解説

検索窓から俯瞰する世界

Overviewing Web Space through Web Search Box

宇津呂武仁

Abstract

本稿では，検索エンジンサジェストを情報源として，Web 情報空間における Web 検索者の関心事項を優先的・選択的に収集した後，集約・俯瞰する方式について述べる．更に，この方式の応用技術として，検索エンジンサジェストと質問回答サイトを併用することによって，より有用性の高いノウハウ知識を収集・集約する技術を紹介する．また，企業名に関する Web 検索者の関心動向を集約することによって，特定の商品ジャンルにおける市場シェアの分析を行う技術を紹介する．これらの方式を用いることによって，ネットユーザの「ネット上に分散している知識を，断片的にではなく，網羅的・集約的に知りたい」という要望を満たす技術が実現可能であることを示す．

キーワード：検索エンジン，検索エンジンサジェスト，トピックモデル，俯瞰，ノウハウ

1. はじめに

各検索エンジン会社においては，Web 検索者の検索ログが蓄積されており，多数の Web 検索者がどのような事柄に関心を持って検索を行ったのかについての情報を集約し，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，本稿では，Web 検索者が入力するクエリ（検索質問）のうち，検索したい事柄の主要な概念を表している部分を「クエリフォーカス」と呼ぶ．また，クエリフォーカスに対して，Web 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，クエリフォーカスに対して詳細な情報を得るために，どのような側面に着目するかを指定している部分を「情報要求観点」と呼ぶ．（図 1 の例では，検索窓に「就活」を入力すると，「メール」，「面接」，「2ch」，「メイク」等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，「就活」がクエリフォーカスであり，「メール」，「面接」，「2ch」，「メイク」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，「就活 AND メール」のように，クエリフォーカスと情報要求観点の AND 検索の形式で表現された検索要求が蓄積されている．）

ここで，検索エンジンにおいて，検索エンジンサジェストとして提示される言葉は，「クエリフォーカス」に対して，多数の Web 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，Web 検索者の関心事項そのものを反映している．これに対して，本稿では，特に，ネットユーザの「ネット上に分散している知識を，断片的にではなく，網羅的・集約的に知りたい」という要望に着目する．そして，この要望を満たす技術の実現を目的として，検索エンジンサジェストを情報源として，Web 情報空間における Web 検索者の関心事項を優先的・選択的に収集した後，集約・俯瞰する方式[(1)]について述べるとともに，その応用技術[(2)～(6)]を紹介する．まず，2. においては，Web 検索者の関心事項である検索エンジンサジェストを収集・集約するとともに，サ

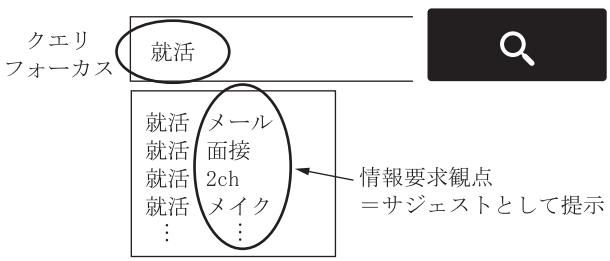


図 1　検索エンジンサジェストにおける情報要求観点の例

宇津呂武仁　正員：シニア会員　筑波大学システム情報系知能機能工学域
Takehito UTSURO, Senior Member (Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba-shi, 305-8573 Japan).
電子情報通信学会誌　Vol.99 No.9 pp.920-927 2016 年 9 月

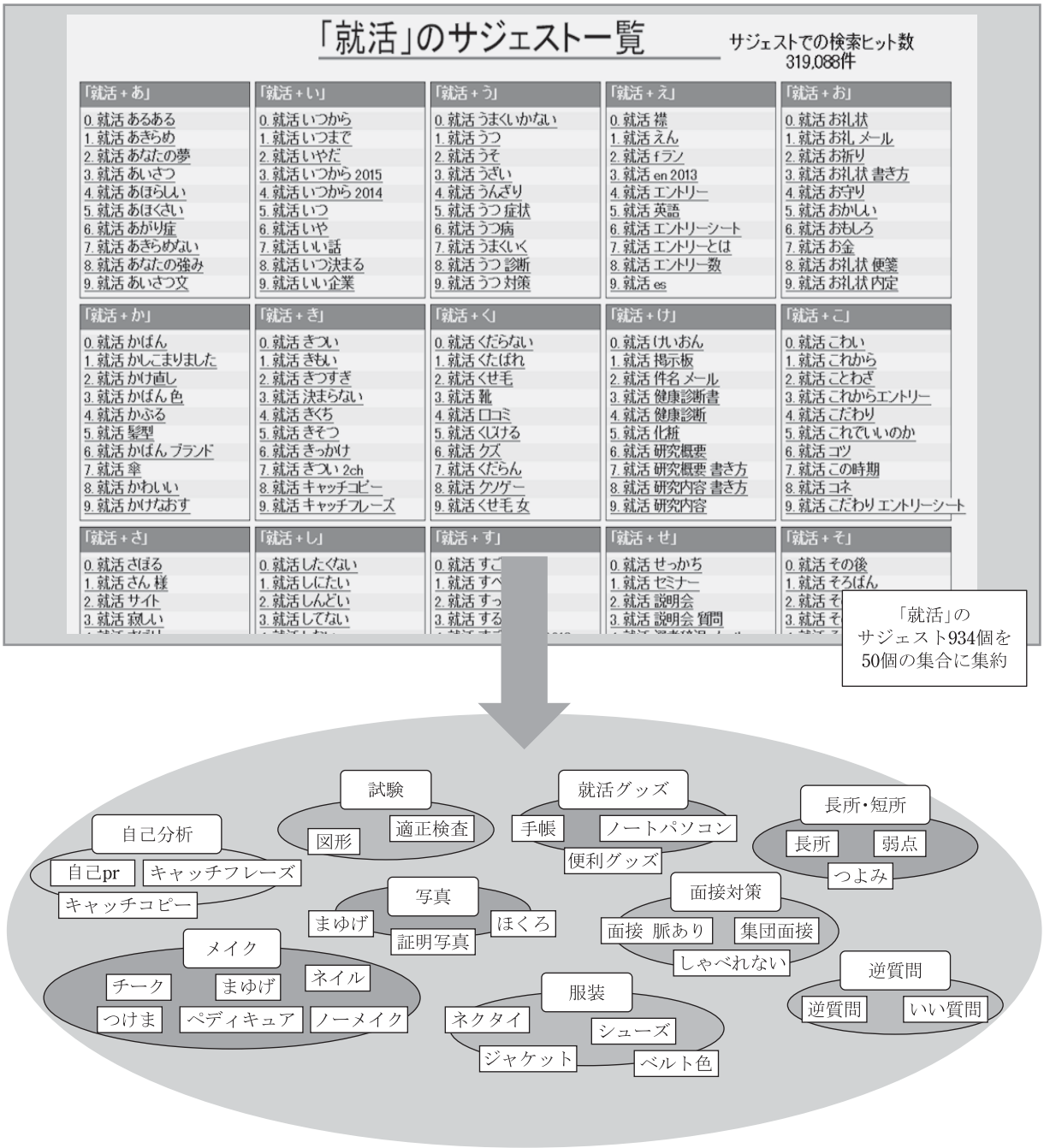


図2　検索エンジンサジェストの集約（クエリフォーカス：「就活」）

ジェストを用いて検索される Web 検索結果を俯瞰する方式を概観する．次に，この方式の応用技術として，3. においては，検索エンジンサジェストと質問回答サイトを併用することによって，より有用性の高いノウハウ知識を収集・集約する技術を紹介する．一方，4. においては，企業名に関する Web 検索者の関心動向を集約することによって，特定の商品ジャンルにおける市場シェアの分析を行う技術を紹介する．

■ 用　語　解　説

潜在的ディリクレ配分法　一つの文書中の話題が複数の潜在的なトピックから構成されるという前提の下で，文書集合のトピックの分布を推定するトピックモデルの一種．

2. Web 検索者の関心動向の収集・集約・俯瞰

本章で紹介する枠組み[(1),(7),(8)]においては，一つのクエリフォーカスに対して，（日本語の場合）最大約 1,000 語のサジェストを収集する[(注1)]．そして，クエリ

表1　提案手法による検索エンジンサジェストの集約結果の例（クエリフォーカス：就活）

人手によりトピックに付与したラベル	トピックに割り当てられたサジェスト（各トピック10サジェストを抜粋）
髪型	“ヘアスタイル　女”，“くせ毛　女”，“写真　髪型”，まとめ髪，おだんご，襟足，ロングヘア，ゆるいパーマ，美容院，シュシュ
身に着けるもの	ネクタイ，シューズ，“ベルト　色”，かばん，ピーコート，シャツ，“パンプス　おすすめ”，“グレー　スーツ”，“ジャケット　ボタン”，防寒
グループディスカッション	グループワークとは，グループディスカッション，“グループディスカッション　テーマ”，評価，グループワーク対策，評価基準，プレゼン，“プレゼン　資料”，グループワーク，能力
自己分析	“長所　真面目”，長所，座右の銘，軸，どうなりたいか，あなたの夢，こだわり，将来の夢，どんな人，なりたい自分
恋愛との両立	“恋愛　両立”，ふられた，恋愛，寂しい，脈あり，結婚，“うまくいかない　彼氏”，“プレゼント　彼女”，わがまま，プレッシャー
メイク	ノーメイク，ビューラー，チーク，化粧，つけま，まつエク，ネイル，まゆげ，“証明写真　メイク”，ペディキュア

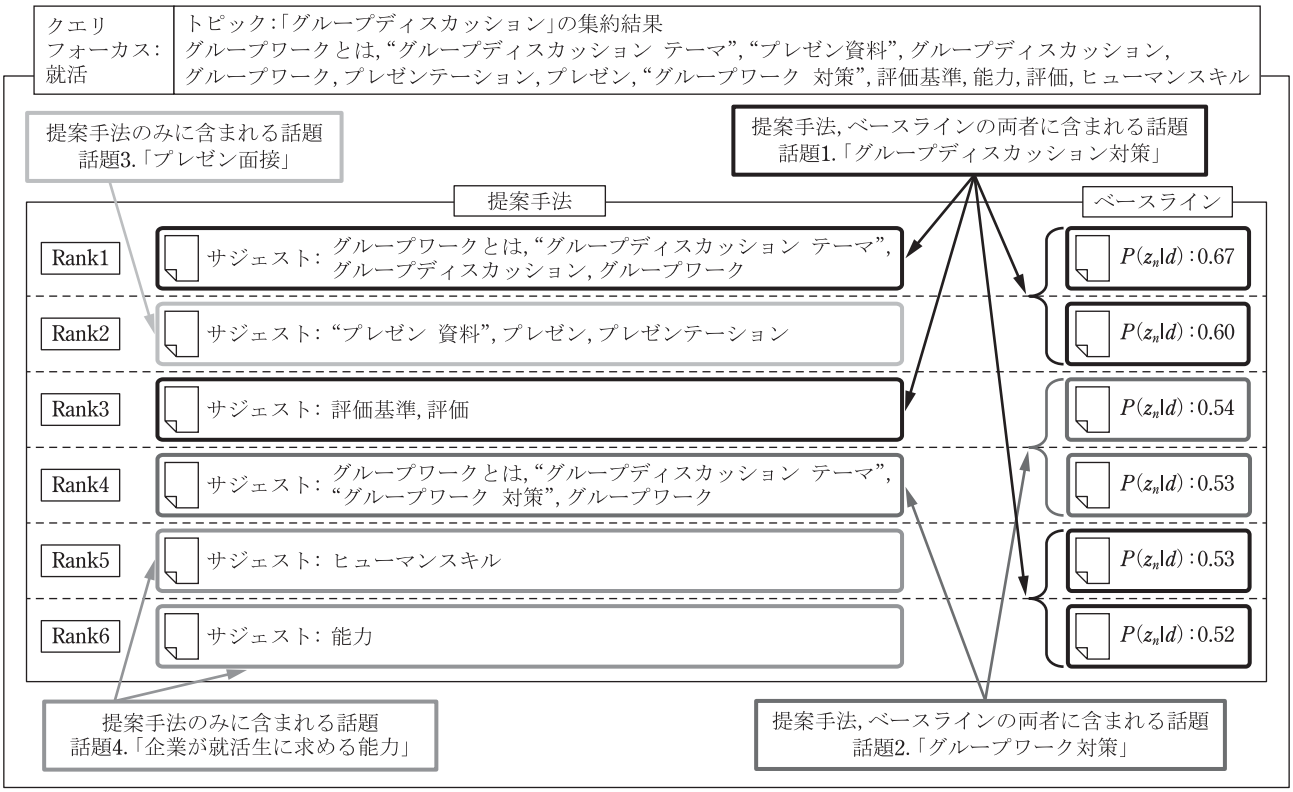


図3　Web検索結果の集約の例（クエリフォーカス：「就活」，トピック：「グループディスカッション」）

フォーカスに加えて一つの検索エンジンサジェストを指定したAND検索によってWebページを収集する．最大約1,000個の検索エンジンサジェストに対してこの方法を用いることにより，あるクエリフォーカスに関する大規模なWebページ集合を収集することができる．しかし，収集されるサジェスト，及び，それらを用いて収集されるWebページ集合では，多くは話題が重複しており冗長である．そこで，以下で述べる方式によって，検索エンジンサジェスト及びWebページ集合の集約・俯瞰を行う．

2.1　トピックモデルを用いた話題の集約

検索エンジンサジェスト及びWebページ集合の集約においては，トピックモデルの一種である潜在的ディリクレ配分法（LDA：Latent Dirichlet Allocation）[9]（用語）を用いる．まず，一つのクエリフォーカス当り最大約1,000語のサジェストを収集し，それらサジェストを用いてWebページの収集を行う．そして収集されたWeb

（注1）　Google検索エンジン（https://www.google.com/）を用いる．

ページ集合に対して，LDA を適用しトピックと呼ばれる話題のまとまりごとに Web ページのクラスタリングを行う．各 Web ページはサジェストを用いて収集されたものであるため，各 Web ページには最低一つ以上のサジェストを対応付けることができる．この対応付けによりサジェストの集約を行う．これにより，約 1,000 語あったサジェストを数十個ほどのまとまりへと集約することができる．クエリフォーカス「就活」に対して収集された 934 個のサジェストが 50 個の集合に集約した結果の概要及び抜粋を図 2 及び表 1 に示す．

2.2　一つのトピック内の話題の俯瞰

ここで，各トピックにおいてサジェストを集約した結果においては，互いに類似するサジェストを用いて Web ページが収集されているため，相互に類似する冗長な Web ページが多数収集されているのが現状である．これらの Web ページ集合を効率良く俯瞰するためには，冗長性をなくしてできるだけ多様な話題を示す Web ページ集合へと集約した上で閲覧する必要がある．そこで，図 3 の例に示すように，できるだけ少ない数の Web ページ集合によって各トピック中のサジェストを 1 回以上提示するという設計の下で，一つのトピック内の話題を俯瞰するインタフェースを開発した．その結果，図 3 の例では，6 種類の Web ページによって 4 個の話題を提示することができた．一方，サジェストを用いず，トピックモデルによる確率分布のみを用いる従来方式では，同数の Web ページによって半数の話題を提示するにとどまった．

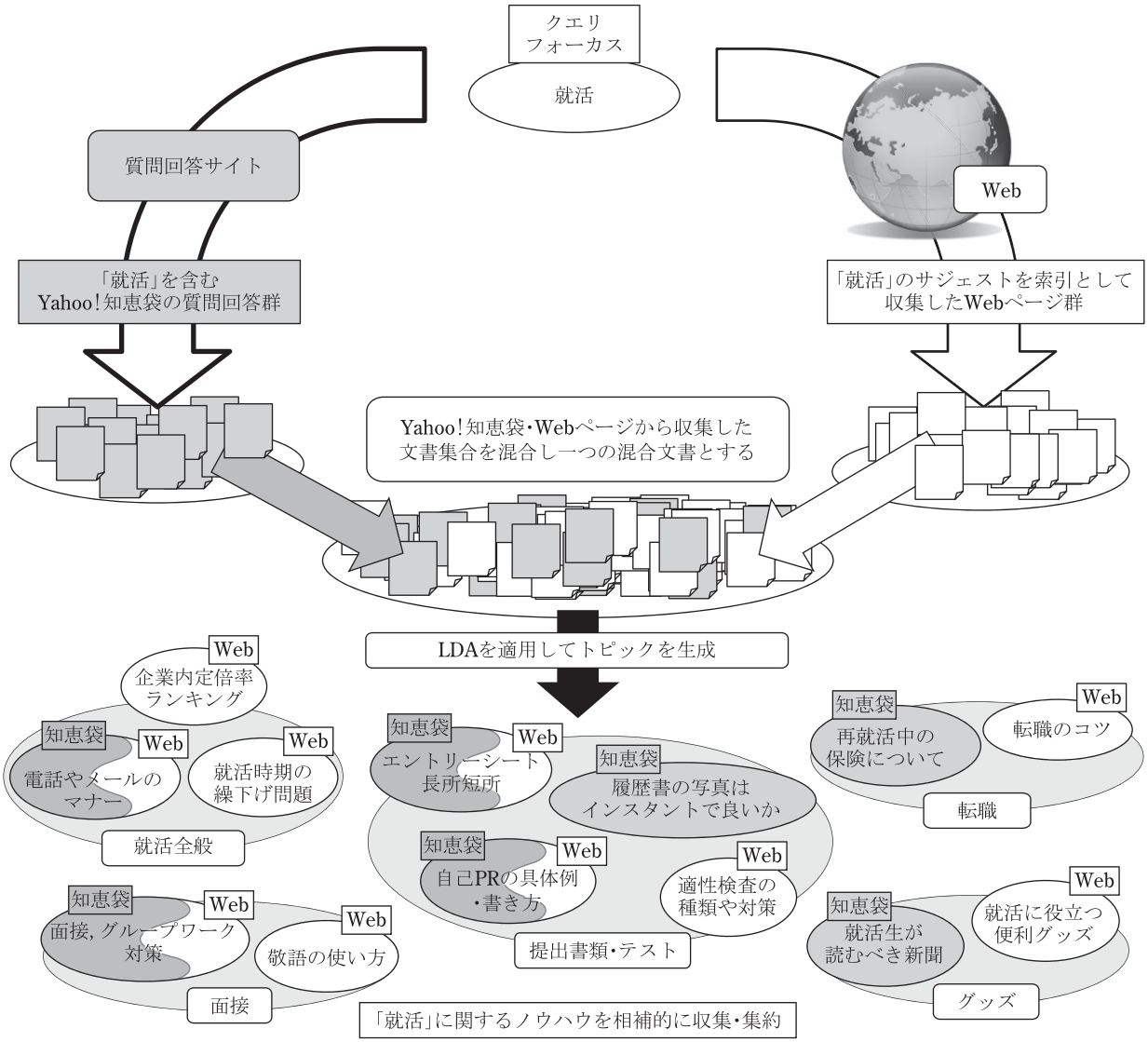


図 4　質問回答サイト・Web からのノウハウ収集・集約の流れ（対象：日本語）

3. 質問回答事例及びサジェストを情報源とするノウハウ知識の収集・集約

次に，前章で紹介した技術の応用技術として，より有用性の高いノウハウ知識を選択的に収集・集約するための戦略として，図4に示すように，Web検索者の関心事項である検索エンジンサジェスト単独でなく，ノウハウ知識を選択的に多く含むことが想定される質問回答サイトを併用する方式について述べる[3),10]．この方式では，質問回答サイト(注2)から収集した質問回答事例，及び，検索エンジンサジェストを用いて検索されたWeb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，図4（クエリフォーカス「就活」に関するノウハウの収集・集約結果）に示す形で話題が集約される．集約された話題は，大別して，質問回答事例のみから得られるノウハウ（履歴書の写真について，就活生が読むべき新聞，等），Webのみから得られるノウハウ（適性検査の種類や対策，就活に役立つ便利グッズ，等），質問回答事例・Web両方から得られるノウハウ（自己PRの具体例，面接・グループワーク対策，等）に分けられる．クエリフォーカスとして，「花粉症」，「結婚」，「就活」を対象として収集・集約したノウハウ知識の話題数を表2に示す．

また，中国語を対象として同様のノウハウ知識収集・集約を行い，その内容を日中間で比較対照分析した結果，中国特有のノウハウ知識として得られたものの抜粋を表3に示す[4),5),11)]．このような日中間の比較対照分析作業は，日中両言語を対象として分析システムを開発するとともに，日本語側・中国語側双方の分析者の間で情報交換をすることによって，比較的容易に実現することが可能である．

4. 企業名に関する関心動向を利用した日中市場シェアの分析

続いて，同様の応用技術例の一つとして，企業名に関するWeb検索者の関心動向を集約することによって，特定の商品ジャンルにおける市場シェアの分析を行う技術[2),6),12)]を紹介する．この方式においては，図5に示すように，日本語を対象として，国内主要電気メーカ10社をクエリフォーカスとして検索エンジンサジェスト及び検索結果のWebページを収集した後，トピック

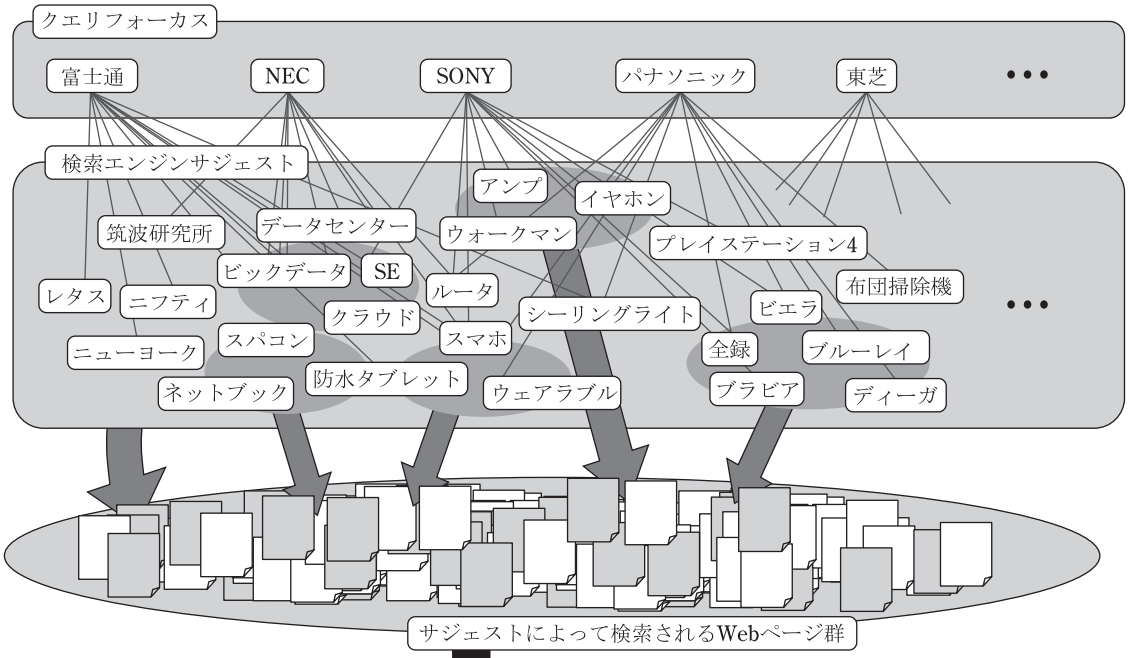
表2　ノウハウ知識の話題数

クエリフォーカス	質問回答サイトのみ	Webのみ	質問回答サイト＋Web	合計
花粉症	6	19	30	55
結　婚	12	7	16	35
就　活	15	17	17	49

表3　中国特有のノウハウ知識の詳細説明

クエリフォーカス	話題	説明
就活	大学に提出する「卒業就職意向表」の書き方	就活生が自分の技術や就職希望企業について「卒業就職意向表」を記述して大学に提出する．大学側は企業による訪問面接の参加のあっ旋をしてくれる．
	就活面接ショー番組	就活生と企業がスタジオで実際に面接をしてその場で選考を行う番組．
	就活用ウィークリーマンション	主に就活生が利用する廉価型のウィークリーマンション．
	不定期個別求職メールの書き方	希望する企業に対して就活生が不定期の時期に自己アピール及び選考依頼のメールを送る際のメールの書き方．
結婚	結婚前の健康診断	中国では，結婚前に健康診断を受ける慣習がある．
	結婚式当日の新居でのいたずら	結婚式当日，親しい友人が新居に集まり新郎新婦にいたずらをする．
	結婚式当日の夜にベッドを装飾する儀式	結婚式当日の夜に，赤い寝具を用いてベッドを装飾するという中国独特の儀式を行う．
	新郎から新婦への結婚保証書	結婚の際，新郎が守るべき約束を結婚保証書に明記し渡す．

（注2）　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を用いた．



サジェストによって検索されるWebページ群

LDAを適用して，各商品ジャンル1〜23のトピックを生成

1.テレビ	2.パソコン	3.カメラ	4.掃除機	5.オーディオ機器
パナソニック　全録 パナソニック　ビエラ SONY　リモート視聴	ASUS　冬モデル2014 NEC　ノートパソコン 軽量 Lenovo g580　増設メモリ	SONY　ピクチャー プロファイル SONY　写真管理 パナソニック　アクション カメラ	東芝　コードレス掃除機 東芝　ルンバ ASUS　ロボット掃除機	SONY　ハイレゾ イヤホン ASUS　ヘッドホン SONY　ワイヤレス マイク
6.通信機器	**7.頭皮エステ用品**	**8.電気素子などを用いた装置**	**9.カメラ**	**10.スマホ**
NEC　無線lan パスワード NEC　ルーター　設定 三菱電機　pon	パナソニック　光エステ 効果 パナソニック　ムダ毛 パナソニック　まつげ くるん	東芝　距離継電器 東芝　変流器 三菱電機　張力制御	SONY　フォトキナ SONY　中判 パナソニック　ズミ ルックス	ASUS　スマホlte 富士通　スマホ SONY　ライフログ
11.液晶ディスプレイ	**12.太陽光発電**	**13.ホームベーカリー**	**14.ゲーム製品**	**15.プリンター**
シャープ　マルチ ディスプレイ ASUS　ピボット 三菱電機　ゲーム用 モニター	三菱電機　太陽電池 三菱電機　ソーラー シャープ　太陽電池 モジュール	パナソニック　ホーム ベーカリー レシピ パナソニック　パン 日立　ベーカリーレンジ	SONY　任天堂 パナソニック　ゲーム機 NEC　ゲームハード	東芝　プリンタ NEC　トナー シャープ　コピー機
16.クラウドサービスなど	**17.工具**	**18.除湿乾燥機**	**19.ロボット研究**	**20.照明器具**
富士通　データセンター 富士通　プライベート クラウド 三菱電機　ビッグデータ	日立　ハンマードリル パナソニック　充電工具 日立　充電工具	富士通　弱冷房除湿 パナソニック　弱冷房除湿 三菱電機　ムーブアイ 衣類乾燥　除湿機	パナソニック　パワード スーツ ASUS　距離センサ 日立　ロボット	東芝　逆富士　led パナソニック　調光 パナソニック　逆富士 蛍光灯

21.医療器具		22.バッテリー		23.データ管理システム	
富士通 三菱電機 NEC	脈拍計測 粒子線治療 病理	ASUS パナソニック Lenovo	急速充電 チャージパッド バッテリー リフレッシュ	富士通 NEC NEC	図面管理 システム 静脈認証 指ハイブリッド

クエリフォーカスの電気メーカ　　サジェスト

図5　日本国内主要電気メーカ10社のサジェストを情報源として生成した23商品ジャンル

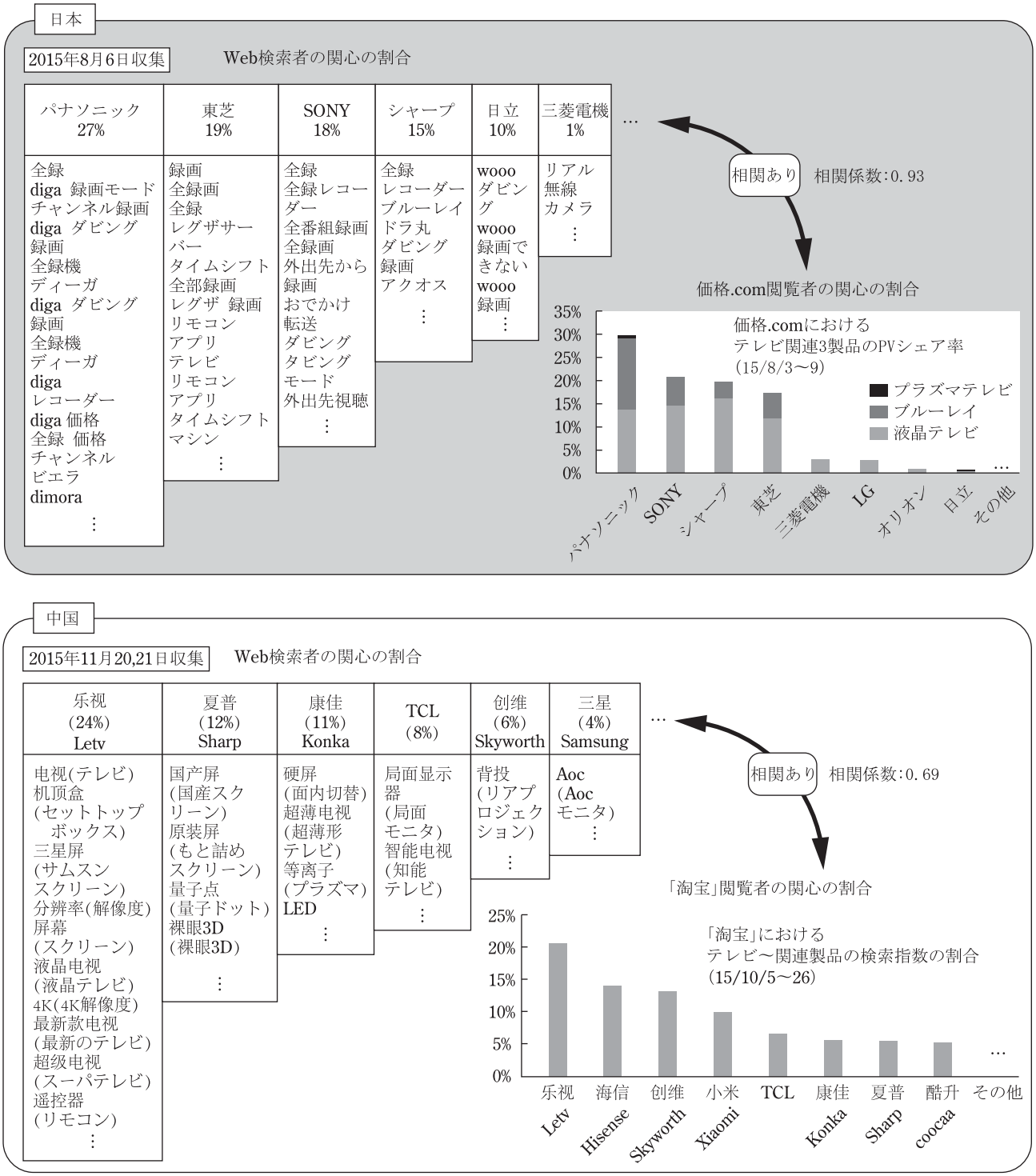


図 6　日中における「テレビ関連製品」分野での Web 検索者の関心の割合と通販サイトにおける関心の割合の間の相関分析

モデルを適用することによって，23 個の商品ジャンルをトピックとして生成することができた．このうちの「テレビ関連製品」のトピックを対象として，サジェストの種類数の統計と通販サイトにおける「テレビ関連製品」のページビュー統計を比較し，相関係数を測定したところ，図 6 上段に示す高い相関が得られた．図 6 下段に示すように，中国語を対象として同様の分析を行った結果においても，一定の相関が得られた．この結果から，企業名に関する Web 検索者の関心動向は，通販サイトにおけるページビュー統計（及び市場シェア統計）との間で一定の相関を持つことが示された．

5. おわりに

本稿では，検索エンジンサジェストを情報源として，Web 情報空間における Web 検索者の関心事項を優先的・選択的に収集した後，集約・俯瞰する方式について述べるとともに，その応用技術を紹介した．これらの方式を用いることによって，ネットユーザの「ネット上に分散している知識を，断片的にではなく，網羅的・集約的に知りたい」という要望を満たす技術が実現可能であることを示した．

なお，本稿で紹介した技術に関連する手法としては，文書集合中の個々の文書に対してラベルの付与を行い，付与されたラベルに基づいて文集の分類と俯瞰を行う手法[13),14)]が提案されている．また，メタ検索エンジンにおいて Web ページ検索結果の上位の記事を対象として，検索結果のクラスタリング及びラベル付けをした結果を提示するサービスとして，Yippy[注3)]が知られている．これらの手法においては，いずれも，閲覧対象の文書集合のみを用いて，文書分類後のクラスタに付与するラベルに相当する情報を抽出している．一方，本稿で紹介した手法においては，その文書集合に対して検索を行った検索者が情報要求観点として指定した語をラベルとして用いる点に大きな特徴がある．今後は，これらの従来方式と比較した場合の性能的優位性について検討を行う必要がある．

謝辞　本稿で紹介した研究の一部は，平成 26～28 年度科研費基盤研究（B）「ウェブ検索における情報要求観点の言語間比較・対照分析システムの研究」及び NII 共同研究における研究成果である．

（注 3）　http://yippy.com/

文　献

（1） Y. Inoue, T. Imada, S. Doi, L. Chen, T. Utsuro, and Y. Kawada, “Selecting Web search results of diverse contents with search engine suggests and a topic model,” Proc. 30th WAINA, pp. 455-460, 2016.
（2） T. Imada, Y. Inoue, L. Chen, S. Doi, T. Utsuro, and Y. Kawada, “An empirical study on optimal correlation between market share and concerns on companies measured through search engine suggests,” Proc. 30th WAINA, pp. 214-219, 2016.
（3） 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.
（4） 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.
（5） 聶　添，陳　磊，今田貴和，宇津呂武仁，河田容英，“検索エンジン・サジェストを情報源とするウェブ検索者の情報要求観点の日中間対照分析，”知能と情報，vol. 27, no. 1, pp. 527-532, 2015.
（6） 宇津呂武仁，徐　凌寒，聶　添，趙　辰，李　佳奇，河田容英，“企業名に関する関心動向のトピックモデリングを用いた日中市場シェアの分析，”第 30 回人工知能学会全国大会論文集，2016.
（7） http://nlp.iit.tsukuba.ac.jp/research/list03-sg-cluster.html
（8） https://youtu.be/YXupu8r8Ufc
（9） D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993-1022, 2003.
（10） http://nlp.iit.tsukuba.ac.jp/research/list03-sg-know-how.html
（11） http://nlp.iit.tsukuba.ac.jp/research/list03-sg-know-how-jc.html
（12） http://nlp.iit.tsukuba.ac.jp/research/list03-sg-social-sensor.html
（13） 馬場康夫，黒橋禎夫，“キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰，”情処学論，vol. 50, no. 4, pp. 1399-1409, 2009.
（14） 戸田浩之，中渡瀬秀一，片岡良治，“特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案，”情処学論：データベース，vol. 46, no. SIG 13(TOD 27), pp. 40-52, 2005.

（平成 28 年 4 月 3 日受付　平成 28 年 5 月 15 日最終受付）

宇津呂　武仁（うつろ　たけひと）（正員：シニア会員）
平 6 京大大学院博士課程了．博士（工学）．奈良先端大，豊橋技科大，京大を経て，平 18 から筑波大．現在，システム情報系教授．自然言語処理，Web マイニング，人工知能の研究に従事．情報処理学会，人工知能学会，言語処理学会，日本音響学会，ACL 各会員．