# 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*

今田 貴和† 小池 大地‡ 守谷 一郎‡ 宇津呂 武仁§ 神門 典子¶
筑波大学理工学群工学システム学類† 筑波大学大学院 システム情報工学研究科‡
筑波大学システム情報系§ 国立情報学研究所¶

## 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[1]．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．ここで，本論文では，検索対象に対して提示されるサジェストのうち，ウェブ検索者の知識の範囲では検索対象との間の関係を推定できないものに着目し，検索対象と検索エンジン・サジェストとの間の関係を提示する枠組みを提案する．具体的には，検索結果上位の文書のタイトル中に頻出する最長文字列を手がかりとすることにより，検索対象とサジェストとの間の関係を提示することが可能であることを示す．

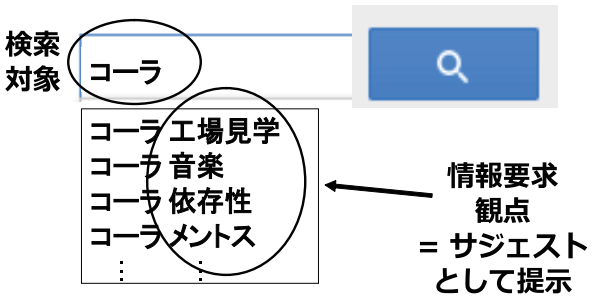


図 1: 検索エンジン・サジェストにおける情報要求観点の例

## 2 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[2] 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．

## 3 検索対象と検索エンジン・サジェストとの間の関係

本論文では，検索エンジン・サジェストを大別し，ウェブ検索者の知識の範囲で検索対象との間の意味関係を推定できるものと，ウェブ検索者の知識の範囲では検索対象との間の意味関係を推定できないものとに分ける．例えば，図 2 (a) の例のように，検索対象が「コーラ」，サジェストが「成分」の場合，コーラの「成分」(実際の成分は砂糖，シナモン，バニラ等) を調べようとした多数のクエリが検索ログに含まれることから出現したと推定可能である．一方，図 2 (b) の例のよう

*Presenting Relation of Web Search Query and a Search Engine Suggest

†Takakazu Imada, College of Engineering Systems, School of Science and Engineering, University of Tsukuba
‡Daichi Koike, Ichiro Moriya, Graduate School of Systems and Information Engineering, University of Tsukuba
§Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba
¶Noriko Kando, National Institute of Informatics

[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．

[2]http://www.google.com/

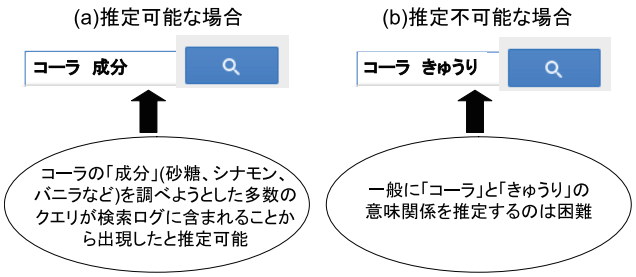


図 2: 検索対象とサジェストとの間の意味関係の推定

に，検索対象が「コーラ」，サジェストが「きゅうり」の場合は，一般に「コーラ」と「きゅうり」の意味関係を推定するのは困難である．本論文では，このように，ウェブ検索者の知識の範囲では，検索対象とサジェストを提示されただけでは，検索対象との間の意味関係を推定できないサジェストに焦点を当てて，それらの意味関係を提示することを目的とする．

ここで，本論文の第一著者がウェブ検索者であるという状況を想定して，10 種類の検索対象について，実際に検索対象とサジェストのみからでは両者の意味関係を推定できないサジェストの割合を調査した結果を表 1 に示す．例えば，検索対象が「犬」の場合，意味関係を推定できない割合は 24.6% であった．ここでは，例えば，サジェストが食べ物を表す語であった場合，その食べ物を犬に与えてよいのか，あるいは，与えてはいけないのかは，検索対象とサジェストのみを見ただけでは判定できない．また，サジェストが「住所」の場合，実際に検索をしてみると，「犬を飼い始めた時などには，住所登録などが必要な地域もある」という意味関係であることが分かるが，検索対象とサジェストのみからその意味関係を推定することは容易ではない．一方，検索対象が「就活」の場合，判定者である本論文第一著者が大学生であることから，すべてのサジェストとの間の意味関係が推定可能であった．検索対象「就活」のサジェストとしては，例えば，「コート」，「エントリーシート」等が含まれていた．

## 4 検索結果上位ページのタイトルの有効性

検索対象とサジェストとの間の意味関係の推定において，検索結果上位ページのタイトルが有効か否かについて述べる．検索対象が「コーラ」，サジェストが「きゅうり」の場合に，検索結果上位のページのタイトル例を図 3 に示す．このタイトル中には「キュウリ味のコーラ」という文字列が含まれており，この文字列を手がかりとすることによって，検索対象「コーラ」

表 1: 判定者が検索対象との間の意味的関係を推定できないサジェストの割合

| 検索対象 | 総サジェスト数 | 判定者が検索対象との間の意味関係を推定できないサジェストの割合 (%) (意味関係を推定できないサジェスト数/判定サジェスト数) |
|---|---|---|
| コーラ | 791 | 24.7 ( 195 / 791 ) |
| 犬 | 939 | 24.6 ( 99 / 403 ) |
| 猫 | 931 | 23.1 ( 106 / 459 ) |
| 尖閣諸島 | 657 | 21.3 ( 140 / 657 ) |
| 海賊版 | 595 | 18.2 ( 108 / 595 ) |
| ニート | 715 | 16.7 ( 65 / 390 ) |
| 卵 | 849 | 16.3 ( 84 / 515 ) |
| 貯金 | 729 | 10.3 ( 75 / 729 ) |
| 京都 | 970 | 0.9 ( 2 / 229 ) |
| 就活 | 906 | 0.0 ( 0 / 507 ) |

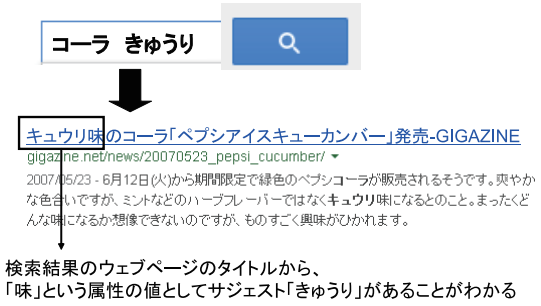


図 3: 検索対象「コーラ」とサジェスト「きゅうり」の関係推定におけるページタイトルの利用

とサジェスト「きゅうり」の間に「キュウリ味」という意味関係を推定することができる．また，タイトルの他に，検索結果上位ページのスニペットについても，関係推定において有効であると考えられる．ただし，本論文の方式によって評価した結果においては，スニペット単独ではタイトルよりも低い評価結果となった．

## 5 検索結果上位ページのタイトルからの関係語収集

本論文では，以下の手順によって，検索対象とスニペットの間の関係を表す語の候補を収集する．

(1) 検索対象とサジェストの AND 検索の検索結果において，順位 $i$ 位のタイトルを $T_i$，順位 $n$ 位まで

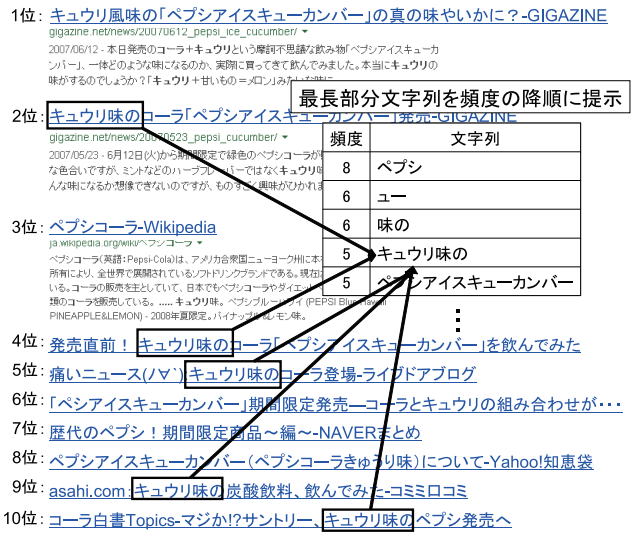


図 4: 検索結果上位ページのタイトルからの関係語収集

のタイトルの集合を $\mathbb{T}_n$ とする．

(2) $\mathbb{T}_n$ 中の 2 文字以上の任意の部分文字列を $s$ とする．タイトル $T_i(\in \mathbb{T}_n, i = 1, \ldots, n)$ のうち，文字列 $s$ を含むものの数を $freq(s)$ と表す．

(3) $\mathbb{T}_n$ 中には，2 文字以上の部分文字列が合計 $m$ 個含まれるとすると，それらの部分文字列 $s$ を $freq(s)$ の降順に並べた結果は，

$$s^1,\ s^2,\ \ldots,\ s^j,\ \ldots,\ s^m \qquad (1)$$

と書ける．ここで、$s^j$ は順位 $j$ 位の文字列を表す．

次節の評価においては，Google における検索順位 $n = 10$ 位までの検索結果のページタイトルを対象として以上の処理を行い，順位 $j$ 位までの文字列の評価を行った．検索対象が「コーラ」，サジェストが「きゅうり」の場合の例を図 4 に示す．この場合，頻度最大の部分文字列は「ペプシ」，頻度の降順 4 位の部分文字列は「キュウリ味の」であり，$freq$(ペプシ) = 8，$freq$(キュウリ味の) = 5 となる．

## 6 評価

一つの検索対象に対して収集したサジェストの集合を $G$，一つのサジェストを $g(\in G)$ とする．ここで，サジェスト $g$ に対して，検索対象とサジェスト $g$ との AND 検索の結果の上位 10 ページまでのタイトルを参照して，いずれかのタイトルに含まれる文字列のうち，検索対象とサジェスト $g$ との関係を表していると考えられる文字列の集合を人手で作成したものを $R(g)$ とする．そして，式 (1) において順位 $j$ 位の文字列を $s^j$

表 2: 検索対象「コーラ」に対して調査対象としたサジェスト 32 個の一覧

| |
|---|
| きゅうり，黒，緑茶，充電，浄水器，浄水，わさび，ブリーチ，染め方，脱色， アイス，バニラアイス，表面張力，メントス，虫，生肉，就職，内定，ルート配送，見学，京都，京都 見学，ミャンマー，掃除，汚れ落とし，洗濯，便器，メッセージ，シークレットメッセージ，ペットボトル 文字，サプライズ，飛行機 |

として，

$$s^j \in R(g)$$

となる $j$ の最小値を $j_{min}(g)$ とする．このとき，以下の二つの割合を用いて検索結果上位ページのタイトルから関係語候補を収集する方式の評価を行う．

(1) 頻度の降順に $s$ を並べた時の順位を $i$ として，$j_{min}(g) \leq i$ となるサジェスト $g$ の割合として次式を用いる．

$$rate_r(i) = \frac{\left|\left\{g \in G \middle| j_{min}(g) \leq i\right\}\right|}{\left|G\right|}$$

(2) 各サジェスト $g$ に対して，順位 $j = j_{min}(g)$ となる時の頻度を $freq(s_g^j)$ とすると，頻度 $f$ において，$freq(s_g^j) \geq f$ となるサジェスト $g$ の割合として次式を用いる．

$$rate_f(f) = \frac{\left|\left\{g \in G \middle| j = j_{min}(g) \text{ として } freq(s_g^j) \geq f\right\}\right|}{\left|G\right|}$$

評価用の検索対象を「コーラ」，検索対象「コーラ」に対する評価用サジェスト 32 個 (表 2) について，$rate_r(i)$，および，$rate_f(f)$ をプロットした結果を図 5，および，図 6 にそれぞれ示す．図 5 より，4 位までの文字列を参照することにより，全サジェストの 60%程度に対して，検索対象との意味関係が推定可能である．一方，図 6 より，頻度 3 以上の文字列を参照することにより，同様に，全サジェストの 60%程度に対して，検索対象との意味関係が推定可能である．

## 7 関連研究

文献 [5] においては，検索エンジンによって提示されるサジェストのうち冗長なものを集約するとともに，関連するサジェストを集約的に俯瞰する枠組みを提案し

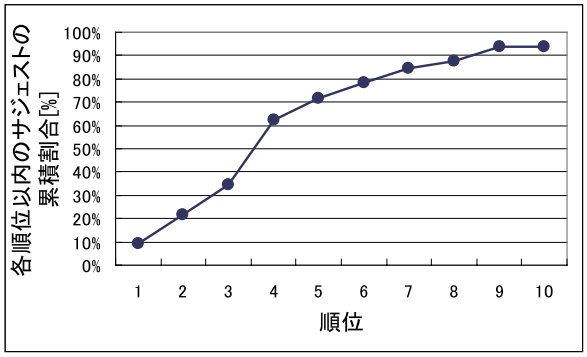

図 5: 評価結果：関係語候補の頻度降順順位に対する累積割合

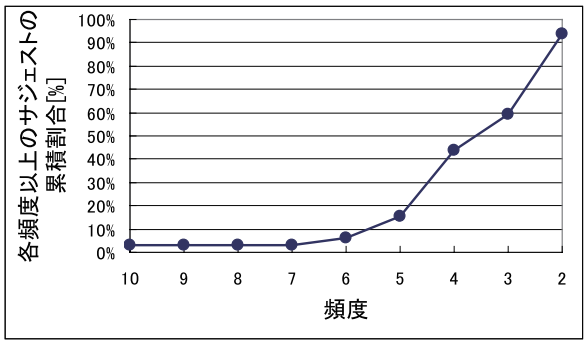

図 6: 評価結果：関係語候補の頻度に対する累積割合

ている．さらに，サジェストに対して収集されるウェブページ集合に対して，話題が重複する冗長なウェブページを集約して俯瞰する枠組みを提案している．本論文は，このサジェスト俯瞰の枠組みの一環として，検索対象とサジェストとの間の意味関係を簡潔に提示する方式として位置付けられる．

このウェブ検索結果俯瞰に関する関連研究としては，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という手法 [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．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[3] が知られている．一方，文献 [6] 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．

その他，文献 [5] に関連して，文献 [7] においては，検索エンジン・サジェストの中でも，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，提示する方式を提案している．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．一方，文献 [7] においては，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を示している．また，検索エンジン・サジェストに関連する研究として，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．

[3] http://yippy.com/

## 8 おわりに

本論文では，ウェブ検索者の関心事項を収集するための情報源として検索エンジン・サジェストに焦点を当て，検索対象と検索エンジン・サジェストとの間の関係を提示する枠組みを提案した．特に，検索結果上位の文書のタイトル中に頻出する最長文字列を手がかりとすることにより，検索対象とサジェストとの間の関係を提示することが可能であることを示した．今後は，検索エンジン・サジェスト集約結果 [5] において本論文の関係提示方式を組み込む手法を確立する予定である．

## 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] 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.

[6] 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.

[7] 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.

[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.