

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

西田 京介[†] 藤村 考[†]

[†] 日本電信電話株式会社 NTT サイバーソリューション研究所

〒239-0847 神奈川県横須賀市光の丘 1-1

E-mail: [†]{nishida.kyosuke,fujimura.ko}@lab.ntt.co.jp

あらまし Q&A コミュニティは新たな情報検索の手段および集合知のリポジトリとして急速に発展している。代表的なコミュニティはカテゴリにより Q&A 文書を整理しているが、現状用いられているカテゴリはユーザの多様な検索意図や知識・疑問に十分対応できていない。そこで、本研究で我々は、文書分類とキーワード抽出の両アプローチにより、文書にカテゴリ・主題・キーワードという意味的階層構造を有するタグを自動的に複数付与可能なシステムを提案する。教えて!goo に実際に投稿された Q&A 文書を用いた実験により、提案システムが文書の内容に即したタグを付与できたことを示す。

キーワード Q&A コミュニティ, 階層的オートタギング, 文書分類, キーワード抽出

Organizing Knowledge of Question Answering Community with Hierarchical Autotagging

Kyosuke NISHIDA[†] and Ko FUJIMURA[†]

[†] NTT Cyber Solutions Laboratories, NTT Corporation

1-1 Hikarinooka, Yokosuka-Shi, Kanagawa, 239-0847 Japan

E-mail: [†]{nishida.kyosuke,fujimura.ko}@lab.ntt.co.jp

Abstract Question answering communities have evolved rapidly as a new generation of information retrieval and the repository of collective knowledge. Representative QA communities have organized question-answer documents with categorization; however, categories are not able to correspond fully to users' various knowledge and interests. In this study, we propose a system that assigns hierarchical tags—category, theme, and keyword tags—to documents by using both document classification and keyword extraction approaches. Experiments with the question-answer documents of *Oshiete! goo* show that our system was able to assign hierarchical tags corresponding to the documents.

Key words Question Answering Community, Hierarchical Autotagging, Document Classification, Keyword Extraction

1. はじめに

教えて!goo^(注1)やYahoo!知恵袋^(注2)に代表される、多数のユーザが質問と回答(Q&A)をウェブサイト上でやり取りするQ&Aコミュニティは、新たな情報検索の手段および集合知のリポジトリとして急速に発展している。

現在、これらの代表的なQ&Aコミュニティにおいては、質問者が質問投稿時に選択するカテゴリによってQ&A文書の整理が行われている。しかし、複数のカテゴリに関連する質問でも1つのカテゴリしか選択できない上に、類似したカテゴリが多数存在するため、質問者が適切なカテゴリを選択できない事

例が多く、集合知のリポジトリとしての品質悪化を招いている。

近年では、カテゴリの限界を打破するものとして、タグと呼ばれる非階層的な短い索引語句を整理対象に複数付与する方法に注目が集まっている[1]。タグによる情報整理は、付与するタグの個数や種類に制限を通常設けないため、新奇な語句による検索や、カテゴリよりも具体的な検索の支援を行える利点がある。この利点を生かして、写真共有サイトのFlickr^(注3)やソーシャルブックマークのDelicious^(注4)などでは、多数のユーザが複数のタグを自由に付け加えていく方法であるフォークソノミー(folksonomy = folks + taxonomy; 民衆による分類法)を採用し、大きな成功を収めている[2],[3]。

(注1) : <http://oshiete.goo.ne.jp>

(注2) : <http://chiebukuro.yahoo.co.jp>

(注3) : <http://www.flickr.com>

(注4) : <http://delicious.com>

Q&A コミュニティにおいても、MSN QnA (旧名 Live Search QnA, 2009 年 5 月 21 日に試験運用終了) がユーザのタギングによる文書整理を採用していた。予め定められた多数のカテゴリの中から、質問を投稿したいカテゴリ、あるいは、興味のあるカテゴリを探さなければならない他の Q&A コミュニティに比べて、ユーザが複数のタグを自由に付与可能な MSN QnA では総称的なタグのみならず具体的なタグも付与され、複数の観点に基づく文書整理が実現されていた [4]。

これらの背景より、我々は、現在の代表的な Q&A コミュニティが蓄積する大量のタグ付けされていない文書に対して適切なタグ付けができれば、集合知のリポジトリとしての品質を大きく向上し、多数のユーザの情報検索や疑問解決に貢献できると考えた。そこで、本研究で我々は、予めカテゴリ分けされた文書集合を基に、文書に対してカテゴリ・主題・キーワードという各意味的階層において複数のタグを付与するシステムを提案する。提案システムは、文書分類とキーワード抽出の両アプローチを使用する特徴を持つ。

なお、自動的にタギングを行う研究については、ブログを対象に研究が進んでおり、文書中のキーワードを抽出してタグとする手法 [5]、協調フィルタリングにより類似エントリに付与されたタグを推薦する手法 [6], [7]、タグの付いたエントリ集合から学習を行って文書内容に適切なタグを推定する手法 [8]~[10] がある。このうち、キーワード抽出以外の手法は、ユーザがタギングを行っていない Q&A コミュニティにおいてはそのまま利用できない。また、これらの従来研究では、本研究のように意味的階層を考慮したタギングは行っていない。

本論文の構成を以下に示す。まず、2. でタギングに関連する Q&A コミュニティの特性について調査した結果を示す。次に、3. にて提案システムについて説明し、4. では提案システムの有効性を評価実験により示す。最後に、5. にて結論を述べる。

2. Q&A コミュニティの特性

本章では、Q&A 文書へのタギングに関係する Q&A コミュニティの特性として、カテゴリ別質問数の偏りと、語のバースト性 (同一文書内で同じ語が繰り返し現れやすい性質) [11], [12] について、総質問数が 450 万件を超える大規模 Q&A コミュニティである教えて！goo を例に挙げ説明する。

2.1 カテゴリ別質問数の偏り

教えて！goo のカテゴリは階層構造になっており、質問者が質問を投稿する際には、投稿先カテゴリが決定するまでサブカテゴリの選択を繰り返す。例えば、『ライフ > 恋愛・人生相談 > 恋愛相談』カテゴリに質問を投稿する際は、13 種類の大カテゴリから『ライフ』を選択し、次に、9 種類の中カテゴリから『恋愛・人生相談』を選択し、最後に 6 種類の小カテゴリから『恋愛相談』カテゴリを選択する (図 1)。投稿先カテゴリの階層の深さは 1 から 3 で、389 種類存在する (2009 年 12 月 27 日時点)。なお、投稿先カテゴリは 1 つしか指定できず、投稿後のカテゴリ変更はできない。

表 1 と図 2 に、2009 年 12 月 27 日に調査したカテゴリ別質問数の統計を示す。質問数が 20 万件を超えるカテゴリがある

恋愛・人生相談 恋愛相談 (220788) 夫婦・家族 (24793) 社会・職場 (16437) いじめ相談 (1819) シニアライフ (587) その他 (34172)	出産・育児 妊娠・出産 (80980) 育児 (17091) その他 (5552)	生活お役立ち 季節の行事 (696) 家事・生活 (13419) マナー・冠婚葬祭 (13658) 防犯・セキュリティ (902) コンビニ・ファーストフード (2003) 正月・年末年始 (533) リサイクル (596) その他 (27249)
結婚 段取り・結婚準備 (9543) 挙式・披露宴 (13139) ハネムーン・新生活 (1555) その他 (22274)	料理・グルメ 素材・食材 (19427) 地方特産・名産 (2335) 料理レシピ (21072) お酒 (5996) お茶・ドリンク (2390) スイーツ (6096) 食器・キッチン用品 (5722) その他 (15615)	家電製品 (37642) その他(ライフ) (118071)
住まい 建築・リフォーム (27474) インテリア・エクステリア (10112) ガーデニング (15061) DIY(日曜大工) (9041) 不動産・引っ越し (33828) その他 (29292)	ペット 犬 (29849) 猫 (3316) 小動物 (1074) 熱帯魚 (1438) その他 (14061)	

図 1 教えて！goo:『ライフ』カテゴリの階層構造。括弧内の数字は各カテゴリの質問数。太字は中カテゴリ、細字は小カテゴリを表す。

表 1 教えて！goo のカテゴリ別質問数の統計。

全カテゴリ (389 カテゴリ)					
合計	平均	標準偏差	最大値	最小値	中央値
4,670,262	12,005.8	20,057.6	220,788	2	5,491
「その他」カテゴリ (57 / 389 カテゴリ)					
合計	平均	標準偏差	最大値	最小値	中央値
1,095,334	19,216.4	22,474.1	118,071	338	11,424

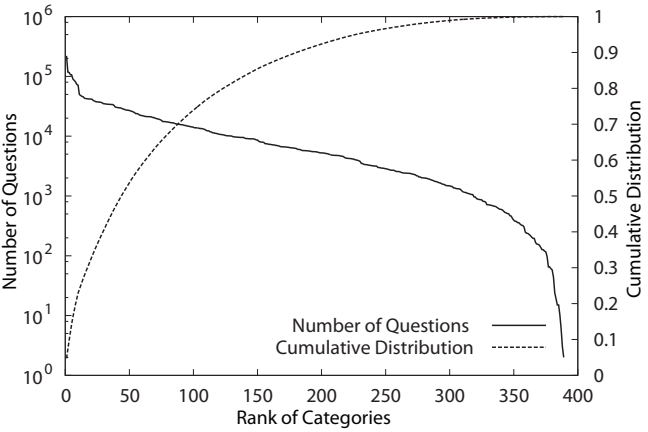


図 2 教えて！goo のカテゴリ別質問数と累積分布。

一方で、半数のカテゴリは質問数が 5,500 件に満たず、質問数が 10 件以下のカテゴリが 4 つあるなど、カテゴリ間の質問数に大きなばらつきがあることがわかる。質問数が非常に多いカテゴリについては、より具体的な内容を表すサブカテゴリを用意すべきと考える。また、全質問中 23.5% の質問が『その他』の文字列が含まれるカテゴリに属しており、ユーザの質問内容に即したカテゴリが十分に用意されていないことが分かる。しかし、カテゴリのさらなる細分化は、質問内容に最も適切なカテゴリの発見を著しく難しくしてしまう。

以上のように、現状のカテゴリ分類は効率よく機能していないと考える。質問者・回答者・閲覧者の多様な検索意図や知識・疑問に応え得る知識のリポジトリを実現するには、様々な意味的抽象度で Q&A 文書を整理する必要がある。そこで、本研究では、意味的階層構造を有するタグを複数付与して Q&A 文書を整理するシステムを提案する。

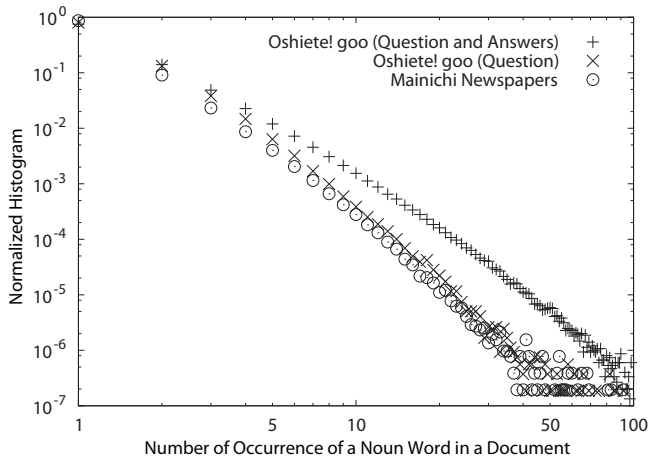


図3 1 文書において同じ名詞語が繰り返し出現する頻度の分布. 教えて!goo における 2009 年 1~3 月期の 218,971 文書と, CD-毎日新聞記事データ集 2008 年版の 73,231 記事から調査.

表2 教えて!goo の Q&A 文書と毎日新聞記事の, 1 文書に含まれる名詞語の総数に関する統計.

教えて!goo (質問文+全回答文)				
異なり語数		総語数		平均の 比率
平均	標準偏差	平均	標準偏差	
84.0	77.0	169.1	169.8	2.01
教えて!goo (質問文のみ)				
異なり語数		総語数		平均の 比率
平均	標準偏差	平均	標準偏差	
27.5	21.1	36.9	32.5	1.34
毎日新聞記事				
異なり語数		総語数		平均の 比率
平均	標準偏差	平均	標準偏差	
70.3	62.5	85.6	86.1	1.22

2.2 語のバースト性

Q&A コミュニティでは, 質問者と回答者が特定の主題について知識や意見を取り交わすため, 単一の著者によって執筆された他メディアの文書よりも, 同一文書内で同じ語が繰り返し出現しやすい性質 (語のバースト性[11]) を持つ. 語のバースト性は, 文書分類と語抽出の両方に影響する重要な特性である.

図3と表2に, 教えて!goo の 2009 年 1~3 月期に質問が投稿された 218,971 文書と, CD-毎日新聞記事データ集 2008 年版の 73,231 記事から調査した, 語のバースト性に関する統計を示す. 教えて!goo については, 質問文 (質問タイトルを含む) のみの場合と, 全文 (質問文+回答文) の両方を調査した. なお, 調査対象の Q&A 文書における平均回答数は 2.82, 標準偏差は 2.34 であった. また, 調査対象の語はタギング時の形態素解析条件に揃え名詞語のみ (連続する名詞語は 1 語とみなす) とし, 抽出には JTAG[13] を用いた.

調査結果が示すように, 教えて!goo の Q&A 文書の質問文のみと新聞記事は同様の傾向を示しているが, 全文 (質問文+回答文) においては, 質問文のみの場合よりも 1 文書中に同じ名詞語が繰り返し出現する確率が高くなること分かる. 本研究では, この特性を考慮して Q&A 文書へタギングを行う.

3. 提案システム

我々は, Q&A コミュニティに蓄積された大量のカテゴリ分けされた文書集合を基に, Q&A 文書に対して「カテゴリ」「主題」「キーワード」という 3 階層でのタギングを行うシステムを提案する. 図4に提案システムの概念図を示す.

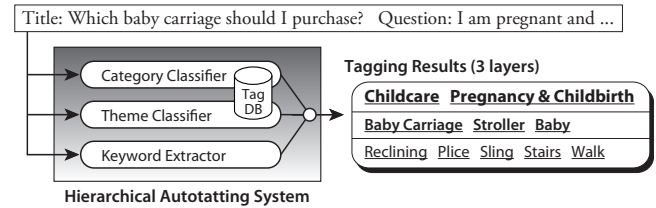


図4 提案システムの概念図

意味的抽象度の高い順にカテゴリ, 主題, キーワードであり, 各階層で付与するタグをカテゴリタグ, 主題タグ, キーワードタグと呼ぶ. 図4の例では, カテゴリタグを2つ (「Childcare (育児)」「Pregnancy & Childbirth (妊娠・出産)」) 主題タグを3つ (「Baby Carriage (ベビーカー)」など), キーワードタグを5つ (「Reclining (リクライニング)」など) 付与している.

なお, 意味的抽象度の高い情報整理を行うためのタグとして適切な語が文書中に常に出現するとは限らないため, カテゴリタグと主題タグについては, 語抽出ではなく文書分類によってタギングを行う. ここで, 文書分類によるタギングを行うためには, 予めタグの候補となる語を文書集合から選定しなければならない.

本章では, まずタグ候補の選定方法について説明する. 次にカテゴリタグ・主題タグの付与方法を示し, 最後にキーワードタグの付与方法を示す.

3.1 タグ候補の選定

まず, カテゴリタグについては, Q&A コミュニティのカテゴリ名をそのまま候補語とする. 次に, 主題タグについては, ブログのようにユーザが付与したタグを参考にできないため, 文書集合から各カテゴリで扱われる主題を表す語をタグの候補として抽出する.

なお, カテゴリ分けされた文書集合から, 文書頻度, カイ二乗統計量, Information Gain, 相関係数, Odds Ratio などの基準を用いて各カテゴリの特徴語を抽出する研究はこれまでも盛んに行われてきた[14], [15]. しかし, これらの基準により抽出される語は, 文書分類の素性としては適しているが, カテゴリの主題を表す語ではないため, タグの候補として相応しく無い.

そこで, 本研究ではカテゴリで扱われる主題を表す語を抽出するために, 文書タイトルに出現する語に着目する. $D_c^*(w)$ をカテゴリ c におけるタイトル中に語 w が出現する文書数, $D_c(w)$ をカテゴリ c を除く全てのカテゴリにおける全文 (タイトルを含む) 中に語 w が出現する文書数の合計, N_c をカテゴリ c に属する文書数, N_c をカテゴリ c を除く全てのカテゴリの文書数の合計としたとき, 提案システムは, Fisher's Exact Test[16] を用いて有意水準 α で片側検定を行い, $D_c^*(w)/N_c$ が

コンピューター > プログラミング > JavaScript (432 文書)	
THEME	画像 HTML JS JQUERY WINDOW FIREFOX 変数 フォーム ラジオボタン プルダウン テキストボックス
TDF	表示 方法 画像 値 取得 リンク JS 文字 HTML IE
IG	ALERT BODY STYLE VALUE ONCLICK DOCUMENT SCRIPTTYPE TEXT INPUTTYPE HTML
CC	DOCUMENT GETELEMENTBYID SCRIPTTYPE ONCLICK ALERT VARI BODY INPUTTYPE ONLOAD FUNCTION
OR	SRCELEMENT ATTACHEVENT VARO TAGNAME VARC CREATEELEMENT ONSUBMIT RETURNFUNCTION SCRIPTWINDOW PARENTNODE
デジタルライフ > ソフトウェア > フリーウェア (1,355 文書)	
THEME	フリーソフト 変換 ITUNES MP3 録音 SKYPE AVIUTL REALPLAYER 解凍 携帯動画変換
TDF	ファイル ダウンロード 変換 ITUNES ファイル 方法 表示 インストール MP3 動画
IG	相手 気 インストール 解凍 OS 事 ファイル 自分 フォルダ 人 編集 変換
CC	フリーソフト ファイル ダウンロード MP3 AUDACITY FLASHGET 仮面舞踏会 IRVINE EXPLZH LAME
OR	ダウンロード支援ソフト SAMURIZE AUTOCOPY AUDIOEDITOR 再配布禁止 アダルトHP LYRICSMaster 画像収集ソフト UNITEMOVIEV2.11 チューチューマウス
ライフ > 恋愛・人生相談 > 社会・職場 (1,583 文書)	
THEME	上司 同僚 先輩 人間関係 後輩 部下 新人 送別会 パワハラ セクハラ
TDF	仕事 上司 会社 女性 方 自分 同僚 私 先輩 長文
IG	同僚 相手 自分 上司 会社 あなた 先輩 私 こと 仕事
CC	上司 元人事労務 仕事 部下 同僚 部署 会社 先輩 社員 人間関係
OR	元人事労務 別職種 A部署 派遣事務職 ダメ上司 人材関連 B部署 承認承諾 相談ライン ロッカー内

図5 教えて!gooの各カテゴリから、提案手法 (THEME)、タイトル文書頻度 (TDF)、Information Gain (IG)、相関係数 (CC)、Odds Ratio (OR) により抽出された語 (スコア上位 10 語)。

$D_c(w)/N_c$ よりも有意に高いことが認められたとき、語 w をカテゴリ c の主題語とする。なお、抽出された主題語の良さを順序付けする際は、カテゴリで高頻度に扱われる主題語に高いスコアを付けるため、検定結果の p 値ではなく $D_c^*(w)$ の値を用いる。Fisher's Exact Test は、カイ二乗検定検定に比べて $D_c^*(w)$ や $D_c(w)$ の値が小さい場合でも正確な検定を実施可能な利点を持つ。

3.1.1 タグ候補選定に関する予備実験

教えて!goo:2009 年 1~3 月期における 218,971 文書から、提案手法 (THEME)、タイトル文書頻度 (TDF; タイトル中に語が出現する文書数)、Information Gain (IG)、相関係数 (CC)、Odds Ratio (OR) により語を抽出した結果を比較する^(注5)、なお、 $\alpha = 0.01$ 、 $D_c^*(w) \geq 5$ と設定した。形態素解析には JTAG [13] を使用し、名詞語 (連続する名詞語は 1 語とみなした) を抽出した。

a) 抽出された語の比較

図5に、各手法によって抽出された語を示す。まず、TDFでは「方法」「私」など、主題ではないがタイトルに高頻度で出現する語が抽出された。次に、IGはカテゴリに強く関連する語と関連しない語の両方を抽出する手法のため、『フリーウェア

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 大腸癌 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

図6 2009 年に最も検索された病名上位 50 語 (goo ランキング)。

表3 教えて!gooの『美容&健康>健康>病気/女性の病気/メンタルヘルス/その他(健康)』カテゴリから、提案手法 (THEME)、タイトル文書頻度 (TDF)、Information Gain (IG)、相関係数 (CC)、Odds Ratio (OR) により抽出された 185 語による 2009 年に最も検索された病名ワードの再現率。

手法	再現率 (上位 30 語)	再現率 (上位 50 語)
THEME	0.867 (26/30)	0.740 (37/50)
TDF	0.600 (18/30)	0.440 (22/50)
IG	0.367 (11/30)	0.240 (12/50)
CC	0.400 (12/30)	0.300 (15/50)
OR	0 (0/30)	0.040 (2/50)

ア』カテゴリにおいて「相手」「自分」などの語が抽出された。また、ORには出現頻度の低い語が多く抽出される問題があり、CCでも『JavaScript』カテゴリにおいて、主題とは関係のない「VARI」という語 (プログラム中の変数名) が抽出された。これらの手法に対して提案手法は、文書のタイトルに出現する語に着目してフィッシャーの正確確率検定を片側検定により行うことで、カテゴリの主題を表す語のみを抽出できた。

b) 検索ランキング上位語の再現率

Q&A 文書の主題語はユーザの疑問や興味を反映しているため、ユーザが良く検索する語と強い相関があると考えられる。そこで、図6に示す「2009 年に最も検索された病名ワード」^(注6)の上位語と、病名に関連する 4 つのカテゴリ『美容&健康>健康>病気/女性の病気/メンタルヘルス/その他(健康)』から抽出した語が一致するかを調査した。

ここで、各カテゴリには病名以外にも「目」「手術」などの主題語が含まれるため、この実験では検索語の再現率によって手法の良さを評価した。なお、提案手法が抽出した主題語の総数は 185 語であり、他手法の抽出語数もこれに揃えた。

表3に示すように、TDFは上位 30 語に比べ上位 50 語の再現率が大きく低下した。これは、ランキング下位の病名よりも「方法」「これ」などの非主題語がタイトルに多く出現したためである。さらに、ORは「ベーチェット病」などランキング下位の病名は抽出できていたが、ランキング上位の語を全く抽出できなかった。提案手法は、これらの手法に比べて有意に高い再現率を実現した (二項検定に基づく対応のある比率の差の検定、上位 30 語: $p < .05$, 上位 50 語: $p < .001$)。

(注5): 相関係数² = χ^2 の関係より、相関係数はカテゴリに強く関連する語のみを抽出するカイ二乗統計量基準と捉えることができるため、カイ二乗統計量は実験対象から除外した。

(注6): 出展: goo ランキング (http://ranking.goo.ne.jp/ranking/n09/n2009-health_keyword/)。goo ヘルスケア (<http://health.goo.ne.jp>) が 2009 年 1 月 1 日から 2009 年 10 月 31 日の期間に集計。

3.2 カテゴリタグ・主題タグの付与

意味的階層構造を有する複数のタグを適切に付与するためには、式 (1) に示す入力文書 d が、カテゴリ $c \in C$ に属し、主題 $t \in T_c$ を扱っている確率 $p(c, t|d)$ が高いカテゴリ c と主題 t をタグとして複数選択すれば良い。

$$p(c, t|d) = \frac{p(d|c, t)p(c, t)}{p(d)} \quad (1)$$

しかし、カテゴリ c に属し主題語 t を含む訓練文書は大量に得られないため、 $p(c, t|d)$ の推定精度は悪い。そこで、比較的高い推定精度が期待できる確率として、入力文書がカテゴリ c に属する確率：

$$p(c|d) = \frac{p(d|c)p(c)}{p(d)} \quad (2)$$

を求め、 $p(c, t|d)$ と $p(c|d)$ を掛け合わせたスコア：

$$S(c, t) = p(c, t|d)p(c|d) \quad (3)$$

の値が高いカテゴリ c と主題 t をそれぞれカテゴリタグ、主題タグとして付与することで、タギング精度の向上を図る。

3.2.1 $p(c|d)$ の推定

式 (2) における $p(c)$ は一様分布とし、 $p(d|c)$ は入力文書 d 中の語 w の出現頻度 $f(w)$ を使用しない多項モデルのナイーブベイズ分類器 [12] により推定する。

$$p(d|c) = p(|d|)! \prod_{w \in W} p(w|c)^{\min\{f(w), 1\}} \quad (4)$$

$$p(w|c) = \frac{\gamma + D_c(w)}{\gamma|V| + \sum_{w' \in V} D_c(w')} \quad (5)$$

ここで、 $D_c(w)$ はカテゴリ c において語 w が出現する文書数、 γ は平滑化パラメータ (> 0) である。

短い Q&A 文書でも精度良く分類するため、式 (4) において確率推定に使用する語の集合 W は、予め各カテゴリの特徴語を選択しておくのではなく、入力文書 d に含まれる K 個の語で構成する。 K 個の内訳は、 d のタイトルに含まれる J 個の語と、 d の本文に含まれる語のうち式 (6) に示すエントロピー $H(w)$ が低い $K - J$ 個の語とする。

$$H(w) = \frac{p(w|c)}{\sum_{c' \in C} p(w|c')} \log \left(\frac{p(w|c)}{\sum_{c' \in C} p(w|c')} \right) \quad (6)$$

式 (6) は、スパムメールフィルタリングにおいて Graham が提案した特徴語の選択手法 [17] を多クラスに拡張したものともみなすことができ、特定のカテゴリに偏って出現する語を抽出できる。

なお、2. にて示したように、Q&A 文書は、1 文書中において同じ語が繰り返し出現しやすい性質（語のバースト性 [11]）を持つ。多項モデルのナイーブベイズ分類器はこのバースト性を考慮していないため、同じ語が繰り返し登場する文書の確率をかなり低く推定してしまう。そこで、提案システムでは、出現頻度を使用しない多項モデルのナイーブベイズ分類器を採用した。また、ナイーブベイズ分類器は逐次学習が可能であり、入力文書ごとに分類時に使用する素性を選択することができる。常に新規文書が生成され、文書長が非常に短い文書が多数含まれる Q&A コミュニティの学習・分類にはナイーブベイズ分類器が非常に適していると考えられる。

3.2.2 $p(c, t|d)$ の推定

次に、 $p(c, t|d)$ については、計算量削減のため、 $p(c|d)$ の値が高い M 個のカテゴリ集合 $C_M \subset C$ についてのみ推定する。なお、式 (1) における $p(c, t)$ は一様分布とし、 $p(d|c, t)$ は以下のように推定する。

$$p(d|c, t) \propto \prod_{w \in W} p(w|c, t) \quad (7)$$

$$p(w|c, t) = \frac{s \cdot p^* + D_{c,t}(w)}{s + N_{c,t}} \quad (8)$$

$$p^* = \frac{\sum_{c \in C_M} \sum_{t \in T_c} D_{c,t}(w)}{\sum_{c \in C_M} \sum_{t \in T_c} N_{c,t}} \quad (9)$$

ここで、 $N_{c,t}$ はカテゴリ c に属し主題語 t を含む文書数、 $D_{c,t}(w)$ はカテゴリ c に属し主題語 t を含む文書のうち語 w が出現する文書数、 s は平滑化パラメータである。式 (8) に示す $p(w|c, t)$ の推定においては $N_{c,t}(w)$ が小さくなるカテゴリ c と主題 t の組み合わせが多数存在するため、 p^* と s による平滑化を行っている。

最終的に、 $S(c, t)$ は計算の簡略化のため c と t に無関係な項を除いて、

$$S(c, t) = p(c, t|d)p(c|d) \propto \prod_{w \in W} p(w|c, t)p(w|c) \quad (10)$$

として求める。

3.3 キーワードタグの付与

TF-IDF 手法に基づき、式 (11) に示す $S(w)$ が高い語 $w \in d$ を、キーワードタグとして付与する。

$$S(w) = \log(1 + f(w) + t(w)) \cdot \log \left(\frac{N}{1 + D(w)} \right) \quad (11)$$

ここで、 $f(w)$ は入力文書 d の全文中に語 w が出現する頻度を表し、 $t(w)$ は語 w がタイトルに出現するとき 1 を、出現しないとき 0 をとる。なお、ここでも語のバースト性を考慮し、対数による出現頻度 $f(w)$ の変換を行った。また、カテゴリタグ、主題タグに語 w が既に含まれている場合は、キーワードタグに含めない。

4. 評価実験

提案システムの有効性を示すため、教えて！goo：2008 年 4 月から 2009 年 3 月の一年間における 775,863 文書を訓練文書集合として、2009 年 4 月からの 3,000 テスト文書に対してタギングを行った（4.3 節の実験のみテスト文書の個数と投稿された期間は異なる）。タギングに関するパラメータは $K = 15$ 、 $M = 10$ 、 $\gamma = 0.001$ 、 $s = 50$ と設定した。なお、訓練文書集合から抽出した各カテゴリの主題語の総数は 18,638 語、異なり語数は 14,103 語であった（ $\alpha = 0.01$ 、 $D_c^*(w) \geq 5$ ）。形態素解析には JTAG [13] を使用し、名詞語（連続する名詞語は 1 語とみなした）を抽出した。

4.1 タギングの例と考察

提案システムにより Q&A 文書^(注7)へカテゴリタグを 2 個、主

(注7) : <http://oshiete1.goo.ne.jp/qa4842983.html>

<http://oshiete1.goo.ne.jp/qa4843047.html>

<http://oshiete1.goo.ne.jp/qa4843483.html>

表4 教えて!goo のQ&A 文書 (上から順に質問番号 4842983, 4843047, 4843483) のタイトルと, 提案システムによるタギング例. 鍵括弧内はユーザが選択したカテゴリ名.

ゴールデンハムスターの死『ライフ > ペット > 小動物』	
カテゴリタグ	小動物 その他 (ペット)
主題タグ	ハムスター 餌 ペットショップ
キーワードタグ	巣箱 頬袋 冬眠 ゴールデンハムスター 暖房
5mm 厚の亚克力板カットと加工について『趣味 > その他 (趣味)』	
カテゴリタグ	DIY (日曜大工) 手芸・ホビークラフト
主題タグ	亚克力板 板木
キーワードタグ	亚克力板カット 厚 5MM 加工 角おとし
青の洞窟の撮り方は?『デジタルライフ > その他 (デジタルライフ)』	
カテゴリタグ	デジタルカメラ カメラ全般
主題タグ	撮り方 デジイチ 三脚
キーワードタグ	青の洞窟 シャッタースピード オート 露光 光量補正

題タグを3個, キーワードタグを5個付与した結果の例を表4に示す. 結果より, 提案システムが「小動物 > ハムスター > ゴールデンハムスター」のように意味的階層構造を有するタグを付与できていることがわかる (質問番号 4842983). また, 質問番号 4843047 (タイトル: 5mm 厚の亚克力板カットと加工について) のように, ユーザが選択したカテゴリ『その他 (趣味)』よりも適切なカテゴリタグ『DIY (日曜大工)』が付与される事例も多く存在した. また, 質問番号 4843483 (タイトル: 青の洞窟の撮り方は?) のように, 内容に関連するカテゴリが複数あると考えられる場合にはマルチタギングの効果が大きい.

ここで, 表4に示す例のうち, 「ハムスター」や「頬袋」は一部のユーザにとっては具体的すぎるかもしれないが, 一方で「ペット」は他のユーザにとって総称的すぎるかもしれないように, 適切なタグの意味抽象度の基準はユーザによって大きく異なる. 従って, 提案システムが階層的にタグを付与することによって, 様々な意味抽象度のタグを用意することは非常に有効であると考えられる. また, 文書分類とキーワード抽出の両アプローチを取ることによって, 文書中に出現しない語と, 過去のどの文書にも出現していない新奇語の両方をタグとして付与できる点は, 従来のブログを対象としたオートタギング手法 [5]~[10] には無い大きな利点であると考えられる.

4.2 カテゴリタグに関する評価

ユーザが選択したカテゴリと, 提案システムが付与したカテゴリタグの一致率について調査し, 提案システムが用いるスコア関数 (式 (10)), 単語選択手法 (式 (6)), 語の出現頻度を使用しないナイーブベイズ分類器 (式 (4)–(5)) について評価する.

4.2.1 スコア関数について

式 (10) に示すスコア $S(c, t) = p(c, t|d)p(c|d)$ を

$$(1) \quad S_1(c, t) = p(c, t|d)$$

$$(2) \quad S_2(c, t) = p(c|d)$$

に変更した場合と比較した.

図7に示す実験結果より, スコアとして $p(c, t|d)$ と $p(c|d)$ の乗算結果を用いる式 (10) を採用することで, どちらか片方の確率をスコアとする場合よりも, カテゴリタグとユーザ選択カテゴリの一致率が高くなった. 特に, カテゴリタグの付与数が1から3のときは, $S(c, t)$ を用いた場合の一致率が, 両ベース

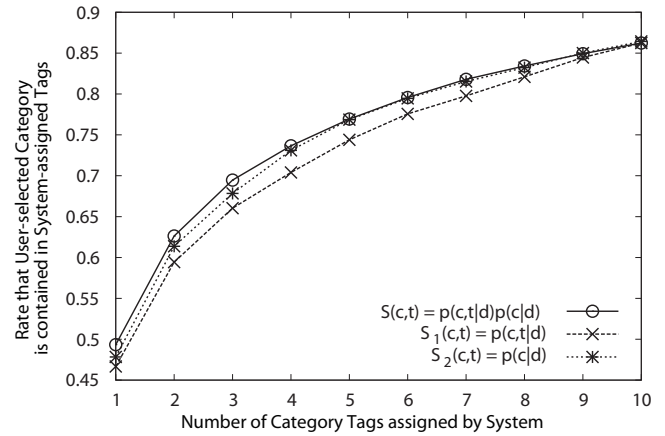


図7 ユーザが選択したカテゴリと提案システムが付与したカテゴリタグの一致率. 提案システムが付与するカテゴリタグの個数を1から10まで変化させた.

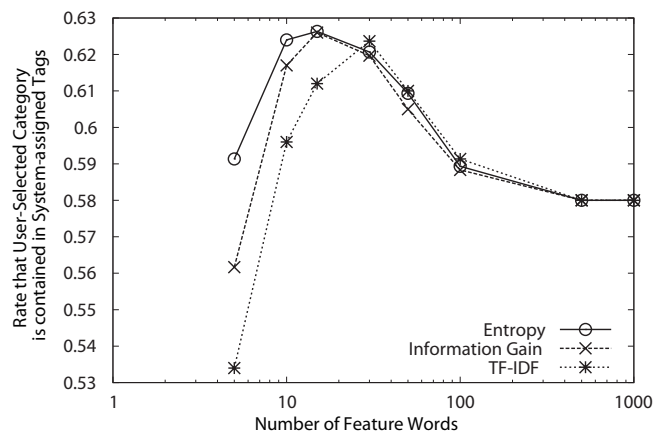


図8 ユーザが選択したカテゴリと提案システムが付与したカテゴリタグの一致率. 入力文書からエントロピー, Information Gain, TF-IDF により K 個の語を選択してタギングを実施した. 付与カテゴリタグ数は2.

インを用いた場合の一致率よりも有意に高かった (二項検定に基づく対応のある比率の差の検定, $p < .001$).

4.2.2 単語選択手法について

次に, $p(c|d)$ と $p(c, t|d)$ の推定における単語選択手法について, 提案システムが用いるエントロピー (式 (6)), Information Gain (IG), TF-IDF を比較した. この実験では, 入力文書から選択する語数 K の値を変えながら実験を行った.

図8に示す実験結果より, 入力文書に出現する全ての語を用いるのではなく, 特徴的な語のみを用いて $p(c|d)$ と $p(c, t|d)$ を推定することにより, カテゴリタグの付与精度が良くなることわかる. 特に, 選択単語数が5のとき, エントロピーを用いた場合のカテゴリタグとユーザ選択カテゴリの一致率が, IG と TFIDF を用いた場合よりも有意に高かった (二項検定に基づく対応のある比率の差の検定, $p < .001$) ことから, エントロピーによる単語選択が, 精度の良い文書分類を実現する語に高いスコアを付与できていることがわかる.

4.2.3 ナイーブベイズ分類器について

最後に, $p(c|d)$ の推定に用いる多項モデルのナイーブベイズ

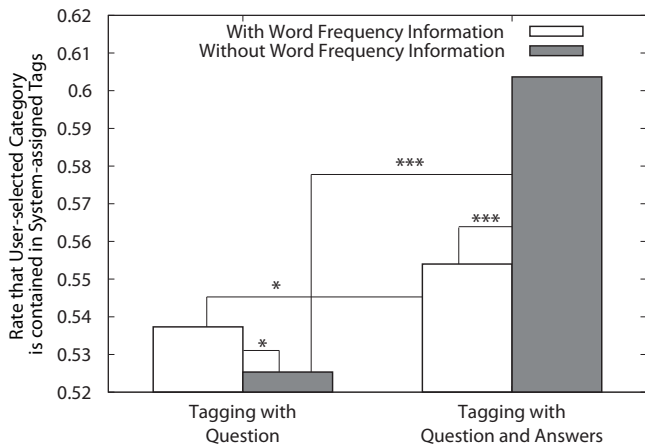


図9 ユーザが選択したカテゴリと提案システムが付与したカテゴリタグの一致率。多項モデルにおける語の出現頻度の使用の有無の影響を比較した。図中の'*'記号は一致率に有意差があることを示す (***) は $p < .001$, * は $p < .05$)。付与カテゴリタグ数は2. $S_2(c, t) = p(c|d)$ を使用。

分類器を、語の出現頻度を使用しない場合（提案システムが使用：式(4)-(5)）と、語の出現頻度を使用する場合（式(12)-(13)）とで比較した。

$$p(d|c) = p(|d|)|d|! \prod_{w \in W} p(w|c)^{f(w)} \quad (12)$$

$$p(w|c) = \frac{\gamma + T_c(w)}{\gamma|V| + \sum_{w' \in V} T_c(w')} \quad (13)$$

ここで、 $T_c(w)$ は、カテゴリ c における語 w の出現頻度である。

この実験では、語のバースト性のナイーブベイズ分類器への影響を調査するため、2009年4月以降のQ&A文書から回答数が5以上のものを3,000文書集め、質問文のみを入力文書 d とする場合と、質問文と回答文の両方を d とする場合でタギングを行った。なお、 $p(c|d)$ の推定の良さに着目するため、スコア関数には $S_2(c, t) = p(c|d)$ を用いた。

質問文のみを入力文書とした場合（バースト性が低い）には、語の出現頻度を使用するナイーブベイズ分類器のほうが有意に一致率が高かった。しかし、質問文と回答文の両方を入力文書とした場合（バースト性が高い）は、語の出現頻度を使用しないナイーブベイズ分類器のほうが有意に一致率が高かった（二項検定に基づく対応のある比率の差の検定、図9参照）。また、質問文と回答文の両方を入力文書とした場合は、質問文のみを入力文書とした場合に比べて、どちらのナイーブベイズ分類器を使用した場合も有意に一致率が高かった。今後の課題として、回答数に応じた語の出現頻度の扱いの変更を検討したい。

4.3 主題タグ・キーワードタグに関する評価

教えて!gooではユーザによるタギングが行われていないため、日本の大規模ソーシャルブックマークであるはてなブックマークのユーザが教えて!gooのQ&A文書に付与したタグ^(注8)と、提案システムが同じQ&A文書に付与した主題タグ・キーワードタグを比較した。

なぜビストバイクなのか『趣味 > 車 > 自転車』(質問番号 4285059)

bicycle(7) 自転車 (4) bike(2) 雑学 (1) ふむふむ (1) 米国 (1) なるほど (1) fixie(1) あとで (1)

カテゴリタグ	自転車 その他 (アウトドア)
主題タグ	ビストバイク ビスト 後輪 ブレーキ クロスバイク
キーワードタグ	メッセンジャー ノーブレーキ ニューヨーク アメリカ

なぜ月にいかないのか?『学問&教育 > 科学』(質問番号 390548)

宇宙 (7) 雑学 (3) 社会 (2) science(2) history(1) 天文 (1) サイエンス (1) technology(1) 経済金融 (1) 科学 (1) 教養 (1) 天文学 (1)

カテゴリタグ	天文学 物理学
主題タグ	宇宙 地球 重力 月 火星
キーワードタグ	ロケット 探査 SELENE アポロ セレーネ

図10 はてなブックマークのユーザによる教えて!gooのQ&A文書へのタギング結果と、提案システムによるタギング結果。括弧内の数字は同じタグを付与したユーザ数を表す。

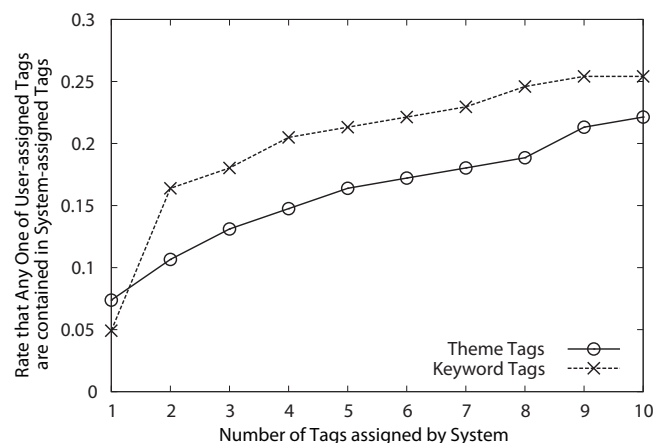


図11 はてなブックマークのユーザが付与したタグが、提案システムが付与したタグに含まれる割合。提案システムが付与するタグの個数を1から10まで変化させた。

図10に示す、はてなブックマークのユーザによる教えて!gooのQ&A文書^(注9)へのタギング例より、はてなブックマークのユーザによるタギング結果には多数のユーザにより分類・整理された結果、「自転車」と「fixie (自転車メーカー名)」、「宇宙」と「雑学」など、提案システムと同様に、様々な意味抽象度かつ複数の観点に基付いたタグが付与されていることが分かる。また、「自転車」「宇宙」など幾つかのタグが提案システムとユーザで一致していることが分かる。

次に、図11に、はてなブックマークのユーザが付与したタグのどれか一つが、提案システムが付与した主題タグ・キーワードタグに含まれる割合を示す。ここでは、2人以上のユーザが併せて3つ以上のタグを付与した122個のQ&A文書を対象に実験を実施した。なお、実験に用いたデータにおける、1文書に付与されたタグの種類数の最小値・第1四分位点は3、中央値は4、第3四分位点は7、最大値は112であった。同様に、1文書にタグを付与したユーザ数の最小値・第1四分位点・中央

(注9) : <http://oshiete1.goo.ne.jp/qa4285059.html>

<http://oshiete1.goo.ne.jp/qa390548.html>

(注8) : <http://b.hatena.ne.jp/> より、2009年2月3日に収集。

値は 2, 第 3 四分位点は 4, 最大値は 265 であった。

評価に用いたデータ中の付与タグ数・付与ユーザ数が少ないため、ユーザが付与したタグと提案システムが付与した主題・キーワードタグが一致する割合の値はカテゴリタグの場合に比べて低いが、主題タグ・キーワードタグのどちらも、タグの付与個数を多くすることによって、ユーザが付与したタグとの一致率が単調に増加していることがわかる。主題タグ・キーワードタグのどちらも、類似したタグを付与してしまう問題（「ピストバイク」と「ピスト」など）が残されており、タグの多様性を向上することが今後の課題である。

5. おわりに

我々は、現在の代表的な Q&A コミュニティにて用いられているカテゴリ分類の問題点を整理し、Q&A 文書にカテゴリ・主題・キーワードという意味的階層構造を有するタグを自動的に複数付与可能なシステムを本論文にて提案した。さらに、カテゴリ分けされた文書集合から、各カテゴリで扱われる主題を表す語を抽出する手法を提案した。通常は非階層的に行うタギングに階層の概念を導入し、意味的抽象度の異なるタグを付与することで、ユーザの様々な検索意図に応えることができる。また、提案システムによって文書に付与されたタグを介して文書間の繋がりが生まれるため、人気の無いカテゴリに投稿された質問に対する回答者の閲覧機会を増やすことができる。提案システムの評価のため、教えて！goo で実際に投稿された Q&A 文書を用いて実験を実施し、各カテゴリから主題語を正しく抽出したことと、Q&A 文書の内容に即したタグを付与できたことを示した。近年、Q&A コミュニティは学術的にも高い注目を集めており [18]~[21]、我々が実施した Q&A コミュニティの特性に関する調査と、Q&A 文書への階層的オートタギングについては新たな成果であり、本研究の学術的貢献度は大きいと考える。今後の課題については、4. にて挙げたものに加えて、同義語・類似語の解決、タギング個数の自動決定、タギング結果への確信度導入を挙げる。

さらに、各カテゴリで扱われる主題を表す語を抽出する手法については、Q&A 文書に限らず、カテゴリ分けされた文書集合であればどのような文書に対しても適用できる。カテゴリの特徴的な語を抽出する他手法と比べて、提案手法は情報検索に適した語を抽出できるため、文書集合の内容を表すタグクラウドなど様々な用途に利用できる。

また、我々は Nelson が提唱した ZigZag インタフェース [22] が Q&A 文書の閲覧に適していると考え、提案システムにより付与した階層的タグを軸とするスムーズな文書閲覧を実現することで、Q&A コミュニティの集合知のリポジトリとしての品質向上へのさらなる貢献を目指している [23]。

文 献

- [1] C. Shirky, "Ontology is overrated: Categories, links, and tags," http://shirky.com/writings/ontology_overrated.html, 2005.
- [2] S.A. Golder, and B.A. Huberman, "Usage patterns of collaborative tagging systems," ACM SIGKDD Explor. Newsletter, vol.32, no.2, pp.198-208, 2006.
- [3] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "Ht06, tagging pa-

- per, taxonomy, flickr, academic article, to read," Proc. 17rd Conference on Hypertext and Hypermedia, pp.31-40, 2006.
- [4] E.M. Rodrigues, N. Milic-Frayling, and B. Fortuna, "Social tagging behaviour in community-driven question answering," Proc. 2008 IEEE/WIC/ACM International Conference on Web Intelligence, pp.112-119, 2008.
- [5] C.H. Brooks, and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering," Proc. 15th International Conference on World Wide Web, pp.625-632, 2006.
- [6] G. Mishne, "AutoTag: A collaborative approach to automated tag assignment for weblog posts," Proc. 15th International Conference on World Wide Web, pp.953-954, 2006.
- [7] S.C. Sood, S.H. Owsley, K.J. Hammond, and L. Birnbaum, "TagAssist: Automatic tag suggestion for blog posts," Proc. International Conference on Weblogs and Social Media, 2007.
- [8] T. Ohkura, Y. Kiyota, and H. Nakagawa, "Browsing system for weblog articles based on automated folksonomy," Proc. WWW 2006 Works. Weblogging Ecosystem: Aggregation, Analysis, and Dynamics, 2006.
- [9] S. Fujimura, K. Fujimura, and H. Okuda, "Blogosonomy: Autotagging any text using blogger's knowledge," Proc. 2007 IEEE/WIC/ACM International Conference on Web Intelligence, pp.205-212, 2007.
- [10] X. Si, and M. Sun, "Tag-LDA for scalable real-time tag recommendation," Information and Computational Science, vol.6, no.1, pp.23-31, 2009.
- [11] S.M. Katz, "Distribution of content words and phrases in text and language modelling," Natural Language Engineering, vol.2, no.1, pp.15-59, 1996.
- [12] K.M. Schneider, "Techniques for improving the performance of naive bayes for text classification," Proc. 6th International Conference on Intelligent Text Processing and Computational Linguistics, pp.682-693, 2005.
- [13] T. Fuchi, and S. Takagi, "Japanese morphological analyzer using word co-occurrence -JTAG-, " Proc. 17th International Conference on Computational Linguistics, pp.409-413, 1998.
- [14] Y. Yang, and J.O. Pedersen, "A comparative study on feature selection in text categorization," Proc. 14th International Conference on Machine Learning, pp.412-420, 1997.
- [15] G. Foreman, "An extensive empirical study of feature selection metrics for text classification," Journal of Machine Learning Research, vol.3, pp.1289-1305, 2003.
- [16] A. Agresti, "A survey of exact inference for contingency tables," Statistical Science, vol.7, no.1, pp.131-153, 1992.
- [17] P. Graham, "A plan for spam," , 2002.
- [18] L.A. Adamic, J. Zhang, E. Bakshy, and M.S. Ackerman, "Knowledge sharing and yahoo answers: everyone knows something," Proceeding of 17th International Conference on World Wide Web, pp.665-674, 2008.
- [19] F.M. Harper, D. Moy, and J.A. Konstan, "Facts or friends?: distinguishing informational and conversational questions in social q&a sites," Proc. 27th International Conference on Human Factors in Computing Systems, pp.759-768, 2009.
- [20] K.K. Nam, M.S. Ackerman, and L.A. Adamic, "Questions in, knowledge in?: a study of naver's question answering community," Proc. 27th International Conference on Human Factors in Computing Systems, pp.779-788, 2009.
- [21] E. Agichtein, Y. Liu, and J. Bian, "Modeling information-seeker satisfaction in community question answering," ACM Transactions on Knowledge Discovery from Data, vol.3, no.2, pp.131-153, 2009.
- [22] T.H. Nelson, "Cosmology for a different computer universe: data model, mechanisms, virtual machine and visualization infrastructure," Journal of Digital Information, vol.5, no.1, 2004.
- [23] 後藤達弥, 甲谷優, 西田京介, 藤村考, "ZigZag インタフェースによる QA 閲覧システムの実現," 第 2 回データ工学と情報マネジメントに関するフォーラム, 2010, A8-5.