カテゴリー分類にもとづくFAQの自動生成

松崎 友洋†　　冨澤 眞樹†
前橋工科大学工学部†

1. はじめに

本研究の目的は，Q&A掲示板からFAQを自動生成することである．本研究で対象とするQ&A掲示板は，スレッドの先頭が質問文で，それに回答を投稿していくものである．また，ここで考えているFAQは，類似文書検索手法に基づきスレッドをカテゴリー分類したものである．スレッドのカテゴリー分類の手順は，まず電子掲示板から各スレッドを単語に分け，統計情報を得る．その統計情報を元にキーワードとなる単語から類似文書検索手法に基づきカテゴリー分類する．本報告では，錦鯉のQ&A掲示板を対象にカテゴリー分類をした結果を示す．

2. スレッドのカテゴリー分類

2.1 対象とするQ&A掲示板

対象とするQ&A掲示板は，コイパーク（URL:http://www.koipark.net/）の『錦鯉 疑問 質問掲示板（Q&A）』にある過去ログを選んだ．スレッドの数は４０７件である．また，全スレッドの投稿数は２２２０件（２００７年１１月現在）である．

スレッドの先頭の投稿が質問文であるものは全体の９割であったため，スレッドの先頭の投稿（以下スレッド質問文とする）は質問文であると仮定して，スレッド質問文についてカテゴリー分類した．

2.2 スレッド質問文からキーワードの取得

キーワードはFAQ作成者が事前に手動でスレッド質問文から取得する．表１に取得したキーワードの一部を示す．キーワードには代表語と関連語があり，合わせて一つのキーワードとする．例えば表１の場合，水槽，金魚鉢，容器，FRP，池を合わせて一つのキーワードで「水槽」とみなす．代表語とは，金魚鉢や容器といった類似した意味の単語の集合を総称する単語を指しており，ここでは水槽となる．関連語は，代表語から総称される単語の集合であり，水槽に総称される単語の集合は金魚鉢や容器となる．関連語を用意しておくことにより，代表語だけでは拾いきれない単語も拾えるのでカテゴリー分類がしやすくなる．

表１　キーワードの一部抜粋

キーワード	
代表語	関連語
水槽	金魚鉢，容器，FRP，池
病名	えら病，眠り病，風邪，カラムナリス
濾材	スポンジ，ウールマット，ドライボール，
酸素	エアーレイション，ぷくぷく，ぶくぶく
成長	でかい，発育

今回取得したキーワードは１０５語（そのうち代表語は２３語，関連語は８３語）である．カテゴリー分類の仕方は，FAQ作成者によって様々であり，キーワードをFAQ作成者が用意することでFAQ作成者の意図をカテゴリー分けに反映できると考えた．

2.3 TF・IDFによる重み付け

キーワードの重み付けとして，TF・IDF[1]を用いる．TF・IDFは，キーワードの重み付け方法として情報検索によく用いられる．TF・IDFを用いると，スレッド質問文dにおけるキーワードtの重みweight(d,t)は次のようになる．

$$weight(d,t) = TF(d,t) \cdot IDF(t)$$

TF(d,t)は，あるスレッド質問文dにおける，キーワードtの頻度である．例えば，スレッド質問文dが「錦鯉は水槽と池のどちらのほうが飼いやすいですか？　」である場合，表１より水槽と池は同じキーワードとして扱うのでキーワードtが水槽である場合，TF(d,t)＝２となる．

IDF(t)は，次のようになる．

$$IDF(t) = \log 10 \frac{N}{DF(t)}$$

ここでNは全スレッド質問文の数，DF(t)はキー

Automatic generation of FAQ based on category classification
† Tomohiro MATSUZAKI, Masaki TOMISAWA
† Dept. of Engineering, Maebashi Institute of Technology.

ワードｔが１回以上出現するスレッド質問文の数である．よって，この方法により１つのスレッド質問文に同じキーワードが多く出現すれば，TF・IDF の値は大きくなる．また，多くのスレッド質問文にキーワードが出現すれば値は小さくなる．

2.4　ベクトル空間法による類似スレッド質問文の取得

ベクトル空間法[2]とは，スレッド質問文中にどのキーワードがどの程度出現しているかをベクトルの形で表現する情報検索手法のひとつである．スレッド質問文ベクトル V1，V2 の類似度 sim(V1,V2)は次の式になる．

$$sim(V1,V2)=\frac{\overrightarrow{V1}\cdot\overrightarrow{V2}}{\left|\overrightarrow{V1}\right|\cdot\left|\overrightarrow{V2}\right|}$$

ここで，スレッド質問文ベクトルとはキーワード数を次元として，各次元の値を TF・IDF から求めたものである．

3.　分類結果

荒らしや本来の趣旨とは関係のないスレッド質問文は遮断するため，スレッド質問文にキーワードが一つもないものは省いた．省かれたスレッド質問文は４０７個中６８個であり，残り３３９個のスレッド質問文についてカテゴリー分類を行った．カテゴリー数を９とした場合のカテゴリー分類結果を表１に示す．

評価方法は，あらかじめスレッド質問文すべてにカテゴリラベルを付加しておき，自動で分類した結果と比べた．カテゴリラベル数はカテゴリー数と同じく９とした．

この分類ではまず，スレッド質問文の数をカテゴリー数とみなして，設定したカテゴリー数になるまで前述したベクトル空間法を用いてもっとも類似したスレッド質問文の組を見つけ，それをあらたなカテゴリーとみなすという作業を続ける．一つのカテゴリーにスレッド質問文が集中しすぎないように，全スレッド質問文数を設定したカテゴリー数で割った値以下に制限した．

各カテゴリー内で最も多い割合のカテゴリラベルをそのカテゴリー自身のラベルとした場合の正解率は，平均３８％，最高値が No４の繁殖で７８％だった．

4.　考察，課題

各カテゴリーが，あるカテゴリラベルに集中することが理想であるが，分類結果では低い精度であったため，FAQ 作成者の意図をキーワードの取得のみで反映することは困難であることが分かった．カテゴリー分類の正解率は，カテゴリーにより差があったため，その精度は，キーワードの取り方により変わってくると考えられる．

今後は，分類精度を上げるために，FAQ 作成者があるスレッド質問文に対し誤ったカテゴリー分類結果を正しいと判断したカテゴリーに加えることで以後処理するスレッド質問文の分類精度を上げる仕組みを持ったフィードバック機能を検討する必要があると考えられる．

表２ カテゴリー分類結果（カテゴリー数＝９）

カテゴリ / ラベル / No	病気	濾過	水質	繁殖	鑑賞	体調	水槽	池	その他	合計
1	27	0	0	0	1	3	0	0	11	42
2	3	2	0	0	2	1	2	2	24	36
3	1	23	2	0	0	2	1	0	11	40
4	8	0	0	7	5	3	0	0	18	41
5	7	1	0	0	8	3	0	0	9	28
6	10	2	1	1	3	1	0	0	22	40
7	17	3	2	0	0	4	1	0	10	37
8	4	17	2	0	0	1	1	1	9	35
9	2	3	2	1	7	1	0	0	24	40
合計	79	51	9	9	26	19	5	3	136	339

5.　参考文献

[1]関洋平，原田賢一：tf/idf 重み付けに基づいく動的文書生成，情報処理学会研究報告デジタルドキュメント，Vol.2001，No.120，pp.25-32(2001).

[2]技術資料 単語意味属性を使用したベクトル空間法
<http://unicorn.ike.tottori-u.ac.jp/murakami/paper/JOURNAL/NLP_2003_04/main/>
(2008/1/8 アクセス)．

[3]鈴木雅実，村松茂樹，松本一則，井ノ上直己，橋本和夫：類似文書クラスタリング手法における新聞記事分類コード推定実験，第 61 回全国大会，データマイニングと情報フィルタリング．