Vol.2017-DPS-172 No.16
Vol.2017-SPT-26 No.16
Vol.2017-EIP-78 No.16
2017/11/30

希望目的地情報の検索におけるトピックドリフト問題の軽減に関する提案

是川俊昭[†1]　児玉英一郎[†1]　王家宏[†1]　高田豊雄[†1]

概要：旅行をしたいと考え目的地を探しているユーザが，Web を利用して情報取得を試みる場合，複数のキーワードを組み合わせサーチエンジンを利用して検索する方法が考えられる．この方法によって，サーチエンジンにて検索を行うと，キーワードの一部が出現しているために検索結果の上位に現れているもの（表層的トピックドリフト）や，キーワードが異なる意味で使用されているのにも関わらず検索結果の上位に出現しているもの（深層的トピックドリフト）といったトピックドリフト問題が発生する．本論文では，このようなトピックドリフト問題を軽減し，ユーザの希望目的地に関する情報を Web 上から効率的に発見可能なシステムのモデルの提案を行う．また，本提案モデルの有用性を確認するために行った評価実験についても報告する．

キーワード：トピックドリフト問題，希望目的地情報の検索

Proposal on Resolving Topic Drift Problems in Retrieving Desired Destination Information

Toshiaki Korekawa[†1]　Eiichiro Kodama[†1]　Jiahong Wang[†1]　Toyoo Takata[†1]

Abstract: Given that someone wants to take a journey, and is making his / her journey plan by searching Web for destination information, it is more than possible that he / she will enter into the search engine a group of keywords representing his / her journey destination. By doing so, however, he / she may suffer from the surface topic drift and the deep topic drift problems. For the former, the ones appearing in the top of the search results have no relation with the entered keywords at all due to that a keyword is included unexpectedly in a long unrelated word. For the latter, the ones appearing in the top of the searching results have no relation with the entered keywords at all due to the incorrect understanding of some of them. In this paper, we propose a system model that can alleviate the topic drift problems and can efficiently provide users with the desired journey destination information satisfying their requirements from the Web. Experiments have been conducted to confirm the effectiveness of the proposed model; we report and discuss the experiment results.

Keywords: Topic Drift Problems, Retrieving Desired Destination Information

1. はじめに

現在の Web サーチエンジンではキーワードによる検索が主流となっている．旅行をしたいと考え目的地を探しているユーザが，Web を利用して情報取得を試みる場合，例えば，“兵庫”，“島”，“美しい”といった複数のキーワードの組み合わせによって検索する．

既存の Web サーチエンジンを利用し，上述の検索キーワードにて検索を行うと，兵庫県ではなく東京にある兵庫島公園のような，公園名内にキーワードの一部が出現しているために検索結果の上位に現れているもの（表層的トピックドリフト）や，兵庫県の NPO 団体がソロモン諸島の清掃活動を行ったという，美しい島に関する情報は確かに含まれているが兵庫県とは関係性の薄い島の情報とマージされて出現しているものなど，検索要求と合致しないが検索結果の上位に出現している（深層的トピックドリフト）といったトピックドリフト問題が発生する．また，“美しい”という単語が含まれておらず，同義語である，“きれい”といった単語で表現されたものは検索結果に出現しないなど，ユーザにとって不便な側面も有している．

そこで，本研究では，ユーザの希望目的地に関する情報を Web 上から効率的に発見可能なシステムのモデルの提案を行う．

2. 関連研究

検索クエリの拡張に関する研究としては，大石らの研究[1]が知られている．大石らの研究では，ユーザから与えられたキーワードで検索後，検索結果の Web ページ内からキーワードに関連する重要語を抽出し，この重要語を用いてクエリの拡張を行うものである．

また，検索クエリの修正に関する研究としては，福地らの研究[2]が知られている．福地らの研究では，動詞を含む

†1 岩手県立大学大学院ソフトウェア情報学研究科
Graduate School of Software and Information Science,
Iwate Prefectural University

Vol.2017-DPS-172 No.16
Vol.2017-SPT-26 No.16
Vol.2017-EIP-78 No.16
2017/11/30

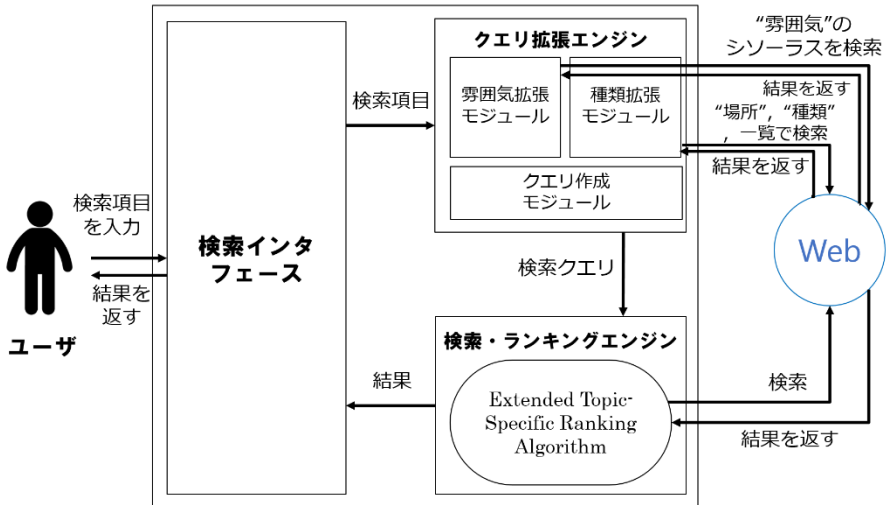

図 1　希望目的地情報検索システムのモデル

クエリに対し，単語間の関係性に注目し，クエリの修正を行っている．

Web ページの再ランキング手法に関する研究としては荒谷らの研究[3]が知られている．荒谷らの研究では，Web ページの内容に基づき，内容が類似した Web ページの順位を上げるものである．このような内容ベースの研究は Web ページ内に十分な量のテキストがない場合には有効に作用しない．そこで本研究では，Web ページ内のテキストの量が十分にない場合も考慮し，トピックドリフト問題を軽減する提案を行う．

また，再ランキング手法に関する他の研究としては，山本らの研究[4]が知られている．山本らの研究では，ユーザの簡単な操作によってランキング結果を再ランキングできるシステムの提案を行っている．

本研究の先行研究として，高杉らの研究[5]が知られている．高杉らの手法は，ユーザから与えられたキーワードに対し類義語を利用してクエリ拡張を行い，それぞれのクエリを検索エンジンで検索後，検索結果を再ランキング，マージし，最終的な検索結果として提示するものである．

高杉らの研究における，表層的トピックドリフトは，例えば“北海道”という単語に含まれる“道”の部分が“でっかい道”といった単語に含まれるために，誤って“でっかい道”という単語を含むページを正解ページとして取得してしまう問題である．この表層的トピックドリフトは 2007 年頃には多くみられたが，年々検索エンジンの改善が行われ，現在では出現頻度が少ないため，この対策に力を入れても効果は少ない．

また，深層的トピックドリフトは，例えば“奈良の面白い博物館に行きたい”という検索要求があった場合に，“博物館”という単語は使われているが，お笑いコンビの笑い飯のネタの一つである“奈良県立歴史民俗博物館”という架空の博物館についての Web ページを誤って取得してしまうような問題である．高杉らの研究では，Web ページ内に出現する“博物館”や“面白い”といった単語の出現回数の差でしか判断しておらず，十分とはいえないと考える．

3. 予備調査

3.1 予備調査の目的

高杉らの研究が行われた頃と現在とでは，Web サーチエンジンから得られる検索結果が異なると考えられるため，現在の検索結果に対し高杉らの手法が有効かどうか調査を行った．また，現在生じている表層的・深層的トピックドリフトの発生頻度についても調査を行った．

3.2 予備調査の方法

調査方法は，“場所”，“種類”，“雰囲気”を示す 3 つの単語の組を検索要求とし，実際に Web サーチエンジン(Google)を用いて検索を行った．検索要求として，10 個のキーワードの組を用意した．各検索要求に対し，検索結果上位 10 件を収集し，人手で正解の判定を行った．本調査では，検索要求を満たす情報のみが紹介されているページは正解ページとし，検索要求を満たさない情報が混在しているようなページは不正解ページとした．その後，これらの Web ページについて，高杉らのアルゴリズムを適用し，表層的トピックドリフト，深層的トピックドリフト，正解の判定を行った．

3.3 予備調査の結果

人手による正解判定の結果，表層的トピックドリフトが 1 件，深層的トピックドリフトが 61 件，正解ページが 35 件，その他が 3 件であった．100 件中の，その他の 3 件を除く 97 件のうち，高杉らのアルゴリズムでは，表層的トピックドリフトが 1 件，深層的トピックドリフトが 73 件，正解ページが 23 件であった．人手で判定した深層的トピックドリフト 61 件のうち，高杉らのアルゴリズムでも深層的トピックドリフトと判断できた件数は 42 件であった．また，高杉らのアルゴリズムにおいて，深層的トピックドリフトと判断すべきページを間違えて正解として捕捉した件数は 18 件，正解にもかかわらず深層的トピックドリフトと判断してしまったページは 30 件であった．

Vol.2017-DPS-172 No.16
Vol.2017-SPT-26 No.16
Vol.2017-EIP-78 No.16
2017/11/30

表 1　並び替え対象の Web ページの集合Wの例

W	Web ページのタイトル	Web ページ内に出現する単語の集合$\overline{w_i}$
w_1	淡路島の美しい景色について	{淡路島，沼島}
w_2	淡路島・岩屋温泉「美湯松帆の郷」	{淡路島}
w_3	淡路島の観光スポット 20 選	{淡路島，成ヶ島，沼島}
w_4	淡路の美しい料理の店	{高島，岩島}
w_5	淡路市今の絵島の美しい岩肌を観光しよう	{淡路島，絵島}

4. 希望目的地情報検索システムの提案

本研究で提案する希望目的地情報検索システムのモデルを図 1 に示す．以下，図 1 の各構成要素の詳細を示す．

● 検索インタフェース

ユーザからの検索要求を取得する．検索要求は，場所(l)，種類(k)，雰囲気(a)を示す 3 つの検索項目から構成される．場所は，希望目的地に関する地名など検索対象の所在地を示す．種類は，島や温泉など検索対象の観光地の種類を示す．雰囲気は，静か，美しいなど検索対象に望む状態を表現した形容詞を示す．ユーザから与えられた検索要求をクエリ拡張エンジンに渡し，検索・ランキングエンジンから得られた結果をユーザに提示する．

● クエリ拡張エンジン

本クエリ拡張エンジンは，以下の 3 つのモジュールから構成される．

(1)　種類拡張モジュール

深層的トピックドリフト問題への対策として，ユーザの検索要求である場所と種類で指定された単語に“一覧”という単語を加えて Web サーチエンジンにて検索を行う．例えば，“兵庫”，“島”という単語に“一覧”を加えて Web サーチエンジンにて検索を行う．検索結果の上位m件の Web ページに含まれる単語のうち，例えば，“淡路島”や“沼島”といった種類として入力されたキーワードkを部分文字列として含む単語を集めた種類拡張集合$T(k)$を生成する．

(2)　雰囲気拡張モジュール

雰囲気についてクエリ拡張を行う．雰囲気として入力されたキーワードaを利用してシソーラスに問い合わせ，その結果として同義語s_i，$1 \le i \le n$を取得する．この同義語s_iと種類で指定されたキーワードkを用いて Web サーチエンジンに問い合わせを行い，検索結果件数が閾値以上であった同義語からなる集合と$\{a\}$との和集合をaの共起同義語集合$C(a)$として取得する．

(3)　クエリ生成モジュール

場所で指定されたキーワードをlとするとき，$ER = \{(l,t,c) \mid t \in T(k), c \in C(a)\}$を拡張検索要求集合として生成する．

● 検索・ランキングエンジン

クエリ生成モジュールで作成された各拡張検索要求$er \in ER$を用い，Web サーチエンジンで検索後，各検索結果$R(er_i)$，$er_i \in ER$の上位p件に対し，下記の Extended Topic-Specific Ranking Algorithm により再ランキングを行う．本処理によりトピックドリフト問題の低減を図る．得られた各検索結果$R(er_i)$に対し，順に$R(er_1)$の 1 位，$R(er_2)$の 1 位，…，$R(er_u)$の 1 位$(u = |ER|)$を取得し，続いて各$R(er_i)$の 2 位を同様に取得する．これを上位q位まで行い，全順序集合を生成する．

表層的トピックドリフト問題，深層的トピックドリフト問題への対策として，以下の Extended Topic-Specific Ranking Algorithm を提案する．

• Extended Topic-Specific Ranking Algorithm

検索結果の並び替え対象の Web ページの集合Wに対し，検索結果の順位を利用して順序を導入する．この順序によりWは全順序集合となる．すなわち，$\forall w_i \in W$に対して，$w_i < w_{i+1}$である．

Input:並び替え対象の Web ページの全順序集合W

Output:並び替え後のW

step1. $w_i \in W$のタイトル内に$c \in C(a)$が出現せず，cを構成する形態素のうち，ストップワード(助詞など)以外のものがタイトル内に含まれているw_iの集合をSTPとする．

step2. $w_i \in W$内に出現する単語の集合$\overline{w_i}$と種類拡張集合$T(k)$とのジャッカード係数$\text{Jaccard}(\overline{w_i}, T(k)) = |\overline{w_i} \cap T(k)|/|\overline{w_i} \cup T(k)|$が 0 となる$w_i$の集合を$DTP$とする．

step3. その他のw_iの集合を$O\big(= W - (STP \cup DTP)\big)$とする．

step4. 各$w_i \in W$に対してiを 1 から$|W|$まで動かす．

$w_i \in O$ならば，w_iはそのままとし，$O = O - \{w_i\}$とする．

$w_i \notin O$ (すなわち$w_i \in STP \cup DTP$)ならば，

$w^* = argmin_{w_j \in O, i<j}\{w_j\}$を求め，$w^*$と$w_i$を入れ替え，

$O = O - \{w^*\}$とする．

$O = \emptyset$となったとき処理を止める．

Vol.2017-DPS-172 No.16
Vol.2017-SPT-26 No.16
Vol.2017-EIP-78 No.16
2017/11/30

5. Extended Topic-Specific Ranking Algorithm の動作例

本動作例の説明においてlを兵庫，kを島，aを美しいとし並び替え対象の Web ページの集合Wは表 1 の通りとする．また，$T(k)$={淡路島，沼島，成ヶ島，絵島}，$C(a)$ ={美しい，きれい}とする．

• 集合STP

w_iのタイトル中に$C(a)$の要素が出現せず“美”の文字が含まれるw_iをSTPの要素とするので，表 1 より，w_2のみが該当し，$STP = \{w_2\}$となる．

• 集合DTP

$w_i \in W$内に出現する単語の集合$\overline{w_i}$と種類拡張集合$T(k)$とのジャッカード係数Jaccard$(\overline{w_i}, T(k))$が 0 となるw_iをDTPとするので，表 1 より，w_4のみが該当し，$DTP = \{w_4\}$となる．

• 集合O

その他のw_iの集合が$O(= W - (STP \cup DTP))$であるため$O = \{w_1, w_3, w_5\}$となる．

これら記号のもと，iを 1 から$|W| = 5$まで動かし，$w_i \in O$ならば，w_iはそのままとし，$O = O - \{w_i\}$とする．$w_i \in STP \cup DTP$ならば，$w^* = argmin_{w_j \in O, i<j}\{w_j\}$を求め，$w^*$と$w_i$を入れ替え，$O = O - \{w^*\}$とする．$O = \emptyset$となったとき処理を止める．

$i = 1$のとき，$w_1 \in O$のためそのままとし，$O = O - \{w_1\}$とする．従って，表 2 の処理後の通りになる．

次に，$i = 2$のとき，$w_2 \in STP \cup DTP$のため，w^*を求め，w_2とw^*を入れ替える．集合Oの要素の中で，jがi =2 より大きく，その中で一番小さいものはw_3となる．従って，表 3 の処理後のようにw_2とw_3を入れ替え，$O = O - \{w_3\}$とする．

次に，$i = 3$のとき，$w_2 \in STP \cup DTP$のため，w^*を求め，w_2とw^*を入れ替える．集合Oの要素の中で，jがi =3 より大きいもので，かつ，その中で一番小さいものはw_5となる．従って，表 4 の処理後のようにw_2とw_5を入れ替え，$O = O - \{w_5\}$とする．ここで$O = \emptyset$となるので処理をやめる．

結果としては表 4 の処理後に示されたWを得る．

表 2　$i = 1$のときの処理前と処理後の変化

処理前		処理後	
W	O	W	O
w_1	$\{w_1, w_3, w_5\}$	w_1	$\{w_3, w_5\}$
w_2		w_2	
w_3		w_3	
w_4		w_4	
w_5		w_5	

表 3　$i = 2$のときの処理前と処理後の変化

処理前		処理後	
W	O	W	O
w_1	$\{w_3, w_5\}$	w_1	$\{w_5\}$
w_2		w_3	
w_3		w_2	
w_4		w_4	
w_5		w_5	

表 4　$i = 3$のときの処理前と処理後の変化

処理前		処理後	
W	O	W	O
w_1	$\{w_5\}$	w_1	$\emptyset$
w_3		w_3	
w_2		w_5	
w_4		w_4	
w_5		w_2	

6. 評価

6.1 評価の目的

本提案モデルの有用性を確認するため，評価を行った．評価項目としては，本提案モデルの評価，本研究独自の Extended Topic-Specific Ranking Algorithm の評価，従来の Web サーチエンジンとの比較，これら 3 つについて評価を行った．

6.2 評価の方法

まず，Extended Topic-Specific Ranking Algorithm の評価を行った．本評価では，被験者 7 名に作成してもらった検索要求 35 件から，ランダムに抽出した 2 件の検索要求を利用した．また，アルゴリズムによって表層的（深層的）トピックドリフト問題と判断したものの数をt_a，人手で決めた表層的（深層的）トピックドリフト問題と判断すべきものをt_h，アルゴリズムによって判断したもののうち，実際に表層的（深層的）トピックドリフト問題であると判断したものの数をt_pとし，次式(1)に示す．T適合率，T再現率という式(1)を本アルゴリズムの評価尺度とした．

$$T\text{適合率} = \frac{t_p}{t_a}$$
$$T\text{再現率} = \frac{t_p}{t_h} \quad (1)$$

次に，本提案モデルの評価を行うため，前述の 2 件の検索要求を用いて実験を行った．本提案モデルに従った検索を行い，適合率，再現率，平均適合率，11 点適合率の算出を行った．また，この際，正解は人手によって判断した．

最後に，従来の Web サーチエンジンとの比較するため，この 2 件の検索要求を用いサーチエンジン(Google)で，検索を行ってもらった．その検索結果上位 10 件の Web ペー

Vol.2017-DPS-172 No.16
Vol.2017-SPT-26 No.16
Vol.2017-EIP-78 No.16
2017/11/30

表 5　検索結果の例

W	Web ページのタイトル
w_1	隠れた魅力がいっぱい！岩手県のオススメ観光スポットランキング TOP15 ...
w_2	岩手県民がオススメする、岩手の穴場観光スポット 20 選！ | RETRIP ...
w_3	岩手に行くならここ！ おすすめ観光スポット 70 選 - skyticket 観光ガイド
w_4	岩手の観光・旅行スポット 10 選！岩手の大自然を満喫出来る必見観光地 ...
w_5	隠れた魅力がいっぱい！岩手県のオススメ観光スポットランキング TOP15 ...

ジに対し，検索意図と合致する Web ページかどうか判断してもらった．その結果を利用し，検索意図と合致しないページについて表層的トピックドリフト問題の発生している Web ページ，深層的トピックドリフト問題の発生している Web ページの割合を求め，平均適合率と 11 点平均適合率を求めた．

6.3 評価に使用したデータ

評価実験で使用した検索要求の例は表 5 の通りである．また，表 6 は，l =岩手，k =観光地，a=有名なという検索要求で検索した際に，本提案システムによって提示される結果の例である．また，表 7 に，拡張検索要求集合ERの要素の例を示す．$T(k)$の決定にあたっては，$m = 10$とした．更に，表 8 にa =“有名な”のときの，同義語s_iと検索結果件数（抜粋）を示す．本評価では，閾値として設定した 3,000,000 件以上の単語を$C(a)$の要素とした．表 8 の中では，“名の高い”のみが該当する．

表 6　実験で使用した検索要求

l	k	a
海外	射撃場	安全な
岩手	観光地	有名な

表 7　拡張検索要求集合ERの要素の例

l	$T(k)$	$C(a)$
岩手	観光地	有名な
岩手	穴場観光地	有名な
岩手	温泉観光地	有名な
岩手	観光地	名の高い
岩手	穴場観光地	名の高い
岩手	温泉観光地	名の高い
岩手	観光地	名の売れた
岩手	穴場観光地	名の売れた
岩手	温泉観光地	名の売れた

表 8　a=“有名な”のときの同義語s_iと検索結果件数（抜粋）

s_i	検索結果件数
有名な	約 1,140,000
著名な	約 1,930,000
名の通った	約 1,070,000
名の知れた	約 1,400,000
誰もが知る	約 585,000
知らない者はいない	約 941,000
名高い	約 430,000
名の高い	約 3,470,000
知名度のある	約 390,000

6.4 評価の結果

Extended Topic-Specific Ranking Algorithm の評価結果は，表 9 の通りとなった．次に，本提案モデルの評価を行うため，本提案モデルによる検索を表 6 に示した検索要求で行い，閾値を 3,000,000，$p = 10, q = 100$とし，実験を行った．本提案モデルに従い，適合率，再現率，平均適合率，11 点平均適合率の算出を行った．結果を表 10 に示す．また，従来の Web サーチエンジンとの比較のため，本提案モデルと既存の Web サーチエンジンについて比較評価を行った．本評価の結果を表 11 に示す．

表 9　Extended Topic-Specific Ranking Algorithm の精度

	DTP	
検索要求	T適合率	T再現率
海外　射撃場　安全な	100%	8.8%
岩手　観光地　有名な	27.2%	72.8%

7. 考察

7.1 誤判定の例・原因分析

誤って取得してしまう Web ページの例として，関東にある“東京ディズニーランド”や“横浜ランドマークタワー”について紹介しているページが挙げられる．

この原因として，種類拡張集合$T(k)$を作成する際に，

Vol.2017-DPS-172 No.16
Vol.2017-SPT-26 No.16
Vol.2017-EIP-78 No.16
2017/11/30

表 10　本提案モデルの精度

	適合率	再現率	平均適合率	11 点平均適合率
海外 射撃場 安全な	65.6%	100%	80.1%	81.7%
岩手 観光地 有名な	89%	40%	87.4%	87.7%
平均	77.3%	70%	83.4%	84.7%

表 11　既存の Web サーチエンジンとの比較

	既存の検索サーチエンジン		本提案モデル	
検索要求	平均適合率	11 点平均適合率	平均適合率	11 点平均適合率
海外 射撃場 安全な	57.1%	60.4%	89.5%	89.5%
岩手 観光地 有名な	88.9%	89.8%	100%	100%

確かに“観光地”という部分文字列を含むが“関東観光地”のように，岩手の観光地の情報と関係のない単語を取得していることが考えられる．また，Web ページのタイトルが“関東の観光スポット”や“富山の観光スポット”のように岩手の観光地の情報とは全く関係がないタイトルであるのにも関わらず，タイトル情報の活用が十分でないため，精度が下がる原因になっていると考えられる．

7.2 取りこぼしの例・原因分析

取りこぼしてしまう Web ページの例としては岩手県にある北上市の観光地情報についての Web ページが挙げられる．

この原因として，この Web ページでは“観光地”という単語は使われておらず，“観光スポット”という似たような意味の単語であったため．取得できなかったことが考えられる．また，北上市は岩手県に属しているため本来は取得するべきページではあるが，北上市が岩手に含まれるといった判断をできないことも考えられる．

また，種類拡張集合$T(k)$を収集する際に，“観光地”という部分文字列を含む単語は考慮しているが，“中尊寺金色堂”のような具体名な地名について考慮されておらず，取りこぼしが発生していると考えられる．

8. おわりに

本稿では，希望目的地情報検索システムのモデルの提案を行った．また，本提案モデルの有用性の確認のため評価を行った．その結果，平均で，適合率 77.3%，再現率 70%，平均適合率 83.4%，11 点平均適合率 84.7%であることを確認した．

参考文献

[1] 大石哲也，蔵元俊介，峯恒憲，長谷川隆三，藤田博，越村三幸：関連単語抽出アルゴリズムを用いた Web 検索クエリの生成，電子情報通信学会論文誌，Vol. J92-D, No.3, pp.281--292 (2009).
[2] 福地大助，山本岳洋，田中克己：動詞クエリの語間の関係性に基づくクエリマイニング，人工知能学会論文誌，Vol.32, No. 1, pp.1--15 (2017).
[3] 荒谷寛和，藤田茂，菅原研二：ウェブページの相互評価による再ランキング手法の改良，日本知能情報ファジィ学会誌，Vol.18, No.2, pp.196--212 (2006).
[4] 山本岳洋，中村聡史，田中克己：ランキング結果閲覧のための柔軟な再ランキングインタフェース，情報処理学会論文誌，Vol.3, No.4, pp.48--64 (2010).
[5] 高杉真理子，児玉英一郎，王家宏，高田豊雄：Web を利用した希望目的地情報の検索に関する研究，平成 25 年度電気関係学会東北支部連合大会講演論文集，2F01 (2013).