2010 Fifth International Conference on Software Engineering Advances

A FAQ Finding Process in Open Source Project Forums

Wei-Chung Hu* Dung-Feng Yu* Hewijin Christine Jiau†

*Institute of Computer and Communication Engineering,
†Department of Electrical Engineering, NCKU, Tainan, Taiwan
*{selain, dfyu}@nature.ee.ncku.edu.tw †jiauhjc@mail.ncku.edu.tw

Abstract— **When people involve with software more in their daily lives, software companies must provide services through handling various operating questions that users request in the forums. However, different from conventional software companies, various types of difficulties, propositions and opinions could be issued by open source software users in addition to the operating questions. These difficulties, propositions and opinions are generally referred as *questions* in the forums. The questions, as valuable knowledge of the open source project, should be systematically managed. To manage the questions, a common strategy is to construct a FAQ in open source projects. The FAQ can reduce the volume of similar questions in the forums and prevent active forum members from wasting time on answering questions which are already handled before. Most previous literature focuses on existing FAQ retrieval instead of finding and constructing FAQ. This study, as a pioneering work, proposes a configurable and semi-automatic FAQ finding process to assist forum managers in constructing the FAQ. Also, two case studies are conducted to evaluate the effectiveness of the proposed FAQ finding process.**

***Keywords*-Open source; FAQ; community management; data mining**

I. Introduction

Software today is capable of doing more things but becomes more complex than before. Commercial software products have to provide different kind of services to respond user requirements in the forums. Most of user requirements on commercial software are operating issues. Less bug reports occur in the user response forums since the commercial software products are usually well tested before release.

Unlike commercial software companies, the open source community usually follows the principle of "release early, release often". Open source developers are encouraged to release the alpha-version of open source software (OSS) as assessable source code. Then the enthusiastic open source community can start to refine and evolve the software rapidly. Assessable source code and the enthusiastic community are the two most important factors for the success of OSS projects. These two factors also make the user response on OSS products different from commercial software products in several aspects. First, the OSS user response is not limited to the software operating issues. OSS users could raise various types of response, including their difficulties, possible solutions, bug reports, usage experience, new feature requests, source code inspection, tool comparison and integration reports. Second, the number of response in OSS forums could increase rapidly in very short period after each release. Response between different releases is mixed in the forums because the release intervals are usually short. Third, the response in OSS forums is handled not only by developers, but also by other community members. These aspects make the responding and answering activities in OSS project forums more like a knowledge sharing and accumulation process [1].

The large amount of response and the variations of response context lead to the problems of efficient knowledge accumulation and retrieval, especially for successful OSS projects with large user community. The variations of response context come from several factors: (1) the OSS project types, (2) users' task types and expectation, (3) other OSS libraries, (4) execution environment, and (5) the deployment of OSS in working context. These factors make the efficient knowledge accumulation and retrieval problems in OSS projects become project-dependent.

Most OSS projects choose the strategy to provide *Frequently Asked Questions* (FAQ) to solve the problem of efficient knowledge retrieval in forums. The "Questions" are referred to usability difficulties, bug reports, and various types of response in the forums. Before the question is issued, the user is suggested to check if the question is already in the FAQ. If the question is in the FAQ, the user can get the answer from FAQ directly. Therefore, the FAQ can reduce the volume of similar questions on the forums [2]. Also, the FAQ can prevent active forum members from wasting extra time on answering questions which are already handled before. To keep the FAQ up-to-date, several forum managers will periodically examine the forums and maintain the FAQ. The forum managers could be core team members, community managers, or senior active forum members. If there are frequently asked questions found, the forum managers will organize them and update the FAQ.

To construct and maintain an up-to-date FAQ do require extra human resource. By the limited human resource, many

978-0-7695-4144-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSEA.2010.46

OSS projects simply provide a short list of FAQ or a pre-defined list of FAQ, which contains only the basic and general question/answer pairs. The short and pre-defined lists of FAQ are usually not constructed from real questions issued by the users through project forums. To organize FAQ from vast real questions, active forum members must take great effort to review questions and discuss with each other to periodically maintain the FAQ. But the cost might delay the real-time FAQ updating. Further, there is no appropriate tool providing quantitative evidence for the forum managers to arrange the FAQ maintenance schedule. With quantitative evidence, the forum managers can be aware of (1) the similarity and association among questions in the forums and (2) the groups of most associated questions. The forum managers can therefore select a group of most similar questions to be organized as a new FAQ based on the quantitative evidence.

Conventional FAQ related studies focus on retrieving constructed FAQ by user queries. This study advocates that tool assistance of FAQ finding and construction is needed in OSS projects. Appropriate tools can provide quantitative evidence and reduce the effort to construct the FAQ. Therefore, this study proposes a FAQ finding process to construct and maintain the FAQ. In the process, the questions in the forum will be clustered into several groups. Each clustered group represents a set of similar or associated questions. Accompanied visualization tools are also provided in the FAQ finding process in this study.

The rest of this paper is structured as follows. The related FAQ retrieval work will be reviewed in Section II. In Section III, the FAQ finding process is defined along with required activities and tools. Two case studies of the proposed process will be presented in Section IV.

II. Related Work

Most FAQ studies in previous literature focus on FAQ retrieval [3], [4], [5]. Very few studies are made in investigating the construction of FAQ. Wu et al. [6] use a probabilistic mixture model to interpret the domain-specific FAQ query and question-answer pairs based on several independent aspects in the medical domain. Through their analysis on medical question data, ten question types are derived by *question stems*. However, this approach can only work in the specified medical domain with the concluded ten question types. In the OSS project domain, there is no general *question stems* across projects. Hence the probabilistic mixture model approach is not applicable for constructing the FAQ in an OSS project forum. Kim et al. [3] build a system FRACT to exploit the concept of weighted terms for FAQ retrieval in a specific domain. The terms are weighted by the analysis of query logs with similar meaning. The FAQ is abstracted by weighted terms, and then clustered to retrieve the required FAQ. With this approach, the requirements of domain knowledge and handcrafted rules can be avoided when retrieving FAQ in a specific domain. This approach is applicable in the OSS project domain if each question in the forum is a query log. However, FRACT is designed for automatic FAQ retrieval with short query logs. Each query log in FRACT has less than 6 words. In the OSS project forums, each question usually has more than 6 words. Ohba et al. propose the Concept Keyword Term Frequency/Inverse Document Frequency (ckTF/IDF) method to efficiently mine *concept keywords* from large size of source code [7]. The ckTF/IDF model is a variation of conventional TF/IDF model which is simplified for mining efficiency. The assumption in ckTF/IDF model is that a keyword in source code occurs in few files with high occurring frequency in each file. However, in OSS project forums, a keyword might occur across many forum questions with low occurring frequency in each forum question. The quantizing threshold introduced in ckTF/IDF model will limit the possibility of finding such keywords. Therefore, the ckTF/IDF model is not applicable in mining keywords from forum questions.

III. The FAQ Finding Process

The shortcomings of applying current FAQ studies in OSS projects motivate the proposition of a FAQ finding process for OSS projects. As illustrated in Figure 1, the FAQ finding process includes four steps: Step-1, Semi-Automatic Keyword Selection. To efficiently manage the vast amount of questions in the forums, *contextual keywords* will be needed to abstract the questions. Each *contextual keyword* is a frequently used term in the project forums. Forum managers are responsible to identify *contextual keywords* of the project. To further ease the burden of identifying *contextual keywords*, several tools and activities are provided for semi-automatic keyword selection. These tools and activities include: (1) automatically prepare question content from *project forum question repository*. (2) automatically apply Porter's stemming [8] to pre-process question content. (3) automatically remove common terms, the preposition, and the article from question content, (4) automatically calculate TF/IDF weight. (5) select *contextual keywords* visually by forum managers. Step-2, Question Abstraction. All questions in the project forums will be abstracted by matching *contextual keywords* within the content of the question. The question abstractions are collected as a *question abstraction matrix*. Step-3, Question Clustering. Based on the *question abstraction matrix*, the questions will be clustered into different groups. Each group is a set of similar or associated questions. Step-4, Clustered Group Visualization. These clustered groups will then be presented to forum managers by suitable visualization methods. Therefore the forum managers can have quantitative and visual evidence to organize clustered groups as the FAQ. The detail of these four steps is described as follows.

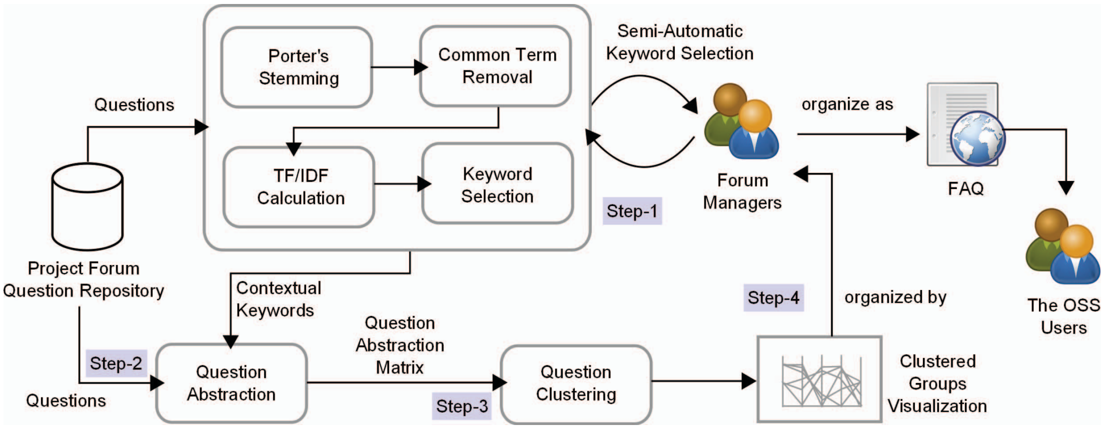


Figure 1. The semi-automatic process to find and construct FAQ.

In different OSS portals or project websites, the forum questions could be preserved in different formats. Most commonly, each forum question is preserved together with HTML tags in the repository. Those content-irrelevant HTML tags or metadata must be removed. Further, *Porter's stemming* is applied to remove the commoner morphological and inflexional endings from words. By applying stemming, word variations, such as "code", "coding", and "codes", will all be eliminated as "code". After the stemming process, common terms, the preposition, and the article will be removed from question content based on a pre-defined term table. This term table can be altered by the forum managers. After removing common terms, the *contextual keyword* candidates will be extracted from the question content. A *contextual keyword* candidate will be a term t_i, and a question will be the document d_j in the TF/IDF model. In this study, a conventional TF/IDF model [9], [10] is selected for the FAQ finding process. The term frequency tf_{ij} in the conventional model is defined as follows.

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{1}$$

Term frequency tf_{ij} gives the *measure of importance* of the term t_i within the particular document d_j. The numerator $n_{i,j}$ is the *number of occurrences* of the considered term t_i in document d_j. The denominator is the sum of $n_{i,j}$ of all k terms in document d_j. The inverse document frequency idf_i is defined as follows.

$$idf_i = \log \frac{|D|}{tf_i}, \text{ where } tf_i = |d_j : t_i \in d_j| \tag{2}$$

The numerator $|D|$ is the total number of documents in the project forums, and the denominator tf_i is the total number of documents with term t_i occurring in the document content. The inverse document frequency idf_i measures the general importance of the term t_i in all documents. The term weight $w(t_i, d_j)$ of a specific term t_i in document d_j is defined as $w(t_i, d_j) = c \times tf_{ij} \times idf_i$. It should be noticed that term wright formula is different from conventional TF/IDF model. The coefficient c is set as 100 to improve the performance of clustering and facilitate the observation of clustered groups. Additionally, the average importance of term t_i in the project forums is defined as follows.

$$w_{avg}(t_i) = \frac{1}{|D|} \sum_{j=1}^{|D|} w(t_i, d_j) \tag{3}$$

The term frequency tf_i, which is the accumulated occurrences of term t_i in all documents, and the average importance $w_{avg}(t_i)$ will both be visually provided to the forum managers for *contextual keyword* selection.

Each question content d_j will be abstracted as a vector v_j. The vector v_j contains $|T|$ elements, where $|T|$ is the total number of *contextual keywords* selected by the forum managers. Each element v_{ji} in v_j is the TF/IDF weight of term t_i in d_j. Therefore v_{ji} is defined as follows.

$$v_{ji} = \begin{cases} w(t_i, d_j) & : \text{if } t_i \in d_j \\ 0 & : \text{otherwise} \end{cases} \tag{4}$$

The value of v_{ji} will be set to zero if the term t_i is not found in question content d_j. The output of the question abstraction process in Figure 1 is a *question abstraction matrix* M. The size of M is $|D| \times |T|$. During the practice of FAQ finding process in various OSS projects, sparse vectors are found in M. The sparse vectors are question abstractions with very few *contextual keywords* inside, which means that almost all v_{ji} will be zero. Such sparse vectors will become the interference to final clustering. Therefore sparse vectors can be removed during question abstraction by specified filtering rules. The filtering rules are specified by the threshold of maximal allowed number of zeros contained in sparse vectors. In this study, sparse vectors that contain more than $|T| - 2$ zeros will be removed.

The Lloyd's K-Means clustering algorithm is used to cluster all question abstractions. Each question abstraction v_j is assumed an observation with $|T|$ dimensions. Through K-Means clustering, the total $|D|$ forum questions will be clustered as $|K|$ groups. The $|K|$ is an important factor that must be decided to apply the K-Means clustering. Unfortunately, there is no general rule to determine the size of $|K|$ in K-Means clustering. In this study, the size of $|K|$ is set by $|K| = \frac{|D|}{S_m}$, where S_m is the appropriate size of questions in each clustered group. Therefore each cluster will have at most S_m questions. The size of S_m is a tuning point in the FAQ finding process for the forum managers.

When $|D|$ is large, the number of *contextual keywords* and the size of $|K|$ might both be large. The large number of *contextual keywords* further result in high-dimensional $|K|$ clustered groups. Each dimension in clustered groups is a *contextual keyword* in FAQ finding process. The forum managers will have difficulties to inspect and compare large size of clustered group data by multiple dimensions. In the FAQ finding process, two visualization methods, the *Parallel Coordinates* [11] and the *Scatter Plot*, are supported to ease the difficulties. The forum managers can view the TF/IDF weights of multiple *contextual keywords* on all clustered groups through *parallel coordinates*. Each coordinate in *parallel coordinates* represents a *contextual keyword*. With *parallel coordinates*, the *contextual keywords* with significant TF/IDF weights in a specific clustered group can be identified easily. However, the *parallel coordinates* method is insufficient to view question distribution between two *contextual keyword* coordinates. Therefore, the *scatter plot*, is used as a complementary method to the *parallel coordinates*. The *scatter plot* can be used to investigate the question distribution in *Cartesian coordinates*.

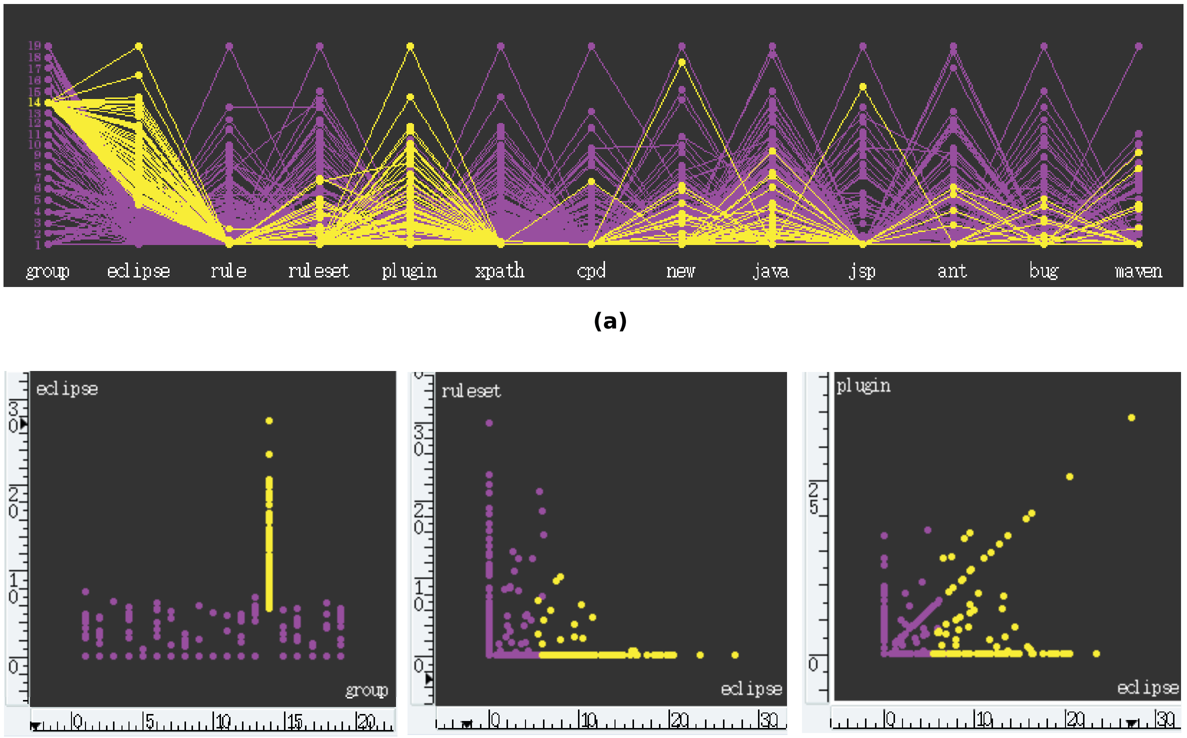
IV. Case Studies

PMD (http://sourceforge.net/projects/pmd) is selected for case studies. The project forums of PMD on SourceForge.net serve as the *project forum question repository* in Figure 1. The forum questions are collected by a web crawler from question pages during March 2010. The web crawler is implemented using Scrapy (http://scrapy.org) framework. There are 986 questions analyzed and clustered in the case studies. Tools are implemented with Python following algorithms introduced in Section III to automatically do the analysis and clustering. A question page on SourceForge.net contains three major parts: (1) the question title, (2) the question body, and (3) the response. The first study investigates the result of using question body in finding possible FAQ clustered groups, and the second study is conducted by using the question title in finding possible FAQ clustered groups. In these two studies, GGobi (http://www.ggobi.org) is integrated into the FAQ finding process to visualize the clustering results.

Following the process introduced in Section III, terms that have $tf_i < 10$ and $w_{avg}(t_i) < 0.01$ are removed from *contextual keywords*. The keyword selection results in 882 *contextual keywords*. Sparse vectors with more than 880 zeros will be removed before clustering. Therefore the number of question abstraction vectors is reduced to 944. These 944 question abstraction vectors are clustered into 19 groups. The visualization of 19 groups is presented in Figure 2(a). The first coordinate "group" is the identification of clustered groups labelled with numbers. Forum managers can highlight a specific group by selecting the identification of clustered groups. Despite the first coordinate "group", other coordinates in Figure 2 represent selected *contextual keywords* and keyword weight in each clustered group. Through the *parallel coordinates* visualization, the forum managers can find the relevance between specific *contextual keywords* and clustered groups. In Figure 2(a), the highlighted "group 14" seems mostly related with keywords "eclipse", "ruleset", "plugin", "new", "java", "bug" and "maven". For example, Figure 2(b) shows that questions in "cluster 14" are highly related to "eclipse" comparing to the questions in other groups. The highlighted spots in Figure 2(b) represent questions in "group 14". The "eclipse" term weight $w(t_i, d_j)$ of questions in "group 14" have distribution from 5.49 to 27.46, and the median weight value is 10.30. However, the *parallel coordinates* method is insufficient when forum managers want to further view the question distribution between "ruleset" coordinate and "eclipse" coordinate in "group 14". Instead, the *scatter plot* can be used to view the question distribution in "eclipse"-"ruleset" coordinates as illustrated in Figure 2(c). Figure 2(c) shows that 8 questions in "group 14" have medium weight ($4 < w(t_i, d_j) < 12$) both on "ruleset" and "eclipse", and most other questions in "group 14" have high weight ($12 < w(t_i, d_j)$) only on "eclipse". Another example in Figure 2(d) shows that several questions have medium weight or high weight both on "eclipse" and "plugin". Questions having medium weight or high weight both on "eclipse" and "plugin" might be various problems associated with *PMD Eclipse plugin module*.

The *contextual keywords* in the first case study are selected by defined filtering rules but not by forum managers. The selection results in 882 *contextual keywords*, which is still too large for efficient clustering. The clustering performance motivates us to conduct the second case study to seek for opportunities in reducing the number of *contextual keywords*. The purpose in the second case study is to use question titles as the input of FAQ finding process instead of question bodies. Question titles usually serve as question body abstractions in the forums. If the question titles do provide meaningful abstractions of the question bodies, it is possible that question titles can be used to find FAQ with smaller number of *contextual keywords*.

In the second study, terms that have $tf_i < 10$ and



(a)

(b)

(c)

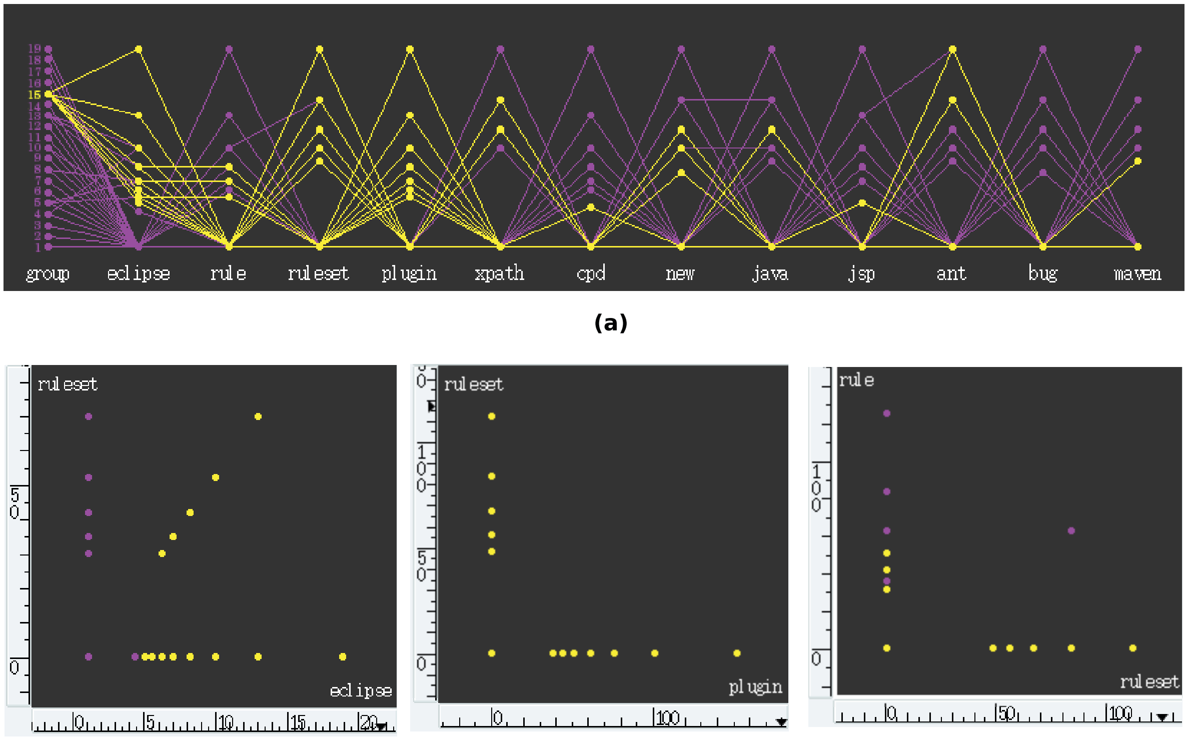
(d)

Figure 2. Visualization of question body based clustering results. (a) *Parallel coordinates* visualization of 19 groups and selected *contextual keywords*. The “group 14” is highlighted. (b) Scatter plot of *(group, eclipse)* in “group 14”. (c) Scatter plot of *(eclipse, ruleset)* in “group 14”. (d) Scatter plot of *(eclipse, plugin)* in “group 14”.

$w_{avg}(t_i) < 0.01$ are removed from *contextual keywords*. The keyword selection results in 43 *contextual keywords*. The number of question abstraction vectors are reduced to 342 after the sparse vectors are removed. The parallel coordinates visualization of all clusters is presented in Figure 3(a). The “group 15” in Figure 3(a), which contains most questions that also occur in “group 14” of Figure 2(a), is highlighted. Figure 3(b) shows the questions distribution in “eclipse”-“ruleset” coordinates, which can be compared with Figure 2(c). Obviously the number of questions containing medium and high weight keywords are much less than the clustering results of the first case study. Also, there are several questions having medium weight of “ruleset” and “plugin” in the first case study, but no question containing medium weight of “ruleset” and “plugin” in Figure 3(c). This finding implies that questions having medium weight of “ruleset” and “plugin” in bodies do not contain these two terms together in titles. Further, Figure 3(d) reveals that “rule” and “ruleset” might be used synonymously in the question titles. Questions having both “rule” and “ruleset” in the bodies might contain only one of these two keywords in the titles. These observations in the second study indicate that the question titles in PMD forums are not good abstraction of the question bodies.

V. Conclusion and Future Work

This study brings out the need of assisting forum managers to find and construct FAQ, which did not gain enough attention previously. Further, a semi-automatic FAQ finding process is advocated for OSS forum managers to use with quantitative and visual evidence. As a preliminary study, common content pre-processing, document clustering and visualization techniques are all developed in the FAQ finding process to exploit the effectiveness of proposed process on OSS projects. Two case studies of the PMD project are presented along with the analysis. The results show that proposed process can discover associated questions as clustered groups. However, the amount of *contextual keywords* in the process could be large if there are lots of questions in the forums. The amount of *contextual keywords* should be further reduced. Also, by applying the process with question titles as input in the second study, we confirm that poor quality of question title limits the possibility to find associated questions as FAQ. Several enhancements of current FAQ finding process are under development. For example, many keywords in questions relate to source code or program errors. By exploiting the semantic association of these keywords in the source code, the efficiency to identify *contextual keywords* can be improved. Besides,



(a)

(b)

(c)

(d)

Figure 3. Visualization of question title based clustering results. (a) *Parallel coordinates* visualization of 19 clusters and selected *contextual keywords*. The “group 15” is highlighted. (b) Scatter plot of *(eclipse, ruleset)* in “group 15”. (c) Scatter plot of *(plugin, ruleset)* in “group 15”. (d) Scatter plot of *(resultset, rule)* in “group 15”.

several variants of the TF/IDF model introduced in [10] will be implemented in the FAQ finding process to validate the effectiveness of model variants in FAQ finding. Further, more larger scale empirical studies will be performed to investigate mismatches between existing FAQ and the clustered FAQ in OSS projects.

REFERENCES

[1] S. K. Sowe, I. Stamelos, and L. Angelis, “Understanding Knowledge Sharing Activities in Free/Open Source Software Projects: An Empirical Study,” *Journal of Systems and Software*, vol. 81, no. 3, pp. 431–446, Mar. 2008.

[2] K. R. Lakhani and E. von Hippel, “How Open Source Software Works: “Free” User-to-User Assistance,” *Research Policy*, vol. 32, pp. 923–943, 2003.

[3] H. Kim and J. Seo, “Cluster-Based FAQ Retrieval Using Latent Term Weights,” *IEEE Intelligent Systems*, vol. 23, no. 2, pp. 58–65, Mar. 2008.

[4] E. Sneiders, “Automated FAQ Answering with Question-Specific Knowledge Representation for Web Self-Service,” in *Proceedings of the Second Conference on Human System Interactions*, May 2009, pp. 295–302.

[5] S.-Y. Yang, “Developing of an Ontological Interface Agent with Template-based Linguistic Processing Technique for FAQ Services,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 4049–4060, Mar. 2009.

[6] C.-H. Wu, J.-F. Yeh, and M.-J. Chen, “Domain-Specific FAQ Retrieval Using Independent Aspects,” *ACM Transactions on Asian Language Information Processing*, vol. 4, no. 1, pp. 1–17, Mar. 2005.

[7] M. Ohba and K. Gondow, “Toward Mining “Concept Keywords” from Identifiers in Large Software Projects,” in *Proceedings of the International Workshop on Mining Software Repositories*, May 2005, pp. 1–5.

[8] M. F. Porter, “An Algorithm for Suffix Stripping,” *Program*, vol. 14, no. 3, pp. 130–137, July 1980.

[9] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” *Information Processing & Management*, vol. 25, no. 5, pp. 513–523, 1988.

[10] S. Lee and H.-J. Kim, “News Keyword Extraction for Topic Tracking,” in *Proceedings of the Fourth International Conference on Networked Computing and Advanced Information Management*, Sept. 2008, pp. 554–559.

[11] A. Inselberg and B. Dimsdale, “Parallel Coordinates: a Tool for Visualizing Multi-Dimensional Geometry,” in *Proceedings of the 1st Conference on Visualization*, Oct. 1990, pp. 361–378.