# ページとクエリの連想確率に基づく希少なWebページの検索

山中 隆広[1]　湯本 高行[1]　新居 学[1]　上浦 尚武[1]

**概要**：近年，情報検索におけるシステムの向上により，入力したクエリに対して一般的に知られている情報を取得することは簡単になった．しかし，クエリに関係するが，一般的に知られていない情報は依然として発見することは難しい．そこで本研究では，そのような情報を希少な情報と定義し，希少な情報を検索することを目的とする．手法としては，ページが与えられた時にクエリを連想できる確率 (関連度) が高く，クエリが与えられた時にそのクエリからページを連想できない確率 (非典型度) が高いページを探す．確率の計算として，ソーシャルブックマークのタグとページ内の語の関係を用いる手法とページ内の語の共起関係に基づく手法の 2 つを提案する．

## 1. はじめに

近年，Web ページの普及に伴って，情報検索や推薦手法の性能の向上により情報収集は容易になっている．たとえば，Google[*1]や Yahoo[*2]などの Web 検索エンジンや個人の趣味・嗜好に合わせた情報を推薦する手法 [1][2] が存在する．これらを用いることによって，入力されたクエリに関係し，一般的に知られている情報，つまり典型的な情報を取得することは簡単になった．実際に，「虫歯」と検索することで虫歯の予防策や治療，虫歯の症状などについてなどの一般的に連想されやすいページが上位に出力されることがわかる．

しかし，典型的な情報ばかりが推薦されると，同じような内容の情報が多くなり，ユーザがクエリに関して新たな知識を得ることが難しくなる．そのため，他の研究では情報検索や推薦手法によって出力される情報の内容に偏りがないようにするため，トピックの多様化 [3][4][5] に関する研究や有用なページを上位に提示する再ランキング手法 [6] の研究などが行われている．また，クエリに関して詳しい人は，検索エンジンで上位に出力される典型的な情報よりも，下位の方で出力されるような一般的には知られていない情報，つまり非典型的な情報を求める．だが，クエリに関係し，非典型的な情報を探すことはユーザにとって大きな負担となる．そこで，本研究ではユーザが入力したクエリに関係し，かつ非典型的な情報が記載されている Web ページを希少な Web ページと定義し，その希少な Web ページを検索することを目的とする．

本研究では希少な Web ページの検索を行うため，ページが与えられた時にクエリを連想できる確率 $P(q|d)$ を関連度，クエリが与えられた時にそのクエリからページを連想できない確率 $P(\bar{d}|q)$ を非典型度と定義する．関連度により関係するページのみの抽出を行い，抽出したページにおいて非典型度を算出し，降順にランキングすることで希少な Web ページの検索を行う．関連度と非典型度を算出する手法として，ソーシャルブックマーク (以下，SBM) のタグとページ内の語の関係を用いる手法とページ内の語の共起関係に基づく手法の 2 通りの手法を提案する．

## 2. 関連研究

### 2.1 SBM サービス

本研究では，関連度と非典型度の算出において，カテゴリと名詞の関係または語の共起関係を表す学習データが必要となる．そのため，学習データとして SBM サービスにおけるページ内の語とページに対するタグの関係を用いる．

SBM サービスとは，Web ブラウザにおけるブックマーク機能と同様に，インターネット上で「ブックマーク」を登録することができ，それらを整理し，他のユーザと共有することができるサービスの 1 つである．日本国内では，はてなブックマーク[*3]，海外では Delicious[*4]などの SBM サービスが多く利用されている．SBM サービスの特徴として，ブラウザのブックマーク機能ではブックマークしたブラウ

[1] 兵庫県立大学
University of Hyogo

[*1] https://www.google.co.jp/
[*2] http://www.yahoo.co.jp/
[*3] http://b.hatena.ne.jp/
[*4] https://delicious.com/

ザでしか見ることができないが，SBM サービスで登録したブックマークは Web 上に保存されているため，どのコンピュータからでも閲覧することが可能である．また，複数のユーザとブックマークを共有して，Folksonomy(フォークソノミー) によりタグ付けを行うことができる．Folksonomy では，複数のユーザが各々の Web ページに対して自由に単語やキーワードなどのタグを付加し，そのページを検索できるように分類することが可能である．さらに，1 つのページに対して複数のタグを付けることも可能である．

SBM サービスに関する研究として，ブックマークした Web ページに対してどのようなタグを使用されているかを調査した Golder ら [7] の研究がある．そのタグの種類の一例を以下に示す．

(1) Web ページのトピックスを表すタグ
(例：“料理”，“スポーツ”)
(2) Web ページの種類を表すタグ (例：“本”，“blog”)

本研究では，カテゴリと名詞の関係を用いる手法において，(1) のトピックを表すタグと (2) の種類を表すタグに着目し，ユーザはブックマークした Web ページをタグによってカテゴリ分類していると考える．これにより，ページとタグの関係を多く取得する．また，多くのページを取得することで，ページ内の語を用いて，語の共起関係についても取得することができる．

### 2.2 非典型的な情報検索

本研究では，クエリに関係し，かつ非典型的である Web ページを検索するため，ページの典型性を考え，クエリからのページの連想のしにくさを数値化する必要がある．同様の目的である，クエリに対して意外な (連想されにくい) 情報を発見する手法として，佃ら [8] の研究がある．この研究では Wikipedia におけるリンク構造を用いて，タイトルとなる語を主題語として，別の主題語へのリンクをグラフとして考える．このグラフを用いることにより，2 つの主題語間のたどり着きにくさをスコアとして表し，意外な主題語の繋がりを発見している．また，主題語がどれだけ知られているかを表す認知度を算出することで，関係のない主題語との繋がりの排除している．この 2 つの指標を用いることで Wikipedia の中から意外な情報を発見する手法を提案している．この手法では，構造化されたデータには適用できるが，構造化されていないデータには対応できない．

そこで本研究では，Web サービスによって構造化されたデータである SBM を用いたページとタグの関係，またはページ内の語の共起関係による指標によりページの連想のしにくさの数値化を行う．

## 3. カテゴリを用いた希少な Web ページの推薦

本研究における先行研究として，Yumoto ら [9] の希少な Web ページを推薦する既存手法がある．この手法では，クエリとカテゴリは同じものとして，タグに使用されている語を入力カテゴリとし，そのカテゴリを用いて希少度の算出を行う．希少度はページがカテゴリに対してどれだけ希少な Web ページであるかを表す指標で，ページからカテゴリがどれだけ連想できるかを表す関連度と，カテゴリからページをどれだけ連想できないかを表す非典型度の積によって算出できる．関連度と非典型度と希少度は 0 から 1 の範囲で算出され，希少度を用いて降順にランキングすることで希少な Web ページの推薦を行っている．しかし，この手法では 2 つの問題点がある．

まず，1 つ目の問題点として，タグに使用された語でなければ入力カテゴリとして使用できないという問題がある．本研究では，この問題を解決するため，クエリとカテゴリは違うものとして考え，入力クエリに対して算出を行い，希少な Web ページの検索を行う．

次に，2 つ目の問題点として，クエリに関係ない Web ページの希少度が大きくなってしまうという問題がある．そこで，関連度に関して分析を行った．その結果，関係する Web ページの確率は 0.1 から 1.0 と広範囲で分布していることがわかった．このことから，関係する Web ページは 0.1 に近い値を算出すると，希少度は小さくなり，関係しない Web ページでは非典型度が大きくなってしまうため，積をとると希少度が大きくなる．そのため，関係しない Web ページが希少な Web ページとして上位に上がってしまう．この問題について我々は関連度のフィルタリングによる方法を提案している [10]．本研究においても，関連度をフィルタリングとして使用することで希少な Web ページを検索する手法を提案する．

## 4. クエリによる希少な Web ページの検索

本研究では，希少な Web ページの検索を行うための関連度と非典型度の算出には，SBM サービスにおけるタグを用いる手法と SBM サービスを用いずにページ内の語の共起関係に基づく手法の 2 通りの手法を提案する．それぞれの提案手法と既存手法の概要を表 1 に示す．表 1 において，$C$ は SBM で使われるタグの集合である．提案手法における算出方法については以下で説明する．

### 4.1 SBM のタグを用いた手法

この手法では，3 章で説明した関連度と非典型度の既存手法を用いる．まず，入力クエリ $q$ のカテゴリ推定を行う．カテゴリ推定の結果から，クエリ $q$ に関係するカテゴリ $c$ をユーザが選択する．選択したカテゴリ $c$ を用いて，入力した文書集合 $\boldsymbol{D}$ において，関連度と非典型度の算出を行う．関連度によるフィルタリングにより，関係する文書を抽出し，抽出した文書集合 $\boldsymbol{D}$' において，非典型度の算出を行う．非典型度の値によって降順に並べ替えることで，希少な Web ページが上位に来るようにランキングし，希

**表 1** 手法の概要

| 手法 | $q$ と $c$ の関係 | フィルタリング | ランキング |
|---|---|---|---|
| 既存手法 | $q \in C$ に限定 | なし | 希少度 (関連度×非典型度) |
| SBM のタグ | $q$ に関係する $c \in C$ を使用 | 関連度 | 非典型度 |
| 語の共起関係 | $C$ には依存しない | 関連度 | 非典型度 |

少な Web ページの検索を行う．以下に，カテゴリ推定の算出手法と既存手法であるカテゴリを用いた関連度と非典型度の算出手法を示す．

#### 4.1.1 カテゴリ推定

入力したクエリ $q$ に関係するカテゴリ候補として，クエリ $q$ を含む Web ページに付けられた名詞 1 つのタグの中で使用されている数の多い上位 $N$ 個を使用する．これにより，カテゴリ候補として不要なカテゴリを減らして，内容を表すタグに限定する．本研究では $N$ として，実験により適切なカテゴリが上位に出力された $N = 20$ を用いる．

次に，このカテゴリ候補にナイーブベイズによる算出式を用いてカテゴリ推定を行う．クエリ $q$ からカテゴリ $c$ を連想できる確率 $P(c|q)$ はベイズの定理により (1) 式の等号が成り立つ．$P(q)$ はクエリ $q$ が変化しないため一定で，$P(c)$ はどのカテゴリ $c$ においても一定であるとすると，(1) 式の比例関係が成り立つ．

$$P(c|q) = \frac{P(c)P(q|c)}{P(q)} \propto \prod_{q_i \in q} P(q_i|c) \tag{1}$$

(1) 式において，$P(q_i|c)$ は SBM サービスにおけるブックマーク数 (以下，BM 数) を用いて算出を行う．$P(q_i|c)$ では，カテゴリ $c$ 内でクエリ $q_i$ が出現するほどクエリ $q$ はカテゴリ $c$ に関係すると考える．つまり，データベース (以下，DB) においてカテゴリ $c$ のタグを含む BM 数 $|BM_c|$ と，本文にクエリ $q_i$ の名詞を含み，かつカテゴリ $c$ のタグを含む BM 数 $|BM_c \cap BM_{q_i}|$ の比をとり，(2) 式のように算出を行う．

$$P(q_i|c) = \frac{|BM_c \cap BM_{q_i}|}{|BM_c|} \tag{2}$$

カテゴリ候補の各カテゴリ $c$ に対して，(1) 式により算出を行い，降順にランキングすることによりカテゴリ推定を行う．ただし，クエリ $q$ を含み，かつカテゴリ $c$ のタグが付いた Web ページの数が 10 件以下のカテゴリは，実験によりクエリ $q$ に関係するカテゴリとして不適切であるため除去する．

#### 4.1.2 カテゴリを用いた関連度

文書 $d$ からカテゴリ $c$ が連想できる確率を関連度 $P(c|d)$ とする．$P(c|d)$ は，文書 $d$ からカテゴリ $c$ を連想できない確率 $P(\bar{c}|d)$ との和が 1 となることが自明であるため (3) 式で表せる．このとき，文書 $d$ の本文に記載されたすべての語 $w_i$ からカテゴリ $c$ を連想できない場合，文書 $d$ はカテゴリ $c$ に関係しないとする．(3) 式の $P(\bar{c}|d)$ は $w_i$ からカテゴリ $c$ を連想できない確率 $P(\bar{c}|w_i)$ の総積によって算出できると定義すると，(4) 式となる．また，$P(\bar{c}|w_i)$ は連想できる確率 $P(c|w_i)$ との和が 1 となることが自明であるため (4) 式は (5) 式で算出できる．これにより，文書中の語 $w_i$ が 1 つでもカテゴリ $c$ に関係すると，総積によって求められる値は小さくなるため，関連度としては値が大きくなる．

$$P(c|d) = 1 - P(\bar{c}|d) \tag{3}$$

$$= 1 - \prod_{w_i \in d} P(\bar{c}|w_i) \tag{4}$$

$$= 1 - \prod_{w_i \in d} \{1 - P(c|w_i)\} \tag{5}$$

(5) 式において，$P(c|w_i)$ は SBM サービスにおける BM 数を用いて算出を行う．$P(c|w_i)$ では，本文中に $w_i$ を含む Web ページのブックマークにカテゴリ $c$ のタグを含む BM 数が多いほど，$w_i$ はカテゴリ $c$ に関係すると考える．つまり，DB において，本文中に $w_i$ を含む Web ページの BM 数 $|BM_{w_i}|$ と，本文中に $w_i$ を含み，かつカテゴリ $c$ のタグを含む BM 数 $|BM_c \cap BM_{w_i}|$ の比をとり，(6) 式のように算出する．

$$P(c|w_i) = \frac{|BM_c \cap BM_{w_i}|}{|BM_{w_i}|} \tag{6}$$

ここで，関連度の算出には (5) 式のように文書 $d$ の本文中の語 $w_i$ を用いる．しかし，すべての語を使用すると，「今日」などのざまざまな分野で記述される語，つまり一般的な語の影響を受ける可能性がある．そこで，本研究では文書 $d$ において重要と思われる語を主要語とし，主要語のみを算出に用いる．主要語の抽出には，$TF\text{-}RIDF$[11] を用いて，$TF\text{-}RIDF$ 値の上位 10 個の名詞のみを使用する．$TF\text{-}RIDF$ の算出には (7) 式を用いる．(7) 式において，$tf(w_i, d)$ は文書中の語 $w_i$ の出現頻度，$df(w_i)$ は DB における $w_i$ が出現した文書数，$|DB|$ は DB における全文書数，$\sum_{d \in DB} tf(w_i, d)$ は DB 内の全文書における $w_i$ の出現数の総和を表す．

$$TF\text{-}RIDF = tf(w_i, d) \times \left\{ \log_2 \left( \frac{|DB|}{df(w_i)} \right) + \log_2 \left( 1 - \exp \left( - \frac{\sum_{d \in DB} tf(w_i, d)}{|DB|} \right) \right) \right\} \tag{7}$$

$TF\text{-}RIDF$ は，文書中の語 $w_i$ の出現頻度である $TF$ 値と，全文書で $w_i$ が出現する文書数の逆数を取った実際の $IDF$ 値から，ポアソン分布により推定された $IDF$ 値を

引いた値である $RIDF$ 値との積を表す．$TF$ 値は本文中に多く出現する語は重要であると考え用いられている．また，$RIDF$ は $IDF$ の多くの文書で使用される語の値は小さくなり，少ない文書で使用される語の値は大きくなる性質に加え，少ない文書で使用され，かつ出現頻度が多い語の値はより大きくなる．これにより，一般語の値は小さくなり，主要語は大きくなるため $TF$-$IDF$ よりも適切な主要語を抽出できる．

#### 4.1.3 カテゴリを用いた非典型度

カテゴリ $c$ から文書 $d$ が連想できない確率を非典型度 $P(\bar{d}|c)$ とする．このとき，文書 $d$ の本文に記載されたすべての語 $w_i$ がカテゴリ $c$ から連想できない場合，文書 $d$ はカテゴリ $c$ 内で典型的でないとすると，非典型度は語 $w_i$ がカテゴリ $c$ から連想できない確率 $P(\bar{w_i}|c)$ の総積で算出できる．また，$P(\bar{w_i}|c)$ は，連想できる確率 $P(w_i|c)$ との和が 1 となることが自明であるため (8) 式で算出できる．

$$P(\bar{d}|c) = \prod_{w_i \in d} P(\bar{w_i}|c) = \prod_{w_i \in d} \{1 - P(w_i|c)\} \tag{8}$$

(8) 式において，$P(w_i|c)$ は SBM サービスにおける BM 数を用いて算出を行う．$P(w_i|c)$ では，カテゴリ $c$ 内で本文中に $w_i$ が出現するほど，$w_i$ はカテゴリ $c$ 内で典型的であると考える．つまり，DB においてカテゴリ $c$ のタグを含む BM 数 $|BM_c|$ と，本文にクエリ $q_i$ の名詞を含み，かつカテゴリ $c$ のタグを含む BM 数 $|BM_c \cap BM_{w_i}|$ の比をとり，(9) 式のように定義する．

$$P(w_i|c) = \frac{|BM_c \cap BM_{w_i}|}{|BM_c|} \tag{9}$$

ここで，非典型度の算出式 (8) 式は Web ページ $d$ の本文中の主要語 $w_i$ を用いる．主要語の抽出には，関連度と同様に $TF$-$RIDF$ 値として (7) 式を用いて算出し，値が大きい上位 10 個の名詞を使用する．

### 4.2 語の共起関係に基づく手法

この手法では，クエリを用いた関連度と非典型度の算出を行い，希少な Web ページの検索を行う．関連度の算出では擬似適合フィードバック [12] を用いた算出手法によって，文書集合 $\boldsymbol{D}$ のフィルタリングを行い，関係する文書の抽出を行う．抽出した文書集合 $\boldsymbol{D}$' を用いて語の共起確率による非典型度を算出し，降順に並べ替えることで希少な Web ページが上位に来るようにランキングし，希少な Web ページの検索を行う．以下に，クエリを用いた関連度と非典型度による算出手法を示す．

#### 4.2.1 クエリを用いた関連度

関連度の算出には擬似適合フィードバックを用いて 2 次検索を行い，クエリ $q$ に関係する文書の抽出を行う．1 次検索として，文書 $d$ からクエリ $q$ の連想確率により各文書ごとに値を算出し，降順にランキングした場合の上位 $k$ 件をクエリ $q$ に関係する文書と仮定する．次に，上位 $k$ 件を 1 つの文書として各文書とのコサイン類似度を算出し，2 次検索を行う．コサイン類似度によって算出した値がしきい値以下の文書を，関係しない文書としてフィルタリングを行い，関係する文書の抽出を行う．本研究では，2 次検索で使用する上位 $k$ 件の文書として，ROC 曲線による AUC が最も良くなった $k = 3$ を使用する．以下には 1 次検索による手法として 2 通りの提案手法を示す．また，2 次検索のコサイン類似度の算出式も示す．

**文書からのクエリの連想確率による 1 次検索**

**・候補文書集合を使用**

文書集合 $\boldsymbol{D}$ において，文書 $d$ からクエリ $q$ を連想できる確率を $P(q|d)$ とする．文書 $d$ がクエリ $q$ に含まれる語 $q_i$ にすべて関係する場合，文書 $d$ はクエリ $q$ に関係すると考える．すると，$P(q|d)$ は (10) 式で表すことができる．また，(10) 式において，$P(q_i|d)$ はスムージングを用いて (11) 式によって求める．

$$P(q|d) = \prod_{q_i \in q} P(q_i|d) \tag{10}$$

$$P(q_i|d) = \frac{tf(q_i, d) + \mu P(q_i|D)}{|d| + \mu} \tag{11}$$

ここで，$tf(q_i, d)$ は文書 $d$ において名詞 $q_i$ が出現する頻度，$|d|$ は文書 $d$ の文書長，すなわち名詞の総数を表す．$\mu$ はスムージングパラメータを表し，ROC 曲線による AUC が最も良くなった 100 を使用する．また，$P(q_i|D)$ は (12) 式により求める．

$$P(q_i|D) = \frac{1}{|D|} \sum_{d \in D} \frac{tf(q_i, d)}{|d|} \tag{12}$$

(12) 式において，$|D|$ は文書集合 $\boldsymbol{D}$ に含まれる文書数を表す．

**・DB の文書集合を使用**

DB よりクエリ $q$ の $tf$ 値が大きい上位 30 件の文書集合 $\boldsymbol{D_q}$ を抽出する．抽出した文書集合 $\boldsymbol{D_q}$ において，文書 $d$ がクエリ $q$ に関係する確率を $P(d|q)$ とすると，ベイズの定理より (10) 式で求めることができる．また，(10) 式において $P(q_i|d)$ はスムージングを用いて (13) 式によって求める．

$$P(q_i|d) = \frac{tf(q_i, d) + \mu P(q_i|D_q)}{|d| + \mu} \tag{13}$$

ここで，$tf(q_i, d)$ は文書 $d$ において名詞 $q_i$ が出現する頻度，$|d|$ は文書 $d$ の文書長，すなわち名詞の総数を表す．$\mu$ はスムージングパラメータを表し，ROC 曲線による AUC が最も良くなった 100 を使用する．また，$P(q_i|D_q)$ は (14) 式により求める．

$$P(q_i|D_q) = \frac{1}{|D_q|} \sum_{d \in D_q} \frac{tf(q_i, d)}{|d|} \tag{14}$$

(14) 式において，$|D_q|$ は文書集合 $\boldsymbol{D_q}$ に含まれる文書数を表す．

**コサイン類似度による 2 次検索**

1 次検索による上位 3 件の文書を 1 つの文書とし，この文書を入力クエリ $q_d$ とする．$q_d$ と文書集合 $\boldsymbol{D}$ の各文書の名詞を用いてそれぞれのベクトルを作成し，名詞は $TF\text{-}IDF$ により重み付けを行う．$q_d$ と $\boldsymbol{D}$ の各文書のコサイン類似度を (15) 式により算出することで関係する文書の抽出を行う．(15) 式において，$d_i$ は文書 $d$，$q_i$ は入力クエリ $q_d$ に含まれる名詞 $w_i$ の重みを表す．

$$cos(\vec{d}, \vec{q_d}) = \frac{\sum_{i=1}^{m} d_i q_i}{\sqrt{\sum_{i=1}^{m} d_i^2}\sqrt{\sum_{i=1}^{m} q_i^2}} \quad (15)$$

ここで，入力クエリ $q_d$ と文書集合 $\boldsymbol{D}$ の各文書 $d$ において，すべての名詞を用いて，算出に使用すると，一般語または不要語の影響を受ける可能性がある．そこで，一般語または不要語を除去するため，ストップワードとして，DB における $df$ 値の上位 $s$ 個を使用する．本研究では $s$ として，F 値が最も良くなった $s = 100$ を使用する．

#### 4.2.2 クエリを用いた非典型度

クエリ $q$ において文書 $d$ が連想できない確率を非典型度 $P(\bar{d}|q)$ とする．文書 $d$ のすべての主要語 $w_i$ がクエリ $q$ と共起されない場合，文書 $d$ はクエリ $q$ に対して典型的でないとすると，非典型度は語 $w_i$ がクエリ $q$ から連想できない確率 $P(\bar{w_i}|q)$ の総積で算出できる．また，$P(\bar{w_i}|q)$ は，連想できる確率 $P(w_i|q)$ との和が 1 となることが自明であるため (16) 式で表せる．(16) 式の $P(w_i|q)$ は (17) 式により算出を行う．(17) 式において，$|D_q|$ は DB において文書中にクエリ $q$ の語を含む文書数を表す．また，$|D_{q_i} \cap D_{w_i}|$ は DB において文書中にクエリ $q$ を含み，かつ主要語 $w_i$ を含む文書数を表す．

$$P(\bar{d}|q) = \prod_{w_i \in d} P(\bar{w_i}|q) = \prod_{w_i \in d} \{1 - P(w_i|q)\} \quad (16)$$

$$P(w_i|q) = \frac{|D_q \cap D_{w_i}|}{|D_q|} \quad (17)$$

ここで，文書 $d$ の主要語 $w_i$ としては文書中の名詞を用いて，(7) 式による $TF\text{-}RIDF$ 値の高い上位 10 個の名詞を使用する．

## 5. 評価実験

評価実験には，提案した関連度による評価と非典型度による評価，関連度と非典型度を組み合わせたことによる希少な Web ページの評価の 3 つの実験を行った．DB には 2011 年 4 月 14 日から 2011 年 10 月 27 日と 2014 年 4 月 28 日から 2014 年 7 月 4 日までに，はてなブックマークよりブックマーク情報を収集したデータを用いて作成を行った．作成した DB の規模を表 2 に示す．なお，本研究で提案した SBM のタグを用いた手法においては，SBM におけるタグと Web ページのデータが必要となるが，語の共起関係に基づく手法においてはページ内の語の共起関係を用いるため，SBM から集められたデータに限定する必要はない．

評価には，表 3 に示す 10 個のクエリを使用し，各クエリに対して 20 件 (関係しない 10 件，関係する 5 件，希少 5 件) の文書集合 $\boldsymbol{D}$ を使用した. 各クエリの文書において典型的な内容と非典型的な内容の一例を表 4 に示す．クエリごとの各文書 $d$ を 2 人の被験者によって，表 5 に示すようにスコア付けを行い評価した．2 人の被験者を A と B として，それぞれが評価した結果の比較を表 6 に示す．

**表 2** DB の規模

| データセット情報 | 件数・種類数 |
|---|---|
| 総 BM 数 | 4,019,427 |
| 総タグ種類数 | 121,270 |
| 総 URL 数 | 158,993 |
| 取得名詞種類数 | 1,108,194 |

**表 3** 評価に使用したクエリ $q$

| | | | | |
|---|---|---|---|---|
| ストレス | スマホ | 虫歯 | 頭痛 | 面接 |
| ダイエット | 花粉症 | 朝食 | 風邪 | 薬 |

**表 4** 各クエリに含まれる文書内容の一例

| クエリ | | 内容 |
|---|---|---|
| ストレス | 典型的 | ストレスの対処法 |
| | 非典型的 | ストレスによる体温上昇のメカニズム |
| 花粉症 | 典型的 | 花粉情報，基本的な花粉症対策 |
| | 非典型的 | ヨガによる花粉症対策 |
| 朝食 | 典型的 | 健康的にダイエットできる朝食 |
| | 非典型的 | 世界のおいしい朝食 |
| 頭痛 | 典型的 | 頭痛の症状や種類 |
| | 非典型的 | ベルギー発の片頭痛緩和ヘッドバンド |
| 風邪 | 典型的 | 風邪の時に食べるといいもの |
| | 非典型的 | 咳止めに効くツボ |
| 面接 | 典型的 | 面接で良く聞かれる質問 |
| | 非典型的 | 子供を面接に強くする方法 |

**表 5** 各文書の評価におけるスコア付け

| スコア | 内容 |
|---|---|
| 2 | クエリに関係し，非典型的である (希少) |
| 1 | クエリに関係し，典型的である |
| 0 | クエリと関係がない |

表 6 より，被験者 A は関係する文書に対して 1 と評価することが多く，被験者 B は関係する文書に対して 2 と評価することが多いことが分かった．すなわち，被験者 A と B では，典型性の判断基準が異なるわけではなく，典型的と非典型的を分けるしきい値が異なると考える．また，$\kappa$ 統計量 (一致率) を算出すると 0.50 となり，ある程度の一致が見られることが分かった．

表 6　各文書の評価におけるスコア付け

| スコア | | 被験者 A | | | 計 |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | |
| 被験者 B | 0 | 85 | 4 | 8 | 97 |
| | 1 | 2 | 34 | 3 | 39 |
| | 2 | 4 | 45 | 15 | 64 |
| 計 | | 91 | 83 | 26 | 200 |

次に，SBM のタグを用いた手法において，カテゴリ推定の結果より被験者が選択するカテゴリ $c$ を使用するため，それぞれの被験者にクエリ $q$ に関連するカテゴリ $c$ を選択してもらった．それぞれの被験者が選択したカテゴリを表 7 に示す．

表 7　被験者が選択したカテゴリ

| クエリ | 被験者 A | 被験者 B |
|---|---|---|
| ストレス | 心理 | 心理 |
| スマホ | スマホ | スマホ |
| ダイエット | ダイエット | ダイエット |
| 花粉症 | 花粉症 | 健康 |
| 虫歯 | 虫歯 | 虫歯 |
| 朝食 | 食 | 食 |
| 頭痛 | 医療 | 健康 |
| 風邪 | 医療 | 健康 |
| 面接 | 面接 | 面接 |
| 薬 | 医療 | 医療 |

### 5.1　関連度の評価

各クエリにおいて，提案手法のカテゴリを用いた関連度とクエリを用いた関連度による 2 通りの手法により算出を行い，文書集合 $\boldsymbol{D}$ より関係する文書が抽出できているか評価を行った．被験者に評価してもらったスコアが 2 と 1 の場合を関係する文書，0 の場合を関係しない文書として評価に使用した．それぞれの提案手法において，(18) 式による F 値の平均が最も高くなるしきい値を調べたところ，カテゴリを用いた関連度では 0.02，クエリを用いた関連度における候補文書集合を使用した場合は 0.15，DB による文書集合を使用した場合は 0.11 で最も高くなった．それぞれのしきい値を用いて F 値を算出し，フィルタリングとして評価した結果を図 1 に示す．また，F 値による平均値を表 8 に示す．

$$
\begin{aligned}
precision &= \frac{\text{スコアが 2 か 1 で抽出した文書数}}{\text{抽出した文書数}} \\
recall &= \frac{\text{スコアが 2 か 1 で抽出した文書数}}{\text{スコアが 2 か 1 の文書数}} \\
F\text{-}measure &= \frac{2 \times precision \times recall}{precision + recall} \qquad (18)
\end{aligned}
$$

図 1 を見ると，ほとんどのクエリにおいてカテゴリを用いた関連度を使用した場合よりも，クエリを用いた関連度を使用した場合に F 値が高いことが確認できる．また，表 8 を見ても，クエリを用いた関連度を使用した方が F 値の

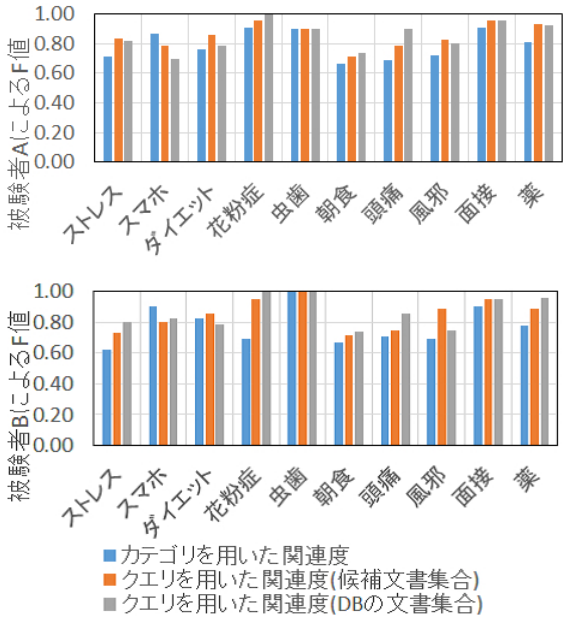


図 1　各被験者による各クエリの F 値

表 8　F 値の平均

| 被験者 | カテゴリを用いた関連度 | クエリを用いた関連度 | |
|---|---|---|---|
| | | 候補文書集合 | DB 文書集合 |
| A | 0.79 | 0.85 | 0.85 |
| B | 0.78 | 0.85 | 0.87 |

平均が高いことがわかる．これは，カテゴリを用いた関連度の算出に，クエリのカテゴリ推定より選択したカテゴリを使用しており，ユーザが選択した表 7 を見ると，多くのクエリで上位語のカテゴリが使用されている．そのため，カテゴリに関係するがクエリに関係しない文書が抽出されてしまい，F 値が低下したと考えられる．次に，クエリを用いた関連度において，1 次検索の算出する文書集合の違いで比べると，F 値で大きな差はないが，DB の文書集合 $\boldsymbol{D_q}$ を用いた方が少し高い事が確認できる．

### 5.2　非典型度の評価

各クエリにおいて，提案手法のカテゴリを用いた非典型度とクエリを用いた非典型度を用いて典型的でない文書 (希少な文書) が上位に出力されるか評価した．評価としては，被験者がスコアを 0 と判断した文書を文書集合 $\boldsymbol{D}$ より除去し，残りの文書集合に対してそれぞれの手法による非典型度を算出し，降順にランキングする．ランキングした上位 5 件のスコアを用いて $nDCG$(Normalized Discounted Cumulative Gain)[13] を算出する．その結果を図 2 に示す．また，$nDCG_5$ による平均値を表 9 に示す．

表 9　非典型度による $nDCG_5$ の平均

| 被験者 | カテゴリを用いた非典型度 | クエリを用いた非典型度 |
|---|---|---|
| A | 0.85 | 0.83 |
| B | 0.83 | 0.81 |

図 2 を見ると，ほとんどのクエリにおいて，手法の違いによる大きな差がないことがわかる．しかし，クエリに

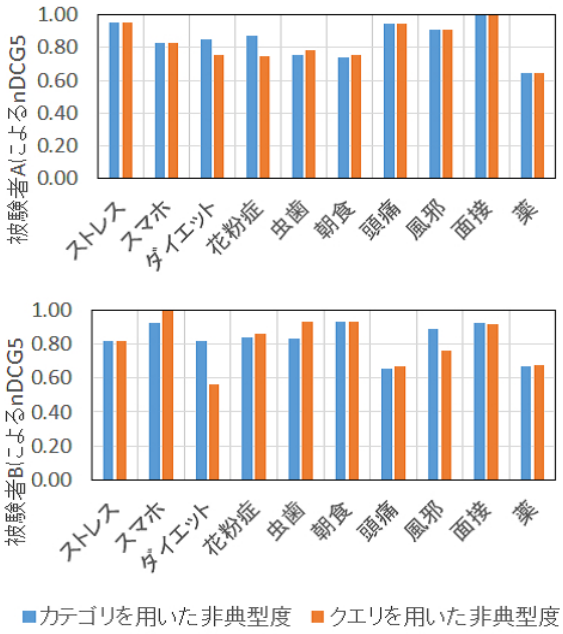


**図 2** 各被験者による非典型度の評価

よっては，カテゴリを用いた非典型度の方が $nDCG_5$ が大きい場合やクエリを用いた非典型度の方が $nDCG_5$ が大きい場合があることが確認できる．次に，表 9 を見ると，どちらの被験者においても，カテゴリを用いた非典型度を使用した方が $nDCG_5$ の平均値が少し高くなっている．

### 5.3 希少な Web ページの評価

最後に，関連度と非典型度を組み合わせて，希少な Web ページが上位に出力されているか評価した．関連度のしきい値は 5.1 節でそれぞれ使用した値を用いて，評価には 5.2 節と同様に，出力された上位 5 件の $nDCG$ を算出する．その結果を図 3 に示し，$nDCG_5$ による平均値を表 10 に示す．また，提案手法と比較を行うため，既存手法である希少度を用いた結果も図 3 と表 10 に示す．

さらに，各関連度のみにより，希少な Web ページの検索が行えるか評価を行った．その結果の平均と組み合わせによる希少な Web ページの評価の平均を図 4 に示す．

**表 10** 希少な Web ページの評価による $nDCG_5$ の平均

| 被験者 | タグを用いた手法 | 語の共起関係に基づく手法 | | 希少度 (既存) |
|---|---|---|---|---|
| | | 候補文書集合 | DB 文書集合 | |
| A | 0.56 | 0.74 | 0.77 | 0.47 |
| B | 0.54 | 0.73 | 0.76 | 0.48 |

図 3 を見ると，ほとんどのクエリにおいて語の共起関係に基づく手法の $nDCG_5$ が大きいことが確認できる．しかし，図 3 の被験者 B の $nDCG_5$ において「ストレス」や「ダイエット」は他のクエリと比べると，どの手法においても $nDCG_5$ が低い．これは，内容としてまったく関係しないがクエリの単語を多く含んでいるため関連度が大きくなってしまうことや，部分的に関係するまたは広義にとらえると関係する文書は関連度では値が大きくなるが，被験者は関係しないと判断してしまうため低下につながっていると考える．実際に，「ダイエット」のクエリでは「PDF ダイエット」という PDF の圧縮ソフトについて書かれたページがあり，ページ内ではダイエットという語が多く出現するため，関連度が大きくなった．しかし，内容としては「ダイエット」ではなく「PDF」の方が関連性が高い．そのため，「PDF」のようにページに関係するクエリとして，$P(q'|d) > P(q|d)$ となるクエリ $q'$ が存在しないか考慮

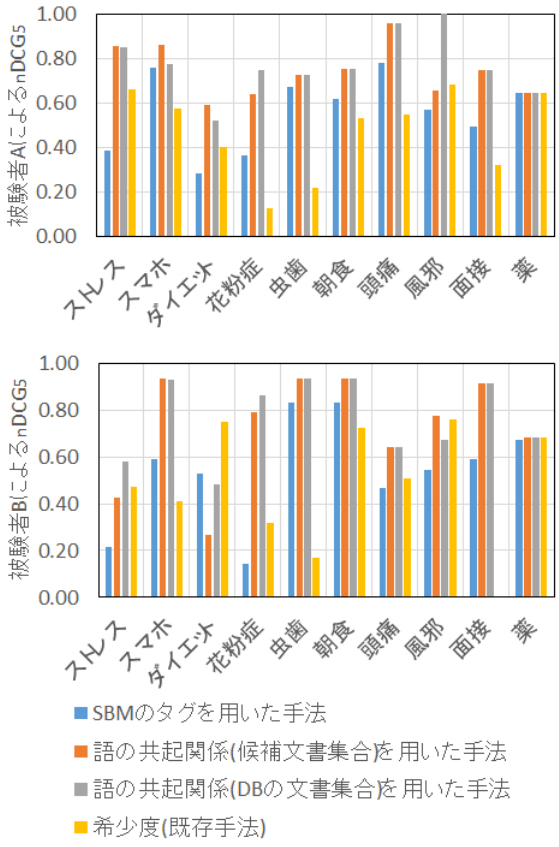


**図 3** 各被験者による希少な Web ページの評価

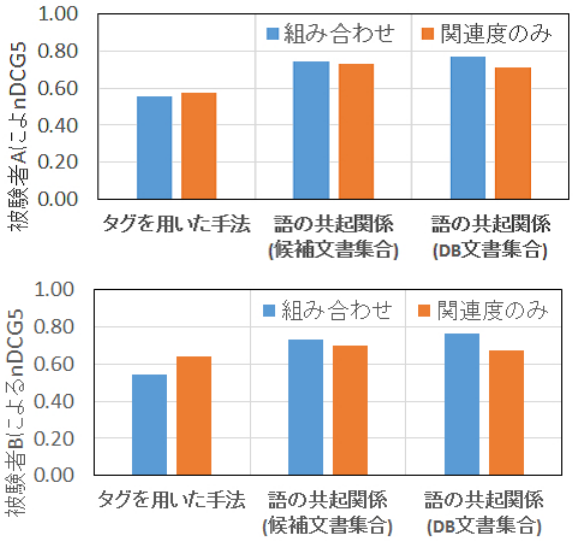


**図 4** 各被験者による関連度のみと組み合わせによる評価の比較

する必要があると考える．

次に，表 10 を見ると，既存手法である希少度を用いた場合よりも提案手法を用いた方が $nDCG_5$ が高くなっていることがわかる．また，提案手法において，SBM のタグを用いた手法よりも語の共起関係に基づく手法の方が $nDCG_5$ が高いことがわかる．これは，タグを用いた手法においては，カテゴリを用いた関連度による関係する文書の抽出がうまくできなかったため，カテゴリを用いた非典型度により関係しない文書の値が大きくなり上位に出力され，$nDCG_5$ が低下したと考えられる．また，語の共起関係基づく手法において，1 次検索として DB の文書集合 $\boldsymbol{D_q}$ から関係する文書を選択した場合に最も高くなった．

最後に，図 4 を見ると，カテゴリを用いた関連度では非典型度と組み合わせた場合よりも関連度のみの方が $nDCG_5$ 高くなった．しかし，クエリを用いた関連度においては，非典型度と組み合わせた方が $nDCG_5$ が高くなった．このことから，関連度だけではなく非典型度と組み合わせることで，希少な Web ページの検索が行えることがわかる．

## 6. おわりに

本研究では，入力されたクエリに対してページからクエリを連想できる確率である関連度とクエリからページを連想できない確率である非典型度を用いて，希少な Web ページの検索を行った．関連度と非典型度の算出には，SBM のタグとページ内の語の関係を用いる手法とページ内の語の共起関係に基づく手法の 2 通りの手法を提案し，それぞれの手法において評価を行った．

まず，関連度の評価では，カテゴリを用いた関連度とクエリを用いた関連度を使用して，文書集合 $\boldsymbol{D}$ より関係する文書が抽出できているか評価を行った．その結果，ほとんどのクエリにおいてクエリを用いた関連度を使用した場合に F 値が高いことが確認でき，1 次検索による算出の違いで比べると，DB の文書集合 $\boldsymbol{D_q}$ を用いた方が高いことが確認できた．

次に，非典型度の評価では，カテゴリを用いた非典型度とクエリを用いた非典型度を使用して，希少な文書が上位に出力されるか評価を行った．その結果，ほとんどのクエリにおいて，2 つの手法による $nDCG_5$ に大きな差はなかったが，平均としてはカテゴリを用いた非典型度の方が $nDCG_5$ が高くなっていた．

最後に，希少な Web ページの評価として，関連度と非典型度を組み合わせて，希少な Web ページが上位に出力されているか $nDCG$ による評価を行った．その結果，ほとんどのクエリにおいて，語の共起関係に基づく手法の $nDCG_5$ が大きいことが確認できた．しかし，図 3 の被験者 B の $nDCG_5$ において「ストレス」や「ダイエット」は他のクエリと比べると，どの手法においても $nDCG_5$ が低かった．これは，内容としてまったく関係しないがクエリの単語を多く含んでいるため関連度が大きくなってしまうことや，部分的に関係するまたは広義にとらえると関係する文書の関連度は値が大きくなるが，被験者は関係しないと判断してしまうため低下した．そのため，ページに関係するクエリとして，$P(q'|d) > P(q|d)$ となるようなクエリ $q'$ が存在しないか考慮する必要があると考える．

今後の課題としては，多様化による手法 [3][4][5] との比較や評価実験に使用する正解データの規模の拡大などにより有用性の評価が必要であると考える．

**謝辞** 本研究の一部は，平成 27 年度科研費若手研究 (B)「情報の詳細関係に基づく Web ページの組織化」(課題番号：24700097) によるものである．ここに記して謝意を表すものとします．

### 参考文献

[1] 清水 拓也，土方 嘉徳， 西田 正吾：発見性を考慮した協調フィルタリングアルゴリズム, 電子情報通信学会論文誌, Vol.J91-D, No.3, pp.538-550, 2008.

[2] 丹羽 智史，土肥 拓生，本位田 真一：Folksonomy マイニングに基づく Web ページ推薦システム，情報処理学会論文誌，Vol.47，No.5，pp1382-1392，2006.

[3] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen: Improving recommendation lists through topic diversification,The 14th International Conference on World Wide Web, pp. 22-32, 2005.

[4] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, Samuel Ieong: Diversifying search results, International Conference on Web Search and Data Mining, pp.5-14, 2009.

[5] Yiqun Liu, Ruihua Song, Min Zhang, Zhicheng Dou, Takehiro Yamamoto, Makoto Kato, Hiroaki Ohshima, Ke Zhou: Overview of the NTCIR-11 iMine Task,Proceedings of the 11th NTCIR Conference, pp.8-23, 2014.

[6] 山家 雄介, 中村 聡史, アダム ヤトフト, 田中 克己：ソーシャルブックマークの特性分析とそれに基づく Web 検索結果の再ランキング手法，情報処理学会論文誌データベース，Vol.1, No.1, pp.88-100，2008.

[7] Scott Golder and Bernardo. A. Huberman : Usage patterns of collaborative tagging systems, Journal of Information Science, vol. 32, no. 2, pp.198-208, 2006.

[8] 佃 洸摂, 大島 裕明, 山本 光穂, 岩崎 弘利, 田中 克己:語の認知度と同意語間の関係に基づく意外な情報の発見, WebDB Forum 2012, B1-2, 2012.

[9] Takayuki Yumoto, Ryohei Tada, Manabu Nii, Kunihiro Sato : Finding Rare Web Pages by Relevancy and Atypicality in a Category, IIAI International Conference on Advanced Applied Informatics, pp.284-288, 2013.

[10] 山中 隆広，湯本 高行，新居 学，佐藤 邦弘：カテゴリ推定を用いた希少な Web ページの推薦，DBS160，2014.

[11] Kenneth W. Church，William A. Gale: Inverse Document Frequency (IDF): A Measure of Deviations from Poisson，The Third Workshop on Very Large Corpora, pp.121-130,1995

[12] Christopher D. Manning，Prabhakar Raghavan，Hinrich Schuetze : Introduction to Information Retrieval，Cambridge University Press，pp.157-172，2008.

[13] Kalervo Järvelin, Jaana Kekäläinen : Cumulated Gain-based Evaluation of IR Techniques, ACM Transactions on Information Systems(TOIS), Vol.20, Issue 4, pp.422-446, 2002.