

QA コンテンツからの観点抽出と それにもとづくウェブ検索結果の再ランキング

山本 岳洋^{†1,†2} 中村 聡史^{†1} 田中 克己^{†1}

我々は膨大な検索結果をさまざまな観点から閲覧し、効率良く多くの方法を閲覧するための仕組みとして、検索結果全体の俯瞰的インタフェース（タームクラウド）と語ベース適合フィードバックにもとづく検索結果の再ランキングシステムを提案してきた。提案してきた俯瞰的インタフェースは有効なものであったが、提示する単語は単純な語の出現頻度にもとづき選ばれていたため、ユーザの再ランキング行為を促進するには不十分であった。そこで本稿では、よりユーザの興味を引く単語を提示するため、“有名な観光地”や“美味しい和菓子屋”といった観点に着目し、そうした観点を QA コンテンツから抽出する手法を提案する。また、本稿では提案手法により得られた観点にもとづくウェブ検索結果の再ランキング手法についても検討する。ユーザ実験の結果、提案手法により得られた観点は、従来のクエリ推薦や頻度にもとづく単語抽出手法にもとづく手法に比べてユーザの興味を引く単語を多く提示可能であることが分かった。

Extracting Aspects from QA contents and its Application to Reranking Web Search Results

TAKEHIRO YAMAMOTO,^{†1,†2} SATOSHI NAKAMURA^{†1}
and KATSUMI TANAKA^{†1}

We have previously proposed a system that reranks Web search results based on users' term-based feedback. The system enables users to explore search results from diverse viewpoints. We have also proposed an interface called Term-Cloud which are generated from frequent terms appear in Web search results. However, these terms are not enough to support users reranking operations. In this paper, we propose a method to extract aspects like “famous spots” or “delicious restaurant” by using QA contents in order to suggest users more interesting terms. In addition, we also propose a method to rerank Web search results based on those aspects. We evaluated the usefulness of our method of extracting aspects by comparing with several baseline methods.

1. はじめに

近年、ウェブ上のコンテンツが膨大になり多種多様な情報が閲覧可能になったことで、人々は些細な調べ事だけではなく、晩ご飯のレシピ探し、物品の購入、旅行の計画にいたるまで、さまざまな目的でウェブ検索を利用するようになってきた。そのため、明日の天気や言葉の意味を調べるなどのように1度の検索だけで求める情報を得るのではなく、何度も反復的に検索しながら情報を収集する検索が増加してきている¹⁾。たとえば、あるユーザが京都観光の計画を立てるためにウェブ検索を利用することを考えてみる。ユーザは京都にどのような観光名所があるのか、何が名物なのか、どこに宿泊するべきなのかといったことに関する情報を、さまざまな観点から繰り返し検索を行いながら情報を収集し、京都観光の計画を具体化していく。しかし、現状のウェブ検索エンジンでこのような検索を支援するには以下のような問題点がある。

- ユーザにとって自らの情報要求を上手く表したクエリを作成することは容易ではない。特に、検索対象に対する知識が乏しい場合は、具体的なクエリを作成することが困難である。たとえば、京都の有名な寺社を訪れたいと思ったとしても、具体的な寺社名を知らなければユーザはクエリを作成することができない。その結果、ユーザが検索エンジンに入力するクエリは短く曖昧なものになってしまうことが多い²⁾³⁾。
- ユーザが“京都 有名 寺社”というクエリを検索エンジンに入力したとしても、現状の検索エンジンは、“京都”、“有名”、“寺社”という語が記載されたページを検索結果として返すだけであるため、ユーザの情報要求を満たす情報であるとは限らない。

結果として、現状の検索システムはユーザが本来興味を持つべき情報への到達を妨げていると考えられる。Käkiらや我々のグループは、ウェブ検索エンジンユーザの多くは検索結果の上位3~5件までしか閲覧していないと報告している³⁾⁴⁾。検索結果の下位に興味を引く情報があったとしても、現状ではそうした検索結果に到達することは困難である。

一方、我々は検索システムを介さず人との対話で情報を探する場合、そこには検索システム特有の制約は存在せず人と人との自由な対話を通じて情報をやり取りする。たとえば、京

†1 京都大学大学院 情報学研究科 社会情報学専攻

Department of Social Informatics, Graduate School of Informatics, Kyoto University

†2 日本学術振興会 特別研究員

JSPS Research Fellow

都観光に関する情報を人に聞く場合は、「有名な寺社は見ておきたいのだけれども、どこに行くべきか?」であるとか「美味しい和菓子をどこか知らないか?」といったように、自らの情報要求をそのままの形で人に伝達し、それを聞いた人は、「有名などころなら金閣寺と清水寺がおすすめ」、「出町ふたばの豆餅が美味しいですよ」といったように回答する。このような、人と人の対話の情報をウェブ検索に生かすことができれば、上記の問題点を解決でき、ユーザは自らの興味に従ってさまざまな観点から情報を収集できると考えられる。

そこで本研究では、人と人による質問と回答のやり取りが行われる、質問応答サイトに着目し、質問応答サイトに蓄積された質問と回答の情報(QAコンテンツ)を利用してユーザのウェブ検索を支援する方法を提案する。QAコンテンツ上の質問にはユーザの情報要求が、そして質問に対する良い回答にはユーザの情報要求を満たす情報が含まれているであろう(図1)。本稿では、こうしたQAコンテンツからの知識を、我々がこれまで提案してきたユーザインタラクションにもとづくウェブ検索結果の再ランキングシステム^{*15)}に利用する。このシステムでは、ユーザは単語単位のフィードバックをシステムに伝達でき、システムはユーザからのフィードバックにもとづき検索結果を再ランキングしユーザに提示する。

我々が提案してきたシステムでは、ユーザの興味を引き出し、より多くの観点からの情報閲覧を促進するため、図2(a)に示すようにウェブ検索結果中に出現する単語の集合を、タームクラウドと呼ばれるタグクラウド形式で俯瞰的に提示することを行っていた⁶⁾。しかし、これまでの手法では単純な語の出現頻度にもとづき単語を提示していたため、たとえば、「京都 観光」というクエリに対して、「情報」や「案内」といった一般的な単語が多く提示されてしまい、ユーザの再ランキングを促進するには不十分であった。また、ユーザが指定した単語を含むか含まないかという基準だけで検索結果の再ランキングを行っており、ユーザの意図を十分に検索結果に反映できていなかった。こうした問題を解決するため、本研究では図1に示す2つの手法を提案する。

- (1) 質問からの修飾語付き観点の抽出
- (2) 回答を用いた検索結果の再ランキング

まず、質問から観点を抽出する手法を提案する。本稿では、ユーザの興味を引く観点として、「有名な寺社」や「美味しい和菓子屋」といった「修飾語付き観点」に着目し、QAコンテンツの質問からウェブ検索に有用な修飾語付き観点を抽出する手法について述べる。図2(b)のように抽出された修飾語付き観点をユーザに明示的に提示することができれば、単

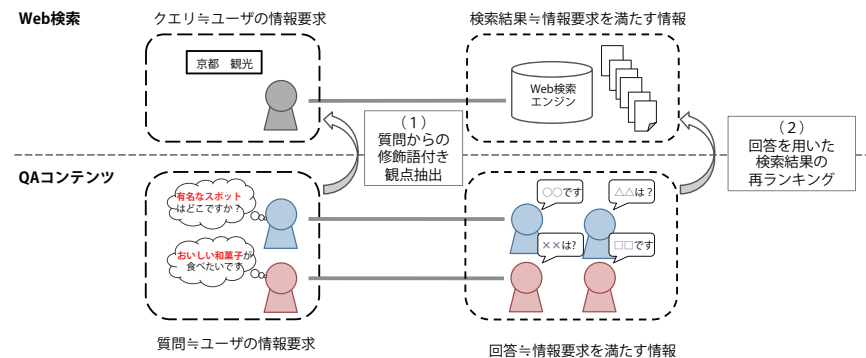


図1 QAコンテンツのウェブ検索への利用

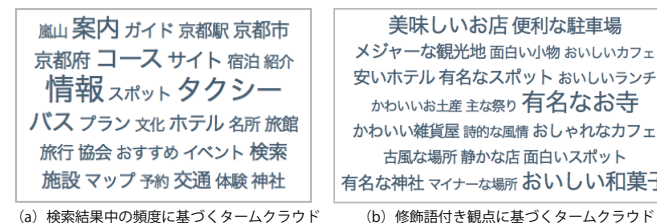


図2 クエリ「京都 観光」における2種類のタームクラウド

純な頻度にもとづくタグクラウドよりもユーザの興味を引き、インタラクションを促進することが可能になると考えられる。

さらに、ユーザがそのような修飾語付き観点を指定した際の検索結果の再ランキング手法についても提案する。これは、ユーザが指定した観点に関する質問に対する回答の情報を利用することで、「有名な寺社」という観点を指定した際に、「金閣寺」や「銀閣寺」といった具体的な語を用いて検索結果を再ランキングする手法である。本稿では、特に修飾語付き観点を抽出手法に注力し、ユーザ実験を通してその有用性を評価する。

以降、2章では関連研究について述べる。3章ではQAコンテンツから観点を抽出する手法について説明し、4章ではユーザ実験の結果と考察を述べる。5章では観点にもとづくウェブ検索結果の再ランキング手法について述べ、最後に本稿をまとめる。

*1 Rerank.jp, <http://rerank.jp>

2. 関連研究

White¹⁷⁾ や Marchionini⁸⁾ は、あいまいな情報要求のもとで、情報の閲覧や検索を繰り返しながら検索対象を探索していく行為を探索的検索と呼び、そうした情報探索行為を支援することの重要性を指摘している。本研究で提案する手法も、ユーザが検索対象に対する知識が乏しく、具体的なクエリの作成が困難である場合に効果的に働くものであり、探索的検索の支援に有用であると考えられる。

クエリ推薦⁹⁾ やクラスタリング型検索エンジン^{*1}を利用することで、ユーザはクエリ作成の手間を省いたり膨大な検索結果から効率的に情報を収集できる。しかし、現状のクエリ推薦手法の多くはクエリログを利用しており、多くのユーザが投入する一般的な語がユーザに提示されることが多く、さまざまな観点からの情報閲覧を支援するには不十分である。また、クラスタリング型検索エンジンでは、各クラスターのラベル名を1語もしくは複数の語で表現している。しかし、一般的であったり、意味が不明瞭であったりするラベル名がユーザに提示されてしまうことも多い。ファセットによるナビゲーションも膨大な検索結果の閲覧を支援した情報閲覧手法である。たとえば、Yeeらはマルチファセットにもとづく画像検索システムを提案している¹⁰⁾。ファセットによるナビゲーションでは多くの場合、ファセットはあらかじめ検索対象に付与されたメタデータから自動的に生成される。そのため、ウェブ検索結果などに適用することが困難である。本稿で提案する修飾語付き観点の抽出手法は、こうしたクエリ推薦やクラスタリング、ファセットによるナビゲーションなどにも適用可能な技術である。修飾語付き観点をクエリとして推薦したり、クラスタリングのラベル名やファセットに利用したりすることで、ユーザの興味を引き、より多くの情報の閲覧を促進できるのではないかと考えている。

ユーザの情報要求を柔軟に伝達可能な検索システムもいくつか提案されている。“オノマトペロリ”はユーザが“シャキシャキ”や“サクサク”といったオノマトペを指定してレシピを検索可能なシステムである¹¹⁾。オノマトペロリでは、レシピに含まれる食材とオノマトペとの関連性を計算することで、ユーザが入力したオノマトペを含んでいないレシピでも、そのレシピを検索結果としてユーザに返すことを可能にしている。また、山家らはソーシャルブックマークを利用することで“便利”や“これはすごい”といったクエリでウェブページを検索可能なシステムを提案している¹²⁾。これらの研究はレシピやソーシャルブ

クマークといったデータを利用することで、ユーザが入力した文字列の有無ではない形でユーザの情報要求を満たす検索を可能にしている。それに対して、我々の研究はQAコンテンツを利用して検索を改善するものである。

ウェブ検索の改善のためにQAコンテンツを利用した研究としてはYoonらの研究がある¹³⁾。Yoonらは、QAコンテンツのカテゴリをユーザのウェブ検索における意図であると捉え、カテゴリとウェブ検索結果を対応付ける手法を提案している。それに対し、我々の手法はQAコンテンツの質問文からユーザの意図の候補を抽出しているという点で、Yoonらの研究とは異なる。

3. QAコンテンツからの観点抽出

本章ではまず、本研究で扱う「修飾語付き観点」について説明する。その後、実際にQAコンテンツからウェブ検索に有用な修飾語付き観点を抽出するための手法について述べる。

3.1 修飾語付き観点

本稿では、ウェブ検索結果を絞り込んだり修正したりするために、ユーザに明示的に提示する語のことを観点と呼ぶ。たとえば、クエリ推薦でユーザに推薦される追加のクエリや、我々のシステムにおけるタームクラウド中の語は観点である。さまざまな観点から情報を収集する必要のある検索において、ユーザの興味を引き、ユーザのインタラクションを促すために提示すべき良い観点の尺度として以下の3つが考えられる。

- **クエリとの適合性：**クエリと関連のない観点を提示しても、ユーザは求める情報に到達できないため、クエリに適合した観点を提示することは最も重要である。
- **観点の意外性：**クエリに関連した観点を提示する際に、ユーザがクエリとして想起するような一般的な観点だけでなく、想起しづらい意外な観点を提示することで、ユーザの興味を引き、より多くの観点からの情報閲覧を促すことができると考えられる。
- **観点の一般性：**たとえば、“京都 観光”というクエリにおいて、“ふたば”という語を提示したとしても、京都に詳しくないユーザには、その語が何を意味するものか分からず、役に立たない。しかし、“和菓子屋”や“有名な和菓子屋”と提示することで、京都に関する知識が乏しいユーザでも、その観点で情報を閲覧すれば、京都の和菓子屋に関する情報が閲覧できることが分かる。このように検索対象に対する知識がないユーザでも、提示された観点の意味が理解できることも重要である。

本研究では、こうした尺度にもとづき、ユーザに提示すべき観点として**修飾語付き観点**に着目する。本稿で扱う修飾語付き観点とは、以下のパターンに当てはまる名詞句である。

*1 Clusty, <http://clusty.jp>

- 形容詞+名詞
- 形容動詞互換+な+名詞

たとえば，“美味しい和菓子屋”や“有名な寺社”といった名詞句が修飾語付き観点となる。普段のウェブ検索において、ユーザが投入するクエリは名詞のキーワードだけで構成されることが多いと考えられる。そのため、キーワードとして想起しづらいと考えられる「修飾語付き観点」は名詞だけの観点に比べてユーザの興味を引き、インタラクションを促すことが期待される。

以降、修飾語付き観点を質問応答サイトに蓄積された QA コンテンツから抽出する手法について述べる。

3.2 質問応答サイト

質問応答サイトはユーザ同士が質問やその回答を投稿し合い、情報をやり取りすることのできるサービスであり、日本では代表的な質問応答サイトとして、Yahoo!知恵袋^{*1}や教えて!goo^{*2}などがある。本研究では、修飾語付き観点を抽出する情報源としてそうした質問応答サイトに蓄積された、質問とそれに対する回答の組からなる QA コンテンツを用いる。1章で述べたように、QA コンテンツの質問にはユーザの興味や情報要求が含まれており、多様な修飾語付き観点が多く存在すると考えられる。また、質問するユーザは、検索対象の知識が少ない場合が多く、検索対象に特有の専門的な言葉を用いずに一般的な言葉で質問を記述するだろう。そのため、質問から観点を抽出することで多くのユーザの興味を引く観点が抽出可能ではないかと考えられる。

3.3 観点抽出の流れ

QA コンテンツから観点を抽出する手法は以下の通りである。

- (1) クエリ q をユーザから受け取る。
- (2) Yahoo!知恵袋の質問検索 API^{*3} を利用し、クエリ q を質問文に含む質問の集合 $Q = \{q_1, \dots, q_k\}$ と、その回答（ベストアンサー）の集合 $A = \{a_1, \dots, a_k\}$ を取得する。ここで a_i は質問 q_i に対する回答である。また、 k は質問検索 API で取得する質問の件数であり、本稿では 1,000 件の質問とその回答を取得する。
- (3) 質問集合 Q 中の文字列を形態素解析し、3.1 節で示したパターンに当てはまる名詞

句を修飾語付き観点として抽出する。なお、形態素解析器には MeCab^{*4}を用いた。

- (4) 抽出した修飾語付き観点の集合 F の各要素を、次節で述べるランキング関数 $Rank$ にてランキングする。
- (5) ランキングされた修飾語付き観点をユーザに提示する。

3.4 観点のランキング

QA コンテンツから修飾語付き観点を単純に抽出しただけでは、クエリと関連がなかったり、意味が不明瞭であったり、“良いアドバイス”といった QA コンテンツに特有な単語が多く抽出されてしまう。そのため、先述の観点抽出の流れにあるように、抽出された修飾語付き観点をランキングし、ウェブ検索に有用な観点を発見する必要がある。本稿では、観点のランキングにあたり、(1) クエリと観点の関連性、(2) 観点と回答の関係の 2 つのアプローチから得られるスコアを組み合わせる。

(1) クエリと観点の関連性

1 つめのアプローチはクエリと観点の関連性の評価である。抽出された修飾語付き観点から、クエリと関連した観点を抽出するためには、何らかの手法を用いてクエリと観点の関連の強さを計算する必要がある。本稿では、QA コンテンツ内におけるクエリと観点の語の共起にもとづいてクエリと観点の関連性を評価する。語の共起にもとづいて単語間の関連性を計算する手法としては、共起頻度、相互情報量¹⁴⁾、 χ^2 検定値など、さまざまな手法が提案されている。たとえば、ある単語 w_i, w_j 間の相互情報量 I は、

$$I(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \quad (1)$$

により算出される。ここで、 $P(w)$ は文書コレクションにおいてある文書が語 w を含む確率、 $P(w_i, w_j)$ はある文書が語 w_i, w_j の両者を含む確率である。 w_i, w_j が同じ文書内で共起する確率が高いほど相互情報量は高い値をとる。相互情報量には、頻度が少なすぎる単語に対して不当に大きな値を与えてしまうという問題点がある。そこで、本稿では相互情報量に $P(w_i, w_j)$ を乗じた Expected Mutual Information¹⁵⁾ を用いる。クエリ q と修飾語付き観点 f の Expected Mutual Information EI は、

$$EI(q, f) = P(q, f) \cdot \log \frac{P(q, f)}{P(q)P(f)} \quad (2)$$

$$P(q) = \frac{n_q}{N}, \quad P(f) = \frac{n_f}{N}, \quad P(q, f) = \frac{n_{qf}}{N} \quad (3)$$

として計算される。ただし、 N は QA コンテンツの全質問数、 n_q, n_f はそれぞれ QA コンテンツの全質問のうち q, f を含む質問の数、 n_{qf} は q と f の両者を含む質問の数を表す。

*1 Yahoo!知恵袋, <http://chiebukuro.yahoo.co.jp/>

*2 教えて!goo, <http://oshiete.goo.ne.jp/>

*3 Yahoo!知恵袋 質問検索 API, <http://developer.yahoo.co.jp/webapi/chiebukuro/>

*4 MeCab, <http://mecab.sourceforge.net/>

本稿では $N = 30,000,000$ とし, n_q, n_f, n_{qf} をそれぞれ “ q ”, “ f ”, “ $q \wedge f$ ” をクエリとして質問検索 API で検索した際のヒット件数とした.

抽出された全ての観点について EI を計算した後, 以下の式にもとづいて最大値が 1 になるように正規化を行い, クエリ q と観点 f の関連性のスコア $Score_{query}$ を求める.

$$Score_{query}(q, f) = \frac{EI(q, f)}{\max_{f \in F} EI(q, f)} \quad (4)$$

(2) 観点と回答の関係

式 (4) によるスコアリングで, クエリと関連した語を上位にランキングできる. しかし, クエリと共起するがウェブ検索に用いる観点としては意味が不明瞭な単語も上位にランキングされてしまう. そこで, 2 つめのアプローチとして, 観点と回答の関係を考慮したスコアリングを行う.

たとえば, “美味しい和菓子屋はどこですか?” という質問がなされた場合, 回答者は “出町ふたば” や “満月” がおすすめですよ” といったように, “美味しい和菓子屋” という修飾語付き観点に対して, “出町ふたば” や “満月” といった, 具体的な語を用いて回答すると考えられる. つまり, ウェブ検索にも有用な観点が登場する質問には, 実際の観光地名や店名といった, 具体的な語が回答に出現しやすいのではないかと考えられる. ここで, “美味しい和菓子屋” に対する “出町ふたば” や “満月” のように, 観点に対する具体的な回答を表した語を本稿ではエンティティと表現する. そして, ウェブ検索に有用な観点にはエンティティが回答に多く出現するという仮説にもとづき, 観点をスコアリングする.

まず, システムは回答集合 A からエンティティの集合 E を抽出する. 上記の例のように, 回答者は質問の回答にふさわしい具体的な語を “「○○」はどうですか?” といったように, 「」や『』などの記号で強調して回答することが多い. そこで, 今回は, 回答の文章中で「」, 『』, “” によって挟まれた, 10 文字以内の文字列のうち, 回答集合 A に 2 回以上出現する文字列をエンティティとして抽出した. 例として, クエリ “京都 観光” で得られた質問集合に対する回答から抽出されたエンティティ集合の一部を表 1 に示す. 表 1 のように, 地物名や店舗名, イベント名といった多様な語がエンティティとして抽出されていることが分かる. しかし, 同時に会話文や, クエリと関係の無い語なども抽出してしまっているのが分かる. 本稿では扱わないが, 今後は抽出されたエンティティとクエリとの関連性の評価や, より洗練された言語パターンなどを利用してエンティティ抽出の精度を向上させる必要がある.

エンティティ集合を抽出した後, 観点 f とエンティティ集合 E との, 質問-回答間の共起

表 1 クエリ “京都 観光” におけるエンティティ集合の例

抽出されたエンティティ集合の一部
嵯峨嵐山, 金閣寺, 鴨川納涼床, ボルガライス, 葵祭り, 丸太町, 八坂神社 鈴虫寺, 半兵衛魁, 蹴上, 辻利, 手をけ弁当, いい古都チケット, 須大社, 南禅寺, 鞍馬の火祭, 祇園四条, 哲学の道, 北野天満宮, よーじや ちよつと早いかな, そうでないところ, 築地, なんば...

度の強さ $EntityCooccurrence$ を下記の式で計算する.

$$EntityCooccurrence(f, E) = \frac{1}{|A_f|} \sum_{a \in A_f} Count(a, E) \quad (5)$$

ここで, A_f は Q の中で観点 f を含む質問に対する回答の集合, $Count(a, E)$ は, 回答 a 中に出現するエンティティの数を表す. ある観点 f を含む質問に対する回答にエンティティが多く含まれる割合が多いほど, $EntityCooccurrence$ は高い値を取る.

抽出された全ての観点について $EntityCooccurrence$ を計算した後, 以下の式にもとづいて最大値が 1 になるように正規化を行い, 観点 f とエンティティ集合 E の共起度のスコア $Score_{entity}$ を求める.

$$Score_{entity}(f, E) = \frac{EntityCooccurrence(f, E)}{\max_{f \in F} EntityCooccurrence(f, E)} \quad (6)$$

修飾語付き観点のランキング関数 $Rank$ を, 式 (4) および式 (6) を結合した以下の式で定義する.

$$Rank(q, f, E) = \alpha \cdot Score_{query}(q, f) + (1 - \alpha) \cdot Score_{entity}(f, E) \quad (7)$$

ここで, α は両者のスコアの重みを決めるパラメータである. 本稿では, 予備実験の結果から $\alpha = 0.7$ を用いた.

4. 実 験

提案手法により得られる修飾語付き観点の有用性を評価するためユーザ実験を行った. 実験の目的は, ウェブ検索をしているユーザの興味を引く単語を提案手法がどの程度提示できるのかを評価することである.

4.1 ベースライン

提案手法の有用性を相対的に評価するため, ベースラインとして次の 3 つの手法を用意した.

WEB (ウェブ検索結果中の修飾語付き観点の出現頻度): 与えられたクエリで Yahoo! ウェ

ブ検索 API^{*1}を用いてウェブ検索結果 1,000 件を取得し、取得した検索結果のタイトルと要約文を形態素解析して、修飾語付き観点を抽出する。そして、それぞれの修飾語付き観点について 1,000 件のウェブ検索結果中の出現頻度を計算し、頻度順に 20 個の修飾語付き観点を出力する手法。ただし、出現頻度が同数の場合は上位の検索結果に出現する修飾語付き観点を優先する。

FQ (ウェブ検索結果中の単語の出現頻度)：上記の手法と同様に取得したウェブ検索結果 1,000 件のタイトルと要約文を形態素解析して、名詞、形容詞、形容動詞を単語として抽出する。そして、それぞれの単語について 1,000 件のウェブ検索結果中の出現頻度を計算し、頻度順に 20 個の単語を出力する手法。

QS (クエリ推薦)：与えられたクエリで Yahoo! 関連検索ワード API^{*2}を用いて関連検索ワードを 100 件取得する。そして、その中から与えられたクエリに単語を 1 つ追加した形の関連検索ワードを上位から 20 件抽出し、追加された 20 単語を出力する手法。

これら 3 つのベースライン手法と、3 章で述べた手法によりランキングされた修飾語付き観点の上位 20 件を用いる手法 (QA) の、計 4 手法を用いて実験を行った。

4.2 実験方法

7 名の被験者を用いて実験をおこなった。被験者はみな男性であり、6 名は大学生、1 名は大学の教員である。実験には表 2 に示す 4 カテゴリからなる計 20 件のクエリを用いた。ただし、病気カテゴリのクエリに関しては、表 2 に示したクエリをそのまま用いると、クエリ推薦にもとづく手法が 20 件の語を出力できなかつたため、“症状”をクエリに含めずに関連検索ワード API で推薦語を取得しユーザに提示した。

実験では、それぞれのクエリに対して被験者に以下のタスクを行ってもらった。

- 被験者に対して簡単な検索意図（京都観光を考えている、デジタルカメラを購入しようと考えているなど）の説明文を提示し、指定したクエリでウェブ検索を行っていると思定してもらい、4.1 節で述べた 4 手法が出力した計 80 個（手法間で重複がある場合は 80 個以下）の単語を 16x5 の表形式にランダムに配置したものを被験者に提示する。そして、被験者は提示された 80 個の単語から「興味を引く単語」を好きなだけ選択する。
- 次に、各手法が出力した 20 個の単語集合を、それぞれ 4x5 の表形式に配置して被験者に提示する。被験者は提示された 4 つの表それぞれに対して、表 3 に示した 3 つの設

表 2 実験に用いたクエリ

カテゴリ	クエリ
旅行	京都 観光, 北海道 観光, 横浜 観光, アメリカ 旅行, ヨーロッパ 旅行
病気	花粉症 症状, インフルエンザ 症状, メニエール病 症状, 二日酔い 症状, ドライアイ 症状
買物	デジタルカメラ, 液晶テレビ, 洗濯機, HDD レコーダー, スマートフォン
料理	豚肉 レシピ, ジャガイモ レシピ, パスタ レシピ, インド料理, イタリアン

表 3 被験者に対して行ったアンケートとその結果。

表中の値は全てのクエリにおけるアンケート結果の平均、() 内は標準偏差を示している。

質問項目	QA	WEB	FQ	QS
1. 検索意図に関連した単語が提示されている。	3.49 (0.95)	2.92 (1.09)	3.52 (0.97)	3.20 (1.26)
2. 検索意図に関連した意外な単語が提示されている。	3.72 (1.03)	3.39 (0.99)	2.35 (0.82)	2.48 (0.96)
3. 難しい単語が提示されている。	2.02 (1.02)	2.35 (1.15)	2.08 (1.12)	2.51 (1.32)

問を、1 (そう思わない) ~5 (そう思う) の 5 段階のリッカート尺度で回答する。なお、被験者に提示する 4 つの表はクエリごとに配置を変更している。

前者のタスクの目的は、ユーザの興味を引く単語がどの程度提示できていたかを手法間で比較すること、後者のタスクの目的は、各手法が出力した単語集合をクエリとの関連性、観点の意外性、観点の一般性の 3 つの尺度で比較することである。

4.3 実験結果

実験結果を図 3 および表 3 に示す。図 3 は各手法から被験者が興味を引くと判定した単語の数をカテゴリごとにまとめたグラフである。図中のエラーバーは標準誤差を表している。また、表 3 はそれぞれの手法から得られた単語集合に対する被験者のアンケート結果をまとめたものである。

図 3 から分かるように、全体として提案手法から最も多くの単語が選択されていた。この結果から、提案手法がベースラインと比べてユーザの興味を引く単語を多く提示できていることが分かる。また、表 3 に示した被験者のアンケート結果では、提案手法による単語集合とクエリ推薦にもとづく手法による単語集合に対して、被験者は同程度ユーザの検索意図に関連した単語が提示されていると回答したことが分かる。さらに、提案手法による単語集合は、他の手法に比べて多くの被験者が意外な単語が提示されていると回答していた。これらの結果は、提案手法で得られた修飾語付き観点をウェブ検索を行っているユーザに提示することで、ユーザの興味を引き、より多くの観点からの情報閲覧を促すことができる可能性を示している。

提案手法は QA コンテンツから修飾語付き観点を抽出したが、表 3 から分かるようにウェブ検索結果から抽出された修飾語付き観点の集合に対しても、多くの被験者が意外な観点が

*1 <http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html>

*2 <http://developer.yahoo.co.jp/webapi/search/assistsearch/v1/webunitsearch.html>

含まれていると回答していた。このことは、修飾語付き観点が通常のウェブ検索ではクエリとして想起しにくい形であり、修飾語付き観点を提示することで意外な観点をユーザに示すことが可能であることを示している。しかし、ウェブから抽出した修飾語付き観点は提案手法に比べてユーザが興味を引く単語であると判断する割合が少なかった。これは、アンケート結果から分かるようにシステムがクエリと関連する修飾語付き観点をあまりユーザに提示できなかったことが原因ではないかと予想される。実際、ウェブ検索結果には修飾語付き観点がほとんど出現しないため、意味のない修飾語付き観点がユーザに提示されてしまっていた。

図3をカテゴリ別で見ると、旅行カテゴリにおいて提案手法は最も有効に働いていることが分かる。表4上部は、実際に“京都 観光”のクエリにおいて提案手法が出力した観点のうち、多くの被験者が興味を引くと判断した観点の上位10件である。たとえば、“うまいランチ”や“有名な観光地”、“古い町並み”といった観点が興味を引く単語として選択されていた。また、他のカテゴリでは、病気カテゴリでは、“軽いめまい”や“激しい頭痛”といった病気の症状を表現した修飾語付き観点が被験者に選択されていた。たとえば、“インフルエンザ”というクエリでは、クエリ推薦にもとづく手法では、“頭痛”や“発熱”といった語がユーザに提示されていた。一方で、我々の提案手法では、“軽い頭痛”や“激しい頭痛”、“急な発熱”といった語がユーザに提示された。このように、修飾語付き観点に着目することで、ユーザは“頭痛”や“発熱”よりも詳細な意図をシステムに伝達することも可能になると考えられる。また、料理カテゴリでは、“ヘルシーな料理”、“簡単なレシピ”、“家庭的な雰囲気”、“辛い料理”といった観点が多くの被験者に選択されていた。

しかし、図3を見てみると買物カテゴリでは他のカテゴリと比べて提案手法がユーザの興味を引く単語を提示できていないことが分かる。この原因として、提案手法の精度の問題があげられる。表4下部はクエリ“デジタルカメラ”において提案手法が出力した観点を10件表示したものである。“綺麗な写真”や“暗い場所”といった観点はデジタルカメラを購入する際の興味を引く観点として被験者に選択されていたが、それ以外の観点はほとんど選択されていなかった。デジタルカメラの例では、観点に対する具体的な語として実際のカメラの製品名や型番などが適していると考えられる。しかし、3.4節で述べたエンティティ抽出手法ではそうした情報が上手く抽出することができていなかった。そのため、ウェブ検索に有用な観点を適切にランキングできなかったのではないかと予想される。実際、3.4節で提案した観点とエンティティの共起にもとづくスコアリングがどの程度有効だったかを調査するため、修飾語付き観点と質問-回答間で共起するエンティティの数と、被験者が興味

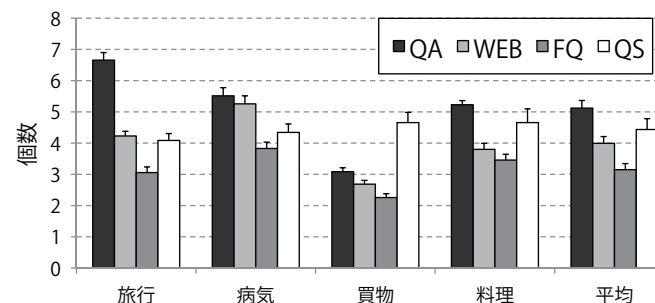


図3 被験者が興味を引くと判定した単語の数

を引くと判定した人数の相関を求めてみると、0.21と正の相関を示しており、観点とエンティティの関係性を考慮することが有用な観点の抽出にある程度有効であることが分かる。しかし、クエリを買物カテゴリに限定した際の相関は0.04であった。このことから、精度良くエンティティを抽出することがより有用な観点の抽出に必要であることが分かる。

今回の実験では、検索対象に対する被験者の知識と、興味を引くと判断する単語の関係について考えなかった。しかし、たとえば京都に関する知識を十分に持っているユーザとそうでないユーザ、また、デジタルカメラを購入した経験のあるユーザとそうでないユーザなどで実験結果は変わってくると考えられる。今後は、幅広い被験者層で実験を行い、検索対象に対する知識が実験結果がどのような影響を与えるのかを分析する予定である。

4.4 考察

実験結果より、提案手法はユーザが興味を引く観点をベースライン手法と比べて多く提示できる可能性があることが分かった。また、ユーザが意外だと感じる、クエリと関連した観点を提示可能であることも分かった。ウェブ検索を行っているユーザに、クエリ推薦や、我々のシステムで提示していた頻度ベースのタームクラウドだけでなく、我々の提案手法で得られた修飾語付き観点をユーザに提示することで、より多くの観点からの情報閲覧を促進できる。

今回は、クエリを含む質問の全てを対象として観点の抽出を試みていた。しかし、実際にQAサイト上に投稿される質問には、単に事実関係を確認したり、相談事をしたりといった、ウェブ検索には有用ではない質問も存在する。提案手法の精度を向上させるためには、そうした質問の種類を考え、対象とする質問を限定することも必要になるだろう。また、高度な専門知識に関する話題のように、QAサイト上で取り上げられないような話題に関する

表 4 提案手法が有効に働いた例と働かなかった例

クエリ	提案手法がユーザに提示した修飾語付き観点の例
京都 観光	うまいランチ, 有名なお寺, おいしいご飯, 有名なスポット, 有名な観光地, 古い町並み, おいしい食べ物, 主な観光地, 美味しいお店, 安い宿
デジタルカメラ	綺麗な写真, 暗い場所, 新しいカメラ, 強い機種, コンパクトなデジタル, 安いカメラ, 主な使用, 可能なデジタル, 古いデジカメ, 無骨なボディ

検索については、提案手法は有効に働かないと考えられる。しかし、今後多くの人々にとって QA サイトが一般的になり、より多様な話題が議論されるようになるにしたがって、提案手法の適用範囲もある程度は広がっていくと考えられる。

提案手法を実際のウェブ検索に適用するにあたり、提案手法には精度の他にも考慮すべき点がある。1 つめは手法の動作時間である。現在では、クエリと各観点の関連度を求めるために質問検索 API を利用している。そのため、実際に結果を出力するまでにウェブアクセスが頻繁に発生してしまう。この問題を解決するためには、API を利用して QA コンテンツを取得するのではなく、提供されているデータベース^{*1}を利用することなどが必要になるだろう。

2 つめは観点の多様性である。現在の手法は単純に観点をランキングをしているため、表 4 の“京都 観光”の例のように、“おいしいご飯”と“おいしい食べ物”といった内容が類似する観点が上位にランキングされる。しかし、たとえば、実際の我々のタームクラウドに対して提案手法を適用する場合は、類似した観点が多く提示されるよりも、多様な観点が多く提示される方が、ユーザはより多くの観点から検索結果を閲覧できる。今後は、実際に観点間の類似度を考慮しながらユーザに提示すべき観点集合を決定することも必要になる。

5. 修飾語付き観点にもとづくウェブ検索結果の再ランキング

5.1 概 要

前章の実験結果より、提案手法でユーザが興味を引くような修飾語付き観点を提示可能ながことが分かった。たとえば、提案手法により得られた修飾語付き観点を、図 2 に示したようなタームクラウドなどでユーザに明示的に提示することで、ユーザの再ランキングを促進することが可能となるだろう。しかし、ユーザが“有名な観光地”を指定した際に検索結果から“有名な観光地”を含む検索結果を上位に再ランキングするだけでは、ユーザの意図を検

索結果に反映できないことも多い。図 4 左図は、クエリ“京都 観光”において「有名な観光地」を含む単語を上位に再ランキングした際のウェブ検索結果例である。しかし、図にあるように“有名な観光地”という文字列が含まれているものの、そのページの内容は京都観光における有名な観光地に関するとは言えない。図 4 右図のように、ユーザが“有名な観光地”を指定した際に“金閣寺”や“銀閣寺”といった、多くのユーザが有名だと考える観光地名を利用して再ランキングすることで、よりユーザの意図に沿った検索結果をユーザに提示できるのではないかと考えられる。一方、3.4 節ではウェブ検索に有用な観点を抽出するために、“有名な観光地”と質問-回答間で共起するエンティティを求めていた。つまり、修飾語付き観点と質問-回答間で共起するエンティティの情報を利用することで、上記のような再ランキングが実現可能ではないかと考えられる。

そこで、本章では、修飾語付き観点にもとづく再ランキング手法の実現にあたり、実際にシステムを実装して得られた考察と、解決すべき課題を述べる。

5.2 考 察

我々は、3.4 節で述べたエンティティ抽出手法に基づいて、観点と質問-回答間で共起するエンティティを元にウェブ検索結果を再ランキングする手法を実際に実装した。これは、修飾語観点をユーザが指定した際に、共起するエンティティの語を含む検索結果を上位に再ランキングするという簡単なものである。この手法でいくつかの修飾語付き観点で再ランキングを実行してみたところ、たとえば、“京都 観光”というクエリにおいて“おいしいカフェ”という修飾語観点を指定して再ランキングすると、“フランソワ”、“ソワレ”といった実際の京都のカフェの店名を含む検索結果が上位に表示された。このように、ユーザが京都のおいしいカフェの具体的な店名を知らなくても、検索意図を検索結果に反映できる可能性を持っていることが分かった。一方で、“有名なお寺”を指定して再ランキングすると、“清水寺”と“鈴虫寺”という二つの寺社名しか再ランキングに利用されず、ユーザの再ランキングの意図を満たしているとは言えなかった。

5.3 現状の課題

現在の手法には、大きく分けて 2 つの課題があると考えている。

1 つめは対象とするデータの規模に関する課題である。現在はクエリを含む 1,000 件の質問とそのベストアンサーをデータの対象としているが、より多くのエンティティを抽出するためには、もっと大規模にデータを対象に集めてくる必要がある。たとえば、先ほど述べたように提案手法では“有名なお寺”と共起するエンティティとして“清水寺”と“鈴虫寺”という 2 つのエンティティしか抽出できなかった。それに対し、“京都 観光 有名なお寺”と

*1 Yahoo! データセット, <http://www.nii.ac.jp/cscenter/idr/yahoo/yahoo.html>

観光タクシー・観光ハイヤー 京都MKタクシー

有名な観光地はもちろん、旬のスポットや京都人だけが知っている穴場にもご案内。...「京都観光ナビ」WEB予約で観光タクシー10%割引...京都観光モラルコースをご紹介します。
http://www.mtc-group.co.jp/kyoto/taxi/kyotoindex.html 元:402

ウィークリーマニション ロイヤル京都駅前

京都駅利用。京都ロイヤルホテルが運営するウィークリーマニション。運営状況、料金等の紹介。...京都駅西口から、ほんのりの有名な観光地も盛り込んだコース。...
http://www.kyotoeki.jp/ 元:303位

レンタル着物 西本

京都をレンタル着物でお楽しみください。足袋や草履、バッグも全てご用意いたしております。...
http://www.okamoto-kimono.com/ 元:424位

京都じゅくり観光金閣寺～御室 散策ガイド

室町幕府第3代将軍、足利義満が築いた山荘を寺にした金閣寺の舍利殿は、ひと目見るとその華やかさに目を奪われます。その他、枯山水の石庭で名高い龍安寺、御室御所と呼ばれる門跡寺院の仁和寺、臨濟宗・妙心寺派の本山で大伽藍の妙心寺と、それぞれに見どころの多いコースです。...

菟野山 清水寺 【公式ホームページ】

京都の清水寺。西国三十三所第十六番札所。「清水の舞台」等の境内の案内、イベントや年中行事情報等。...
http://www.kiyomizu-dera.jp/ 元:303位

京都観光(廻り寺)～哲学の道～南禅寺

このページは京都の観光コースである廻り寺～哲学の道～南禅寺をWEB上でじっくり散策できるページです。...
http://www.kyoto-tour.com/ 元:424位

図 4 修飾語付き観点を直接利用した再ランキングと、質問-回答間で共起するエンティティを用いた再ランキング

いうクエリで質問検索 API で検索して再度質問を集め、また、ベストアンサーだけではなく通常の回答も対象にエンティティを抽出したところ、“清水寺”、“南禅寺”、“金閣寺”、“銀閣寺”、“高台寺”、“東寺”、といったようにより多くの、また多くのユーザが有名だと考える寺名を抽出することができた。現在は API を利用しているが、システムがデータベースを内部に持つことで対象とするデータを増やすことで、より精度良く、また、より多くのエンティティを再ランキングに利用できるだろう。

2 つめはエンティティの抽出手法である。今回提案したエンティティ抽出手法は単純な言語パターンにもとづく手法であった。そのため、クエリと関係の無いエンティティが多く抽出されてしまっていた。クエリと関連したエンティティを精度よく抽出するためには、3.4 節で行ったような単語間の関連性評価手法などで、抽出してきたエンティティとクエリの関連性を評価することが必要である。また、今回の手法では地物名などはある程度精度良く抽出可能であるが、実際にはそうでないエンティティも多い。たとえば、病気に関する質問で“軽い頭痛”といった観点に関連する回答情報としては、治療薬に関する情報や、直し方に関する情報などが回答に出現する情報であると考えられる。このような多様な情報をエンティティとして抽出するための手法が必要になってくると考えられる。

今後は、こうした課題を解決してよりユーザの意図に沿った再ランキング手法を実現し、実際の我々のシステムに適用していきたいと考えている。

6. まとめ

本稿ではウェブ検索においてユーザがより興味を引く単語を提示するために、“有名な観光地”や“美味しい和菓子”といった修飾語付き観点到に着目し、そのような観点を QA コンテンツから抽出しランキングする手法を提案した。ユーザ実験の結果、提案手法はベースライン手法に比べてよりユーザの興味を引く語が提示できていることや、旅行に関するカテゴ

リについて本提案手法の有効性が高いことを示した。

今後は、本稿であげた課題を解決し、よりユーザの意図に沿った、修飾語付き観点にもとづく検索結果の再ランキング手法について取り組む予定である。そして、これらの手法を我々のシステムに適用したものを実装し、修飾語付き観点を提示することがユーザのウェブ検索行為にどのような影響を与えるのかを調査していく予定である。

謝辞 本研究の一部は、科学研究費補助金特別研究員奨励費「ユーザインタラクションに基づく情報検索」(研究代表者:山本岳洋, 課題番号 09J55302), グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」, 文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」, 計画研究「情報爆発に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者:田中克己, 課題番号 1809041) によるものです。ここに記して感謝の意を表します。

参考文献

- 1) White, R. and Roth, R.: Exploratory Search: Beyond the Query-Response Paradigm, *Synthesis Lectures on Information Concepts, Retrieval, and Services*, Vol.1, No.1, pp.1-98 (2009).
- 2) Cui, H., Wen, J., Nie, J. and Ma, W.: Probabilistic query expansion using query logs, *Proceedings of the 11th International Conference on World Wide Web*, pp. 325-332 (2002).
- 3) Käki, M. and Aula, A.: Controlling the Complexity in Comparing Search User Interfaces via User Studies, *Information Processing & Management*, Vol.44, No.1, pp.82-91 (2008).
- 4) Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Tezuka, T., Oyama, S. and Tanaka, K.: Trustworthiness Analysis of Web Search Results, *Research and Advanced Technology for Digital Libraries*, pp.38-49 (2007).
- 5) 山本岳洋, 中村聡史, 田中克己: Rerank-By-Example: 編集操作の意図伝播によるウェブ検索結果のリランキング, *情報処理学会論文誌 (トランザクション) データベース*, Vol.49, No.SIG7(TOD37), pp.16-28 (2008).
- 6) Yamamoto, T., Nakamura, S. and Tanaka, K.: TermCloud for Enhancing Web Search, *Proceedings of the 11th International Conference on Web Information Systems Engineering*, pp.159-166 (2009).
- 7) White, R., Kules, B. and Drucker, S.: Supporting Exploratory Search, *Communications of the ACM*, Vol.49, No.4, pp.36-39 (2006).
- 8) Marchionini, G.: Exploratory Search: from Finding to Understanding, *Communications of the ACM*, Vol.49, No.4, pp.41-46 (2006).

- 9) Koenemann, J. and Belkin, N.: A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness, *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, pp.205–212 (1996).
 - 10) Yee, K., Swearingen, K., Li, K. and Hearst, M.: Faceted Metadata for Image Search and Browsing, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.401–408 (2003).
 - 11) カンウイパーラートサムルアイパン, 渡辺知恵美, 中村聡史: オノマトペロリ: オノマトペを利用した料理推薦システム, 情報処理学会第73回デジタルドキュメント研究会, Vol.2009-DD-73, No.6, pp.1–7 (2009).
 - 12) Yanbe, Y., Jatowt, A., Nakamura, S. and Tanaka, K.: Towards Improving Web Search by Utilizing Social Bookmarks, *Proceedings of the 7th International Conference on Web Engineering*, pp.343–357 (2007).
 - 13) Yoon, S., Jatowt, A. and Tanaka, K.: Intent-Based Categorization of Search Results Using Questions from Web Q&A Corpus, *Proceedings of the 11th International Conference on Web Information Systems Engineering*, Springer, pp.145–158 (2009).
 - 14) Church, K. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational linguistics*, Vol.16, No.1, pp.22–29 (1990).
 - 15) Croft, B., Metzler, D. and Strohman, T.: *Search Engines: Information Retrieval in Practice*, Addison-Wesley Publishing Company, USA (2009).
-