研究指導　中澤真　准教授

Q&A サイトにおけるユーザの行動履歴を用いた質問推薦アルゴリズム

吉澤　杏子

1 はじめに

近年,「Yahoo!知恵袋[1]」や「教えて!goo[2]」をはじめとするQ&A サイトが情報収集の手段の1つとして活発に利用されている. Q&A サイトは質問も回答も自然言語文によるやり取りで成り立つため, ユーザ同士のコミュニケーションや, 知識の正しさや意見の妥当性を確認するなどの利用に適している[1]. また, その規模は拡大しており, 例として「Yahoo!知恵袋」では, サイトに投稿される質問の総数が現在約 5400 万件を超え, わずか 1 時間で約 7000 件もの回答がユーザから投稿されるという巨大サービスとなっている. しかし, このような膨大な質問が投稿されるようになったことで, ユーザが回答可能と判断できる質問を見逃す機会が増え, 回答率の低下に繋がるということが懸念される. また, 専門知識を持つユーザなどに質問を見逃されてしまえば, 回答の質の低下にも繋がると考えられる.

そこで本研究では, 回答率及び回答の質を向上させるために, 膨大な質問の中からユーザごとに回答可能であると判断される質問のみを推薦するシステムを提案する. まず, ユーザがどのような事柄に興味関心があるのかを見出すため, ユーザごとに回答履歴から過去に回答したことのある質問文を用いて, 頻出キーワードの抽出を行った. そして, そのキーワードを多く含む質問ならば回答可能である確率が高いと考えた. しかし, 複数の事柄に興味関心があり, 回答可能な質問が多岐にわたるユーザの場合, 全く関連性のないキーワード同士を同じ事柄に関するものであると誤って判断しかねない. この誤認を防ぐためには, 頻出キーワードのグループ分けを行う必要がある. そこで, 共起分析を行いキーワード間の関連性を見出すことに取り組み, より精度の高いユーザの行動履歴を用いた新たな質問推薦システムを目指した.

[1] http://chiebukuro.yahoo.co.jp/
[2] http://oshiete.goo.ne.jp/

2 Q&A サイトの現状

Q&A サイトとは, オンラインのユーザが質問を投稿し、それに対して他のユーザが回答を投稿することで悩みや疑問を解決し合うソーシャルメディア[3]の一つである. 本節ではQ&A サイトの利用状況と, 利用者の増加によって起きる問題点について論じる.

2.1 Q&A サイトの利用状況に対する考察

今日における情報検索の手段としては,「Yahoo!JAPAN[4]」や「Google[5]」といった検索エンジンサイトを利用することが一般的である. こういった通常の検索エンジンは, インターネット上に存在する膨大な数のWebページの中から瞬時に目的に合ったWebページのみを表示することが出来るため, あるキーワードに関連する情報を網羅的に集めたい場合に有効な手段である. しかし, 単純にキーワードを組み合わせただけでは表すことのできない複雑な質問においては, 検索することが難しいというデメリットがある. 一方 Q&A サイトは, 検索エンジンほどのレスポンスの速さは無いものの, 人間同士のやり取りのため, 自然言語文により細かいニュアンスを伝えたり, 意図を汲み取ったりすることが可能である. このような利点から, Q&A サイトは検索エンジンが苦手とする自然言語文による情報検索の手段として多く用いられている.

実際, OCN が 2010 年 2 月 25 日～3 月 1 日に実施した調査によると, 6 割以上のユーザが Q&A サイトの利用経験があると回答しており[2], その利用率はかなり高い. また, 既存の代表的な Q&A サイトである「Yahoo!知恵袋」や「教えて!goo」では, 図 1 に示す通り 2006 年 3 月から 2008 年 3 月にかけて利用者数が年々増大しており[3], Q&A サイトの需要が高まってきていることがうかがえる.

[3] Web 上で提供されるサービスのうち, ユーザの積極的な参加によって成り立ち, ユーザ間のコミュニケーションをサービスの主要価値として提供するサービスの総称.
[4] http://www.yahoo.co.jp/
[5] http://www.google.co.jp/

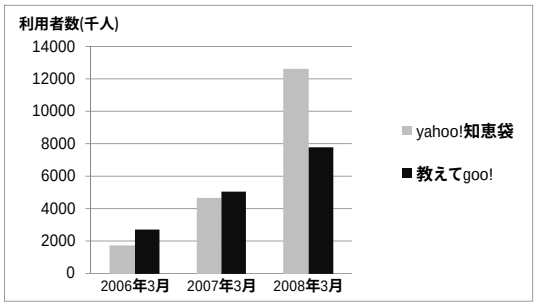

図 1 「Yahoo!知恵袋」「教えて！goo」の利用者数推移

これらのデータから，ユーザは検索エンジンだけでは得ることの出来ない情報について，Q&A サイトを介して他のユーザに問いかけるという手段を選択することが一般的になりつつあるといえる．すなわち，それだけ多くの質問が，ユーザによって投稿される場面が増えたことを意味している．

2.2 既存の Q&A サイトにおける問題点

Q&A サイトのニーズの増加に伴い投稿される質問の総数も増加し，その中からユーザ自身が答えられそうな質問を効率的に探し出すことが難しくなりつつある．

一般に Q&A サイトを回答者として利用する場合には，キーワード検索かカテゴリによる絞り込みによって，回答可能な質問を探すことになる．しかし，前者の方法ではありふれたキーワードや少ないキーワードで検索した場合うまく絞り込みが出来ず，検索結果が膨大に表示されてしまう．そのため，ユーザはより適切なキーワードの組み合わせを考えなければならない．しかし，株式会社メディアインタラクティブの調査[4]によると，ユーザの 92%が情報を検索する際に 2 つ以下のキーワードを入力している．このことから，ユーザがその都度 3 つ以上のキーワードの組み合わせを考え，絞り込んだ情報を表示させようとすることは容易ではないといえる．また，後者の方法では既存サイトのカテゴリ分けが細分化されていないため，1 つのカテゴリ内に膨大な数の質問が表示されてしまう．例えば「Yahoo!知恵袋」における最小カテゴリ[6]の一つ「恋愛相談」では，約 1 万 3000 件もの質問が登録されており[7]，回答者はこの膨大な質問の中から自分に適したものを探す作業を強いられる．結果としてどちらのケースにおいても，回答者は自分に適した質問を手作業によって絞り込む必要が生じ，効率的な回答を妨げる要因となっている．

質問の効率的な検索が困難になると，ユーザが本来ならば答えられる質問を見逃すケースも増え，結果として回答率の低下を招いてしまう．また，回答率の低下は回答の質も低下させてしまうであろう．そのため，これらの問題に対する解決策を検討する必要がある．

3 質問を効率的に検索するための既存のアプローチ

回答ユーザが自分に適した質問を探し出す作業負担の問題に対し，情報推薦機能を用いて問題の解決を図る試みがなされている．ここでは既存のシステムとして「Yahoo!知恵袋」と「BIGLOBE なんでも相談室[8]」の 2 つの Q&A サイトで導入されている推薦機能について検証し，その問題点について検討する．

「Yahoo!知恵袋」における質問推薦システムでは，ユーザが直近に回答した質問と同一のカテゴリの質問を一覧表示するという単純な仕組みであり，ユーザの特性などを考慮して質問を抽出するような高度な処理はなされていない．この方法では，膨大な数の質問の中からユーザ自身が答えられそうな質問とそうでない質問を手動で取捨選択しなければならないという問題点を解決できていない．

一方の「BIGLOBE なんでも相談室」における質問推薦システムでは，それぞれの質問ごとに類似した質問を 3 つ表示するという機能が導入されている．しかし，この質問推薦システムを利用するためには，まずユーザが回答可能な質問を手動で探さなくてはならない．そのため，ユーザの作業負担の問題は改善されていないといえる．

以上の結果から，過去に回答した質問のカテゴリ情報だけでは，ユーザの特性を十分に把握することができないということが考察できる．また，質問を提示する際にカテゴリ単位で提示したのでは，絞り込みに寄与しないということが分かった．さらに，質問文に対して類似質問を提示する方法では，ユーザの作業負担は軽減されないということも明らかになった．以上を踏まえ，カテゴリ以外の情報を用いてユーザの特性を把握し，そのユーザが回答できる質問の

[6] 最も細分化されたカテゴリ．Yahoo!知恵袋では，生き方と恋愛，人間関係の悩み＞恋愛相談，人間関係の悩み＞恋愛相談と細分化されている．

[7] 2011 年 2 月 2 日現在の調査データ．

[8] http://soudan.biglobe.ne.jp/

みを的確に推薦できるシステムを提案する.

4 問題点を改善した推薦システムの構築

ユーザに適した質問を自動的に推薦するシステムを構築するためには, どのような事柄について興味関心があり, 知識を有しているのかといったユーザの特性を把握する仕組みが必要である. しかし, 過去の回答履歴のカテゴリ情報を用いるだけではユーザの特性を十分に把握することができない. そのため, どのような内容の質問に回答したのかという中身にまで踏み込む必要がある. 本節では, 頻出キーワードを用いて質問の中身を分析し, 抽出したキーワード同士の関連性を見出すための取り組みについて述べる.

4.1 頻出キーワードの抽出によるユーザの特性分析

本研究では, ユーザ個人の特性を捉えるための情報として, 頻出キーワードを用いることにした. ここでいう頻出キーワードとは, ユーザが過去に回答したことのあるすべての質問文において, 出現頻度が高いキーワードのことである. すなわち, このキーワードはユーザが興味関心の高い言葉であり, このキーワードと関連性の高い質問を推薦することで, ユーザにとって回答可能な質問を推薦することができると考えられる. なお, 回答文の頻出キーワードを用いない理由は, 質問文に比べて文の量が少ないため, 文章内のキーワード数が少なくなり, 頻出キーワードの出現頻度に大きな差が出ないことからユーザ個人の特性を捉えづらいと判断したためである.

今回は, 実際に頻出キーワードを用いた質問の絞り込みの有効性を検証するために, 「Yahoo!知恵袋」を事例として分析を行う. 分析対象とするユーザは「Yahoo!知恵袋」にアカウントを持ち, 過去の回答数が 30 件以上のユーザ 100 人とした[9]. また, 頻出キーワードの抽出はフリーソフトである「EKWords[10]」を用いて行った.

まず, ユーザごとに過去に回答したすべての質問文中から出現回数の多いキーワードを抽出した. 次に, このキーワードを出現頻度の高い順に並び替え, 上位 3 位までのキーワードを全て含む質問が, 「Yahoo!知恵袋」内に何件あるのか調査した. 同様に, 上位 4 位までのキーワードを用いた場合と, 上位 5 位までのキーワードを用いた場合も検証し, 質問の絞り込みに適しているキーワードの数を検討した. その結果, 0～10 件の質問数に絞り込むことができたユーザが, どの場合においても最も多いことが明らかになった (表 1). この数は 2 節で説明した最小カテゴリ内の質問数に比べて桁違いに少ない数であり, この方法が質問数の絞り込みに有効であることを裏付ける結果となっている. なお, この場合の 0 件というのは現在回答を受け付けている質問がない状態を示しており, 必ずしも Q&A サイト内にキーワードを全て含む質問が一つもないという意味ではない. そこで, それぞれの場合において, Q&A サイト内の全ての質問を対象として絞り込み処理をした場合に, 該当する質問の数が 0 件となるユーザの数を調査した (図 2). すると, 上位 5 つのキーワードを用いた場合においては, 約半数が Q&A サイト内にキーワードを全て含む質問がないという結果になった. つまり実際に質問推薦システムを導入した場合に, 約半数のユーザへ推薦する質問が一つもないということになる. このことから, 質問の絞り込みに用いるキーワードの数は 3 つから 4 つが妥当であると判断した.

表 1 頻出キーワードを含む質問の絞り込み結果

絞り込みの結果 (件)	上位3つのキーワードを用いた場合 (人)	上位4つのキーワードを用いた場合 (人)	上位5つのキーワードを用いた場合 (人)
0～10	63	81	92
11～20	9	7	6
21～30	3	4	2
31～40	2	0	0
41～50	3	1	0
51～60	3	1	0
61～	17	6	0

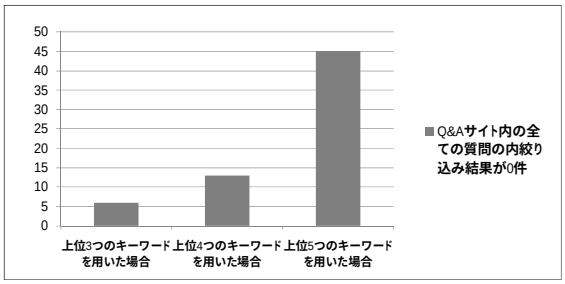

図 2 キーワードを含む絞り込み結果が 0 件のユーザ

4.2 共起分析による複数カテゴリの質問推薦

ここまで, 頻出キーワードを用いることにより推薦すべき質問を絞り込めることを示したが, 抽出した頻出キーワード同士が全て同じカテゴリに属するとは限らない. 例

[9] 調査時期は 2010 年 9 月.
[10] http://www.djsoft.co.jp/products/ekwords.html

えば，頻出キーワードの上位 4 つが「漢字検定」「勉強法」「アニメ」「ドラえもん」というユーザの場合，「漢字検定」と「勉強法」は「資格・習い事」というカテゴリに属するものであり，「アニメ」と「ドラえもん」は「アニメ・コミック」という異なるカテゴリに属するものである．このように，ユーザにとって回答可能な質問のカテゴリが複数ある場合，上述の方法ではそれを判断することが不可能である．そこで，ユーザが興味を持っている複数のカテゴリを機械的に判別するために，頻出キーワードをその関連性の有無によってグループ分けし，カテゴリの境界線を判断する方法を提案する．

4.2.1 頻出キーワードの共起関係

頻出キーワード同士の関連性を見つけるため，キーワード間の共起関係に注目した．共起とは，ある 2 つのキーワードが１つの文章内で同時に出現することである．共起関係を見ることで，キーワード間に繋がりが有るか無いかを判断することができる．また，キーワード同士が共起していれば，その二つは同じカテゴリに属する可能性が高いと考えられる．そこで，一人のユーザが回答した全ての質問文中から，共起関係がある頻出キーワードの組み合わせを抽出した．また，この結果を基にキーワード同士の関係を表現するため，グラフ[11]を用いた．このグラフでは，キーワードをノードとし，一度でも共起関係にあるキーワード間をエッジで結ぶことで，関連性の有無を表している．その結果，複数のカテゴリに興味があるユーザのグラフは，図 3 に示すようにそれぞれ独立した部分グラフを形成する傾向にあることが明らかになった．また，このようなグラフを形成するユーザの割合が，全体の 4 分の 1 を占めていることも明らかになり，単純に出現頻度の高いキーワードを組み合わせただけの質問の絞り込み方法では，これらのユーザに対して誤った判断をしていたということが判明した．この誤認を防ぐためには，頻出キーワードの抽出を行った上で，更に共起分析を行うべきである．

しかし，この場合どのキーワードを質問の絞り込みに用いるかということが問題になる．そこで，本研究では完全グラフ[5]の考えを用いてキーワードの選択を行う．

[11] 要素同士のつながり方を，「ノード」と「エッジ」で分析するグラフ理論[5]を用いている．

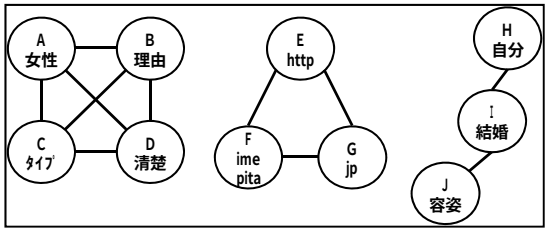

図 3 カテゴリが複数あるパターンのグラフ

4.2.2 完全グラフを用いたキーワードの選出

完全グラフとは，どのノードも他のすべてのノードとエッジで結ばれているグラフのことをいう．つまり，図 3 ではＡＢＣＤの集合とＥＦＧの集合，またＨＩの集合とＩＪの集合が完全グラフである．完全グラフを形成しているキーワードの集合は，その結びつきが強く一つの概念を表している可能性が高いと考えることができる．一方，ＨＩＪ の集合のような不完全グラフは，必ずしも共通のカテゴリに属するキーワードのみで形成されているとは言えないため，これらをユーザの特性を表すキーワードとするのは相応しくない．

この点を確認するために，複数のカテゴリに興味があるすべてのユーザのグラフにおいて，不完全グラフを形成しているキーワードの組み合わせのみを用いて質問の絞り込みをそれぞれ行った．その結果，絞り込まれた質問の内，平均して半数以上の質問が異なるカテゴリに属するものであった（図 4）．この結果から，不完全グラフのキーワードの組み合わせを質問の絞り込みを行うことは，有効でないと考えられる．

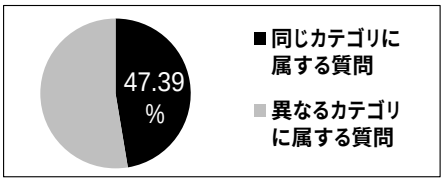

図 4 不完全グラフを用いた場合のカテゴリの偏り

さらに，完全グラフを形成するキーワードであっても，図 3 のＨＩやＩＪのような２つのノードのみで形成されているものも適切でないことが明らかとなった（図 5）．

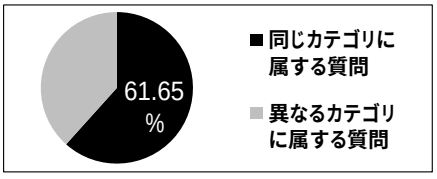

図 5 2つのノードで形成されるグラフを用いた場合

そこで，質問の絞り込みに使用するキーワードは以下の 2 つの条件を満たすものに限定するようにした．第一に完全グラフとなる共起関係を形成すること．第二に共起関係のグラフのノード数が 3 以上であること．これらの条件を満たすキーワードの集合を用いて絞り込まれた質問を，回答者に推薦するシステムとした．

実際に，この条件で再び質問の絞り込みを行なった結果，異なるカテゴリに属する質問の割合は，6 分の 1 程度に抑えられることが明らかとなった（図 6）．さらに，それぞれのユーザに推薦した平均質問数についても検証した．その結果，表 2 に示したように，回答可能なカテゴリが一つに偏っているユーザには，平均 14 件の質問が提示された．一方，回答可能なカテゴリが複数あるユーザには，1 カテゴリあたり平均 2 件の質問が提示された．これらの値は，十分に質問の絞り込みができていることを示している．また，それぞれのユーザに対して，Q&A サイト内の全ての質問を対象として絞り込み処理をした場合に，該当する質問の数が 0 件となるユーザの数を調査した．この調査では，カテゴリが一つに偏っているユーザの 6 人がこれに当てはまり，カテゴリが複数あるユーザにおいては一人も当てはまらないという結果になった．この値は，キーワードによる過剰な絞り込みを回避できていることを示している．

これらの検証結果から，本研究で提案するシステムは，既存の質問推薦システムよりも質問の絞り込みやユーザの特性を考慮している点において優れているということがいえる．

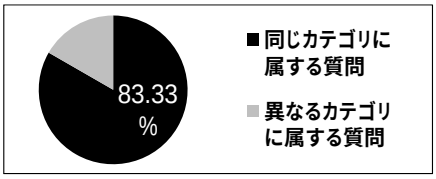

図 6 条件を満たす完全グラフを用いた場合

表 2 それぞれのユーザに対する質問の平均提示数

	絞り込んだ質問の数
カテゴリが単独のユーザ	14件
カテゴリが複数のユーザ（1カテゴリあたり）	2件

5 むすび

本研究では，Q&A サイトを回答者として利用する際，ユーザが膨大な質問の中から回答可能なものだけを手動で取捨選択しなければならないという問題点を改善するため，頻出キーワードの抽出と共起分析を用いた新たな質問推薦システムを提案した．まず，頻出キーワードの抽出によってユーザごとの特性を捉え，推薦する質問数を既存サイトの最小カテゴリよりもはるかに小さく絞り込むことに成功した．また，共起分析を用いることでキーワード間の関連性を探り，ユーザが興味を持っている複数のカテゴリを見つけることが可能になった．このシステムの導入により，ユーザの特性に適した質問を推薦できることが期待できる．

しかし，本研究で提案するシステムは，抽出した頻出キーワードが同じカテゴリに属するものであるかといった判断をする際，共起頻度が 1 回の場合でも強い共起関係があり，同じカテゴリに属する可能性が高いと判断している．より高い精度で質問推薦をするためには，共起頻度に応じてキーワード間の関係を分析し，抽出した頻出キーワードが，同じカテゴリに属する可能性が高いということをより正確に判断する必要がある．そのため，共起頻度に重点を置いてシステムの改良を図ることが，今後の課題として挙げられる．

参考文献

[1] Yahoo!JAPAN，チエブクロのなかみ～知識共有コミュニティ「Yahoo!知恵袋」に集約されるデータとは何か～，Yahoo!JAPAN ネット生活予測レポート，http://i.yimg.jp/images/docs/report/pdf/006.pdf

[2] OCN ブリエ，暮らしの疑問を解決する「Q&A サイト」の利用実態，OCN 大人の趣味生活，http://briller.ocn.ne.jp/hakusho/archives/report/061.htm」

[3] 日経 BP 社，「Yahoo!知恵袋」「教えて!goo」の利用者数，ページビュー数推移，ITpro，http://itpro.nikkeibp.co.jp/article/Research/20080423/299899/

[4] 株式会社メディアインタラクティブ, 第 7 回「検索エンジンのニーズと利用」に関する調査（下）, Web マーケティングガイド, http://www.e-research.biz/profile/prosem/003197.html

[5] 浜田隆資, 秋山仁, グラフ論要説, 槇書店, 1982.

[6] 甲谷優, 川島晴美, 藤村考, “QA コミュニティの成長パターンに基づく回答者への質問推薦, ”日本データベース学会論文誌, Vol.8, No.1, pp.89–94, 2009.

[7] 甲谷優, 岩田具治, 塩原寿子, 藤村考, “QA コミュニティにおける複数情報源を用いた効果的な質問推薦, ”情報処理学会論文誌, Vol.3, No.4, pp.34-47, 2010.

[8] 甲谷優, 川島晴美, 藤村考, “QA サイトにおける質問応答グラフの成長パターン分析, ”日本データベース学会論文誌, Vol.7, No.3, pp.61–66, 2008.