情報処理学会論文誌　データベース　Vol.6　No.1　29–40 (Jan. 2013)

階層的オートタギング技術とその応用

西田 京介[1,a)]　星出 高秀[1]　藤村 考[1,†1]　内山 匡[1]

受付日 2012年6月20日, 採録日 2012年10月3日

概要：本論文では，カテゴリ・主題・キーワードという異なる意味抽象度のタグを文書へ自動的に付与する階層的オートタギング技術について論じる．本技術は，タグ未付与のあらかじめカテゴリ分けされた入力文書集合から，各カテゴリの主な話題を表す主題語を自動的に発見し，各入力文書に対して，文書分類によるカテゴリ・主題タグの付与と，単語抽出によるキーワードタグの付与を行う．本論文では，Q&A コミュニティ「教えて！ goo」と Q&A 検索サービス「QA.ON/OFF」への本技術の導入事例を紹介する．そして新たに，文書構造を考慮したキーワードタグ抽出法の提案と，主題語の抽出精度の評価，応用事例の利用者を対象に行った被験者実験によるタギング精度の評価を実施し，各技術要素が従来手法に比べて優れた性能を示したことを報告する．

キーワード：タギング，文書分類，キーワード抽出

Hierarchical Auto-tagging and Its Application

Kyosuke Nishida[1,a)]　Takahide Hoshide[1]　Ko Fujimura[1,†1]　Tadasu Uchiyama[1]

Received: June 20, 2012, Accepted: October 3, 2012

Abstract: We present our hierarchical auto-tagging method that assigns three different levels of tags to documents: category, theme, and keyword. Our method consists of a classification method for assigning category and theme tags, a new proposed keyword extraction method that considers the structure of documents, and a method for selecting theme tag candidates from each category. We introduce its applications to a Q&A community service and a Q&A search service. We also report the new experiment concerning the selection of theme tag candidates and the user evaluation on the tagging accuracy of our method.

Keywords: tagging, text classification, keyword extraction

1. はじめに

Web 上で億単位の文書を扱うようになった現代においても，文書整理の主たる道具はカテゴリ分類である．たとえば，新たな集合知レポジトリとして発展し続けている Q&A コミュニティでも，「Yahoo! Answers」「教えて！ goo」などの代表的なコミュニティでは，カテゴリによる文書整理（質問者が質問を投稿する際に，数百のカテゴリの中から適切なカテゴリを 1 つ選択する）を採用している．しかし，これらのコミュニティにおいて文書が適切に整理されているとはいい難い．たとえば，『クマのぬいぐるみの洗い方』という質問文書は「掃除・洗濯」と「おもちゃ」のどちらのカテゴリに投稿すべきであろうか？　現状のカテゴリによる文書整理には，カテゴリ体系が非排他的・非網羅的であることなどの問題がある．

これらのカテゴリがかかえる問題を解決するため，西田らは，階層的オートタギング技術：TagHats（Hierarchical Auto-Tagging System に由来）を提案した [1]．本技術は，あらかじめカテゴリ分けされた（タグ未付与の）入力文書集合から，各カテゴリの主な話題を表す主題語を自動的に発見し，各入力文書に対して，文書分類によるカテゴリ・主題タグの付与と，単語抽出によるキーワードタグの付与を行う．タグに基づく文書整理では，カテゴリのように個

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation, Yokosuka, Kanagawa 239–0847, Japan
[†1] 現在，大妻女子大学社会情報学部
Presently with School of Social Information Studies, Otsuma Women's University
[a)] nishida.kyosuke@lab.ntt.co.jp

数や抽象度の制約がないため，複数の観点による整理や新奇概念の整理が容易に実現できる．さらに，抽象度が異なるタグによって，異なる検索意図を持った多数のユーザのための文書整理が実現できる．なお，本技術は，ユーザによるタギング体系の構築や，ユーザによる各文書へのタギングをいっさい必要とせず，Q&A 文書やニュース記事など様々な文書メディアで汎用的に利用できる点が優れている．

本論文では，階層的オートタギング技術を Q&A コミュニティ「教えて！ goo」と Q&A 検索サービス「QA.ON/OFF」へ導入した事例について紹介する．さらに，キーワードタグ抽出法に関する新提案を含む技術内容，各技術要素に関する評価実験について論じる．

本論文の貢献：先行文献 [1] からの本論文の貢献は，(1) 階層的オートタギング技術の実サービス導入事例と得られた知見の紹介，(2) 文書構造を考慮したキーワード抽出手法の提案，(3) タギング精度に関する追加評価実験の実施，の 3 点にある．なお，貢献 2（3.4.2 項）と貢献 3 の一部（4.1 節，4.3 節）は文献 [2], [3] に報告済みのものを含む．

本論文の構成：2 章でカテゴリによる文書整理の問題点について論じる．3 章で技術内容について説明する．4 章にタギング精度に関する評価実験の結果を示す．5 章で，本技術の応用事例を紹介する．また，応用事例の利用者を対象に行ったタギング精度の評価実験の結果をあわせて示す．6 章に関連研究を示し，7 章に結論を示す．

2. カテゴリによる文書整理の問題点

本章では，2 つの Q&A コミュニティ「Yahoo! Answers」（以下，YA と略す）と「教えて！ goo」（以下，OG と略す）の 2010 年 3 月 31 日時点のデータを例に，カテゴリによる文書整理について解析した結果を示す．

2.1 抽象的なカテゴリ

YA と OG は，カテゴリのツリーによって文書を整理している．ここで，**図 1** と**表 1** に示す YA と OG の各投稿先カテゴリの質問文書数の統計から明らかなように，各カテゴリに投稿された文書数は偏っている．YA では総文書数の 27.1% が 948 カテゴリ中の 10 カテゴリに集中し，OG でも同様に 22.8% の文書が 389 カテゴリ中の 10 カテゴリに集中している．YA の「Singles & Dating」(5,675,514 文書；6.37%）や OG の「恋愛相談」(230,688 文書；4.75%）のように多数の文書が属するカテゴリは，回答者や閲覧者が望む文書を発見しやすいように，より具体的なサブカテゴリに分割すべきである．

2.2 MECE でないカテゴリ体系

文書は複数の話題を含む場合があるため，Q&A コミュニティのように 1 つの文書が複数のカテゴリに所属することを許さないシステムでは適切な文書整理を行いにくい．また，文書が単一の話題しか含まないとしても，カテゴリ体系が MECE（mutually exclusive and collectively exhaustive）でないため，最も適切なカテゴリの発見は難しい．

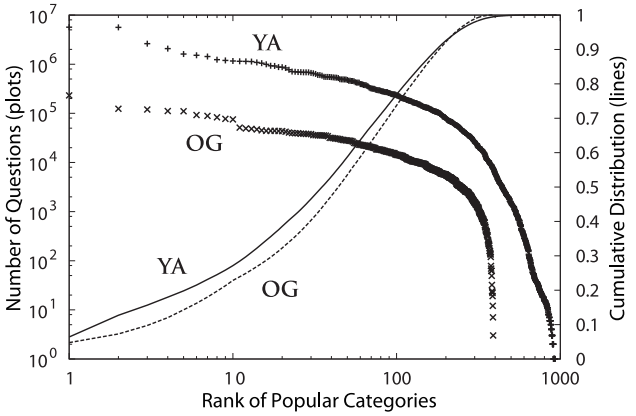

図 1「Yahoo! Answers」（YA）と「教えて！ goo」（OG）の各カテゴリの質問数と，その累積分布

Fig. 1 Numbers of questions for each category in “Yahoo! Answers” (YA) and “Oshiete! goo” (OG) and their cumulative distributions.

表 1 各投稿先カテゴリの質問数の統計．“Qu.” は四分位点を，“N” はカテゴリ数を表す

Table 1 Statistics of number of questions for each category in YA and OG.

	Min.	1st Qu.	Median	3rd Qu.	Max.
YA	0	27.5	2,046	54,122.5	5,675,514
OG	3	663	5,746	77,138	230,688
	N	Mean	Std Dev.	Sum.	
YA	948	93,950.4	336,605.3	89,064,955	
OG	389	12,482.6	20,811.4	4,855,723	

2.2.1 相互排他性（mutually exclusive）

まず，質問投稿時に単一のカテゴリを選択することの難しさの例を示す．**表 2** は，YA における，「clean（掃除）」と「stuffed animals（ぬいぐるみ）」の両方の語がタイトルに含まれる 52 の質問が投稿された 9 つの投稿カテゴリである．この例では，「Cleaning & Laundry」カテゴリが最も多く選択されていたが（35 個の質問が投稿された），「Toys」カテゴリも文書内容に合致しており，投稿先のカテゴリとして適切である．このような事例は，2 つのカテゴリが相互に排他でないことが原因で生じる．

さらに我々は，カテゴリ選択の難しさを検証する行動実験を OG の質問文書を利用して行った．実験では，2 人の被験者（28 歳，25 歳の日本人男性）に，ランダムに選択した 200 個の文書を，実際の質問投稿者と同様の条件で分類させた．**表 3** に示すように，被験者と実際の質問者の一致率はどちらも 5 割に満たず，23.5% の質問文書では 2 人の被験者と実際の質問者がすべて異なるカテゴリを選択

表 2　YA における，“clean（掃除）” と “stuffed animals（ぬいぐるみ）” が両方含まれたタイトルの投稿カテゴリ．N は Yahoo! Answers API で抽出された質問数

Table 2　Submission categories of questions whose titles contain both “clean” and “stuffed animals” in YA. N means the number of the questions.

Categories	N	Example of Question Title
Cleaning & Laundry	35	How do I clean stuffed animals?
Toys	7	What can I use to clean stuffed animals?
Newborn & Baby	3	How do you clean baby toys (stuffed animals)???
Dogs	2	any suggestions on cleaning stuffed animals...
Do It Yourself (DIY)	1	how can I clean my stuffed animals? they have stains?
Hunting	1	how do you clean stuffed animals (deer, ram)?
Adolescent	1	I'm cleaning out my closet: Old stuffed animals and dolls?
Other - Pets	1	how to clean stuffed animals?
Toddler & Preschooler	1	Best way to clean my daughter's dolls and stuffed animals

表 3　2 人の被験者（s_1 と s_2）と，実際の質問投稿者（a）の投稿先カテゴリの一致率

Table 3　Agreement rates of submission categories among the actual asker, a, and the two subjects, s_1 and s_2.

Users	Rates
$s_1 = a$	0.475
$s_2 = a$	0.410
$s_1 = s_2$	0.470
$s_1 = s_2 = a$	0.295
$s_1 = a$ or $s_2 = a$ or $s_1 = s_2$	0.765

表 4　YA と OG の各「その他」カテゴリの質問数の統計

Table 4　Statistics of number of questions for each “other” category in YA and OG.

	Min.	1st Qu.	Median	3rd Qu.	Max.
YA	8	7,093	53,823	205,364	919,946
OG	386	4,976	11,854	29,123	123,444
	N	Mean	Std Dev.	Sum.	
YA	83	150,901.5	210,124.6	12,524,823	
OG	53	19,941.9	23,343.2	1,136,686	

した．

2.2.2　完全網羅性（collectively exhaustive）

YA と OG は，「その他」のカテゴリを設けることで擬似的に完全網羅性を実現している．**表 4** に示す，カテゴリ名に「Other」「その他」が含まれる投稿先カテゴリの質問文書数の統計から，OG では 23.4% の文書が，YA では 14.1% の文書が「その他」のカテゴリに投稿されていることが分かる．しかし，回答者や閲覧者にとっては「その他」のカテゴリにどのような内容の質問文書が投稿されているか不明であり，有益な文書整理になっていない．

2.3　本解析のまとめ

代表的な Q&A コミュニティである「Yahoo! Answers」と「教えて！ goo」は以下の特性を持つことが分かった：(i) 抽象的すぎるカテゴリが存在し，各カテゴリの文書数が強く偏っている．(ii) カテゴリ体系が MECE でない．このため，(1) 質問者が，質問文書の投稿時に適切な（1 つの）カテゴリを選択できない，そして (2) 回答者と閲覧者が，カテゴリに投稿されている質問文書の具体的な内容を把握できないという問題がある．これらの問題は，Q&A コミュニティに特有のものではなく，他メディアでも同様に生じていると考える．

ここで，付与個数や内容・抽象度に関する制約がないタグを利用すれば，複数の観点で，カテゴリよりも具体的な文書整理が可能となり，カテゴリがかかえる問題を解決できる．ソーシャルタギング [4], [5] ではすでに蓄積された莫大な文書の整理が難しいことと，異なる知識レベルと検索意図を持った多数のユーザの存在を考慮すれば，我々は異なる意味抽象度のタグを自動的に文書に付与可能な階層的オートタギングが有用であると考える．

3.　階層的オートタギング技術：TagHats

階層的オートタギング技術（TagHats）について，形態素解析方法，主題タグ候補の選択方法，そして，カテゴリ・主題・キーワードの 3 層のタギング方法について示す．

3.1　概要

TagHats は，あらかじめカテゴリ分けされた，タグが付与されていない文書集合を基に学習を行い，各文書に対してカテゴリ・主題・キーワードという 3 つの階層のタグを付与する（**図 2**）．

ここで，総称的な（上位概念の）タグに利用可能な語句は文書中には出現しない場合が多いため，TagHats は文書分類によってカテゴリ・主題タグを付与する．しかし，文書分類ではあらかじめ分類するタグ候補を選定する必要があるため，直近に出現した語句や具体的な語句を扱うことが難しい．そこで，TagHats では入力文書からのキーワード抽出についても行い，具体的なタグを付与する．

なお，主題タグは文書分類，キーワードタグは単語抽出，

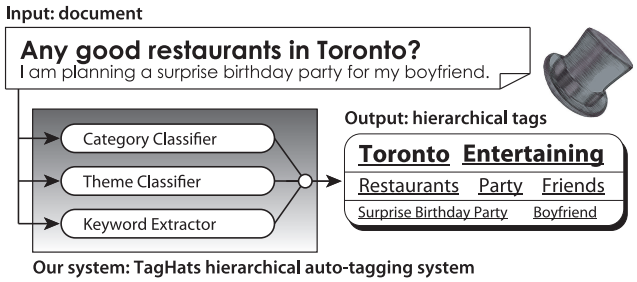

図 2　階層的オートタギングシステム TagHats の概念図
Fig. 2　Concept image of our hierarchical auto-tagging system, TagHats.

という別の手法に基づいて付与するため，これらの間の階層関係は理論的には保証されない．主題タグ候補語には特定のカテゴリの文書タイトルに特徴的に出現する語のみを用意し（3.3 節），キーワードタグには様々なカテゴリに属する複数個の文書間にリンクを生成する具体的な語を文書から抽出する（3.4.2 項）ことで，全体として様々な検索意図に対応できる多様なタグを付与する．

3.2　形態素解析

TagHats は，入力として与えられた文書を以下の処理により語句の集合に分割する．

(1) 日本語 Wikipedia の見出し・転送語の形態素解析辞書への追加
(2) 各文書について形態素解析を実施
(3) 得られた語句集合について：
　(a) 連続する名詞語（カタカナ，数値，アルファベット，接尾辞など含む）の連結
　(b) 名詞以外の語の削除

上記の処理により，たとえば『お酒などによる逆流性食道炎で咽頭がんや食道がんになるリスクはありますか？』という文章から「お酒」「逆流性食道炎」「咽頭＋がん」「食道がん」「リスク」という語句集合を得る（‘+’ 記号は名詞の連結を表す）．本研究では，上記のように取得した語句を，(i) 主題タグの候補語（3.3 節），(ii) 文書分類（カテゴリ・主題タグの付与）の素性（3.4.1 項），(iii) キーワードタグの候補語（3.4.2 項），として共通利用する．

上記の処理 (1)〜(3)(b) は，有益なタグの候補語の選出と，文書分類（カテゴリ・主題タグの付与）の精度向上のために実施した．たとえば「逆流性食道炎」という単語は，「逆流」「性」「食道」「炎」の 4 単語ではなく，「逆流性食道炎」のまま 1 単語として扱う方がタグとして役立つため，処理 (1) と処理 (3)(a) を実施した．また，1 つのまとまりとして扱うべき語が細かい単語に分割されることによる分類精度への悪影響が従来研究にて指摘されており [6]，処理 (1) と処理 (3)(a) は分類精度の向上にも役立つ．また，形容詞・動詞にはタグとして有益なものも含まれるが，文書分類の精度に悪影響が出るため，提案手法では語句集合から削除した（処理 (3)(b)）．数値については，名詞＋数値の語から「エクセル 2007」などの具体的な絞り込みに役立つタグを得られるため，連結対象の語に含めた．なお，TagHats は形態素解析器に JTAG [7] を用いるが，他の形態素解析器でも上記処理は実施できる．

3.3　主題タグ候補の選択

文書分類に先だって，タグ（分類先）の候補を決定する必要がある．まず，カテゴリタグの候補については，与えられたカテゴリ名集合をそのまま使用する．次に，主題タグの候補を，文書の内容を理解するために重要な語句が含まれている「文書タイトル」に着目して，各カテゴリから自動的に選択する．ここでは，フィッシャの正確確率検定（片側検定，有意水準 α）[8] を用いて，ある語句 w が，特定のカテゴリ c の文書タイトル中に，他のカテゴリ $\bar{c}$ の全文中に比べて統計的に有意に多く出現するとき（$p < \alpha$ のとき），w をカテゴリ c の主題語として選択する．ただし，カテゴリ c のタイトル中出現頻度 $T_c(w)$ が N_{low} に満たない語句 w は，主題タグの候補から外す．

3.4　タギング方法

本節では，文書分類によるカテゴリ・主題のタグの付与方法と，単語抽出によるキーワードタグの付与方法をそれぞれ示す．

3.4.1　カテゴリ・主題タグ

入力文書 d がカテゴリ c に属し主題 t を扱う確率 $p(c,t|d)$ が高い c と t を，d に付与すべきタグの組とする．ここでは多数の主題タグ候補を扱うため，$p(c,t|d)$ を直接推定することは難しい．そこで，$p(c,t|d) = p(c|d) \cdot p(t|c,d)$ のように分解し，$p(c|d)$ と $p(t|c,d)$ をそれぞれ推定する．まず，ベイズルールを用いて $p(c|d)$ を書き換える．

$$p(c|d) = \frac{p(c)p(d|c)}{p(d)} \tag{1}$$

次に，d に含まれる各単語 $w_1, w_2, \ldots, w_n \in \mathcal{W}$ の出現は，クラス c が与えられたとき他の単語と独立と仮定すると，$p(d|c)$ は以下のとおり推定できる．なお，$f(w)$ は入力文書中に w が出現する回数である．

$$p(d|c) = \prod_{w \in \mathcal{W}} p(w|c)^{f(w)} \tag{2}$$

そして，$p(w|c)$ は，以下のように推定する．

$$p(w|c) = \frac{\beta + F_c(w)}{\beta|V| + \sum_{x \in V} F_c(x)} \tag{3}$$

式 (3) の β はスムージングの強さ，$F_c(w)$ は訓練文書集合中のカテゴリ c における語句 w の総出現回数を表す．$V = \{w \,|\, D(w) \geq N_{\mathrm{cutoff}}\}$ は，訓練文書集合の語彙集合を表す．なお，$D(w)$ は訓練文書集合中における語句 w の出現文書数，閾値 N_{cutoff} は稀にしか出現しない語句を無視

するためのパラメータである [9].

それから，入力文書からカイ二乗統計量 [10] を用いて m 個の特徴語集合 $\mathcal{FW}$ を選択する．事前確率 $p(c)$ は訓練文書集合から推定可能ではあるが，各カテゴリに含まれる文書数の偏りを考慮して $p(c)$ は無視する．さらに，式 (1) の分母 $p(d)$ も c に非依存なため無視する．

また，$p(t|c,d)$ は $p(c|d)$ と同様に推定可能である．

$$p(w|c,t) = \frac{\gamma + F_{c,t}(w)}{\gamma|V_c| + \sum_{x \in V_c} F_{c,t}(x)} \tag{4}$$

ここで，γ はスムージングの強さ，$F_{c,t}(w)$ はカテゴリ c に属し主題 t を扱う訓練文書集合中における w の総出現回数，$V_c = \{w \,|\, D_c(w) \geq N_{\text{cutoff}}\}$ は訓練文書集合中のカテゴリ c における語彙集合，$D_c(w)$ は訓練文書集合中のカテゴリ c における語句 w の出現文書数を表す．なお，「主題 t を扱う文書」は「タイトル・本文に主題語 t が含まれる文書」として定義している．

以上に基づき，TagHats は入力文書 d に対して，カテゴリ・主題タグ $\{c,t\}$ を，式 (5) に示す対数尤度比 $L(c,t)$ の値が高い順に N 組付与する（あるいは，$L(c,t)$ が閾値 θ を越える $\{c,t\}$ の組を付与する）．ここで $\{c,\bar{t}\}$ はカテゴリ c における主題 t を除く全主題を表す．

$$L(c,t) = \log\left(\prod_{w \in \mathcal{FW}} \frac{p(w|c)^{f(w)}\, p(w|c,t)^{f(w)}}{p(w|\bar{c})^{f(w)}\, p(w|c,\bar{t})^{f(w)}}\right) \tag{5}$$

式 (5) は，先行研究 [1] のカテゴリ・主題付与方法と比べ，対数尤度比を採用した点が異なる．式 (5) で対数尤度比を利用することで文書長のタグスコアへの影響が軽減され，閾値 θ を利用して確信度の高いタグのみを付与する，といったタギング制御がより容易になる．なお，付与されるカテゴリ・主題タグの精度・傾向に関する先行研究 [1] との明確な違いはないことから，本論文では式 (5) による精度改善に関する主張・評価は行わない．

3.4.2　キーワードタグ

構造化された複数のセクションを持つ文書に対しては，入力文書を含む多くの文書中で複数のセクションに出現する語句がキーワードタグとして適切であると我々は考える．たとえば，Q&A 文書であれば，質問文のみに出現する語句よりも，質問文と回答文の両方に出現する語句が重要と考えられる．この仮説に基づき，我々は，残差 IDF（RIDF；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）[11] とセクション頻度（SF）を組み合わせて語句の重要度を計算する手法を新たに提案する．

$$\begin{aligned}\text{SF-RIDF}(w) = s(w) \cdot \Bigg[&\log_2\left(\frac{D+1}{D(w)+1}\right) \\ &+ \log_2\left\{1 - \exp\left(-\frac{S(w)+s(w)}{D+1}\right)\right\}\Bigg]\end{aligned} \tag{6}$$

ここで，セクション頻度 $s(w)$ は，入力文書 d の中に語句 w が出現したセクション数，$S(w)$ は語句 w が出現した訓練文書集合中の総セクション数，D は全訓練文書数である．なお，RIDF は文書の内容を現す重要な語句ほど 1 つの文書中に繰り返し出現する特性を利用した指標で，一般語に対してはスコアが低くなる．

従来技術である TF-IDF に比べて SF-RIDF が優れている点は，(1) 単一のセクションに固有なノイズ語句（誤用・乱用・タイプミスなど）にロバストであること，(2) 多くの文書に出現している語句（低 IDF）でもキーワードとして抽出可能，の 2 点である．なお，タグを検索用途として用いる際は，1 つの文書にのみ付与されるタグよりも，複数個の文書に付与され，文書間にリンクを産み出すタグの方が重要である．TF-IDF では，多くの文書に付与されるような語よりも，IDF 値の高い語句をより多く抽出してしまうという問題がある．

4.　評価実験

本章では，階層的オートタギング技術（TagHats）の各技術要素についての評価を行う．訓練文書集合には，「教えて！ goo」の 2008 年 4 月～2009 年 3 月の 777,266 文書（389 カテゴリ）を用い，各文書を 3 つのセクション：質問タイトル・質問本文・回答文に分割した．

4.1　タギング例

表 5 は，「教えて！ goo」の Q&A 文書（質問番号：4842983, 4536688, 335041）[*1] に対して TagHats と TF-IDF によりタグを付与した結果の例である（カテゴリ・主題タグ：$L(c,t) > 0$ となる $\{c,t\}$ の組を 2 つまで，キーワードタグは 3 つまで）．TagHats は入力文書に対して，「小動物

表 5　「教えて！ goo」の文書に対するタギング例
Table 5　Examples of auto-tagging for Q&A documents of Oshiete! goo.

	ゴールデンハムスターの死（質問番号：4842983）
カテゴリ	その他（ペット） 小動物
主題	ハムスター 餌
キーワード	ゴールデンハムスター 冬眠 暖房
TF-IDF	巣箱 暖房 風邪 冬眠 頬袋 死亡 餌
	円周率の求め方（質問番号：4536688）
カテゴリ	数学
主題	円周率 π
キーワード	アルキメデス テイラー展開 円周
TF-IDF	atn 239 式 1 アルキメデス π 99
	ユキピタスって…（質問番号：335041）
カテゴリ	その他（インターネット接続） その他（デジタルライフ）
主題	ネットワーク
キーワード	ユビキタス パソコン 言葉
TF-IDF	ユキビタス いつ BIQ ユキビタスネットワーク ユキピタス ブロードバンド 利用例

[*1] http://oshiete.goo.ne.jp/qa/(4842983|4536688|335041).html

表 6　「教えて！ goo」における質問数上位カテゴリの主題タグの抽出結果（p 値の低い順に 10 個）

Table 6　Top 10 theme tag candidates (10 lowest p-values) extracted from each popular category in Oshiete! goo.

カテゴリ	主題タグ抽出結果
恋愛相談	告白 彼氏 彼女 男性 片思い 脈 元彼 復縁 デート 元彼女
Windows XP	タスクバー sp3 言語バー アプリケーションエラー ムービーメーカー フリーズ 文字入力 media player デスクトップアイコン 起動時
法律	交通事故 相続 遺産相続 相続放棄 傷害事件 著作権 裁判員制度 名誉毀損 強制執行
その他（デジタルライフ）	ニコニコ動画 ipod youtube 動画 yourfilehost veoh itunes 2 ちゃんねる ipodnano 地デジ
Office 系ソフト	セル エクセル excel マクロ word ワード vba access excel2007 excel2003

＞ハムスター＞ゴールデンハムスター」のように階層的にタグを付与できた．また，TF-IDF では同一セクション内で繰り返し出現する語句（高 TF）の影響により「atn」「239」などのタグとして不適切な語句が抽出された．さらに，TF-IDF では「ユキビタス」「ユキピタス」などの誤用・タイプミスであるノイズ語句（IDF 値が高い）が抽出されたが，TagHats が用いる SF-RIDF では正しい語である「ユビキタス」のみをキーワードタグとして抽出できた．これは，RIDF 値の利用により，多くの文書のセクションで出現するような重要な語句に高いキーワードスコアが与えられるためである．

4.2　主題タグ候補抽出結果の評価

本節では，各カテゴリからの主題タグ候補の抽出例と，検索ランキング上位語を用いた主題タグ候補抽出に関する評価結果を示す．

4.2.1　主題タグ候補の抽出例

3.3 節に示した提案手法により，訓練文書集合から 34,604 個（26,438 種類）の主題タグ候補を抽出した（$\alpha = 0.01$, $N_{\mathrm{low}} = 3$）．**表 6** に示す質問数上位のカテゴリにおける抽出結果より，本技術が各カテゴリに関連する重要な語句を抽出できたことが分かる．

ここで，「デジタルライフ＞その他（デジタルライフ）」の主題タグ候補抽出結果からは，2 つのサブカテゴリを発見することができる．1 つ目は，動画共有サイトに関するもので，ニコニコ動画，YouTube，YourFileHost，veoh などが抽出されている．「教えて！ goo」では動画共有サイトに関するカテゴリを 2010 年 3 月当時有していなかった[*2]ため，「その他（デジタルライフ）」に動画共有サイトに関する質問が投稿されたからと考える．2 つ目は，音楽プレイヤに関するもので，iPod，iTunes，iPod nano などが抽出

[*2] 現在は，「デジタルライフ＞動画サービス」というカテゴリが存在する．

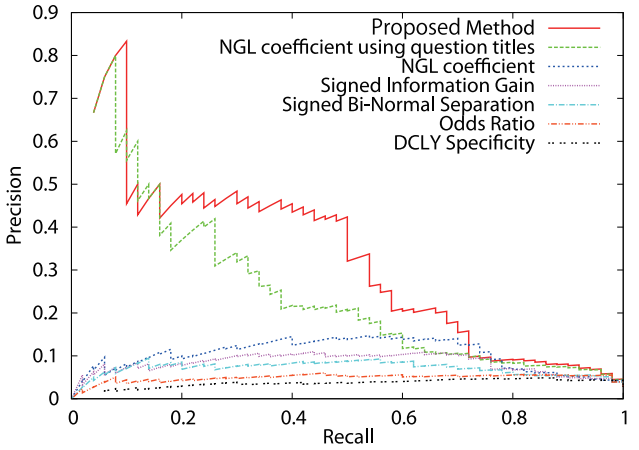

図 3　「病気」カテゴリからの主題タグ候補の抽出結果に関する識別率–再現率曲線

Fig. 3　Precision-recall curves with the proposed and conventional methods for extracting Japanese Disease theme tags.

された．「教えて！ goo」では，「趣味＞AV 機器＞iPod・携帯音楽プレイヤー」カテゴリが用意されているが，トップレベルカテゴリの「デジタルライフ」と「趣味」が相互に排他でないため，「iPod・携帯音楽プレイヤー」カテゴリを発見できなかったユーザが「デジタルライフ＞その他（デジタルライフ）」に音楽プレイヤに関する質問を投稿していたと考える．

4.2.2　病気に関する検索語による評価

TagHats は文書検索を目的としてタグを付与するため，ユーザが高頻度で検索するような語を主題タグ候補とすることが望ましい．そこで，「美容&健康＞健康＞病気」カテゴリ（訓練文書集合に 13,050 文書含まれる）から提案手法（3.3 節）と従来手法により抽出した語を，2009 年の病気に関する検索ランキング上位 50 語[*3]と比較した．

従来手法には，DCLY specificity [12]，NGL coefficient [13], [14]（符号付きカイ二乗統計量），質問タイトルを利用した NGL coefficient（NGLT），符号付き Information Gain [15]，符号付き Bi-Normal Separation [9]，オッズ比 [16] を利用した．DCLY specificity は質問トピックを発見するための指標として提案されたもので，他の手法は変数選択手法として広く用いられているものである．なお，カテゴリ c において $T_c(w) < 3\,(= N_{\mathrm{low}})$ を満たす語は抽出候補から除いた．また，質問タイトルを利用した NGL（NGLT）は，以下のとおり求めた．

$$\frac{\{T_c(w)(D_{\bar{c}} - D_{\bar{c}}(w)) - D_{\bar{c}}(w)(D_c - T_c(w))\}\sqrt{D}}{\sqrt{D(w)(D - D(w))D_c D_{\bar{c}}}} \quad (7)$$

ここで，D_c はカテゴリ c の全訓練文書数である．

図 3 に示す識別率–再現率曲線のとおり，提案手法は識別率と再現率の両面で最も良い結果を示した．NGLT も良

[*3] http://ranking.goo.ne.jp/ranking/n09/n2009_health_keyword/

表 7　2009 年の病気に関する検索ランキング上位 50 語（goo ランキング提供）と，提案手法による「病気」カテゴリからの上位 50 個の主題語抽出結果．太字の 21 語は検索ランキング上位 50 語に含まれる

Table 7　Top 50 searched disease names of 2009 in Japan and top 50 theme tag candidates extracted from Disease category by the proposed method.

検索ランキング上位 50 語
インフルエンザ ヘルペス 糖尿病 卵巣腫瘍 高血圧 子宮筋腫 帯状疱疹 咳 甲状腺 膀胱炎 膠原病 腰痛 めまい 湿疹 痔 頭痛 大腸がん 骨折 貧血 風邪 肺炎 脳梗塞 下痢 便秘 ヘルニア 口内炎 子宮頸がん 腹痛 肩こり 関節リウマチ 痛風 更年期障害 鼻血 不整脈 耳鳴り 肺がん 結核 子宮内膜症 アルツハイマー 血便 てんかん 子宮体がん リウマチ うつ病 自律神経失調症 リンパ 気管支炎 肋間神経痛 じんましん ベーチェット病
提案手法による抽出結果（上位 50 語）
痛み **インフルエンザ** **痔** **頭痛** **咳** **糖尿病** しこり 何科 花粉症 しびれ 腫れ **血便** 手術 できもの **耳鳴り** **帯状疱疹** 喉 緑 内障 椎間板ヘルニア **膀胱炎** **脳梗塞** B 型肝炎 **腰痛** 潰瘍性大腸炎 **リウマチ** 痰 アトピー **ヘルペス** 喘息 **腹痛** **めまい** 斜視 胸痛 **口内炎** 乳がん 水虫 **痛風** 口唇ヘルペス 逆流性食道炎 **不整脈** **下痢** 肝炎 吐き気 **じんましん** 微熱 名医 大腸がん 副鼻腔炎 突発性難聴 **高血圧**

い結果を示していることから，質問タイトルに着目することが重要であることが分かる．ここで，「病気」カテゴリで特徴的に出現する語であっても，そのすべてがカテゴリの主題を表す語ではないため，タイトルに着目しない変数選択手法では高い識別率を実現することができない．

表 7 に，2009 年の病気に関する検索ランキング上位 50 語と，提案手法による「病気」カテゴリからの上位 50 個の主題語抽出結果の比較を示す．提案手法は抽出した 50 語のうち 21 語が検索ランキング上位 50 語に含まれていた．さらに，上位 50 語に含まれていない語もすべて医学に関する語であり，主題語を正確に抽出できていたことが分かる．なお，検索ランキング上位語と抽出語を比較すると，提案手法では癌（卵巣腫瘍や大腸がん）や精神的問題（うつ病）の関連語が抽出されていない．これは，「教えて！goo」が「癌」や「メンタルヘルス」という，「病気」カテゴリよりも専門的なカテゴリを有しているため，これらのカテゴリに関する語が「病気」カテゴリでは上位の主題語として抽出されなかったためである．

図 4 は，「病気」カテゴリに含まれる語が主題語であるかについてのフィッシャの正確確率検定の p 値（3.3 節）と，タイトル中出現頻度（$T_c(w)$）の関係である．他のカテゴリにも多く含まれる語（首，足，顔など）は，タイトル中出現頻度が高いが，p 値は 1.0 に近い値であるため主題語としては抽出されない．なお，検索ランキング上位 50 語中，47 語において $p < .05$，45 語において $p < .01$ であり，提案手法が検索ランキング上位 50 語を主題語として正確に抽出できたことが分かる．

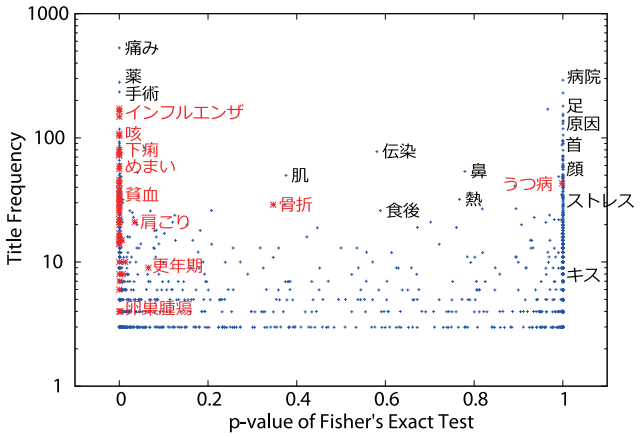

図 4　フィッシャの正確確率検定の p 値とタイトル中出現頻度の関係．赤字は検索上位 50 語を示す

Fig. 4　p-value of Fisher's Exact Test vs. title frequency. Red-colored words are contained in top 50 searched diseases.

4.3　タグ付与結果の評価

本節では，カテゴリ・主題・キーワードタグの付与精度について，それぞれ評価する．パラメータ値は，$\beta = 0.1$，$\gamma = 0.1$，$N_{\mathrm{cutoff}} = 3$，$m = 100$ と設定した．

なお，カテゴリ・主題の組について十分な量の教師データを得ることが難しいため，カテゴリタグの評価（教師データ：ユーザの投稿先カテゴリ）と，主題・キーワードタグの評価（教師データ：はてなブックマークのユーザが「教えて！ goo」の Q&A 文書に対して付与したタグ）に分けて，評価実験を実施した．

4.3.1　カテゴリタグの評価

まず，式 (5) に示す（TagHats が使用する）階層的な分類 $L(c,t)$ によって付与されるカテゴリタグについて評価する．ここでは，単純なカテゴリ分類 $p(c|d)$ と，全カテゴリ・主題の組数のカテゴリ分類 $p(c,t|d)$ をベースラインとして用いた．なお，先行研究 [1] のカテゴリ・主題付与方法 $S(c,t)$ から，対数尤度比をとる式 (5) に変更した目的はタグスコアへの文書長の影響の軽減にあるため（3.4.1 項参照），$S(c,t)$ は本精度評価に加えていない．

図 5 は，実際の質問者が選択したカテゴリと，システムが付与したカテゴリタグのうちどれか 1 つの一致率を表す（テストデータには，2009 年 4 月からの 10,000 文書を用いた）．$p(c,t|d)$ によるフラットなカテゴリ・主題の推定は，すべてのカテゴリと主題について学習するために十分なデータがないため精度が非常に悪い．$L(c,t)$ による階層的分類は最も良い結果を残した．特に，カテゴリタグの付与数が 1 から 4 個の際は，階層的分類は，フラットなカテゴリ分類 $p(c|d)$ に比べて有意に一致率が良かった（$p < .01$，符号検定）．これは，カテゴリ間で $p(c|d)$ の値に有意な差がない場合に，$L(c,t)$ ではカテゴリ中の主題によりマッチするカテゴリに高いスコアを与える点が，精度向上につながったと考える．

4.3.2　主題・キーワードタグの評価

次に，TagHats により付与した主題・キーワードタグと，TF-IDF により付与したタグについて，はてなブックマークのユーザが「教えて！ goo」の Q&A 文書に対して付与したタグを用いて比較評価を行う．本実験では，少なくとも 2 人のユーザによって 3 つ以上のタグが付与された 122 文書[*4]をテストデータとした．

図 6 は，TagHats が付与したタグのうちどれか 1 つが，はてなブックマークのユーザが付与したタグに含まれる割合を表す．本論文で提案したキーワードタグ抽出手法 SF-RIDF は，従来手法である TF-IDF を凌駕する結果を得た．特に，タグ付与数が 4 個以上のとき TF-IDF よりも有意に良い結果を得た（$p < .01$，符号検定）．これは，IDF では多くの文書に出現する語にペナルティを与えてしまうが，RIDF では多くの文書のセクションに出現する重要語に高いスコアを付与できるからと考える．また，より総称的なタグである主題タグも，付与数 1 個のときは TF-IDF より良い結果を得た．このことから，TagHats のように主題タグとキーワードタグをあわせて付与することにより，ソーシャルタギングで付与される多様なタグをより広くカバーできると考える．

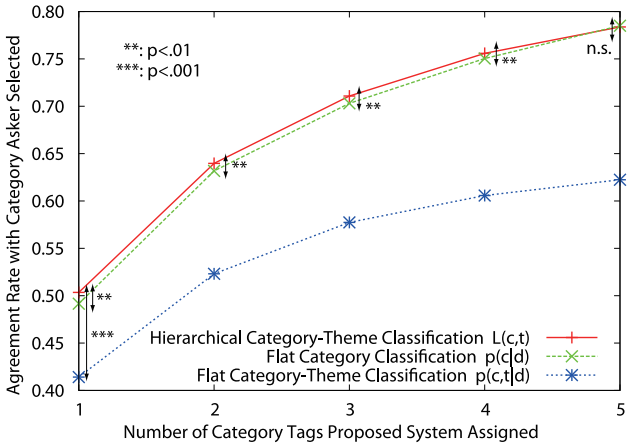

図 5　システムが付与したカテゴリタグのうちどれか 1 つが，「教えて！ goo」のユーザが実際に選択したカテゴリと一致した割合

Fig. 5　Agreement rates between category selected by asker and any of category tags assigned by our system.

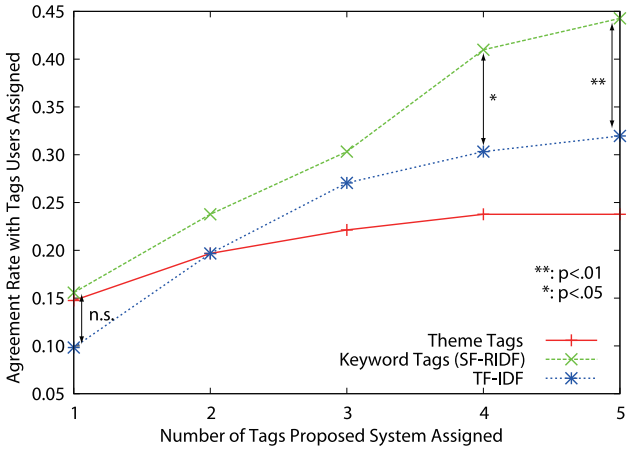

図 6　システムが付与した主題・キーワードタグのうちどれか 1 つが，はてなブックマークのユーザの付与タグに含まれる割合

Fig. 6　Rates that any of the theme or keyword tags that our system assigned are contained in the tags that the Hatena social bookmarking service users assigned.

[*4] http://b.hatena.ne.jp から 2010 年 2 月 3 日にクロール取得．

5.　応用事例と被験者実験

実サイト上での階層的オートタギング技術（TagHats）の応用事例として，(1)「教えて！ goo」の Q&A 文書へのタグリンク挿入と，(2) Q&A 検索サービス「QA.ON/OFF」での利用の 2 つがある．本章では，これらの応用事例を説明するとともに，各事例の利用者を対象としたタギング精度評価実験の結果について示す．

5.1　「教えて！ goo」の Q&A 文書へのタグリンク挿入

2010 年 9 月 1 日から 2011 年 3 月 31 日まで，TagHats を用いて「教えて！ goo」の Q&A 文書にタグ（主題タグ 2 つとキーワードタグ 3 つ）のリンクをスコア（式 (5)，式 (6)）の高い順に付与し，Web ページ上にキーワード検索リンクとして表示した（**図 7**）．なお，表示スペースの制約と，各 Q&A ページにはあらかじめカテゴリタグに相当するリンクが存在するため，「分類キーワード（タグ）」欄にカテゴリタグは含めなかった．

我々は，TagHats により付与されたタグのクリック数を測定し，Q&A 文書の内容に関連したタグが付与できたか否かを評価した．なお，本実験では「文書内容に関連したタグほど，ユーザにクリックされる傾向が強い」という仮説をおいている．また，表示位置の影響が出ないように，表示順序はランダムに決定した．

図 8 にクリック数の測定結果を示す．2 つの主題タグのクリック比率に偏りがあるかについて χ^2 適合度検定を実施したところ，スコア順位別に有意な偏りが確認できた

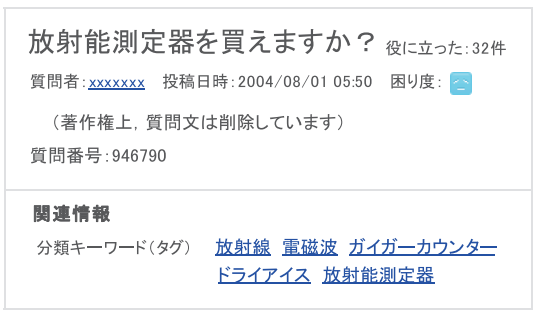

図 7　「教えて！ goo」の Q&A 文書へのタグ付与例（主題タグ：放射線，電磁波．キーワードタグ：ガイガーカウンター，ドライアイス，放射線測定器）

Fig. 7　An example of assigning tags to the Q&A document of Oshiete! goo.

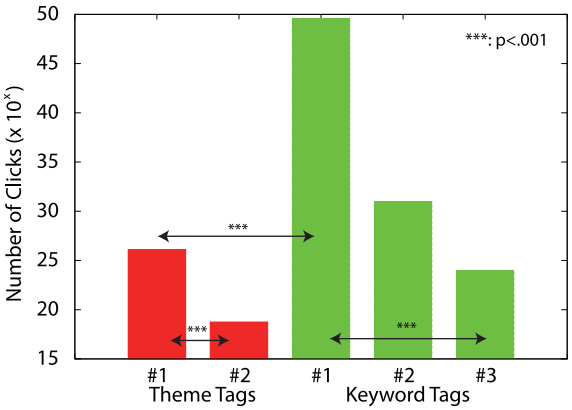

図 8　「教えて！ goo」の Q&A 文書におけるタグ種別クリック数．スコア上位順に主題タグ 2 つ，キーワードタグ 3 つを文書に付与

Fig. 8　Number of clicks of two theme and three keyword tags that have highest scores.

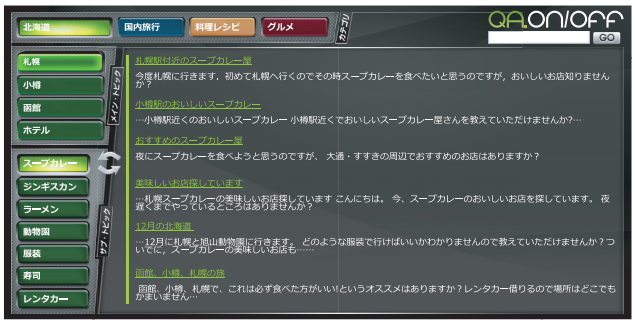

図 9　「QA.ON/OFF」のナビゲーション画面（「スープカレー」検索時）．3 層のタグ（カテゴリ，主題，キーワード）ボタンが画面左側に配置され，オン状態のカテゴリ（例では「北海道」）に関する文書が画面右側に表示される．ボタンのオンオフによって検索カテゴリの変更や，検索語の追加/削除が可能

Fig. 9　A screenshot of the service “QA.ON/OFF”. Turning the buttons representing hierarchical tags on and off enables users to find Q&A documents quickly.

($p < .001$)．同様に，3 つのキーワードタグのクリック比率に偏りがあるかについて χ^2 適合度検定を実施したところ，スコア順位別に有意な偏りが確認できた（$p < .001$）．以上のことから，TagHats が付与するタグのスコアが高いほどユーザが多くクリックすることが確認できた．

5.2　Q&A 検索サービス「QA.ON/OFF」

2011 年 1 月 31 日から 12 月 16 日まで，TagHats により付与したタグを利用した Q&A 検索の実証実験サービス「QA.ON/OFF」を実施した（図 9）[*5]．本サービスでは，画面右側の検索結果の Q&A 文書にそれぞれ付与された 3 層のタグを集約して画面左側にナビゲーションボタンとして配置する．検索結果はカテゴリごとにまとめられ，ユーザは，カテゴリタグボタンのオンオフによって，検索結果の視点の切替えが可能になる．また，主題タグ（「メイン・トピック」ボタン）・キーワードタグ（「サブ・トピックボタン」のオンオフによる検索語の追加・削除により，検索内容の絞り込み・拡大が可能になる．

[*5] http://qaonoff.goo.lab.ne.jp/（現在は実験終了）

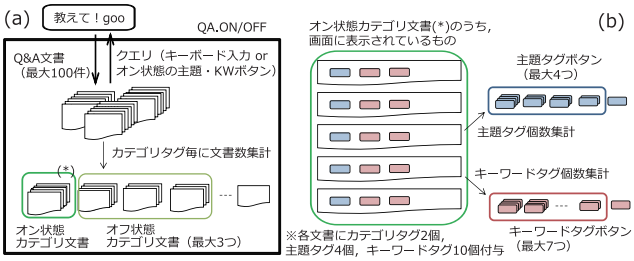

図 10　「QA.ON/OFF」における階層的オートタギングで付与したタグの利用方法．(a) カテゴリボタン，(b) 主題・キーワード（KW）ボタンの決定方法

Fig. 10　How to use hierarchical tags on “QA.ON/OFF”. (a) category buttons. (b) theme and keyword buttons.

5.2.1　階層タグによる検索ナビゲーション

図 10 に，「QA.ON/OFF」における TagHats で付与したタグの利用方法を示す．

まず，QA.ON/OFF では，あらかじめ TagHats により「教えて！ goo」の Q&A 文書に対してカテゴリタグを 2 個，主題タグを 4 個，キーワードタグを 10 個付与しておく．

ユーザが検索を行うと（検索語の入力，あるいは，主題・キーワードボタンのオンオフ操作），「QA.ON/OFF」では「教えて！ goo」の API を用いて，検索語（入力クエリ，あるいは，オン状態にある主題・キーワードボタンのタグ名）が含まれる文書を 100 件取得する．次に，取得文書集合のカテゴリタグを集計し，件数が多い順に最大 4 つのカテゴリを取得する．これらの 4 つカテゴリのうち，最も件数の多いカテゴリ（表示カテゴリ）をオン状態のボタンとして表示し，他のカテゴリはオフ状態のボタンとして表示する（図 10 (a)）．

画面右側には，表示カテゴリがタグとして付与された文書を表示する．このとき，画面内に表示された文書に付与された主題・キーワードタグを集計し，出現回数が多い順に 4 つの主題タグボタンと 7 つのキーワードタグボタンを取得する（図 10 (b)）．なお，QA.ON/OFF はユーザによる検索画面のスクロール（画面に表示される文書の変更）に応じて，主題・キーワードタグボタンを再選択する．この再選択により，タグボタンは検索用途以外にも，検索結果全体の内容を把握するために役立つようになる．

5.2.2　タグボタンによる再検索

ユーザが検索エンジンを利用する際に，1 度の検索で望ましい検索結果が見つからない場合は，検索語の変更・追加・削除などを繰り返すことで目的とする文書を探そうとする．我々は，「QA.ON/OFF」でのユーザの行動を調査して，TagHats によって付与された多数のタグが，ユーザの再検索行動の手助けになっていたかについて評価した．

図 11 に，カテゴリ・主題・キーワードタグボタン（オン

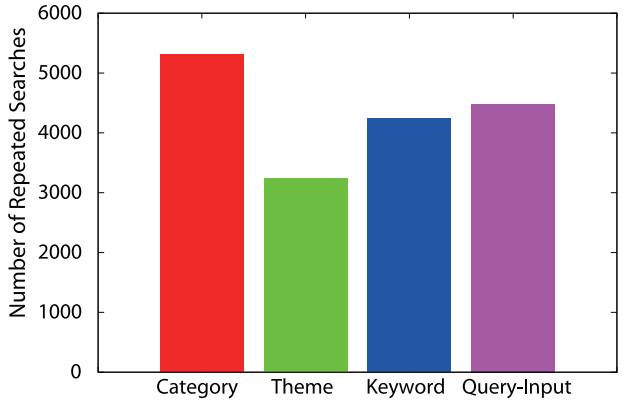

図 11　タグボタン（カテゴリ・主題・キーワードタグのオンオフ操作）と検索窓からのクエリ入力による再検索回数

Fig. 11　Number of repeated searches by turning tag-buttons on/off and inputting queries.

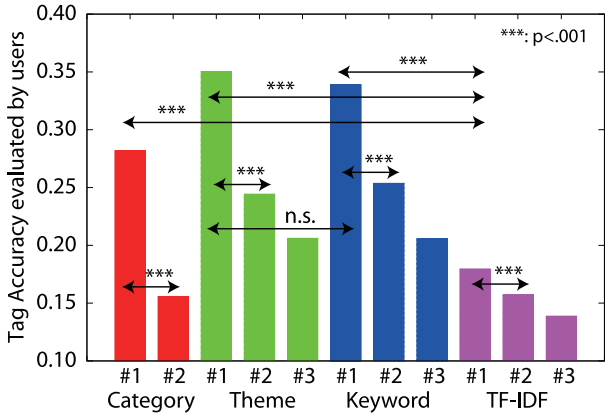

図 12　提案手法（カテゴリ・主題・キーワードタグ）と TF-IDF 法に関するユーザ評価タギング精度

Fig. 12　Tag accuracy evaluated by users of the proposed method (category, theme, and keyword tags) and common TF-IDF method.

オフ操作の合計）による再検索と，検索窓からの入力（トップ画面・ナビゲーション画面（図 9 右上）の合計）による再検索の回数を示す．すべてのタグボタンによる再検索回数の合計は，検索窓からの入力に比べ，約 3.07 倍となった．この結果は，TagHats によって付与したタグが，ユーザが自ら検索語の修正を行う負担を低減させることができたことを表していると考える．

5.2.3　タギング精度の評価

5.1 節の評価実験では，直接的なタギング精度の評価は実施できていなかった（タグが Q&A 文書の内容に関連していない場合でも，ユーザの興味によってはそのタグをクリックするケースがある）．そこで，我々は被験者実験を行い，タギング精度について直接評価した．本実験では，被験者 2,500 人（平均年齢 42.37 歳（標準偏差 11.83），男性 1,441 人，女性 1,059 人）に対して，初めに Q&A 文書を提示し，その後，文書内容に適切なタグをすべて選択させる操作を 10 回繰り返した．Q&A 文書は被験者ごとにランダムに 10 個選択し，各文書に対して，提案手法によりカテゴリタグ 2 つ，主題タグ 3 つ，キーワードタグ 3 つ，TF-IDF によりタグを 3 つ付与した．これらのタグは，ランダムな順で画面上に提示し，被験者の回答を，各タグのスコア順位ごとに集計した．

図 12 に結果を示す．各タグのスコア 1 位と 2 位の間で 2 群の比率の差の検定を行ったところ，タグのスコア別の適合度に有意差が認められた（$p < .001$）．また，スコア 1 位のタグどうしで比較を行ったところ，提案手法により付与されるカテゴリ・主題・キーワードタグのいずれも TF-IDF により付与されるタグより適合度が有意に高かった（$p < .001$）．

また，キーワードタグと主題タグの間にはタギング精度に有意差は認められなかった．このことから，「教えて！goo」におけるタグ種別クリック数（図 8）で得られた「キーワードタグが主題タグよりもクリック数が有意に多かった（$p < .001$）」知見については，具体的な内容を表すタグの方がユーザのクリックを誘発しやすい特徴があるからと考える．なお，図 11 でも同様に，キーワードタグボタンの方が，主題タグボタンよりも再検索回数が多かった．

6.　関連研究

従来のオートタギング，特にブログ記事に対してタギングを行う研究には，キーワード抽出 [17]，協調フィルタリング [18], [19]，文書分類 [20], [21]，トピックモデル [22] など，様々なアプローチがある．これらは，キーワード抽出 [17] を除いて，ユーザが付与したタグを基に学習するものであり，タグがいっさい付与されていない文書集合に対してタギングを可能にした階層的オートタギング技術の貢献は大きい．また，階層的にオートタギングを行うという観点も，本技術が初めての試みである．また，Heymann らは，ソーシャルタギングシステムのタグ集合を階層的なタクソノミに変換するアルゴリズムを提案した [23]．この研究は，ユーザが付与したタグが利用可能なメディアを対象としているので，ユーザがタグを付与していない Q&A 文書から適切な主題タグ候補を自動的に発見することができる本技術とは異なる．

また，階層的オートタギング技術の応用先である Q&A コミュニティは利用者が作成する新たな情報基盤の 1 つとして成長を続けており，Q&A コミュニティに関する様々な研究が行われている．Adamic らは，「Yahoo! Answers」について解析を行い，Q&A コミュニティにおける文書の特徴，ユーザの行動に関する分析を行った [24]．Nam らは，Naver Knowledge-iN コミュニティにおける回答者の行動を解析し，回答数の上位 1,000 人のうち，52% のユーザが単一のカテゴリでのみ回答を行っていることを明らかにした [25]．これは，同一の質問内容が複数のカテゴリに分散するようなカテゴリ体系の場合，質問者が回答を

受ける機会が大きく失われることを意味する．Harper らは，文書分類アプローチにより，「Yahoo! Answers」の質問文を informational なものと conversational なものに分類し，質問タイプとカテゴリとの関係を明らかにした [26]．また，Morris らは，Facebook や Twitter におけるステータスメッセージにおける Q&A 行動を分析した [27]．

7. おわりに

本論文では，カテゴリ・主題・キーワードという異なる意味抽象度のタグを文書へ自動的に付与する階層的オートタギング技術について論じた．まず，本技術を「教えて！goo」と Q&A 検索サービス「QA.ON/OFF」へ導入した応用事例を紹介した．これらの事例からは，具体的な内容を表すキーワードタグの方が，より抽象的な主題タグに比べて，ユーザのクリックを誘発しやすいという特徴があるという知見を得た．次に，文書構造（たとえば，Q&A 文書であればタイトル・質問文・回答文の 3 つのセクションから構成される）を考慮した新たなキーワード抽出法 SF-RIDF の提案を行い，提案手法が従来法の TF-IDF よりも検索用途のタグとして有益なキーワードを抽出できることを示した．そして，新たな評価実験として，主題タグ候補の抽出精度と，応用事例の利用者を対象に行った被験者実験によるタギング精度の評価を実施し，各技術要素が，従来手法に比べて文書内容に適切なタグを付与できることを示した．今後は，カテゴリに含まれる文書の傾向が急激に変化した場合に，タギング精度が悪くなってしまうという問題（コンセプトドリフト [28]）について解決したい．

参考文献

[1] 西田京介，藤村　考：階層的オートタギングによる Q&A コミュニティの知識整理，日本データベース学会論文誌，Vol.9, No.1, pp.1–6 (2010).

[2] 西田京介，星出高秀，藤村　考：階層的オートタギングによる文書整理，*GNWS*, IPSJ Symposium Series, Vol.8, pp.51–56 (2010).

[3] Nishida, K. and Fujimura, K.: Hierarchical Autotagging: Organizing Q&A Knowledge for Everyone, *CIKM*, pp.1657–1660 (2010).

[4] Golder, S.A. and Huberman, B.A.: Usage patterns of collaborative tagging systems, *ACM SIGKDD Explorations Newsletter*, Vol.32, No.2, pp.198–208 (2006).

[5] Marlow, C., Naaman, M., Boyd, D. and Davis, M.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read, *Hypertext*, pp.31–40 (2006).

[6] Rennie, J.D.M., Shih, L., Teevan, J. and Karger, D.R.: Tackling the Poor Assumptions of Naive Bayes Text Classifiers, *Proc. 20th Int'l Conf. Machine Learning*, pp.616–623 (2003).

[7] Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Co-occurrence, *COLING*, pp.409–413 (1998).

[8] Agresti, A.: A Survey of Exact Inference for Contingency Tables, *Stat. Sci.*, Vol.7, No.1, pp.131–153 (1992).

[9] Forman, G.: An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Reserach*, Vol.3, pp.1289–1305 (2003).

[10] Yang, Y. and Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization, *ICML*, pp.412–420 (1997).

[11] Church, K.W. and Gale, W.A.: Inverse Document Frequency (IDF): A Measure of Deviations from Poisson, *Proc. 3rd Works. Very Large Corpora*, pp.121–130 (1995).

[12] Duan, H., Cao, Y., Lin, C.-Y. and Yu, Y.: Searching Questions by Identifying Question Topic and Question Focus, *ACL-HLT*, pp.156–164 (2008).

[13] Ng, H.T., Goh, W.B. and Low, K.L.: Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization, *SIGIR*, pp.67–73 (1997).

[14] Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol.34, No.1, pp.1–47 (2002).

[15] Zheng, A., Wu, X. and Srihari, R.: Feature selection for text categorization on imbalanced data, *ACM SIGKDD Explorations Newsletter*, Vol.6, No.1, pp.80–89 (2004).

[16] Mladenic, D. and Grobelnik, M.: Feature Selection for Unbalanced Class Distribution and Naive Bayes, *ICML*, pp.258–267 (1999).

[17] Brooks, C.H. and Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering, *WWW*, pp.625–632 (2006).

[18] Mishne, G.: AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts, *WWW*, pp.953–954 (2006).

[19] Sood, S.C., Owsley, S.H., Hammond, K.J. and Birnbaum, L.: TagAssist: Automatic Tag Suggestion for Blog Posts, *ICWSM* (2007).

[20] Ohkura, T., Kiyota, Y. and Nakagawa, H.: Browsing System for Weblog Articles based on Automated Folksonomy, *Proc. WWW 2006 Works. Weblogging Ecosystem: Aggregation, Analysis, and Dynamics* (2006).

[21] Fujimura, S., Fujimura, K. and Okuda, H.: Blogosonomy: Autotagging Any Text Using Blogger's Knowledge, *WI*, pp.205–212 (2007).

[22] Si, X. and Sun, M.: Tag-LDA for Scalable Real-time Tag Recommendation, *Information and Computational Science*, Vol.6, No.1, pp.23–31 (2009).

[23] Heymann, P. and Garcia-Molina, H.: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems, Technical Report 2006-10, Computer Science Department (2006).

[24] Adamic, L.A., Zhang, J., Bakshy, E. and Ackerman, M.S.: Knowledge sharing and yahoo answers: Everyone knows something, *WWW*, pp.665–674 (2008).

[25] Nam, K.K., Ackerman, M.S. and Adamic, L.A.: Questions in, knowledge in?: A study of naver's question answering community, *CHI*, pp.779–788 (2009).

[26] Harper, F.M., Moy, D. and Konstan, J.A.: Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites, *CHI*, pp.759–768 (2009).

[27] Morris, M.R., Teevan, J. and Panovich, K.: What Do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&A Behavior, *CHI*, pp.1742–1748 (2010).

[28] Žliobaitė, I.: Learning under Concept Drift: an Overview, Technical report, Vilnius University (2009).

西田 京介　（正会員）

2004 年北海道大学工学部情報工学科卒業．2006 年同大学大学院情報科学研究科修士課程修了．2008 年同博士課程修了．同年日本電信電話（株）入社．以来，データマイニングの研究開発に従事．現在，NTT サービスエボリューション研究所所属．博士（情報科学）．電子情報通信学会，日本データベース学会各会員．

星出 高秀　（正会員）

1993 年九州大学大学院総合理工学研究科修士課程修了．同年日本電信電話（株）入社．以来，e ラーニングシステム，マイニングの研究開発に従事．現在，NTT サービスエボリューション研究所所属．

藤村 考　（正会員）

1984 年北海道大学工学部電気工学科卒業．1986 年同大学大学院修士課程修了．1989 年同大学院博士課程修了．1989～2012 年日本電信電話（株），2001～2011 年電気通信大学大学院情報システム学研究科客員教授を経て，2012 年より大妻女子大学社会情報学部教授．ソーシャルメディアからの知識抽出，情報可視化，インタラクティブメディア等の研究開発に従事．工学博士．電子情報通信学会，日本データベース学会各会員．

内山 匡　（正会員）

1985 年名古屋大学理学部物理学科卒業．1987 年同大学大学院修士課程修了．同年日本電信電話（株）入社．1998～2001 年 NTT コミュニケーションズ，2004～2006 年 NTT レゾナントにてポータルサービスの開発等に従事．2007 年より NTT サイバーソリューション研究所（現サービスエボリューション研究所）所属．行動モデリングの研究開発に従事．電子情報通信学会，日本応用数理学会各会員．

（担当編集委員　鈴木 伸崇）