Yahoo!知恵袋の質問文分類のための質問文分析

劉舒暢†，伊東栄典‡，中島幸子*，廣川佐千男‡

†九州大学システム情報科学研究府，‡九州大学情報基盤研究開発センター
* 梅花女子大学

lsc19920813@gmail.com　ito.eisuke.523@m.kyushu-u.ac.jp　s-nakajima@baika.ac.jp　hirokawa@cc.kyushu-u.ac.jp

1. はじめに

図書館では，来館者は特定の目的なしに本を読む場合や，特定分野の調査や学習を目的として資料を探して学習する場合もある。図書館のリファレンスサービスは，調査や学習を目的とする利用者のために，資料の検索・提供と口頭回答で利用者を支援している。

公立図書館の利用者は一般の人が多く，学習や研究資料の検索に慣れていない。また，自分が求める情報が，何の分野かを言葉で表現できない。例えば，認知症になった同居者のための対処方法の調査などでは，医療問題や住居構造，行政支援などの社会制度など，調査範囲が広い。そのためリファレンス係では，資料提供の前に利用者が何に困っているのか，どの分野を調査したいのかを，質問を繰り返しながら明確にしている。

我々は，多数の人が持つ課題の質問には類型があると想定している。多数の質問文を統計処理することで質問類型を抽出可能と考えている。日本最大の質問・回答サイトである「Yahoo!知恵袋」には多数の質問と回答がある。しかも多くの質問文は非専門家による曖昧さを含む質問文であろう。そこでYahoo!知恵袋のデータを，質問文の類型抽出に用いる。

なお，Yahoo!知恵袋のデータを用いた研究には立石らの，ライフイベントに関するユーザ属性抽出がある[1]。しかし質問文の類型を調査した研究は無いと思われる。

2. Yahoo!知恵袋データ第2版

国立情報学研究所は，情報学研究リポジトリと名付けた，研究用のデータ集合を提供している。この中で，ヤフー株式会社は国立情報学研究所を通じて「Yahoo!知恵袋」のデータを日本の研究者に提供している[1]。「Yahoo!知恵袋」は，質問したい人と回答したい人をむすび，知恵と知識を参加者同士で共有することを目的として2004年4月からヤフー社が提供している知識検索サービスである。「Yahoo!データセット」には，「Yahoo!知恵袋」サービスに投稿された質問および回答のデータが含まれている。

本データセットのデータ件数等を表1に示す。

表1　Yahoo!データセット件数等

項目	内容
総質問件数	16,098,580
全カテゴリ数（小カテゴリ数）	444

表2に，質問件数が多い上位10カテゴリのカテゴリ名と，質問件数を示す。

表2　カテゴリと質問件数（上位10カテゴリ）

	カテゴリ	質問件数
1	恋愛相談	556,418
2	恋愛相談、人間関係の悩み	444,521
3	病気、症状、ヘルスケア	442,925
4	Yahoo!知恵袋	380,200
5	政治、社会問題	367,043
6	プロ野球	294,603
7	パソコン	292,657
8	アダルト	263,043
9	Yahoo!オークション	255,615
10	Windows全般	248,811

表2に示すように，恋愛相談の質問件数が最も多い。健康（病気、症状、ヘルスケア）も3位で多い。パソコンに関する質問は7位と10位であり，この分野の質問も多い。

3. 質問文の分類

我々は，Yahoo!知恵袋に投稿される質問文として，その分野を専門としない人が解を求める質問や，その分野についての調査を求める文章を質問文として想定していた。実際にYahoo!データセットに含まれるYahoo!知恵袋の質問文を見ると，解を求めるものや，調査に関する質問だけではないことが分かった。我々はYahoo!知恵袋の質問文は，

表 3 に示す５つの型があると考えている。

表 3　質問文の大分類

番号	内容
A	求解
B	共感
C	調査
D	釣り，ネタ，笑い
E	その他

A の求解型は,具体的な答えを探す質問である。パソコンおよび Windows 全般カテゴリの質問文は，文字コードの変換方法，エクセルの操作方法など，具体的な答えを探す質問が多い。これらの質問文は A 型である。A 型の例を表 4 に示す。

表 4　A: 求解型の例（パソコン）

質問タイトル: 入力モードがひらがなの時に、カタカナの言葉を入力したいのですが、例えば「ケラチナミン」と打つと漢字 質問本文: 入力モードがひらがなの時に、カタカナの言葉を入力したいのですが、例えば「ケラチナミン」と打つと漢字混じりで変換されてしまいます。入力モードをカタカナにせずに、一発でカタカナにするにはどうしたらいいですか？

B の共感型は恋愛や健康相談に多い。恋愛相談の質問文の多くは，具体的な回答ではなく，読者の共感を求めているものが多い。恋愛は状況が様々であるため，答えが決まらない。そのため共感や事例を求める依頼文になる。B 型の例を表 5 に示す。

表 5　B:共感型の例（恋愛相談）

質問タイトル: 周りがどんどん結婚して… 質問本文: しょーもないことなんですけど、聞いて下さい。私は今年２０歳になったばっかり（中略）周りがどんどん結婚していくので、なんでかわからないですが焦る気持ちが出てきているのです。。変な文章になりましたが、こういう気持ちになるっておかしいですか??

C の調査型は，世論調査に類する質問である。食べ物の嗜好，プレゼントに適した物，パソコンや携帯の機種，などを広く調べるものである。C 型の例を表 6 に示す。

表 6　C:調査型の例（健康）

質問タイトル: 禁煙したいです。 質問本文: 嫁が妊娠をしました。待望の妊娠なので、生まれてくる赤ちゃんの為…（中略）…禁煙に成功した方、どうやって禁煙したかご伝授して下さい。宜しくお願いします。

D 型の釣りや笑いを誘うネタ型の質問文も多い。具体例を表 7 に示す。なお，以下の質問は恋愛相談カテゴリで最も回答数の多い質問でもある。

表 7　D:釣り型の例（恋愛相談）

質問タイトル: 助けてください！！！！ 質問本文: はい～！！釣られましたね！！釣られたついでに回答してってください。絶対！！絶対にですよ！！！！好きな飲み物はなんですか？

我々が質問類型の分類対象として想定するのは A の求解型か，C の調査型の質問文である。求解型であれば，手掛り語や分野の用語から，質問の類型を抽出することが出来ると考えている。

なお，表 3 に示した５つの型は網羅的な調査に基づいていない。将来詳細な調査と評価で変更する可能性がある。

4. 質問文の分析

Yahoo!知恵袋における，恋愛相談，健康，パソコンの３カテゴリの質問文について，いくつかの分析を行った。なお，健康は「病気、症状、ヘルスケア」である。質問件数の時間推移，質問文の長さ（バイト数），質問文への回答数，質問から解決までの期間について調査した。

4.1. カテゴリ選択の理由

健康は，以前からの調査対象として選択した。健康カテゴリには求解・調査型の質問の他に，共感型の質問も多い。恋愛相談は，表 2 に示すように質問件数の最多カテゴリであるため，選択した。恋愛相談カテゴリには，共感型や釣り型の質問が多く，それらの特徴を抽出しやすいと考えている。三番目のパソコンは比較対象として選択した。パソコンは我々の専門分野であるため内容を理解出来る。また，パソコンカテゴリは求解型や調査型の質問が多いため，その特徴を理解しやすい。

4.2. 質問投稿数の時間推移

３つのカテゴリの質問投稿数を月毎に数えた。図 1 に各月の質問投稿数を示す。

恋愛相談は 2006 年 9 月以降にしか存在しない。健康とパソコンについては，2005 年 11 月に一度減少している。減少の理由は特定できていない。当時，カテゴリの分割が有ったのかもしれない。

健康と恋愛相談は，増加傾向にある。一方，パソコンの質問件数は多少の増減はあるものの，大きな変化は無い。

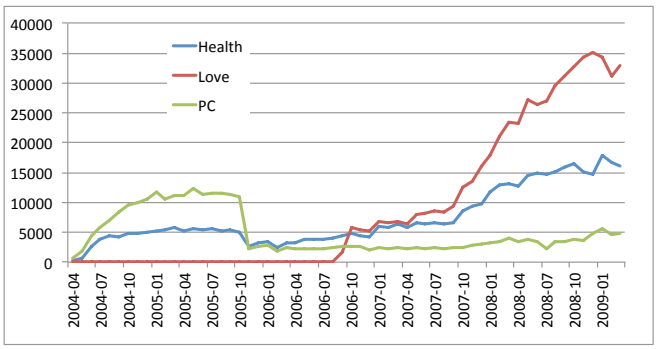

図 1　月毎の質問投稿数

4.3. 質問に対する回答数

回答が多い質問もあれば，少ない質問もある。回答件数に対する質問件数を図 2(健康)，図 3(恋愛)，図 4（パソコン）に示す。図 2〜4 の 3 とも両対数尺度である。

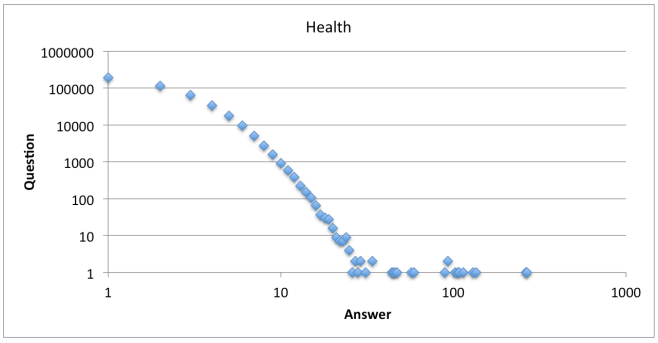

図 2　横軸の回答数を持つ質問件数（健康）

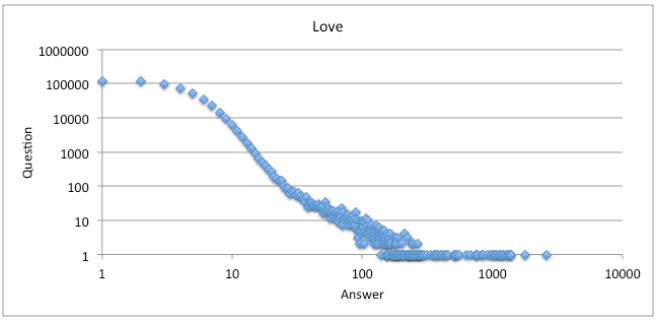

図 3　横軸の回答数を持つ質問件数（恋愛相談）

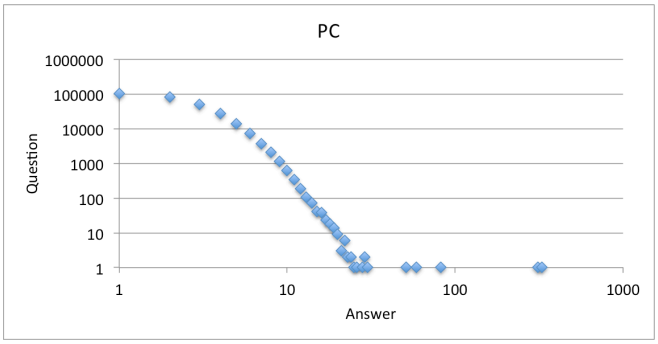

図 4　横軸の回答数を持つ質問件数（パソコン）

図 2〜4 を見ると，どの分野でも回答数が 1 件や 2 件の質問が多い。なお，回答数 0 件の質問文はデータセットに含まれていない。

健康とパソコンで，回答数の分布は対数正規分布に似た曲線を描いている。恋愛相談では，回答件数が少ない（1〜10 件）の部分は対数正規分布に似た曲線であるが，回答件数 20 件以上の部分は冪分布に似た直線を描いている。

健康とパソコン分野では，回答数の少ない質問は求解型が多い。求解型の質問に対し，具体的な解が回答されると，他の回答は出なくなる。回答が多数になる質問は，曖昧で議論を起こしやすいものが多い。表 8 に健康カテゴリにおける質問文タイトルと回答数を示す。

表 8　健康・質問タイトルの例

回答数	質問タイトル
226	5 歳の娘にプチ整形することは異常ですか？
15	タバコって妊娠にどのていどひびくものなんでしょうか？妊娠したらやめるじゃ手遅れ？
1	もし、大腿骨頸部骨接合術をする場合骨に穴を開けると思うのですが、抜釘術後その穴は塞がりますか?
1	精神障害者手帳を持っていると、県営、市営などの住宅に、応募できるそうですが、自分が住んでいる地区以外

表 9 に恋愛相談カテゴリにおける質問文タイトルと回答数を示す。

表 9　恋愛相談・回答件数と質問タイトル

回答数	質問タイトル
1771	「死ね」と言われて一番ベストなかえしかたを教えてください
1043	不倫は悪くない！
1152	初デートで彼がユニクロだったので別れたいです！！！
80	絶対回答して！★キスは浮気になりすか？
80	女の幸せ。皆さんはどう思いますか？
1	恋人と直接会った時の会話量と、電話での会話量と、メールでの文章量は、比例しますか？
1	彼氏の誕生日プレゼントについてです。今考えているもの二つがあるのですが、いまいちパッとしないし、

恋愛相談では，回答数 1000 件以上の質問文は，多くは D の釣り型そのものか，釣りを意図させる文章である。また，質問文投稿者は議論や共感を得るために投稿した質問であるかもしれないが，実際には釣りやネタと思われて回答が発散しているもの見受けられる。また，回答件数が少ないものでも求解型の質問は少ない。

4.4. 質問文の長さ（バイト数）

質問文の長さ（バイト数）の分布を調べた。分布を図 5（健康），図 6（恋愛），図 7（パソコン）に示す。図 5〜7 とも横軸は対数尺度である。

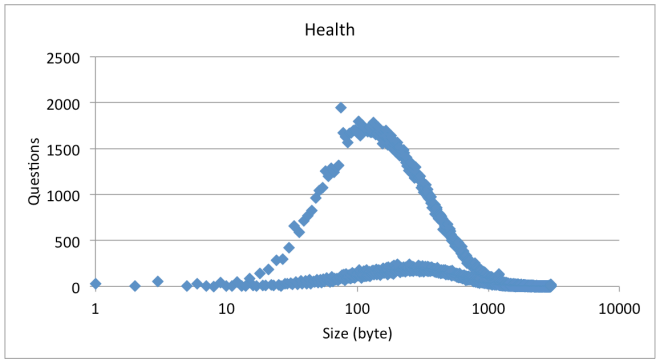

図 5　横軸のサイズ(byte)を持つ質問件数（健康）

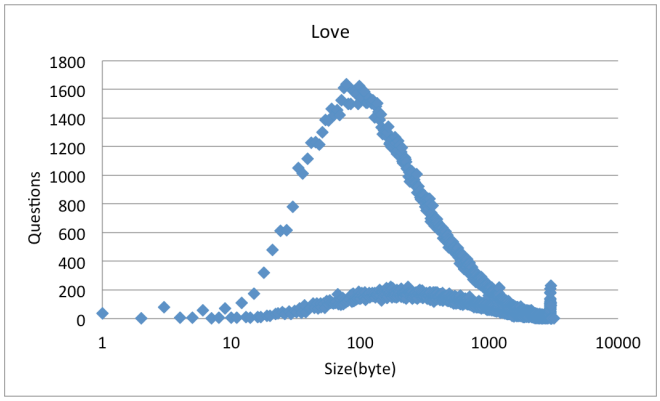

図 6　横軸のサイズ(byte)を持つ質問件数（恋愛相談）

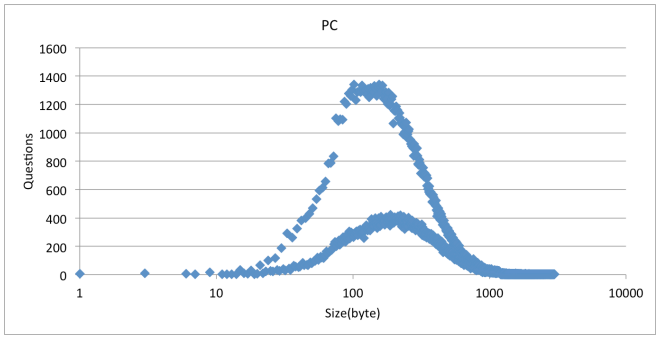

図 7　横軸のサイズ(byte)を持つ質問件数（パソコン）

質問文の長さについては，健康・恋愛相談・パソコンの 3 カテゴリとも，100 文字（300byte）程度の質問が多いことがわかる。また，3 つのカテゴリとも対数正規分布を示している。

3 カテゴリとも，二つの対数正規分布を持つ。上の件数の多いものは UTF8 コードで一文字 3byte になる日本語文字のみの質問文である。件数の少ない分布は 1byte の英数字や記号を含む質問文である。パソコン分野では，1byte の英数字や記号を含む質問文が多いが，健康や恋愛相談では日本語文字のみが多い。

図 6 のでは 3000byte の質問文が少し多い。Yahoo!知恵袋の質問文は最長 1000 文字の制限がある。1000 文字まで書くと 3000byte になる。恋愛相談では制限値までの長文質問が多い。一方，健康やパソコンのカテゴリでは，制限値までの長文質問は少ない。

5. おわりに

我々は求解型や調査型の質問文の類型分類を目指している。そのために，本研究では Yahoo!知恵袋の質問と回答データを持つ Yahoo!データセットについての分析を行った。Yahoo!知恵袋に投稿される質問は，求解型・調査型ではない質問も多い。質問文分類のための 5 つの型を提示した。次に，恋愛相談，健康，パソコンの 3 つのカテゴリについて，統計的な分析を行った。また，幾つかの定性的な分析を行った。

今後，5 つの型で質問文のフィルタリングを実現したい。そのために，型ごとの特徴を調査する予定である。5 つの型へのフィルタリングが実現したら，求解型と調査型の質問文を集め，それぞれの特徴を詳細に分析する予定である。最後は質問文の類型から，質問文に対する回答の類型を求めることで，専門家ではない人からの質問について，回答を支援するシステムを実現していきたい。

文　献

[1] 国立情報学研究所, Yahoo!ジャパン社: Yahoo!データセット: http://www.nii.ac.jp/cscenter/idr/yahoo/yahoo.html, (accessed at Dec.12, 2014).

[2] 立石健二，宮崎林太郎，長田誠也: Yahoo!知恵袋を用いたライフイベントに関するユーザ属性抽出，信学技報 114(211), pp.53-57, 2014.