Vol.2014-DBS-160 No.12
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.12
2014/11/18

カテゴリ推定を用いた希少なWebページの推薦

山中 隆広[1]　湯本 高行[1]　新居 学[1]　佐藤 邦弘[1]

概要：近年，情報検索におけるシステム向上により，入力したクエリに対して一般的に知られている情報を入手することは簡単である．しかし，一般的な情報ばかりでは既知の情報ばかりが推薦されてしまう問題がある．本研究では，クエリに関係し，かつ一般的にあまり知られていない情報，つまり希少な情報を推薦することを目的とする．手法として，クエリのカテゴリ推定を行い，ユーザが選択したカテゴリを用いて，クエリに関係するおおまかな文書集合を抽出する．その文書集合で所属確率と非典型度の二つの指標を組み合わせて希少な Web ページの推薦を行う．それぞれの指標の算出には確率モデルを使用する．その結果，クエリに対して同義語のカテゴリでは平均適合率が 0.81 であった．

1. はじめに

近年, Web ページの普及と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．例えば, Google[*1]や Yahoo[*2]などの Web 検索エンジンを用いた情報検索や個人の趣味・嗜好に合わせた推薦手法が存在する [1]．これらを用いることにより，入力されたクエリに関係する一般的に知られている情報，つまり典型的な情報は容易に取得できる．実際に，「iphone」と検索すると最新機種の製品情報やその料金プランなどの典型的な情報が検索結果の上位に出力される．

しかし，典型的な情報ばかりが推薦されてしまうと，同じような情報ばかりで，ユーザの知識は偏ってしまい，ユーザが満足する情報を得ることができない．そのため，他の研究では情報検索や推薦手法によるトピックの多様化に関する研究が行われている [2][3]．また，情報検索をよく使用する上級者にとっては典型的な情報ばかりが推薦されると，既知の情報ばかりが推薦されてしまうため，未知の情報である典型的でない情報を求めるようになる．だが，典型的でない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，かつ典型的でない情報，つまり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 ページを推薦することを目的とする．

本研究では，希少な Web ページを推薦するため，入力したクエリのカテゴリ推定を行う．カテゴリ推定によりユーザがクエリに関係するカテゴリを選択し，そのカテゴリを用いて，クエリに関係する大まかな Web ページの集合 (以下，文書集合) を抽出する．抽出した文書集合に対して所属確率と非典型度の算出を行い，2 つの指標の組合せにより，希少な Web ページの推薦を行う．所属確率は Web ページがどれだけ関係しているかを表す指標であり，非典型度は Web ページがどれだけ典型的でないかを表す指標である．カテゴリ推定と所属確率，非典型度の算出にはソーシャルブックマークサービス (以下，SBM サービス) を用いて算出する．SBM サービス内でブックマークされている Web ページ中の名詞と Web ページを分類するためのタグの関係を学習データとして使用し算出を行う．

2. ソーシャルブックマークサービス

本研究では，カテゴリ推定および所属確率と非典型度の算出には，カテゴリと名詞の関係を表す学習データが必要となる．しかし，Web ページ上に出現する名詞に対して，人が評価したデータを収集することは非常に困難である．そこで，学習データとして SBM サービスにおける Web ページと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

[1] 兵庫県立大学大学院工学研究科
[*1] https://www.google.co.jp/
[*2] http://www.yahoo.co.jp/
[*3] http://b.hatena.ne.jp/
[*4] https://delicious.com/

Vol.2014-DBS-160 No.12
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.12
2014/11/18

コンピュータからでも閲覧することが可能である．また，多くのユーザとブックマークを共有して，コメントやタグなどの情報を付加できることが挙げられる．タグは Web ページに対して付加できる単語やキーワードであり，タグを用いて，Web ページを分類することができる．また，1 人のユーザが複数のタグを付けたり，複数のユーザが 1 つの Web ページにタグを付けることが可能である．

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 [4] の研究がある．そのタグの種類を以下に示す．

(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”)
(3) Web ページの著者，サイトを表すタグ
(4) 共起しているタグが指す範囲を限定するタグ
　(例：(バージョン情報) “3.5”)
(5) Web ページの評価や感想を表すタグ
　(例：“funny”，“これはすごい”)
(6) Web ページにどのように関与しているか表すタグ
　(例：“mycomment”，“mystuff”)
(7) ユーザが Web ページをどう扱うか表すタグ
　(例：“あとで読む”，“jobsearch”)

本研究では，タグの種類の中でも Web ページの内容を表すタグとして，(1) のトピックを表すタグと (2) の種類を表すタグ，(3) の Web ページの著者，サイト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，SBM サービスよりタグと Web ページの関係を大量に取得することができ，Web ページ中の名詞とタグの関係を学習データとして用いることで，カテゴリ推定および所属確率と非典型度をそれぞれ算出する．

3. 先行研究

希少な Web ページを推薦する既存手法として Yumoto ら [5] の研究がある．既存手法の概要を図 1 に示す．この手法では，ユーザがクエリとしてタグに使用されている語を入力し，その語を用いて希少度の算出を行う．希少度は Web ページがクエ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，Web ページがクエリに対してどれだけ関係するかを表す所属確率と，Web ページがクエリに対してどれだけ典型的でないかを表す非典型度の積によって算出できる．所属確率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クエ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手法として，クエリに対するカテゴリ推定を用いる．

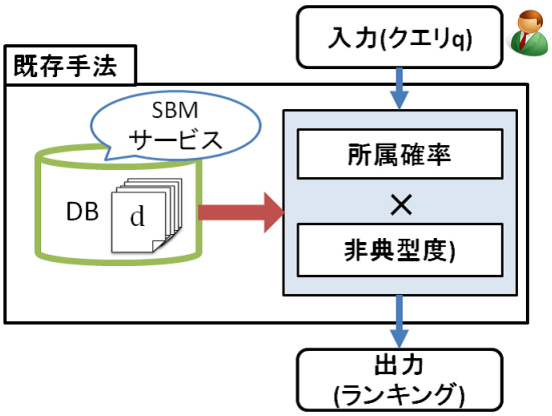

図 1 既存手法

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，所属確率に関して分析を行った．クエリとして，「政治」，「テレビ」の 2 つのクエリを使用し，それぞれのタグが付いた Web ページと付いていない Web ページを 15 件ずつ，計 30 件の Web ページに対して所属確率を算出する．その Web ページを人手によりクエリに関係するかしないかの 2 値に分類し，所属確率の分布を箱ひげ図により図 2 と図 3 に示す．

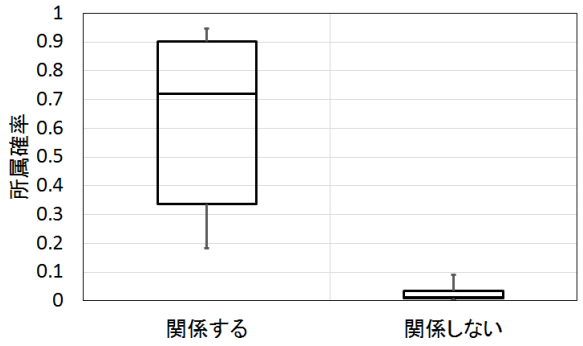

図 2 「政治」の所属確率の分布

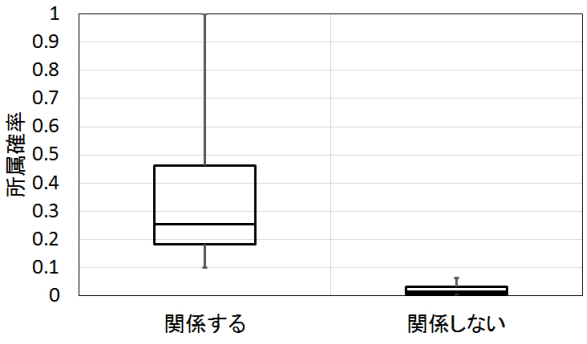

図 3 「テレビ」の所属確率の分布

図 2 と図 3 より，所属確率は関係する Web ページの確率

を 0.1 から 1.0 と広範囲で分布していることがわかる．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 ページは非典型度が大きくなってしまうため，積をとると，関係しない Web ページの希少度が大きくなり，希少な Web ページとして上位に上がってしまうと考えられる．そこで，本研究では，所属確率をフィルタリングとして使用することで希少な Web ページの推薦する手法を提案する．

4. 希少な Web ページの推薦

本研究における提案手法の概要を図 4 に示す．まず，入力したクエリ 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 とクエリ q を用いて，クエリ q に関係する大まかな文書集合 $\mathbf{D}$ を抽出する．選択したカテゴリ c により，抽出した文書集合 $\mathbf{D}$ の所属確率と非典型度を算出し，所属確率の値が 0.1 以下の Web ページを関係の無い Web ページとして除去し，フィルタリングを行う．残った文書集合 $\mathbf{D}$ を非典型度を用いて，降順に並び替えることで，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 ページの推薦を行う．カテゴリ推定と所属確率，非典型度の算出手法を以下に示す．また，使用するデータベースについても示す．

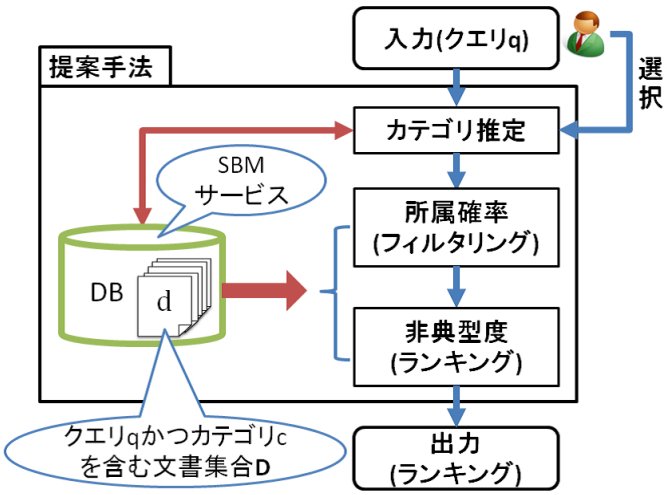

図 4　提案手法

4.1 カテゴリ推定

まず，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 20 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．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 がカテゴリ c に関係する確率 $P(c|q)$ はベイズの定理により (1) 式で表せる．(1) 式より $P(q)$ はクエリ q は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) 式は (2) 式により算出できる．

$$P(c|q) = \frac{P(c)P(q|c)}{P(q)} \tag{1}$$

$$\propto \prod_{i=1}^{n} P(q_i|c) \tag{2}$$

(2) 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，データベースにおいてカテゴリ c のタグを含む BM 数と，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 数の比をとり，(3) 式のように算出を行う．

$$P(q_i|c) = \frac{q_i \text{ を含みカテゴリ } c \text{ を含む BM 数}}{\text{カテゴリ } c \text{ を含む BM 数}} \tag{3}$$

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クエリ q に関係するカテゴリとして不適切であるとみなし，除去する．

4.2 所属確率

Web ページ d がカテゴリ c に関係する確率を所属確率 $P(c|d)$ とする．$P(c|d)$ は，Web ページ d がカテゴリ c に関係しない確率 $P(\bar{c}|d)$ との和が 1 となることが自明であるため (4) 式で表せる．このとき，Web ページ d の本文に記載された語 w_i がすべてカテゴリ c に関係しない場合，Web ページ d はカテゴリ c に関係しないとする．すると，(4) 式は (5) 式となる．また，語 w_i がカテゴリ c に関係しない確率 $P(\bar{c}|w_i)$ は関係する確率 $P(c|w_i)$ との和が 1 となることが自明であるため (5) 式は (6) 式で算出できる．

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

$$= 1 - \prod_{i=1}^{n} P(\bar{c}|w_i) \tag{5}$$

$$= 1 - \prod_{i=1}^{n} \{1 - P(c|w_i)\} \tag{6}$$

(6) 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 のタグを含むブックマークが多いほど，w_i はカテゴリ c に関係すると考える．つまり，データベースにおいて，本文中に w_i を含む

Web ページの BM 数と，本文中に w_i を含み，かつカテゴリ c のタグを含む BM 数の比をとり，(7) 式のように算出する．

$$P(c|w_i) = \frac{w_i \text{ を含みカテゴリ } c \text{ を含む BM 数}}{\text{名詞 } w_i \text{ を含む Web ページの BM 数}} \quad (7)$$

ここで，所属確率の算出には (6) 式のように Web ページ 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 Web ページ 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，$TF\text{-}RIDF$[6] 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 (8) 式を用いる．(8) 式において，TF は文書中の語 w_i の出現頻度，DF はデータベースにおける w_i が出現した Web ページ数，$ALLURL$ はデータベースにおける全 Web ページ数，$ALLTF$ はデータベース内の全 Web ページにおける w_i の出現数の総和を表す．

$$TF\text{-}RIDF = TF \times \left\{ log_2\left(\frac{ALLURL}{DF}\right) + log_2\left(1 - exp\left(-\frac{ALLTF}{ALLURL}\right)\right)\right\} \quad (8)$$

$TF\text{-}RIDF$ は，文書中の語 w_i の出現頻度である TF 値と，Web ページ集合全体で w_i が出現する Web ページ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，$RIDF$ は多くの Web ページで使用される語の値は小さくなり，少ない Web ページで出現頻度が多い語の値は大きくなる．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．

4.3 非典型度

カテゴリ c において Web ページ d が典型的でない確率を非典型度 $P(\bar{d}|c)$ とする．このとき，Web ページ d のすべての語 w_i がカテゴリ c 内で出現しない場合，Web ページ d はカテゴリ c 内で典型的でないとすると，非典型度は (9) 式により算出できる．また，語 w_i がカテゴリ c 内で出現しない確率 $P(\bar{w_i}|c)$ は，出現する確率 $P(w_i|c)$ との和が 1 となることが自明であるため (9) 式は (10) 式で算出できる．

$$P(\bar{d}|c) = \prod_{i=1}^{n} P(\bar{w_i}|c) \quad (9)$$

$$= \prod_{i=1}^{n} \{1 - P(w_i|c)\} \quad (10)$$

$$P(w_i|c) = \frac{w_i \text{ を含みカテゴリ } c \text{ を含む BM 数}}{\text{カテゴリ } c \text{ を含む BM 数}} \quad (11)$$

(10) 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，データベースにおいてカテゴリ c のタグを含む BM 数と，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 数の比をとり，(11) 式のように算出を行う．

ここで，非典型度の算出式 (10) 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，所属確率と同様に $TF\text{-}RIDF$ 値として (8) 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．

4.4 データベース

カテゴリ推定や所属確率，非典型度の算出に使用するデータベースとして，はてなブックマークよりブックマーク情報 (ユーザ ID，URL，タグ，ブックマーク日時，コメント) を収集したデータを用いる．さらに，URL の Web ページより HTML ファイルを取得する．データベースには，表 1 に示すようにテーブルとフィールドを作成する．

取得した HTML ファイルの本文抽出には，ExtractContent[*5]を使用する．ExtractContent は HTML タグを用いて，本文テキスト以外の広告などを取り除いて本文抽出を行うモジュールである．また，本文中の名詞抽出には MeCab[*6]を用いる．MeCab はオープンソースの形態素解析エンジンで，入力した文章を形態素に分割し，「名詞」や「動詞」などの品詞を抽出することができる．MeCab の辞書に 2014 年 6 月 20 日時点の Wikipedia のタイトルリスト[*7]を追加し，形態素解析を行うことで Web ページごとに名詞の抽出を行い，データベースを作成する．

5. 評価実験

評価実験として，カテゴリ推定における評価と $TF\text{-}RIDF$ による主要語の評価，希少な Web ページの推薦における評価の 3 つの実験を行った．カテゴリ推定と希少な Web ページの推薦における評価実験には，同じクエリ q を入力として使用した．使用したクエリを表 2 に示す．また，データベースには 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．作成したデータベースの規模を表 3 に示す．

5.1 カテゴリ推定の評価

入力されたクエリ q に対して，カテゴリ候補を抽出し，(2) 式によりカテゴリ推定を行う．(2) 式により算出された

*5 http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html
*6 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
*7 http://dumps.wikimedia.org/jawiki/latest/

表 1 データベースにおける，テーブル，フィールドの一覧

テーブル名	フィールド名	フィールドの説明
url	id	URL の ID
	title	URL のタイトル
	lemma	id に対応する URL
user	id	ユーザアカウントの ID
	lemma	id に対応するアカウント名
word	id	名詞，タグの ID
	lemma	id に対応する名詞，タグ
bm	id	ブックマークの ID
	user_id	ユーザアカウントの ID
	url_id	URL の ID
bm_tag	bm_id	ブックマークの ID
	tag_id	bm_id のブックマークに付けられたタグの ID
url_noun	url_id	URL の ID
	noun_id	url_id の Web ページ本文に出現する名詞
	count	noun_id の名詞が url_id のページに出現する回数
df	noun_id	名詞の id
	df_count	noun_id の名詞が出現する url 数
	all_tf	noun_id の名詞が全 url に出現する回数

表 2 評価実験に使用したクエリ

Google	NHK
youtube	アマゾン
Twitter	サッカー
ユニクロ	進撃の巨人
東京	地震

表 3 データベースの規模

データセット情報	件数・種類数
総ブックマーク数	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 4 クエリごとのカテゴリ推定の評価結果

クエリ	AP	個数
Google	0.92	3
NHK	0.85	4
アマゾン	1.00	2
サッカー	1.00	5
ユニクロ	1.00	3
進撃の巨人	1.00	6
東京	0.64	2
Twitter	1.00	2
youtube	1.00	3
地震	1.00	3
平均	0.94	3

値によりカテゴリ候補を降順にランキングしたとき，上位にクエリ q に関係するカテゴリ c が推定されているか評価を行った．クエリに関係するカテゴリとしては，上位語，下位語，同義語を正解とし，評価には平均適合率 (Average Precision) を用いて，上位 10 個のカテゴリを用いて評価を行った．平均適合率 AP は (12) 式により算出する．

$$AP = \frac{1}{n}\sum_{k=1}^{10} precision(k)r(k) \tag{12}$$

(12) 式において，n は上位 10 個での正解の数を示し，$precision(k)$ は順位 k までの適合率を示す．$r(k)$ は順位 k のカテゴリが関係するカテゴリなら 1，関係しないなら 0 の 2 値の値をとる．つまり，正解のカテゴリが出現した順位すべての適合率の平均値を算出する．平均適合率は 0 から 1 までの範囲で算出され，上位に正解が多いほど 1 に近い値となる．その結果と関係するカテゴリの数を表 4 に示す．

表 4 の結果より，どのクエリにおいても平均適合率が高く，上位に関係するカテゴリが出力されていることがわかる．また，関係するカテゴリの数として，平均 3 個のカテゴリが推定されている．このことから，提案したカテゴリ推定によりクエリに関係するカテゴリの推定が行えていることがわかる．しかし，クエリが「東京」では他と比べると平均適合率が低いことがわかる．これは，他のクエリでは上位 5 個の内に関係するカテゴリが多く出力されていたので評価が高くなったが，「東京」では関係するカテゴリの 2 つのうち 1 つは 1 番目に出力されたが，もう 1 つは 7 番目に出力され，2 番目から 6 番目では「グルメ」や「旅行」などの関係はありそうだが，上位語，下位語，同義語としては当てはまらないカテゴリが出力されたため平均適合率が低くなった．

5.2 主要語の評価

データベースより 10 件の URL をランダムに抽出し，それぞれの URL の Web ページ中の名詞の $TF\text{-}RIDF$ 値を

Vol.2014-DBS-160 No.12
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.12
2014/11/18

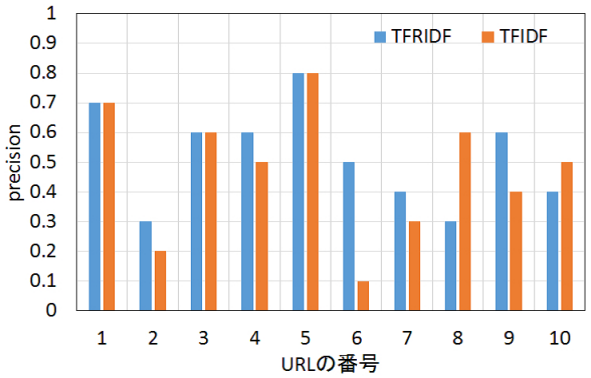

図 5 主要語の評価結果

表 5 主要語の評価結果の平均

	TFRIDF	TFIDF
平均	0.52	0.47

(8) 式により算出する．そのときの上位 10 個にどれだけ重要と思われる語が含まれているか，適合率を算出し評価を行った．正解データとして，1 名の被験者がそれぞれの Web ページを閲覧し，形態素解析を行った名詞一覧から選んだ，重要と思われる名詞 10 個を使用する．また，比較として *TF-IDF* を用いた場合の評価も同様に行った．*TF-IDF* の算出には (13) 式を用いた．*TF-RIDF* と *TF-IDF* の評価結果を図 5 と表 5 に示す．なお，抽出した Web ページにそれぞれ 1 から 10 の番号を振り分けた．

$$TF\text{-}IDF = TF \times log_{10}\left(\frac{ALLURL}{DF}\right) \qquad (13)$$

図 5 より，10 件の Web ページの内 5 件において *TF-IDF* と比べて，*TF-RIDF* を用いた場合に主要語を多く抽出できていることがわかる．また，表 5 の平均値を比較すると，*TF-RIDF* が 0.52，*TF-IDF* が 0.47 となり，*TF-IDF* を用いた場合よりも *TF-RIDF* を用いた方が主要語を多く抽出していることが分かる．このことから，*TF-IDF* よりも *TF-RIDF* による主要語の抽出手法が優れていることがわかる．

5.3 希少な Web ページの評価

次に，入力したクエリのカテゴリ推定を行い，推定されたカテゴリを用いて，希少な Web ページの推薦が行われるか評価した．

5.3.1 上位に出力されたカテゴリによる評価

5.1 節の結果から最も上位に出力されたクエリ q に関係するカテゴリ c を用いて，クエリごとに希少な Web ページの推薦を行った．入力したクエリ q と選択したカテゴリ c より文書集合 $\mathbf{D}$ を抽出し，$\mathbf{D}$ の各 Web ページ d に対して (6) 式により所属確率を算出する．所属確率が 0.1 以下の Web ページを除去し，残った各 Web ページ d の非典型度を算出する．算出した非典型度の値に応じて降順にランキングした．このとき，クエリ q に対して希少な Web ページが上位に推薦されるか評価を行った．評価としては，カテゴリ推定における評価と同様に，平均適合率を用いて上位 10 件の Web ページを使用した．正解データとして，1 名の被験者が上位 10 件の Web ページを閲覧し，希少な Web ページであれば 1，そうでなければ 0 とした．また，比較として先行研究の手法である所属確率と非典型度の積の値により降順にランキングした場合の評価も同様に行った．その結果を表 6 と図 6 のグラフに示す．なお，本研究における所属確率によるフィルタリングを用いた結果を「提案」，先行研究における所属確率と非典型度の積を用いた結果を「既存」として表している．また，提案手法において出力された Web ページ数を表 7 に示す．

表 6 希少な Web ページ推薦の評価結果

クエリ	カテゴリ	提案	既存
Google	google	0.76	0.77
NHK	nhk	1.00	0.71
アマゾン	amazon	0.88	0.92
サッカー	サッカー	0.79	0.98
ユニクロ	ユニクロ	1.00	1.00
進撃の巨人	進撃の巨人	1.00	0.86
東京	東京	0.96	0.88
Twitter	Twitter	0.99	0.93
youtube	YouTube	0.00	0.46
地震	地震	0.73	0.72
平均		0.81	0.82

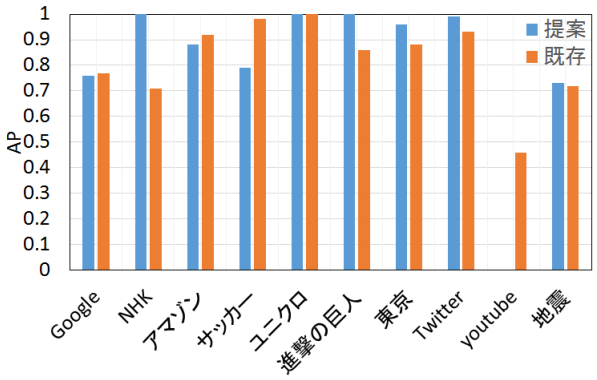

図 6 評価結果の比較

表 6 と図 6 より，既存手法の平均適合率に比べて提案手法の平均適合率が向上しているクエリもあれば，提案手法の平均適合率が低下しているクエリもあり，10 個のクエリの内 5 個は既存手法に比べて提案手法が向上した．また，表 6 の平均の値を見ると，既存手法では 0.82，提案手法では 0.81 と既存手法と提案手法でどちらも平均適合率が高く，あまり差がないことがわかる．このことから，提案手法は既存手法と比べて平均適合率が高い場合もあるが低い場合もあり，それぞれ同等の性能を持つことがわかる．しかし，表 7 を見ると，提案手法の「NHK」，「ユニクロ」，

表 7 提案手法の出力 Web ページ数

クエリ	カテゴリ	Web ページ数
Google	google	29
NHK	nhk	2
アマゾン	amazon	27
サッカー	サッカー	30
ユニクロ	ユニクロ	5
進撃の巨人	進撃の巨人	4
東京	東京	25
Twitter	twitter	29
youtube	YouTube	1
地震	地震	24

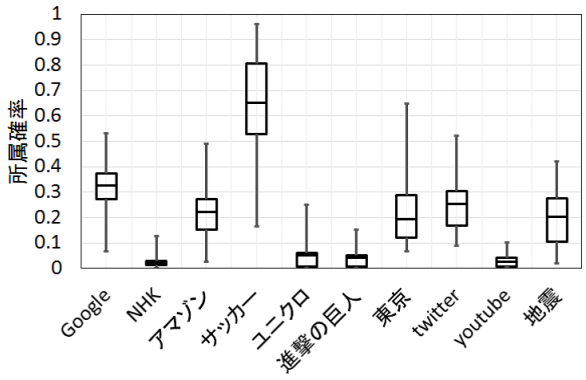

図 7 クエリごとの所属確率の分布

表 8 カテゴリごとの BM 数

カテゴリ	BM 数
google	42065
nhk	1869
amazon	10497
サッカー	7933
ユニクロ	612
進撃の巨人	196
東京	15693
twitter	37941
YouTube	240
地震	8433

「進撃の巨人」,「youtube」のクエリにおいて，出力された Web ページの数が 10 件以下で少ないことがわかる．これは，カテゴリによっては，Web ページの所属確率の値が小さな範囲で算出されているため，所属確率のフィルタリングにより多くの Web ページが除去されたと考えられた．そこで，それぞれのクエリにおける所属確率の分布を箱ひげ図により図 7 に示す．また，データベースに含まれるカテゴリごとの BM 数を表 8 に示す．

図 7 より，出力された Web ページ数の多いクエリでは，所属確率が広く分布しており，0.1 から 1.0 の範囲で分布しているクエリもあれば，0.0 から 0.6 の範囲で分布しているクエリもあることがわかる．しかし，出力された Web ページ数の少ないクエリでは，0.0 から 0.2 の小さな範囲で分布しており，文書集合 **D** としてはクエリとカテゴリに関係する Web ページが多く抽出されていたが，Web ページの多くが 0.1 以下で算出されていることがわかる．その原因として，データベースに含まれる BM 数が少ないカテゴリでは，所属確率の算出における (7) 式の分子が小さくなり，全体としても所属確率が小さくなりやすかったため，分布として小さな範囲で算出されたと考えられる．実際に，表 8 より，出力された Web ページ数の少ないカテゴリでは，2000 以下の BM 数しかなく，他のカテゴリでは，7000 以上の BM 数があることがわかる．そのため，所属確率ではカテゴリの BM 数の違いにより，フィルタリングを行うしきい値を考慮する必要がある．または，所属確率ではなく別の手法を用いて，クエリまたはカテゴリに関係する文書の抽出を行う方法も考えられる．

表 9 上位語のカテゴリによる評価

クエリ	カテゴリ	AP
進撃の巨人	漫画	0.57
NHK	テレビ	0.61
youtube	動画	0.11
サッカー	スポーツ	0.99
地震	災害	0.71
平均		0.60

5.3.2 上位語による評価

5.3.1 項ではカテゴリとして，最も上位に出力されたカテゴリを使用しているため，実験ではクエリに対して同義語のカテゴリのみを使用した．そのため，同義語では平均適合率として高い値が出力され，良い結果が得られた．そこで，クエリに対して広い意味で考えたときのカテゴリとして，上位語を用いた場合の評価を行った．カテゴリ推定により，上位に出力された上位語のカテゴリ c が推定されたクエリ q のみを用いて，上位語のカテゴリにより希少な Web ページの推薦を行ったときの評価を行った．今回のクエリでは，カテゴリ推定の結果より 5 つのクエリで上位語のカテゴリが上位に出力された．評価には，5.3.1 項と同様に平均適合率により上位 10 件の Web ページの評価を行った．その結果を表 9 に示す．また，上位語と同義語の結果を比較するためそれぞれの結果を図 8 のグラフに示す．

表 9 を見ると，平均適合率の平均として 0.60 と低くなっている．また，図 8 を見ると，クエリの「NHK」,「地震」,「進撃の巨人」において同義語のカテゴリを用いた場合よりも上位語のカテゴリを用いた場合の方が，評価が低くなっていることがわかる．これは，クエリに関係するカテゴリとして，上位語のカテゴリでは広い意味で文書集合 **D** を抽出してしまうため，カテゴリに関係するがクエリに関係のない Web ページが抽出され，上位に出力されてしまったため，平均適合率が低くなったと考える．そのため，文書集合 **D** としてクエリにより深く関係する Web ページを

Vol.2014-DBS-160 No.12
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.12
2014/11/18

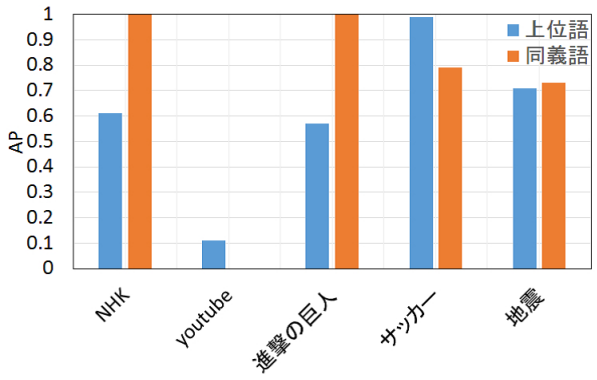

図 8 同義語との比較

抽出するための手法が必要であると考える．

6. おわりに

本研究では，カテゴリ推定を用いて，クエリ q に関係するカテゴリ c をユーザが選択し，選択したカテゴリ c を用いて希少な Web ページの推薦を行う手法を提案した．また，提案手法のカテゴリ推定の評価と *TF-RIDF* による主要語の評価，希少な Web ページの推薦における評価を行った．

その結果，カテゴリ推定においては平均適合率の値は高く，上位にクエリに関係するカテゴリが推定されていることが確認できた．また，*TF-RIDF* を用いた主要語の評価では，*TF-IDF* に比べて適合率が高く，重要と思われる語の抽出数が多いことが確認できた．

次に，希少な Web ページの推定において，最も上位に推定されたカテゴリを用いて提案手法と先行研究の既存手法の評価を行った．その結果，既存手法と提案手法の評価として 10 個のクエリの内 5 個のクエリは提案手法により向上し，平均が 0.81 と高い結果が得られた．しかし，クエリによっては文書集合の関係する Web ページが多く除去され，出力される Web ページ数が極端に少なくなった．これは，カテゴリによってはデータベースに含まれる BM 数が少なく，所属確率の値が小さい範囲で出力されたため，BM 数の少ないカテゴリでは多くの Web ページが除去されてしまったためであると考えられる．

最後に，上位語のカテゴリを用いた場合の評価を行った．その結果，平均適合率の値は平均が 0.60 となった．これは，上位語では同義語よりクエリに関係するカテゴリとして広い意味を持つため，上位語のカテゴリではカテゴリに関係する Web ページだがクエリに関係しない Web ページを文書集合として抽出され，上位にクエリに関係のない Web ページが推薦されるため平均適合率が低くなった．そのため，文書集合 **D** としてクエリ q に関係のある Web ページを多く抽出できるように改善する必要があると考える．

今後の課題として，所属確率のフィルタリングとカテゴリ c の上位語の問題を改善するため，文書集合 **D** としてクエリ q と Web ページ d の関係性を表す指標を用いて，関係性の高い Web ページを抽出する手法を検討する．これにより，所属確率によるフィルタリングを行うのでは無く，文書集合によりクエリに関係する Web ページのみを抽出する．抽出した文書集合に非典型度を用いることで，希少な Web ページの推薦を行う．

謝辞 本研究の一部は，平成 26 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] 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.

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

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

[5] T. Yumoto, R. Tada, M. Nii, K. Sato : Finding Rare Web Pages by Relevancy and Atypicality in a Category, IIAI International Conference on Advanced Applied Informatics, pp.284-288, 2013.

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