ウェブ検索者の情報要求観点の集約*

小池 大地† 鄭 立儀† 今田 貴和‡ 守谷 一郎† 井上 祐輔‡
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† 筑波大学理工学群 工学システム学類‡
(株) ログワークス§ 国立情報学研究所¶

1 はじめに

近年インターネットの普及により，多くのユーザはウェブページ上から情報を得ている．情報を収集する手段は，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うのが一般的である．各検索エンジン会社においては，検索者が入力した検索語のログが蓄積されており，多数の検索者が検索した検索語に対して，強い関連を持つ語を検索エンジン・サジェストとして提示するシステムを提供している．ここで，本論文では，検索者が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，より詳細な情報を得るために，AND 検索の形で二つ目以降に入力する語を**「情報要求観点」**と呼ぶ (図 1)．情報要求観点には，ウェブ検索者の関心事項そのものが反映されており，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の関心事項を集約，俯瞰する手法を確立することを目的とする．

本研究においては，まず，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する．具体的には，一つの検索対象に対して，最大約 1,000 語のサジェストを収集する．ただし，収集されるサジェストの多くは話題が重複し冗長である．これを改善するために，冗長性を考慮してサジェストの集約を行う．具体的には，各サジェストを用いた検索によって収集されるウェブページのスニペットをサジェストに付与し，これをクラスタリングすることにより，冗長なサジェストを集約する．この手法を用いることにより，サジェストが示す話題を考慮し，類似する話題ごとに集約してサジェストを提示することが可能となる．閲覧者が検索対象に関する前提知識をほとんど

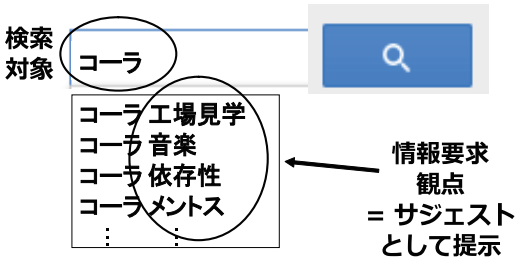


図 1: 検索エンジン・サジェストにおける情報要求観点の例

持たない場合には，より詳細な情報を得るための情報要求観点を自身で思いつくことが難しい．しかし，本研究の手法によって提示されるサジェストの集約結果を参照することにより，検索対象に関して収集された膨大な数の情報要求観点を容易に俯瞰することができ，情報を効率よく収集することができる．本研究では，以上の考え方に基づき，集約したサジェストをクラスタごとに一覧で提示し，閲覧者があるクラスタを選択すると，そのクラスタに分類されたサジェストと関連性の強いウェブページの一覧を提示するインタフェース (図 3 参照) を作成し，その有効性を示す．

2 ウェブ検索者の情報要求観点の集約

2.1 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[1] 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「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提示されるので，それらの収集を行う．

*Aggregating Viewpoints of Web Search Information Needs
†Daichi Koike, Liyi Zheng, Ichiro Moriya, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takakazu Imada, Yusuke Inoue, College of Eng. Sys., School of Science and Engineering, University of Tsukuba
§Yasuhide Kawada, Logworks Co., Ltd.
¶Noriko Kando, National Institute of Informatics

[1] https://www.google.com/

表 1: 評価用の検索対象ごとのサジェスト数，および，参照用クラスタの数

検索対象	コーラ	猫	犬
サジェスト数	791	931	939
参照用クラスタの数	100	146	155

2.2 情報要求観点の集約

ウェブページの収集においては Yahoo! Search BOSS API[2] を用い，検索エンジン API に対して，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 クエリ当たり最大 1,000 件のウェブページ，およびスニペットを取得することが可能である．また，検索クエリについては，検索対象と収集したサジェストの AND 検索，および，サジェスト単体の 2 通りの方法で作成した．そして，作成した各検索クエリごとに検索を行うことで，最大 1,000 件のウェブページ，および，スニペットを取得する．そして，これらのスニペットから内容語の頻度ベクトルを作成し，これを各サジェストの文書ベクトルとする．次に，二つのクラスタの間で，全てのサジェスト組の間の類似度が下限値以上となる場合のみ，二つのクラスタを併合するボトムアップクラスタリングにより各サジェストの文書ベクトルのクラスタリングを行う．以上の手順により出力されたクラスタリングの結果の例の一部を表 2 に示す．

2.3 評価

サジェストのクラスタリング結果の評価においては，参照用クラスタを用いる．評価尺度としては，以下の再現率，適合率を用いる．ただし，サジェスト数が 2 以上となるクラスタのみを評価対象とする[3]．全検索対象に対する再現率・適合率の推移をプロットした結果を図 2 に示す．

$$\text{再現率} = \frac{\text{出力された各クラスタに含まれるサジェストのうち，参照用クラスタに含まれるサジェストの和}}{\text{各参照用クラスタに含まれるサジェスト数の和}}$$

$$\text{適合率} = \frac{\text{出力された各クラスタに含まれるサジェストのうち，参照用クラスタに含まれるサジェストの和}}{\text{出力された各クラスタに含まれるサジェスト数の和}}$$

検索対象とサジェストの AND 検索，および，サジェスト単独の二通りの評価結果を比較すると，最適な類似度下限値における評価結果の間には大きな違いはなかった．ただし，クラスタリングされたサジェストの結果を見ると，検索対象とサジェストの AND 検索でのみ得られるクラスタが存在する．例えば，検索対象が「コーラ」の場合，サジェスト単独での検索結果においては,「工場見学」と「京都 見学」の検索結果は全く異なるため，これらのサジェストが同一クラスタにクラスタリングされることはないが，検索対象とサジェストの AND 検索の場合には，これらのサジェストはどちらも「京都にあるコーラの製造工場の見学」という，共通の話題のウェブページが検索結果となるため，同一クラスタにクラスタリングされるという結果となった．

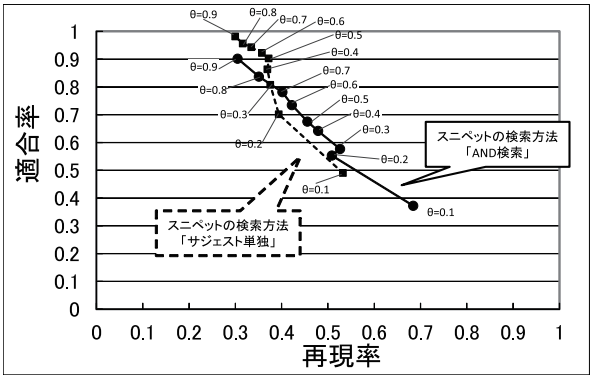


図 2: ウェブ検索者の情報要求観点の集約: 評価結果 (全検索対象分のマクロ平均)

3 情報要求観点およびウェブ検索結果の俯瞰

図 3(a) に示すように，収集したサジェスト全てをそのまま一覧で提示した場合，全体でいくつの話題の情報要求観点が提示されているかを俯瞰することは困難である．また，サジェストを用いて検索を行う際には，話題が重複する冗長なサジェストを指定した検索を繰り返し行なうなどの非効率的な検索を余儀なくされることが予測され，できるだけ多様な話題の情報を効率よく収集する場合には大きな障害となる．この問題を解決するために，本研究のインタフェースにおいては，各サジェストをクラスタに集約し，各クラスタ内のサジェストをリスト形式で閲覧する仕様とした．これにより，閲覧者は，話題が類似するサジェストをまとめて俯瞰することができるようになり，この機能によって情報要求観点の俯瞰を実現した．また，図 3(b) に示すように，収集されたウェブページについても，話題が重複するウェブページを集約した上で [5]，クラスタに分類されたサジェストとの関連性の強いウェブページを一覧で提示した．これにより，話題が重複する冗長なウェブページをスキップするとともに，話題

[2] http://developer.yahoo.com/search/boss

[3] クラスタリング手法の評価を行うにあたり，最初に，参照用クラスタと，出力されたクラスタとの対応付けを行った．対応付けにあたり，上記の計算方法により，評価対象となる出力クラスタに対して全参照用クラスタとの F 値を算出し，その値が最も高いクラスタを探索することによって，1 対 1 の対応付けを行った．これを全出力クラスタに対して行うことで，クラスタリング結果の評価値を算出した．

表 2: 提案手法による検索エンジン・サジェストのクラスタリング結果の例

検索対象	人手により付与したラベル	スニペットの検索方法	
		検索対象とサジェストとの AND 検索	サジェスト単独での検索
コーラ	京都の工場見学	見学，工場見学，京都 見学	京都，京都 見学
	料理に使う	肉，肉 柔らかく，レシピ，料理 レシピ，料理，鶏肉，料理 鶏肉，スペアリブ	肉，肉 柔らかく，牛肉，ステーキ
			料理 鶏肉，レシピ 公開，レシピ，料理 レシピ，煮，鶏肉，料理，レシピ オークション，鶏肉
	お酒と混ぜる	テキーラ，ウォッカ，ブランデー	(該当クラスタなし)
	爆発させる	ラムネ 実験，実験，表面張力	ラムネ，ラムネ 実験
猫	体調に関して	食欲旺盛，食欲，元気が無い，ぐったり	(該当クラスタなし)
	行動範囲	行動，行動範囲，マーキング	行動，行動範囲
	品種に関して	種類 日本，種類，種類 性格，性格，模様 種類，特徴，柄 性格	性格，種類 性格
			模様 種類，模様
犬	餌に関して	野菜，にんじん，手作りおやつ，おやつ，ジャーキー，チーズ，レバー，生肉	おやつ，手作りおやつ
	しつけ	散歩 しつけ，トイレのしつけ，しつけ教室，トイレで寝る，粗相，リーダーウォーク，行動，問題行動，マウンティング，散歩 歩かない，しつけ，散歩，ジャンプ 教え方	散歩 しつけ，しつけ教室，トイレのしつけ，しつけ
	ペットグッズ	ベッド おしゃれ，ベッド，ベッド 夏，ベッド 通販，プレゼント グッズ，グッズ，グッズ 通販，通販，用品，便利グッズ，プレゼント，お出かけ，おもちゃ，ピュアラ，ペット用品，サークル，ケージ，ゲージ	グッズ，グッズ 通販，便利グッズ
	与えられない餌	チョコレート，たまねぎ，中毒，玉ねぎ，チョコ 致死量，チョコレート 症状	チョコレート，チョコレート 症状

が関連するウェブページを集約的にまとめて提示することによって，ウェブ検索結果の俯瞰を実現した．

4 関連研究

関連研究として，Web ページの検索結果を分類し，各分類に対して適切な要約文を付与する手法 [4]，検索された個々の Web ページ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分類を行う手法 [1, 3, 9]，階層的なトピックの体系を推定する手法 [2] 等が提案されている．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，ファセット体系およびファセットラベルに相当する情報を抽出している．また，メタ検索エンジンにおいてウェブページ検索結果の上位 200 記事程度を対象にして，検索結果のクラスタリングおよびラベル付けをした結果を提示するサービスとして，Yippy [4] が知られている．一方，文献 [7] においては，与えられた文書集合の話題を俯瞰するタスクにおいて，Wikipedia を知識源として，検索された文書集合全体にわたる分野や話題の粒度にまで抽象化されたファセット体系を用いる手法を提案している．これらの先行研究においては，いずれも，与えられた文書集合における話題の広がりを俯瞰することに焦点が当てられている．その他，文献 [8] においては，検索エンジンの検索ログを情報源として語の意味関係に関する多様な知識を獲得する方式を紹介している．

[4] http://yippy.com/

5 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，その集約を行う手法を提案した．今後の課題としては，分類語彙表 [6] や Wikipedia 等から収集可能な既存の語彙知識を用いることにより精度を改善することが挙げられる．

参考文献

[1] 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. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1399–1409, 2009.

[2] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS'03*, 2003.

[3] W. de Winter and M. de Rijke. Identifying facets in query-biased sets of blog posts. In *Proc. ICWSM*, pp. 251–254, 2007.

[4] 原島純, 黒橋禎夫. PLSI を用いたウェブ検索結果の要約. 言語処理学会第 16 回年次大会論文集, pp. 118–121, 2010.

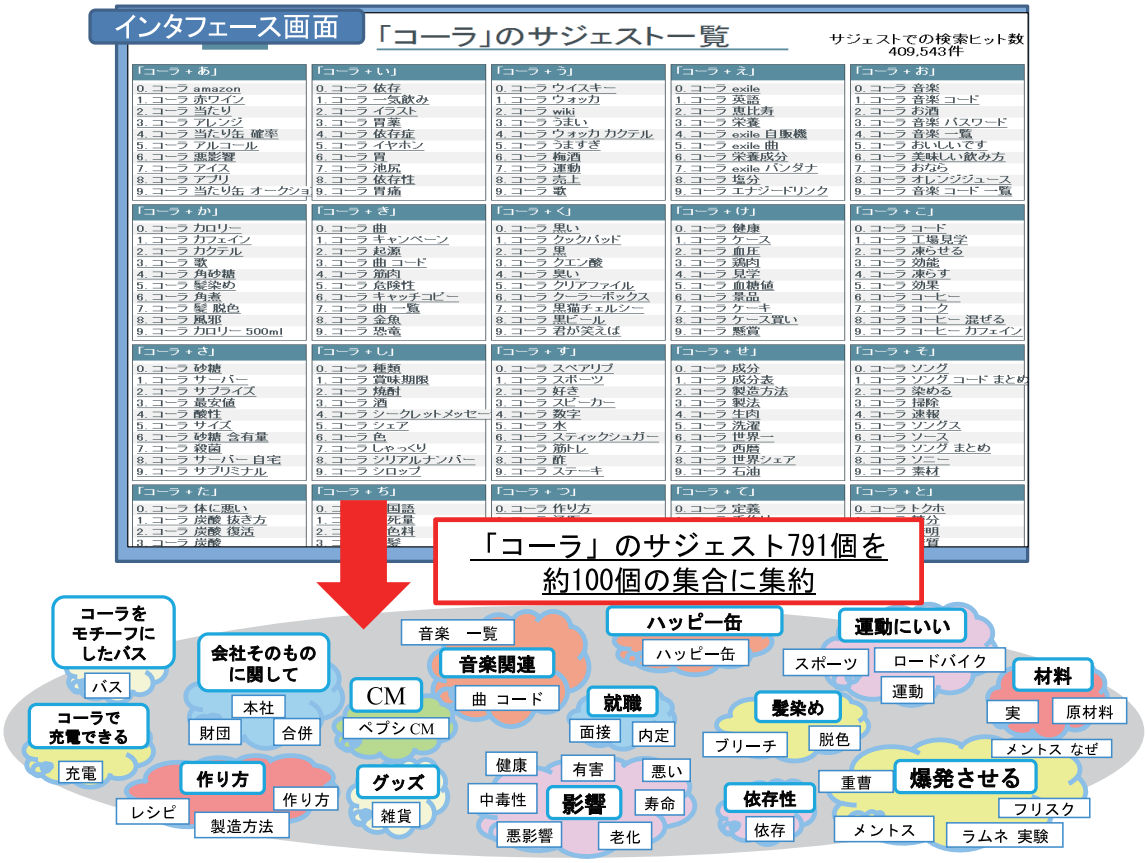
[5] 井上祐輔, 今田貴和, 守谷一朗, 陳磊, 宇津呂武仁, 河田容英, 神門典子. 冗長な情報要求観点の集約によるウェブ検索結果の集約. 第 28 回人工知能学会全国大会論文集, 2014.

[6] 国立国語研究所: 分類語彙表 (増補改訂版). 国立国語研究所資料集 14. 大日本図書, 2004.

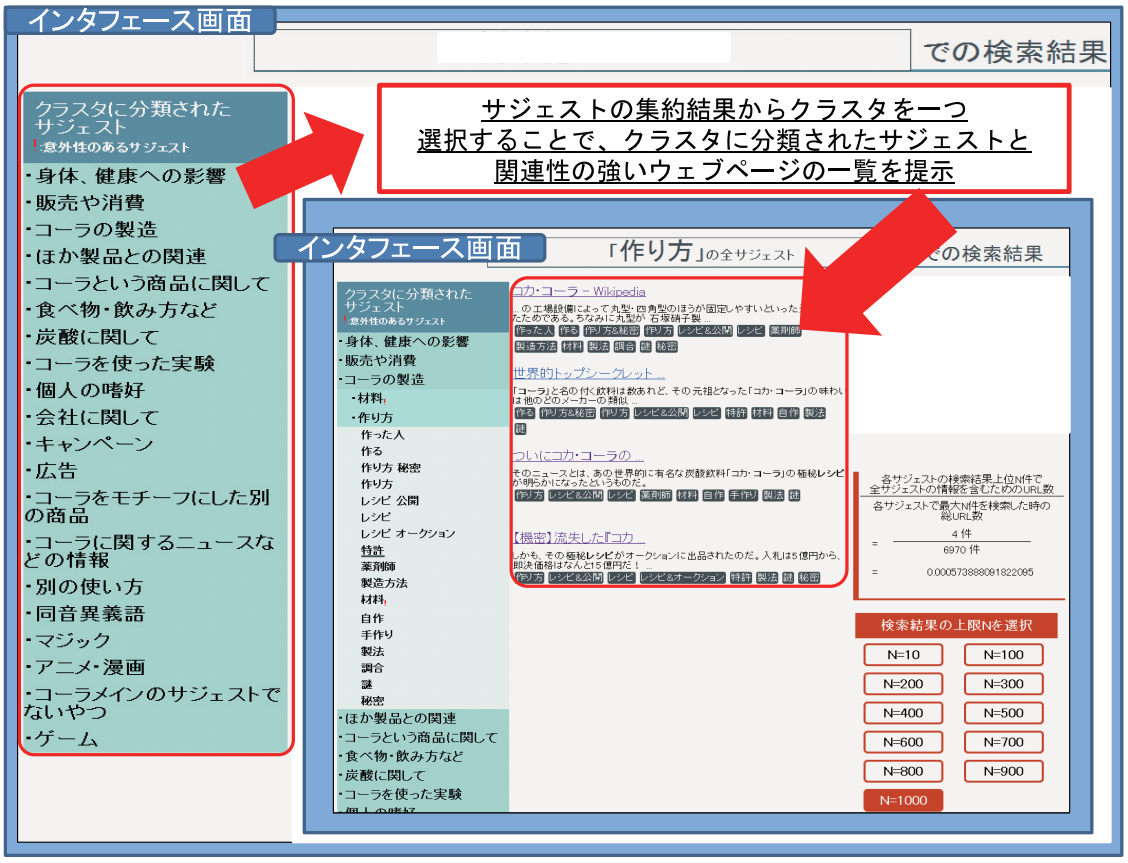
[7] 牧田健作, 鈴木浩子, 小池大地, 宇津呂武仁, 河田容英. Wikipedia を知識源とする分野トピックモデルの推定と分析. 情報処理学会研究報告, Vol. 2012–DBS–155, , 2012.

[8] M. Pasca. Web search queries as a corpus. In *Tutorial at Proc. 25th ACL*, 2011.

[9] 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. 情報処理学会論文誌：データベース, Vol. 46, No. SIG 13(TOD 27), pp. 40–52, 2005.



(a) サジェストの集約結果



(b) ウェブ検索結果の俯瞰

図 3: サジェスト集約結果・ウェブ検索結果の俯瞰インタフェース