

Q&A サイトに見られるコンピュータ問題解決支援分析

松川 晃大[†] 菱沼 千明[†]

東京工科大学コンピュータサイエンス学部[†]

1. 背景と目的

単なる検索エンジンは知識を生み出さず、既存の知識を検索するだけであったが、利用者が知識を出し合い、内容を創っていくメディアである CGM (Consumer Generated Media, ユーザー参加型のコンテンツ形成システム) の一つの Q&A サイトの利用率が年々高まってきている。[1]

本研究では、ネット Q&A サイトに見られる質問・回答状況からソフトウェアの開発に係る問題点を探る。そこで、コンピュータを専門とする企業関係者が Q&A のテキストから日々、業務上でどのように困ったことが解かれているか、またどういった質問が多いかを分析し発見することで今後のコンピュータサイエンスの教育やソフトウェア開発支援ツールの開発に反映しソフトウェアを開発する上での産業発展に貢献する事を目的とする。

2. 手法

本研究では職場で最も利用されている Q&A サイト、OKwave を分析対象とする。コンピュータを専門とする問題を調べたいので分析対象となるデータは、「技術者向けコンピューター」カテゴリ内の「ハードウェア」「OS」「データベース」「プログラミング」「開発」「業務ソフトウェア」「運用管理」「ネットワークセキュリティ」の 9 つである。[2]

カテゴリ別に「質問件数」「回答率」を出し、どのカテゴリの質問が多いのか、また、質問者が納得の良く回答を貰えた比率の高いカテゴリではどのような質疑応答が行われているのかを解析する。

また、OKwave 「技術者向けコンピューター」内のカテゴリでは「OS」「データベース」「プログラミング」「開発」「業務ソフトウェア」にサブ項目が含まれている。

サブ項目内を解析すると、質問者の相談内容に違いがあり、全てのデータを解析しても、どのように困ったことを解決しているのか求めることが出来ない。

そこで予め、サブ項目内でデータの選定をし、分析対象を決めてから「良回答率」を出すことにより、ジャンル別に質問者が納得の良く回答を貰える可能性が高いジャンルを割り出すことが出来る。

また、この手段を取ることで、別のサブカテゴリといった不必要なテキストをマイニングする必要がなくなり作業の効率化につながる。

そして、「良回答率」を求めてからそれに関しての特徴を見出し、そこから分析対象となるデータの頻度分析、共起分析により、得られたデータの繋がりからコンピュータ関連で起きた問題の中で解決した一例としてはどのような質問が多かったかを読み取る。

3. データ抽出結果

(1) 質問数と回答率

カテゴリ別に質問件数のデータを取得する。質問データの取得日数は OKwave 開始時の 2001 年 1 月～2009 年 10 月までのものである。

取得したデータを図 1 に示す。

	質問数 (件)	回答率 (%)
プログラミング	104,749	99.7
OS	30,610	98.4
業務ソフト	22,399	98.6
ネットワークセキュリティ	17,435	98.1
データベース	14,698	98.4

図 1. 質問数と回答率のデータ

これらのカテゴリに対し質問が多い。尚、回答率とは投稿者の質問に対し、一度でも返答のあった質問を割り出した物であり、技術者向けコンピュータ内では、どのカテゴリに対しても質問を出すと 98%以上の確率でなんらかの回答が帰ってくるという事が特徴である。

An Analysis supported on the Q&A site for Computer Technology

[†]Akihiro Matsukawa, [†]Chiaki Hishinuma, School of Computer Science, Tokyo University of Technology

(2) サブ項目内解析

小カテゴリごとにそれぞれ質問数が一番多いものを分析対象とし、本大学の学習対象として、カテゴリ OS から「Linux」プログラミングから「Java」の 2 つを加えて分析を行う。尚、最も質問の多かったプログラミングの小カテゴリ比率を図 2 に示す。

	質問数(件)	回答率(%)
C++	16,638	98.5
Visual basic	14,106	99.5
HTML	13,641	99.3
PHP	10,446	99.2
Java	9,818	99.2

図 2. 小カテゴリ別のデータ

プログラム言語別の求人件数は 2007 年と 2008 年のデータで Java と C がずば抜けて多く [3]

プログラミング内では上位 5 つの分野で質問数の 6 割を占めており、これらの分野に対しニーズが高いことがわかる。

4. 分析結果

(1) 良回答率

カテゴリ別に回答者が質問者に納得の良く回答を与えた割合を示す。選定したサブカテゴリ別にデータを図 3 に示す。

	質問数(件)	良回答の質問数(件)	良回答率(%)
C++	16,638	12,782	76.8
Graphic soft	13,217	10,072	76.2
MySQL	2,580	1,916	74.3
Java	9,818	7,120	72.5
Linux OS	8,601	6,224	72.4
ネットワークセキュリティ	17,045	11,014	64.6
運用管理	7,152	4,345	60.7
WindowsOS	14,986	8,844	59.0
その他	33,163	18,242	55.0
ハードウェア	8,713	4,785	54.9

図 3. サブカテゴリを含む良回答率データ

(2) 頻度分析、共起分析

良回答率の高いものを分析対象とし、良回答率の最も高い「C++」を分析対象とした。頻度分析をすると、「初心者」という単語が他のカテゴリに比べて多く見受けられ、利用者には初心者が多いように見受けられる。共起分析では極めて頻度の高い単語同士のつながりが見受けら

れず、このことから C++ に関しては専門家や経験者の丁寧な返答から質問者の様々な悩みをわかりやすく回答していることが分かる。

5. 考察

質問件数から見ると「プログラミング」「OS」「業務ソフト」「ネットワークセキュリティ」「データベース」これらのカテゴリに関して質問の需要が高い事がわかる。また、質問を出すと 98%以上の確率で返答を貰えることから回答者が真摯に答えている事は注目に値する。

質問件数に関してはプログラミングが他に比べ圧倒的に多く、Java、C++の頻度分析では「初心者」と言う単語が多く見受けられた。

それだけ質問者はプログラミングに対して悩みを抱えている人が多いということである。

また、Okwave 全体の良回答率が 85%に対し、今回選定したカテゴリ内のデータでは良回答率 85%を超えるものが一つもなく、技術者向け「コンピュータ」では専門的な事を聞く以上質問者も納得のいく回答をもらえる確率がどうしても平均より下がってしまうことが把握できた。

6. むすび

質問文からコンピュータ関連の問題を読み取るとすると業務関連者から回答することは多くても質問する事は少なかったように思われ質問者は初心者の利用が多いようである。

また、質問件数や良回答率が大きかった関係上 C++や画像編集ソフトに対するニーズが高いため、C++言語や画像編集ソフトの支援が望まれる教育プログラムが作成されるとユーザーの悩みも少しは解消に繋がるのではないと思われる。

また、本論分では簡単のため今後の課題として、時系列で質問件数を把握できると時代ごとに質問傾向の違いがわかるので Q&A の投稿日を記録しておく必要があった。

共起分析、頻度分析を行う際に注意点として、質問文を直接、テキストマイニングするとソースコード等が混入して分析の支障となるので質問者が最も聞きたい内容であるタイトルを対象として分析を行い、プログラム内容についての分析が出来ずじまいであった事が反省点である。

参考文献

- [1] ネットレイティングス プレスリリース 2008-12-24
- [2] Okwave (<http://okwave.jp/>)
- [3] (株) ワークポート 2009-2-13