして，トピックモデルによって集約された話題の情報を，百科事典における実際の記述内容として要約して提示するための自動要約手法を確立することが不可欠であり，今後取り組む予定である．

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] D. Eklou, Y. Asano, and M. Yoshikawa. How can the Web help Wikipedia? a study of information complementation of Wikipedia by the Web. 情報処理学会論文誌：データベース, Vol. 5, No. 3, pp. 64–74, 2012.

[3] 藤井敦, 三條場旭彦. Wikipedia を用いた用語説明のモデル化と事典的検索への応用. 人工知能学会研究会資料, SIG-SWO-A803-01, 2009.

[4] 藤井裕也, 藤井敦, 徳永健伸. Wikipedia 記事構造のモデル化による用語説明の自動編集. 言語処理学会第 18 回年次大会論文集, pp. 1059–1062, 2012.

[5] 今田貴和, 小池大地, 守谷一朗, 宇津呂武仁, 神門典子. 検索対象と検索エンジン・サジェストとの間の関係の提示. 言語処理学会第 20 回年次大会論文集, 2014.

[6] 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.

[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, 2014.

[9] 鄭立儀, 小池大地, 聶添, 今田貴和, 陳磊, 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日中間対照分析. 言語処理学会第 20 回年次大会論文集, 2014.