Contents

REGULAR PAPERS

Evaluating a Lightweight Forum-based Tool: Empirical Studies on Requirements Elicitation Process
Han Lai and Yuze Ni

A Data Analysis Method and Its Applications in EXCEL
Jinlei Qin, Yuguang Niu, and Zheng Li

Measuring Developers’ Design Contributions in Evolved Software Projects
Maen Hammad, Mustafa Hammad, Hani Bani-Salameh, and Ebaa Fayyoumi

An Improved Face Recognition Technique Based on Modular Multi-directional Two-dimensional Principle Component Analysis Approach
Xiaoqing Dong and Hongcai Chen

Designing and Implementation of an Online System for Electronic Contract Negotiation Based on Electronic Signature
Juntao Gu and Xiaodong Zhu

Lexical-semantic SLVM for XML Document Classification
Jun Long, Luda Wang, Zude Li, Zuping Zhang, Huiling Li, and Guihu Zhao

Novel Metrics for Bug Triage
V. Akila and Dr. G. Zayaraz

Implementing Six Sigma to Reduce Online Petitions on the E-petition System in Taiwan Power Company
Pei–Shih Chen, Chang–Hsien Hsu, Shih–Yun Hsu, and Chun–Ming Yang

Energy Constrained Target K-coverage Algorithm in Heterogeneous Wireless Sensor Networks
Yong-sheng Wang

Research on Deep Web Query Interface Clustering Based on Hadoop
Baohua Qiang, Rui Zhang, Yufeng Wang, Qian He, Wei Li, and Sai Wang
Evaluating a Lightweight Forum-based Tool: Empirical Studies on Requirements Elicitation Process

Han Lai¹,²
¹Chongqing Key Lab. of Electronic Commerce & Supply Chain System, Chongqing Technology and Business University, Chongqing, China
²School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing, China
Email: laihan_ctbu@126.com

Yuze Ni³,*
³State Key Lab. of Software Engineering, Computer School, Wuhan University, Wuhan, China
Email: niyuze@whu.edu.cn

Abstract—The objective of this paper is to empirically evaluate SKLSEForum, a lightweight forum-based tool, for predicting the likelihood of acceptance of the tool in requirements elicitation practice and finding directions for improvement. We analyzed three data sets from a family of experiments, deriving from two replicated controlled experiments and a survey of expert panel. Results showed that SKLSEForum can improve the quantity of posts and judge satisfaction of requirements. In addition, the tool was perceived as useful and easy to use by the participants, who also expressed their intention to use SKLSEForum in the future. Furthermore, some insights about the improvement of SKLSEForum also have been found.

Index Terms—requirements engineering, requirements elicitation process, forum, expert panel, controlled experiment

I. INTRODUCTION

Open source and distributed software development communities already take advantage of Internet-based communication and coordination technologies to operate effectively. As social media has gained adoption, opportunities for creating software in new ways have risen to enhance and augment the old [1].

Ahmed, et al. [2] analyzed a dataset consisting of 1880 open source software projects covering a broad range of categories in this investigation. The results show that online forums play a significant role in managing software defects, implementation of new requirements and providing support to the users in open source software and have become a major source of assistance in maintenance of the open source projects.

However, many problems have emerged in the forum-based requirements elicitation [3], such as the disorder of post format, incompleteness of requirements information, low participation and enthusiasm of stakeholders, etc.

SKLSEForum [4] is a lightweight tool based on forum to support the forum-based requirements elicitation process. The tool includes five specialized features, i.e., thread template, reward topic, online chatting, thread state identification, and score exchange, with which it can complete and clarify the requirements’ content, incentive the desire of participation.

This paper presents the empirical study, involving controlled experiment and survey of expert panel, defined and conducted to evaluate the SKLSEForum. We discuss the quantitative and qualitative findings of our empirical study and their potential for improving the SKLSEForum. These findings indicate that SKLSEForum could be considered as a promising tool for requirements elicitation.

This paper is organized as follows: Section 2 gives an overview of the five specialized features of our lightweight forum-based tool. Section 3 presents the evaluation method based on controlled experiments and a survey of expert panel, and how we conducted the empirical experiments. Section 4 gives the results of the experiments. In Section 5 we analyze the threats to the experiment validity. Section 6 presents the related work. Finally, Section 7 gives the conclusions and presents the future work.

II. SKLSEFORUM

P. Laurent and J. Cleland-Huang [3] explored and evaluated the forum-based requirements gathering and prioritization processes adopted by vendor-based open source software projects. They pointed that almost all of the forums they surveyed did a very poor job in managing the status of each feature request.
TABLE I
SKLSEFORUM’S FIVE FEATURES

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Requirements Activity</th>
<th>Functional Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thread Templates</td>
<td>User can create a new opinion topic according to the guidance of requirements description template; and the topic must pass the template-based verification</td>
<td>Requirements Creation</td>
<td>Standardize the topic format, and improve the judge satisfaction of requirements</td>
</tr>
<tr>
<td>Reward-topic</td>
<td>The forum administrator can submit the reward topic to reward Q &amp; A, and general users can answer the questions to prompt their reputation and increase scores</td>
<td>Requirements Improvement</td>
<td>Increase the reward topic’s attention to gain more comments or suggestions, leading to higher judge satisfaction</td>
</tr>
<tr>
<td>Online Chatting</td>
<td>Users can chat with each other online in forum</td>
<td>Requirements Decision</td>
<td>Promote the communication and negotiation between users, leading to requirements with more judge satisfaction</td>
</tr>
<tr>
<td>Thread State Identification</td>
<td>The opinion topic has the marked status (e.g. New, Suggestion collected, Locked, and Unlocked, etc.). The transition flow among these states is based on authority and rules.</td>
<td>Requirements Management</td>
<td>Help the management of requirements, and instruct users to identify the status, leading to users’ more participation on requirements decision to get higher requirements judge satisfaction</td>
</tr>
<tr>
<td>Score Exchange</td>
<td>General user can exchange the prize by their scores</td>
<td>The Whole Process</td>
<td>As an incentive mechanism to increase the users’ participation and enthusiasm</td>
</tr>
</tbody>
</table>

Moreover, the stakeholders’ participation is low because of the lack of two-way conversations which can engage more stakeholders in the requirements elicitation process. So it is necessary to provide a channel for stakeholders to communicate the process and decisions, seek clarification, etc.

To solve the problems, P. Laurent and J. Clelland-Huang proposed, we established a framework ReqForum (see Fig. 1) to define the meta-model of the requirement elicitation forum, and developed a lightweight forum-based tool SKLSEForum with five specialized features (see Fig. 2) introduced in the requirements elicitation process [4]. The five features and their functional goals are introduced in Table I.

```
<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Requirements Activity</th>
<th>Functional Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thread Templates</td>
<td>User can create a new opinion topic according to the guidance of requirements description template; and the topic must pass the template-based verification</td>
<td>Requirements Creation</td>
<td>Standardize the topic format, and improve the judge satisfaction of requirements</td>
</tr>
<tr>
<td>Reward-topic</td>
<td>The forum administrator can submit the reward topic to reward Q &amp; A, and general users can answer the questions to prompt their reputation and increase scores</td>
<td>Requirements Improvement</td>
<td>Increase the reward topic’s attention to gain more comments or suggestions, leading to higher judge satisfaction</td>
</tr>
<tr>
<td>Online Chatting</td>
<td>Users can chat with each other online in forum</td>
<td>Requirements Decision</td>
<td>Promote the communication and negotiation between users, leading to requirements with more judge satisfaction</td>
</tr>
<tr>
<td>Thread State Identification</td>
<td>The opinion topic has the marked status (e.g. New, Suggestion collected, Locked, and Unlocked, etc.). The transition flow among these states is based on authority and rules.</td>
<td>Requirements Management</td>
<td>Help the management of requirements, and instruct users to identify the status, leading to users’ more participation on requirements decision to get higher requirements judge satisfaction</td>
</tr>
<tr>
<td>Score Exchange</td>
<td>General user can exchange the prize by their scores</td>
<td>The Whole Process</td>
<td>As an incentive mechanism to increase the users’ participation and enthusiasm</td>
</tr>
</tbody>
</table>
```

Figure 1. Forum-based requirement elicitation framework. [4]

In the forum-based requirement elicitation framework, stakeholders are classified into General Users and Management Users. The stakeholders who use the forum to give feedback on the open source communities’ opinions or suggestions are deemed as General Users. Requirements analyst and the forum system administrator are deemed as Management Users. Each type of stakeholders has their own Identity, Rights, Reputation, Capability and Score in the forum. Different type of stakeholders has the different Rights to use forum’s functions; the stakeholder’s Reputation represents the trusted degree and also represents its prestige in the forum; the stakeholder’s Capability represents the degree of mastering domain knowledge, e.g. how to use the open source system’s function or the applied domain of open source system used; the stakeholder’s Score represents its active degree which participate in the forum’s discussion. The Requirements are represented as threads. A requirement is corresponding to a Topic and multiple Posts. A general Topic consists of title and body. The Req. Relation represents the relationship, which exists between the requirements. The S-ROperation represents the stakeholders can perform various operations on the forum’s topic and post to reflect the activity of different stakeholders to get involved in requirement elicitation. The S-SOperation represents the operations between Stakeholders, e.g. management user rewards score to general users in the forum to discuss, stakeholders communicate with each other by sending and receiving message etc.

Different from the ordinary forum, a requirement elicitation process can be supported by SKLSEForum. The process is shown in Fig. 3. Firstly, the stakeholders can create requirements by using the thread templates; secondly, in the phrase of requirements improvement, the stakeholders comment on the posts, and the Analysts can set reward topic to draw more general user’ attention and participate the feedback. Then, during the requirements decision, the online chatting provides a platform for stakeholders to communicate with each other in the synchronous setting. In addition, the thread state identification can help manage the requirements and prompt the awareness about the requirements’ disposition. The score exchange can stimulate stakeholders to participate the discussion.

By using the five specific features in SKLSEForum, we can join these features into the SECI model [5]. The thread templates transform the stakeholders’ ideas into formative requirements and make other stakeholders understand more about the requirements, thus it facilitates the stakeholders’ knowledge Externalization and
III. SKLSEFORUM EVALUATION

This section describes the empirical validation of SKLSEForum. This was done by conducting two replicated controlled experiments and a survey of expert panel.

A. Evaluation Design

The evaluation design refers to the guidelines proposed by Wohlin [6]. According to the Goal-Question-Metric (GQM) template [7], the goal of the experiment is to test the SKLSEForum with respect to its performance and perception for predicting the likelihood of acceptance of the tool in requirements elicitation practice and find directions for improvement.

The research questions addressed by the experimentation are:

**RQ1:** Can the performance of requirements elicitation be improved by applying SKLSEForum?

**RQ2:** Is SKLSEForum perceived as both easy to use and useful?

**RQ3:** Is there an intention to use SKLSEForum in the future?

**RQ4:** What improvement can be made for SKLSEForum?

The first research question (RQ1) was addressed by defining the following hypotheses:

- **H1**: SKLSEForum will not improve the quantity of posts.
- **H2**: SKLSEForum will not improve the judges' satisfaction of requirements.

The second research question (RQ2) was addressed by the formulation of the following hypotheses:

- **H3**: SKLSEForum is perceived as not useful.
- **H4**: SKLSEForum is perceived as difficult to use.

The third research question (RQ3) was addressed by defining the following hypotheses:

- **H5**: There is no intention to use SKLSEForum in the future.
Table II lists all the dependent variables in the experiments. The chosen Perceived-based dependent variables are based on the adapted Davis’s TAM [8][9]. The TAM is one of the most widely applied theoretical models to study user acceptance and usage behavior of emerging information technologies, and it has received extensive empirical support through validations and replications [10][11].

Table II

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity of Posts</td>
<td>Count</td>
</tr>
<tr>
<td>Judge Satisfaction</td>
<td>Calculated as the mean of the grades from two teachers</td>
</tr>
<tr>
<td>Perceived Usefulness</td>
<td>Calculated as the mean of the items obtained from the questionnaire</td>
</tr>
<tr>
<td>Perceived Ease of Use</td>
<td>Calculated as the mean of the items obtained from the questionnaire</td>
</tr>
<tr>
<td>Self-predicted Future Use</td>
<td>Calculated as the mean of the items obtained from the questionnaire</td>
</tr>
</tbody>
</table>

B. Questionnaire Design

To obtain the opinions of experts and students in the experiment groups about the perceived usefulness, perceived ease of use and self-predicted future use of the whole SKLSEForum, we designed the questions in the questionnaire based on the adapted Davis’s TAM [12], which has six items in perceived usefulness, six items in perceived ease of use and two items in self-predicted future use, as shown in Table III. The questionnaire also included five semi-open questions to inquire the subjects whether the five specific features could achieve their functional goals in Table I, and one open question to collect ideas, advice or suggestions to SKLSEForum for improvement in the future.

C. Participants and Training

28 fourth-year Computer Science undergraduate students at the Chongqing Technology and Business University participated in the controlled experiments. The participants are volunteers and we provide a movie ticket as the reward of participation. In order to avoid persistence effects, the students are chosen from those who had never done similar experiments before. The students were randomly divided into 4 groups, (i.e. two experiment groups, and two control groups), each composed of 7 members.

The subjects respectively received 80 minutes training on SKLSEForum or ordinary forum (i.e. Discuz 1, the most widely used forum application in China and is the SKLSEForum’s prototype). The training was consisted of 40 minutes presentation for explaining the main characteristics of SKLSEForum or Discuz, and 40 minutes for tool demo. They did not receive further teachers’ assistance while they were using the forum.

We invited 16 experts from different professional backgrounds to participate our survey by email and 14 accepted (representing a take-up rate of 88%), including requirements engineering (RE) researchers, experienced open source community (OSC) forum managers, active forum users and open source software consultants. In addition, we also included experts from collaborative computing. Table IV shows the profile of the experts who participated in the study.

Table III

<table>
<thead>
<tr>
<th>Code</th>
<th>Measurement standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>Using this collaborative tool in my job, I would be able to accomplish requirements elicitation tasks more quickly.</td>
</tr>
<tr>
<td>U2</td>
<td>Using this tool would improve my performance on requirements elicitation tasks.</td>
</tr>
<tr>
<td>U3</td>
<td>Using this tool for requirements elicitation tasks would increase my productivity.</td>
</tr>
<tr>
<td>U4</td>
<td>Using this tool would enhance my effectiveness on requirements elicitation tasks.</td>
</tr>
<tr>
<td>U5</td>
<td>Using this tool would make it easier to do requirements elicitation tasks.</td>
</tr>
<tr>
<td>U6</td>
<td>I would find this tool useful to perform requirements elicitation tasks.</td>
</tr>
<tr>
<td>E1</td>
<td>Learning to operate this tool would be easy for me.</td>
</tr>
<tr>
<td>E2</td>
<td>I would find it easy to get this tool to do what I want it to do to perform most of the requirements elicitation tasks.</td>
</tr>
<tr>
<td>E3</td>
<td>My interaction with this tool would be clear and understandable.</td>
</tr>
<tr>
<td>E4</td>
<td>It would be easy to become skillful in using this tool.</td>
</tr>
<tr>
<td>E5</td>
<td>It would be easy to remember how to perform various requirements elicitation tasks using this tool.</td>
</tr>
<tr>
<td>E6</td>
<td>I would find this tool easy to use.</td>
</tr>
</tbody>
</table>

Table IV

<table>
<thead>
<tr>
<th>Roles</th>
<th>Years of Experience</th>
<th>Research Area</th>
<th>Position / Relevant Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>6</td>
<td>RE Associate Professor</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>8</td>
<td>OSC and RE Professor</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>4</td>
<td>RE and CC Associate Professor</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>3</td>
<td>RE Associate Professor</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>10</td>
<td>OSC and RE Professor</td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>5</td>
<td>OSC Forum Manager</td>
</tr>
<tr>
<td>7</td>
<td>P</td>
<td>4</td>
<td>OSC Forum Moderator</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>7</td>
<td>OSC and RE IT Consultant</td>
</tr>
<tr>
<td>9</td>
<td>P</td>
<td>9</td>
<td>OSC IT Consultant</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>6</td>
<td>OSC IT Consultant</td>
</tr>
<tr>
<td>11</td>
<td>P</td>
<td>5</td>
<td>OSC Business Analyst</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>3</td>
<td>OSC and RE Forum Moderator</td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td>4</td>
<td>OSC and RE Forum Manager</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
<td>3</td>
<td>OSC and CC Active User</td>
</tr>
</tbody>
</table>

A: Academic P: Practitioner B: Both CC: Collaborative Computing

The experts are given some materials prepared with the questionnaire. The prepared materials were an instructional video describing the main characteristics of SKLSEForum, and the procedure for applying it.

1 Discuz: http://www.discuz.net
D. Controlled Experiment & Survey

Following the general plan of evaluation, we carried out two controlled experiments and a survey, as is shown in Fig. 4.

The evaluation process involves controlled experiment and survey, depending on the type of participants and the experimental design employed.

**Controlled experiment:** Undergraduate students were divided into four groups of 7 members. Each experiment consisted of two groups, and the second experiment is a strict replication of the first.

**Survey:** The survey consists of the 14 experts.

In order to discover potential problems in the questionnaire design and application, we ran a pilot study involving three researchers in the fields of open source community and RE. Through analyzing the pilot test responses, we made changes as a result of the feedback. We also used the pilot study to assess the time required to complete the questionnaire which we estimated to be 10 min.

In Controlled experiment, we conducted two replicated controlled experiments under the same conditions (strict replication), changing only the students [13]. The control groups use Discuz as ordinary forum, while experiment group use SKLSEForum. Each of the teams was asked to design and document software architecture of a “university library web portal” for three days. The second experiment was performed on the same three day as the first one. This application scenario was deemed sufficiently complex such that students would be able to generate requirements for three days and also because it is highly familiar to college students. The teams were asked to use the tools in the process of requirements elicitation, to capture the key business scene, case, questions and so on. During the experiment, the method of role play [14] was adopted and the students used the forum by means of remote access.

After three days, we collected the requirements the students proposed, and the quantity of posts each proposed in four groups is sorted and counted; the judge satisfaction value of requirements is graded by teachers. Moreover, the two experiment groups were asked to finish the questionnaires after the experiment. In order to prevent the potential bias, the students were told that their answers would be treated anonymously, and they need not please the experimenters by favorable judgments of SKLSEForum. Moreover, before the experiment, the students were also informed that their performance in the experiment would not affect them or their grade.

In the survey of expert panel, we handed out the questionnaires by E-mail, and collected the feedback from the experts about the perceived usefulness, perceived ease of use and self-predicted future use of SKLSEForum.

IV. RESULTS

This section reports the analysis result of the data collected during the studies. First, we assess whether the SKLSEForum can improve the quantity of posts and judge satisfaction of the requirements elicited. Then, we investigate perception-based variables (i.e. perceived useful, ease of use, self-predicted future use) the results from the students’ operation experience and experts’ knowledge.

The quantitative analysis was performed by using the SPSS v19 statistical tool and $\alpha=0.05$.

A. Performance-based Measures

After three days of controlled experiments, we collect the whole requirements elicited from the four group students. Table V shows the mean quantity of posts by each student contributed in the four experiments.

**TABLE V**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Statistical analysis results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>1st Control Group</td>
<td>8.29</td>
<td>2.498</td>
<td>$p=0.010$ (t-test)</td>
</tr>
<tr>
<td></td>
<td>1st Experiment Group</td>
<td>12.71</td>
<td>3.592</td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>2nd Control Group</td>
<td>8.71</td>
<td>1.704</td>
<td>$p=0.005$ (t-test)</td>
</tr>
<tr>
<td></td>
<td>2nd Experiment Group</td>
<td>13.57</td>
<td>3.867</td>
<td></td>
</tr>
</tbody>
</table>

The data from Table V highlight that the students in experiment group has generated more posts that represent the key business scenarios, use cases, and problems than control group. The data are normally distributed, therefore, one-tailed two independent sample t-test are adopted.

The comparison of the mean quantity of posts highlights that the SKLSEForum can easily promote the effectiveness of the stakeholders’ participation and stimulates their enthusiasm. By analyzing the posts elicited from the experiment groups, it seems that stakeholders is encouraged to generate more requirements in order to get redeem gifts, and based on the online chatting, more opinions and ideas are captured.

The two teachers performed a general analysis of the requirements, to determine the judge satisfaction value. Table VI represents the judge satisfaction value of the requirements elicited from the four group students.
It is obvious that the judge satisfaction of requirements in experiment groups (i.e., the 1st experiment group and 2nd experiment group) got higher values than the control groups (i.e., the 1st control group and 2nd control group).

In conclusion, the analysis of the data collected during the experiment indicate that using SKLSEForum can improve the requirements’ judge satisfaction during the requirements elicitation process in a global environment. By analyzing the requirements elicited from the experiment groups, it seems that stakeholders are more likely to submit requirements with more judge satisfaction by using SKLSEForum.

### B. Perception-based Measures

The construct validity of the metrics with the perceived usefulness, perceived ease of use and self-predicted future use was evaluated by using inter-item correlation analysis. The evaluation was based on two criteria: convergent and discriminant validity. Low divergent validity values indicate high discriminant validity. According to Campbell and Fiske [15] an item’s convergent validity value must be higher than its divergent validity value; otherwise the data on the item should not be used. The results of inter-item correlation analysis (see Table VII for an example) for the three data sets were:

- 1st Experiment Group: E3 and E6 did not pass the validity test. E3 and E6 were therefore excluded from the analysis. With the exclusion of this item, the average convergent validity was 0.60 for perceived usefulness, 0.43 for perceived ease of use, and 0.77 for self-predicted future use.
- 2nd Experiment Group: S2 did not pass the validity test. With the exclusion of this item, the average convergent validity was 0.49 for perceived usefulness, 0.40 for perceived ease of use, and 0.53 for self-predicted future use.
- Expert panel: E2 and U3 did not pass the validity test. With the exclusion of this item, the average convergent validity was 0.52 for perceived usefulness, 0.45 for perceived ease of use, and 0.64 for self-predicted future use.

The use of multiple items to measure a same construct requires the examination of the reliability or internal consistency of the questionnaire. We calculated the Cronbach's alpha for each set of questions in the questionnaire. For this analysis, the items that did not pass the validity test were excluded from their data sets. Table VIII indicates that almost all the questions have an alpha value greater than 0.7, which is a common reliability threshold [16], the least reliable one is ‘Perceived Ease of Use’ of 1st Experiment Group.

Table IX shows descriptive statistics about perceived usefulness, perceived ease of use and self-predicted future use of SKLSEForum provided by the experiment group students and expert panel. The result shows that we corroborated empirically the students and experts perceived SKLSEForum as being useful and easy to use, and that there is an intention to use it in the future.

In order to analyze the actual usefulness and ease of use of five specific features in SKLSEForum, the students were asked to write down their opinions based on the actual using experience in SKLSEForum.

The students’ response showed some disagreement that score exchange feature is useful. One reason could be that potential stakeholders could not realize the function of the scores, and are regardless of the exchange gift or virtual rewards. Another one could be that since the period of experiment is short, thus the stakeholders cannot accumulate many scores in just three days to exchange gift or virtual rewards. The few opportunity for stakeholders to use the score exchange feature lead to the students’ disagreement with the usefulness of the feature.

Some students indicated that the thread templates feature and thread state identification feature are not easy to use as expected. For the thread templates feature, a factor that might reduce the ease of use is that the usability of the thread templates is restricted to the experienced users only. Templates can conduct the mandatory and optional items for the requirements information, but for the novice, the feature is not easy to apply especially when they set up simple requirements. And a mass of templates will cause a waste of time and aggrandize redundancy of requirements information, making the thread templates feature not easy to use. On the other hand, the thread state identification feature is mainly applied by forum administrators or project managers to add or alter the requirements states. Similarly for the novice, it is not easy to distinguish or identify the requirement states, let alone conducting the requirement elicitation process according to the requirements states.

In the structured questionnaire conducted during the expert panel, the experts were asked to feedback their views and suggestions about the SKLSEForum.

Some experts pointed out that by applying the SKLSEForum, it is unavoidable to increase the dependence degree on stakeholders’ participation. Despite this conforms to our intention of promoting the effectiveness of the stakeholders’ participation, one expert said “This will increase the management cost of the open forum”. One practitioner also remarked that “In open forum, the requirements provided by users are imperfect and still need project manager to collect and consolidate”. Another manager said that “The topic status is not necessary for users, but is more important for site manager”.

Some experts went further and suggested other features to incentive the stakeholders’ participation: “You can show the users ranking based on the contribution

---

**Table VI**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Group</th>
<th>T1</th>
<th>T2</th>
<th>Judge Satisfaction Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>1st Control Group</td>
<td>62</td>
<td>70</td>
<td>66</td>
</tr>
<tr>
<td>1st</td>
<td>1st Experiment Group</td>
<td>81</td>
<td>87</td>
<td>84</td>
</tr>
<tr>
<td>2nd</td>
<td>2nd Control Group</td>
<td>59</td>
<td>67</td>
<td>63</td>
</tr>
<tr>
<td>2nd</td>
<td>2nd Experiment Group</td>
<td>76</td>
<td>85</td>
<td>80.5</td>
</tr>
<tr>
<td>T1: 1st Teacher</td>
<td>2nd: 2nd Teacher</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
that their answers would be treated anonymously, and their performance in the experiment would not affect them or their grade.

V. THREATS TO VALIDITY

In the following, we explain how we dealt with the validity threats.

Construct validity: By using the convergent and discriminant validity, we eliminated the invalid data in the feedback of questionnaire. The internal consistency of the questions was verified by using the Cronbach’s α which is over 0.7 for all items instead for one, which is higher >0.6 and that is valid for exploratory studies [17].

Internal validity: The experts had enough maturity level to participate in the expert panel due to their previous knowledge and experience of Open Source Community, Requirements Engineering and Cooperative Computing, thus the responses provided by the experts on the SKLSEForum can be biased by the personal experience and domain knowledge. The teachers who evaluate the requirements quantity of posts and judge satisfaction face the same question, so we employed two teachers who come from third-party to avoid the bias. Moreover, the results may be influenced by the students’ ability and energy kept during the experiment period. We have reduced the students’ bias by informing the students

<table>
<thead>
<tr>
<th>Table VII</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTER-ITEM CORRELATION ANALYSIS FOR THE FIRST EXPERIMENT GROUP</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Perceived Usefulness</th>
<th>Perceived Ease of Use</th>
<th>Self-predicted Future Use</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>1.00</td>
<td>0.65</td>
<td>0.47</td>
<td>Y E S</td>
</tr>
<tr>
<td>U2</td>
<td>0.44</td>
<td>1.00</td>
<td>0.47</td>
<td>Y E S</td>
</tr>
<tr>
<td>U3</td>
<td>0.65</td>
<td>0.81</td>
<td>0.71</td>
<td>Y E S</td>
</tr>
<tr>
<td>U4</td>
<td>0.65</td>
<td>0.75</td>
<td>0.70</td>
<td>Y E S</td>
</tr>
<tr>
<td>U5</td>
<td>0.68</td>
<td>0.00</td>
<td>0.45</td>
<td>Y E S</td>
</tr>
<tr>
<td>U6</td>
<td>0.84</td>
<td>0.37</td>
<td>0.63</td>
<td>Y E S</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>S1</th>
<th>S2</th>
<th>CV</th>
<th>DN</th>
<th>Valid?</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>0.55</td>
<td>-0.40</td>
<td>0.35</td>
<td>0.23</td>
<td>0.31</td>
<td>0.76</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>0.68</td>
<td>0.60</td>
<td>0.59</td>
<td>0.51</td>
<td>0.60</td>
<td>0.62</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td>0.42</td>
<td>0.00</td>
<td>0.54</td>
<td>0.63</td>
<td>0.82</td>
<td>0.50</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E4</td>
<td>0.40</td>
<td>0.44</td>
<td>0.04</td>
<td>0.05</td>
<td>0.44</td>
<td>0.73</td>
<td>0.34</td>
<td>0.42</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>0.26</td>
<td>0.71</td>
<td>0.36</td>
<td>0.26</td>
<td>-0.44</td>
<td>0.27</td>
<td>0.06</td>
<td>0.66</td>
<td>-0.54</td>
<td>-0.26</td>
<td>1.00</td>
</tr>
<tr>
<td>E6</td>
<td>0.59</td>
<td>0.78</td>
<td>0.76</td>
<td>0.59</td>
<td>0.00</td>
<td>0.71</td>
<td>0.54</td>
<td>0.87</td>
<td>0.00</td>
<td>0.00</td>
<td>0.76</td>
</tr>
</tbody>
</table>

| | S1 | -0.12 | -0.09 | -0.27 | 0.19 | 0.00 | 0.20 | 0.00 | 0.00 | -0.11 | 0.31 | 0.00 | 1.00 | 0.77 | 0.12 | Y E S |

<table>
<thead>
<tr>
<th>Table VIII</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>QUESTIONNAIRE CRONBACH’S ALPHA</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Item</th>
<th>Cronbach’s α</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Experiment Group</td>
<td>Perceived Usefulness</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td>Perceived Ease of Use</td>
<td>0.694</td>
</tr>
<tr>
<td>2nd Experiment Group</td>
<td>Perceived Usefulness</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>Perceived Ease of Use</td>
<td>0.755</td>
</tr>
<tr>
<td></td>
<td>Self-predicted Future Use</td>
<td>0.789</td>
</tr>
<tr>
<td>Expert panel</td>
<td>Perceived Usefulness</td>
<td>0.725</td>
</tr>
<tr>
<td></td>
<td>Perceived Ease of Use</td>
<td>0.812</td>
</tr>
<tr>
<td></td>
<td>Self-predicted Future Use</td>
<td>0.732</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table IX</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DESCRIPTIVE STATISTICS FOR QUESTIONNAIRES</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data set</th>
<th>Item</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
<th>Statistical analysis results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Experiment Group</td>
<td>Perceived Usefulness</td>
<td>2.67</td>
<td>4.67</td>
<td>3.74</td>
<td>0.751</td>
<td>P=0.020 (t-test)</td>
</tr>
<tr>
<td></td>
<td>Perceived Ease of Use</td>
<td>3.17</td>
<td>4.50</td>
<td>4.00</td>
<td>0.451</td>
<td>P=0.000 (t-test)</td>
</tr>
<tr>
<td></td>
<td>Self-predicted Future Use</td>
<td>3.00</td>
<td>5.00</td>
<td>4.14</td>
<td>0.748</td>
<td>P=0.004 (t-test)</td>
</tr>
<tr>
<td>2nd Experiment Group</td>
<td>Perceived Usefulness</td>
<td>3.50</td>
<td>4.33</td>
<td>4.01</td>
<td>0.343</td>
<td>P=0.013 (t-test)</td>
</tr>
<tr>
<td></td>
<td>Perceived Ease of Use</td>
<td>3.33</td>
<td>4.33</td>
<td>4.02</td>
<td>0.390</td>
<td>P=0.002 (t-test)</td>
</tr>
<tr>
<td></td>
<td>Self-predicted Future Use</td>
<td>3.50</td>
<td>4.50</td>
<td>4.00</td>
<td>0.500</td>
<td>P=0.011 (t-test)</td>
</tr>
<tr>
<td>Expert panel</td>
<td>Perceived Usefulness</td>
<td>3.21</td>
<td>4.59</td>
<td>4.29</td>
<td>0.501</td>
<td>P=0.001 (t-test)</td>
</tr>
<tr>
<td></td>
<td>Perceived Ease of Use</td>
<td>3.14</td>
<td>4.48</td>
<td>4.17</td>
<td>0.467</td>
<td>P=0.010 (t-test)</td>
</tr>
<tr>
<td></td>
<td>Self-predicted Future Use</td>
<td>3.42</td>
<td>4.71</td>
<td>4.30</td>
<td>0.658</td>
<td>P=0.006 (t-test)</td>
</tr>
</tbody>
</table>

**External validity:** Since we have evaluated the applicability of the SKLSEForum with different kind of stakeholders, i.e. students, researchers and practitioners, our results can have general validity. According to Tichy [18], the students are appropriately trained and the data is used to establish a trend, thus here we consider the students’ ability for requirements elicitation is comparable to that of typical novice requirements analyst [19]. However, we used the “university library web portal” to conduct the experiment and collect the data. To show that our findings are true for other cases we should...
test the applicability of the SKLSEForum to other people in other industrial contexts.

**Conclusion validity:** An important threat to the conclusion validity of our study is the relatively small number sample that participated to the studies. In respect to the random heterogeneity of participants, the participants have received the same training about SKLSEForum. We should organize other study to have a bigger data sample to draw our conclusions.

VI. RELATED WORK

There are several researches related our work. One of them is studies on forum based requirements elicitation. Cleland-Huang et al. explore and evaluate the forum-based requirements gathering and prioritization processes by vendor-based open source software projects [3]. The study reports a number of interesting lessons that can be learned and applied in forum-based requirements elicitation processes. Then they proposed a method to improve recommendations by enhancing stakeholder profiles [20], utilized recommender systems to support software requirements elicitation [21], applying a recommender system for requirements elicitation in large-scale software projects [22], and provided automated support for managing feature requests in open forums [23]. They highlighted the methods like data mining and machine learning techniques which can give aid to the software requirements elicitation. However they didn't provide an integration tool to solve the problems occurred in open forums. In addition, none of the mentioned studies has been evaluated in practical industry context.

In the field of requirements engineering case tools, many requirements elicitation and management tools have been proposed from academic, like specialized KAOS’ support tool Objectiver [24]. Requirements engineering specific wikis [25]. One drawback of these tools is that they general lack proper empirical evaluations and very few have had widespread use in practice. Even in the implement of empirical evaluation, they have given a comprehensive description of their usage context, study design, and discussed their validity, but their usage contexts are most in academia. If further evaluation can be carried out in industry, it will increase the possibility of transferring it to industry context [26][27]. Here we adopted on the expert panel to capture their perceived usefulness, perceived ease of use and self-predicted future use in industry context.

VII. CONCLUSION AND FUTURE WORK

The main goal of this study is to evaluate SKLSEForum for supporting the requirements elicitation process and understand the demands for adjusting and extending that tool.

Through the two replicated controlled experiments, we have evaluated that the application of SKLSEForum can improve both the quantity of posts and judge satisfaction of the requirements elicited. Furthermore, by analyzing the responses from the expert panel and the students in experiment groups, we evaluated the perceived usefulness, perceived ease of use and self-predicted future use of SKLSEForum. The results have found SKLSEForum is being useful and easy to use, and that there is an intention to use it in the future.

In addition, the qualitative analysis indicated that the feature with more potential is the score exchange. On the other hand, the students indicated that users may have some difficulties using SKLSEForum, because SKLSEForum is a prototype tool that still needs improvement in its usability. The thread templates and thread state identification are considered as hard features to use. Moreover, the questionnaire results and attitudes of the experts towards SKLSEForum are likely to be of interest to a wide range of specialists in both the research community and the software industry.

The evaluation show us the improvements can be made to SKLSEForum, specifically the ones related to improve the ease of use of SKLSEForum. In our future work, we plan to improve our SKLSEForum, and conduct experiments with practitioners in industry contexts to acquire more empirical results. Other research directions we will investigate how to increase the perceived usefulness, perceived ease of use and self-predicted future use about integrate our tool as a plug-in into industry-level requirements engineering case tool and the usability in real world. For the thread templates provided by SKLSEForum, we plan to find a method to improve the templates structure for easier knowledge externalization, to inspire the expression of stakeholders, like the templates based on story [28]. Furthermore, we will introduce the argumentative discussions based on IBIS [29] and SIOC [30] to promote the stakeholders’ discussions in the SKLSEForum.

ACKNOWLEDGMENT

The research was supported by the National Natural Science Foundation of China under Grant No. 61170026, the Chongqing Key Lab. of Electronic Commerce & Supply Chain System Special Fund (2012ECSC0210).

REFERENCES

elicitation in large-scale software proje


Han Lai was born in 1981, is currently a lecturer in Chongqing Technology and Business University, Chongqing, China. He received master degree from Jilin University, Changchun, China, in 2005. He is currently pursuing doctor degree at Computer School in Wuhan University. His research interests include requirement engineering, software process improvement and empirical software engineering.

Yuze Ni was born in 1991. She received bachelor degree in Wuhan University in 2012. She is now a master student in Wuhan University. Her research interests include requirement engineering and empirical software engineering.
A Data Analysis Method and Its Applications in EXCEL

Jinlei Qin
Information and Network Management Center, North China Electric Power University, Baoding, China
Email: jlqin717@163.com

Yuguang Niu and Zheng Li
State Key Laboratory of Alternate Electric Power System with Renewable Energy Sources, North China Electric Power University, Beijing, China
Department of Computer Science and Technology, North China Electric Power University, Baoding, China
Email: nyg@ncepu.edu.cn, yeziperfect@163.com

Abstract—Data analysis is a quite important process and has been extensively employed in many areas. Especially in statistics, the distribution type test of data often needs to be handled. This paper presents that linear regression as a type of universal method can be applied to the distribution type test. In reliability engineering due to the failure data is commonly of non-linear relationship so that the linear regression method can not be directly employed. This difficulty can be resolved through linear transformation. Aimed for the linearization procedure of four kind of typical distributions, e.g., exponential distribution, normal distribution, logarithmic normal distribution and two-parameter Weibull distribution, their transformations are respectively different. After linearization transformations, correlation coefficients had been used as a criterion to choose the most matched distribution type. This method can be conveniently operated in MS EXCEL. Several samples and experiments illustrated the detailed transformation and the efficiency than other methods.

Index Terms—data analysis, failure data, linear regression, correlation coefficient, distribution type test

I. INTRODUCTION

Data analysis is a quite important process and has been extensively employed in many fields such as energy, software, communication and various engineering etc [1-3]. Some related methods had been researched by many scholars. Sun et al. [4] provides a robust data-recovery method based on functional data analysis to enhance the reliability of multichannel sensor system. Dauck et al [5] have developed an industrial data analysis (IDA) platform that automates the data analysis process to a large extend. The IDA platform uses fuzzy knowledge bases to match user requirements to features of analysis methods and to select, configure and execute IDA processes automatically. Lu et al [6] introduce a new lifetime evaluation method that can accomplish rapidly the long lifetime evaluation by using short-term aging data. In Ref [7], the author developed an efficient data analysis package for personal computer use in response to growing needs of the wind industry. Kovac et al [8] present a data analysis software package ‘AFV-SOFT’ that has been developed for the evaluation of alternative fuel vehicle performance. Jiang [9] has simulated the characters of diode by the NI Multisim platform and plotted conveniently voltage-ample characteristic curve in MS EXCEL according to the data analysis results. The curves plotted by EXCEL are more accurate and intuitive than the traditional hand-drawn curves. Shi [10] works out a EXCEL tool by taking advantage of the VBA language to accomplish the judgment function of data analysis. Some data analysis methods were also applied to software performance prediction and evaluation [11, 12]. The mimic situations arise frequently in reliability engineering and the rules or relationships among data can frequently arouse researcher’s interests. Among a group of given failure data, the most concerned issue is which the distribution type belongs to.

Hypothesis test method is a traditional way for this issue. The general methods include Kolmogorov-Smirnov (K-S) test, Chi-square $\chi^2$ test, Anderson-Darling (A-D) test and Crammer-von Mises (C-M) test [13-19]. However among those above mentioned methods, their computational process is always so complicated that more probability and statistical knowledge requires grasp for the data analyst. Furthermore, the computational results of hypothesis test may lead to not unique because that several distribution types can satisfy the conditions of hypothesis test. This situation often confuses the researchers and results in the imprecise judgment which can bring severe aftermath. Although some commercial software, e.g., MatLab and SAS, provide some mimetic functions, but more disk spaces are required meanwhile demanding pre-trainings are also necessary for the data analyzers.

Linear regression method had been put forward by scholar Goodman in the 1950s [20] and then been widely employed in many applications [21, 22]. The EXCEL as a kind of office software has been so broadly used in various fields that a number of researchers and analyzers...
can operate it expertly. The linear regression method as a tool had been embedded in EXCEL and the results of data analysis can be easily displayed on the chart by some simple operations like clicking on some correspond command buttons. In this paper, the applications of linear regression method had been developed for the judgment of failure data distribution type. Through the comparison of correlation coefficient in the EXCEL, the most matched distribution type can be quantitatively chosen so that the non-unique above mentioned would be avoid.

This remaining will be organized as follows. Section II presents the principle of linear regression and the property of correlation coefficient. In section III, we will show how to linearize respectively several typical distributions in order to apply the method in EXCEL. In section IV, a number of experiments are conducted in EXCEL to demonstrate the linear regression method and then to compare the results. Finally, the conclusions drawn from this study are given in Section V.

II. LINEAR REGRESSION METHOD DESCRIPTION

Linear regression method is often used to describe the linear relationship among some random variables. The method can be depicted as following. Given random variables \( X \) and \( Y \), their sample values \((x_i, y_i), i = 1, 2, \ldots, n\) are independent respectively. Then the following formula

\[
y_i = a + bx_i + e_i \quad (i=1,2,\ldots,n)
\]  

(1)
can reveal the laws between variables \( X \) and \( Y \). In (1), \( a, b \) are unknown parameters and named as regression coefficient. The random variable \( e_i \) subjects to s-Normal distribution \( N(0,\sigma^2) \) and stands for the errors on \( y_i \). The main object of regression analysis is to calculate the estimation value \( \hat{\alpha}, \hat{\beta} \) of regression coefficient \( \alpha, \beta \) according to the experimental data so that the value of \( Y \) can be predicted on a given value of \( x \).

Given \( x \), the following formula

\[
\hat{y} = \hat{a} + \hat{b}x
\]

(2)
can be seen as estimation of \( y = a + bx \). Equation (2) is also called as linear regression equation \( \mu(x) \) and whose plots are named as regression lines. Once the estimation value \( \hat{\alpha}, \hat{\beta} \) could be found out, we will get the explicit form of (2) which could be used to calculate the value of variable \( \hat{y} \) on the given variable \( x \).

The least square method (LSM) is often employed to get the estimation value of regression coefficient. Generally speaking, we have \( y_i \neq \hat{y}_i \) because the random variable \( e_i \) is not always equal zero. The approach degrees of theoretical value \( y \) and factual value \( \hat{y} \) can be represented by the following formula

\[
Q(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 .
\]

(3)

So we may naturally think that the less value of \( Q(\hat{\alpha}, \hat{\beta}) \), the more high of fit degree. Consequently, according to the principle of LSM, we will get the estimation value of regression coefficient when equation (3) reaches its minimum. It can be described by

\[
Q(\hat{\alpha}, \hat{\beta}) = \min Q(a, b) .
\]

(4)

According to the principles of derivative, extreme values can be fetched by resolving the following equations

\[
\begin{align*}
\frac{\partial Q}{\partial a} &= -2\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\
\frac{\partial Q}{\partial b} &= -2\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0
\end{align*}
\]

(5)

After the simplification of equation (5) and we will get the following equations

\[
\begin{align*}
\hat{b} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} , \\
\hat{a} &= \bar{y} - \hat{b}\bar{x}
\end{align*}
\]

(6)

where \( \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \), \( \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \). We may introduce the following notations

\[
\begin{align*}
S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \\
S_{xy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 \\
S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)
\end{align*}
\]

(7)

Then equation (6) can be changed into the following forms

\[
\begin{align*}
\hat{b} &= \frac{S_{xy}}{S_{xx}} , \\
\hat{a} &= \bar{y} - \hat{b}\bar{x}
\end{align*}
\]

(8)

As a consequence of the above analysis, the estimation values of regression coefficient can be gotten by the method of LSM.

Correlation coefficient is the most general index to measure the degrees of linear relationship among variables that are linearly related. Correlation coefficient is defined by

\[
\rho = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} .
\]
\[ \gamma = \frac{S_{\alpha}}{\sqrt{S_{\alpha} \times S_{\beta}}}. \]  

\( \gamma \) is a dimensionless statistics and its absolute value is less than or equal 1. When \( |\gamma| = 1 \), it shows that all the sample data locate a straight line. Then we can come to a conclusion that the variables \( x, y \) are of linear relationship at the probability value 1. If \( |\gamma| \neq 1 \), then the bigger of the value of \( |\gamma| \), the better of the linear related degree among the variables; the less of the value of \( |\gamma| \), the worse of the linear related degree among the variables. In practical engineering, the square value of \( \gamma \) is often used in order to compute conveniently.

III. TRANSFORMATION OF TYPICAL DISTRIBUTION

A. Data Preparation

The most important thing needed to do is the variable transformation so that the transformed variables are of linear correlation. In reliability engineering, a lot of failure level and failure data (such as time) are not of linear relationship. On the contrary, there are some nonlinear relationships, e.g., exponential distribution, normal distribution, lognormal distribution and Weibull distribution, between variables failure level and failure time. Consequently we can’t directly employ the correlation coefficient to judge which distribution type the law of failure data belong to. After variable transformation the relationships between failure level and failure time can be converted into the linear relationship. The detailed transformation of each distribution type will be treated in the next subsection B. Then correlation coefficient can be computed in EXCEL to judge which distribution type is more matched than the others by the comparison of correlation coefficient. Then the errors came from personal factor can be easily avoided, furthermore the quantification precision to choose the most matched distribution type will be dramatically improved.

The second important matter is how to determine the value of the failure level corresponding the given failure time. According to the random property of failure, failure data can be seen as a random variable \( T \). Order all the failure data by ascendant sequence, e.g., \( t_1 \leq t_2 \leq \cdots \leq t_i \leq \cdots \leq t_n \). Each failure data \( t_i \) has a corresponding failure level \( F(t_i) \) which can be described by the approximate median ranks formula [23]

\[ F(t_i) = P(T \leq t_i) = \frac{i-0.3}{n+0.4} \quad (n \leq 20). \]  

Then above (10) is also named as experiential distribution function then the value of failure level can be determined for a given value of failure time \( t_i \). The data points \( (t_i, F(t_i)), i = 1, 2, \cdots, n \), will be properly transformed into \( (\sigma(t_i), \xi(F(t_i))), i = 1, 2, \cdots, n \), so that the latter are of linear relationship. Then using the transformed failure data, correlation coefficient can be calculated and compared quantitatively so that a most matched distribution type will be chosen.

B. Transformation of Typical Distribution Type

As the above mentioned analysis, some data points \((t_i, F(t_i))\) belonging to a certain kind of distribution types aren’t of linear relationship and special variable transformation need to be done. Aiming to different distribution type the linearization’s method is also respectively different. Then what require do is to analyze different distribution type and find out its respective linearization method. The following will address detailedly the linearization process.

The cumulative distribution function (CDF) of exponential distribution is often depicted as follow

\[ F(t) = 1 - e^{-\frac{t-\varphi}{\eta}}, (t \geq \varphi). \]  

Parameters \( \varphi \) and \( \eta \) are the unknown and parameter \( \varphi \) is named as location parameter. Obviously, these data points \((t_i, F(t_i))\) aren’t of linear relationship. However the linear equation (12) can be obtained by taking logarithm of (11) and simplifying:

\[ \ln \frac{1}{1-F(t)} = \frac{t-\varphi}{\eta}. \]  

For (12), construct converted variables as follow

\[ y_i = \ln \frac{1}{1-F(t_i)}, \quad x_i = t_i. \]  

From (12) and (13), a conclusion that the variables \((x_i, y_i)\) is of linear relationship can be drawn.

Normal distribution’s CDF can be written as follows

\[ F(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{0}^{t} \frac{e^{-(x-\mu)^2/2\sigma^2}}{2\sigma} dx, t \geq 0. \]  

The unknown parameter \( \mu \) represents for expectation and \( \sigma \) for standard deviation. According to the relationship between normal distribution and standard normal distribution, any normal distribution \( F(t) \) can be changed into standard normal distribution \( \Phi(y) \) whose form is as same as (14) with the parameters \( \mu = 0, \sigma = 1 \) by the following linear transformation

\[ y = \frac{1}{\sigma}t - \frac{\mu}{\sigma}. \]  

Then \( F(t_i) = \Phi(y_i) \) can be obtained on the base of (15). For each \( t_i \), if supposing \( x_i = t_i \), then \( x_i \) has a linear relationship with variable \( y_i \) based on the (15). The
value of $y_i$ can be gotten by looking for standard normal distribution table.

The distinction between logarithmic normal distribution and normal distribution is whether or not the variable has been taken logarithmic operation. Its CDF can be written down

$$F(t) = \int_0^t \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{\ln x - \mu}{2\sigma^2}} dx.$$ (16)

Be similar to the linearization of normal distribution, the $F(t_i) = \Phi(y_i)$ will be finished by the following variable transformation

$$y = \frac{1}{\sigma} \ln t - \frac{\mu}{\sigma}.$$ (17)

From (17) it can be shown that data points $(x_i, y_i)$ are of linear relationship by setting $x_i = \ln t_i$.

Weibull distribution is widely used to model the variability in the fracture properties of ceramics and metals where the concept of the weakest link has been applied. For the two-parameter Weibull distribution, its CDF can be depicted as following [24]

$$F(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^m}.$$ (18)

In (18), $\eta$ is the scale parameter and $m$ is the Weibull modulus alternatively referred to as the shape parameter. Taking logarithm of (18) twice, it yields a linear equation

$$\ln \ln \frac{1}{1 - F(t)} = m \ln t - m \ln \eta.$$ (19)

Setting

$$y_i = \ln \ln \frac{1}{1 - F(t_i)}$$

$$x_i = \ln t_i,$$ (20)

then the data points $(x_i, y_i)$ are of linear relationship based on the (19).

IV. APPLICATION METHOD IN EXCEL

Supposing there are 20 failure data in the following Table I. They have been arranged from small to large and from left to right in the table I.

<table>
<thead>
<tr>
<th>Table I</th>
<th>FAILURE DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.8</td>
<td>11.3</td>
</tr>
<tr>
<td>19.4</td>
<td>20.6</td>
</tr>
<tr>
<td>26.6</td>
<td>28.5</td>
</tr>
<tr>
<td>34.5</td>
<td>37.0</td>
</tr>
</tbody>
</table>

For each data in the Table I, it has a corresponding failure level according to the (10). Their failure levels have been listed in the following Table II.

<table>
<thead>
<tr>
<th>Table II</th>
<th>FAILURE LEVEL OF EACH FAILURE DATA IN TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.034314</td>
<td>0.083333</td>
</tr>
<tr>
<td>0.132353</td>
<td>0.181373</td>
</tr>
<tr>
<td>0.230392</td>
<td></td>
</tr>
<tr>
<td>0.279412</td>
<td>0.328431</td>
</tr>
<tr>
<td>0.377451</td>
<td>0.426471</td>
</tr>
<tr>
<td>0.47549</td>
<td>0.52451</td>
</tr>
<tr>
<td>0.573529</td>
<td>0.622549</td>
</tr>
<tr>
<td>0.671569</td>
<td>0.720588</td>
</tr>
<tr>
<td>0.769608</td>
<td>0.818627</td>
</tr>
<tr>
<td>0.867647</td>
<td>0.916667</td>
</tr>
<tr>
<td>0.965686</td>
<td></td>
</tr>
</tbody>
</table>

For exponential distribution, according to the (13) the values of variables $x_i, y_i$ can be calculated by combining the data of Table I and Table II in EXCEL. The results (E and F column) have been shown in the following Fig. 1.

Figure 1. Linear transformation for exponential distribution

Complying with the procedures of Charting in EXCEL, scatter diagram can be depicted for the E and F columns. First clicking the command “append trend line” in the chart menu, and then choosing “linear type” furthermore setting the display of formula. The result can be shown in the Fig. 2.

Figure 2. Scattered diagram and trend line for exponential distribution

The parameter $R^2$ is equal to the square of $\gamma'$. In other words the $R^2$ represents the linear correlation degree. At the same time, the linear regression equation is also shown in the Fig. 2.
For normal distribution, the first thing needed to do is to get the value of \( y_i \). Using the \( F(t_i) = \Phi(y_i) \), \( y_i \) can be gotten by the inverse function of standard normal distribution. In EXCEL the function is NORMINV whose syntax format is NORMINV (probability, mean, standard_dev).

The meanings of those parameters are as follows: probability: the probability value of normal distribution. mean: the arithmetic average of normal distribution. standard_dev: standard deviation of normal distribution.

From the (15), we can easily know that the value of \( x_i \) is equal to the \( t_i \). The values of \( x_i \) (E column) and \( y_i \) (F column) have been calculated and shown in the following Fig. 3.

![Figure 3. Data after transformation for normal distribution](image)

The scatter diagram and trend line shown in the following Fig. 4 can be depicted in the similar manner.

![Figure 4. Scattered diagram and trend line for normal distribution](image)

For the logarithmic normal distribution, the analysis method is similar to the normal distribution. The only distinguish lie in the value of \( x_i \) is equal to \( \ln t_i \) according to the (17). The data after transformation can be also depicted as the following Fig. 5.

![Figure 5. Data after transformation for logarithmic normal distribution](image)

The scatter diagram and trend line for logarithmic normal distribution can be depicted by the similar manner in the following Fig. 6.

![Figure 6. Scattered diagram and trend line for logarithmic normal distribution](image)

For two-parameter Weibull distribution the data can be transformed according to the (20) in the similar method and the values of \( x_i \) (E column) and \( y_i \) (F column) can be depicted as the following Fig. 7.

![Figure 7. Data after transformation for two-parameter Weibull distribution](image)
The scatter diagram and trend line can be shown in the following Fig. 8 by the similar method.

For the sake of convenience, the above computational result of the data analysis can be listed in the following Table III.

<table>
<thead>
<tr>
<th>Distribution Type</th>
<th>Linear Regression Equation</th>
<th>Correlation Coefficient ($R^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$y = 0.0766x - 1.0754$</td>
<td>0.9248</td>
</tr>
<tr>
<td>Normal</td>
<td>$y = 0.0849x - 2.2526$</td>
<td>0.9892</td>
</tr>
<tr>
<td>Logarithmic Normal</td>
<td>$y = 1.9793x - 6.3006$</td>
<td>0.9669</td>
</tr>
<tr>
<td>Two-parameter Weibull</td>
<td>$y = 2.4895x - 8.469$</td>
<td>0.9989</td>
</tr>
</tbody>
</table>

From the comparison table, we can draw a conclusion that the most matched distribution of failure data is subjected to the two-parameter Weibull distribution because the correlation coefficient is the highest. At the same time, the linear regression equation had been found out which can be used to estimate the related parameters of the two-parameter Weibull distribution according to the (19) and (20). Because the correlation coefficient is computed quantitatively, the personal observation error will be avoided efficiently. The failure data can be treated as the two-parameter Weibull distribution for prediction or some other intents.

V. CONCLUSIONS

In this paper, the linear regression principle had been put forward as a type of data analysis method for the distribution type test. The method can be employed for the distribution type inference of failure data. For each general distribution, the linear regression method can’t be implemented directly. First linearization needs to do and the detailed procedures are also respective different. The correlation coefficient had been examined for the comparison of several probable distribution types.

The detailed steps of the application for the linear regression in EXCEL have been illustrated. A group of data has been used for the distribution type test by the comparison of the correlation coefficient. The result has shown that the linear regression method is validated and convenient for the analyst. Furthermore, the method is also adapted for developing related software.

ACKNOWLEDGMENT

The authors wish to thank the reviewers and the editor for their constructive comments that have helped to improve this article. This work was supported in part by a grant from Key Program of National Nature Science Foundation of China (51036002), National Basic Research Program of China (973 Program) (2012CB215203) and Hebei Province Natural Science Foundation (F2014502081).

REFERENCES


Jinlei Qin received his B. S. degree and M. S. in department of computer science and technology at north china electric power university. He is now a Ph. D. graduate student in the school of control and computer. His research interests include condition-based maintenance, system reliability analysis and life data analysis.

Yuguang Niu obtained his Ph. D. degree in the department of thermal energy and power engineering at north china electric power university. He is a professor and doctoral tutor in the school of control and computer. He works as a deputy director of the state key laboratory of alternate electric power system with renewable energy sources. His research interests include modeling and control of industry process, fault detection and diagnosis and system reliability. He has participated and completed several research projects of national level and ministry level and obtained recognition of peers and experts.

Zheng Li received her B. S. degree and M. S. in department of computer science and technology at north china electric power university. She is now a Ph. D. graduate student in the school of control and computer. Her research interests include the research and application of intelligence algorithms in electric power system.
Measuring Developers’ Design Contributions in Evolved Software Projects

Maen Hammad
Department of Software Engineering, The Hashemite University, Zarqa, Jordan
Email: mhammad@hu.edu.jo

Mustafa Hammad
Department of Information Technology, Mu'tah University, AL Karak, Jordan
Email: hammad@mutah.edu.jo

Hani Bani-Salameh, Ebaa Fayyoumi
{Department of Software Engineering, Department of Computer Science and Applications}
The Hashemite University, Zarqa, Jordan
Emails: {hani, enfayyoumi}@hu.edu.jo

Abstract—The work presented in this paper measures the contribution of developers towards evolved structural design of software systems. Measuring the contribution of developers is useful for project managers who manage the development process. Project managers can empirically identify developers who made changes to the structural design and compare among them based on their contributions. The proposed measures help to understand the nature of developers’ code changes activities. The process of calculating the measures is based on the historical code changes committed by developers. Specifically, code changes that affect the corresponding UML class diagram representation of the source code. Both type and amount of previous changes to class elements are used to measure the design contributions. The proposed measures are helpful for open source projects where no detailed information is available about various developers involved in the development process. A tool has been developed to automatically measure the contributions based on archived historical code changes. The measures have been applied on two open source projects. Results showed that only small number of developers have the major design contributions.

Index Terms—Software development and design, software measurements, software evolution.

I. INTRODUCTION

The development process of software systems is difficult to manage and control. Many developers with different skills, experiences, and roles are involved in the development process. Developers usually maintain software system via a set of commits, which are committed over time. As a result, not all developers have the same contributions in the evolved software projects. They vary in the number and the type of code changes that they committed. Some developers make major contributions, while others’ contributions are minor and limited to small and specific code changes.

Measuring the contributions of developers helps project managers to get more information about their teams and about the nature of their coding activities. Analyzing design contributions of developers helps in identifying who are the senior developers and estimating the amount of their contributions to the evolved structural design.

Some software developers focus on developing test cases, while others focus on updating documents. In this paper, we focus on developers’ contributions that are related to the design of software systems. The structural design of software systems is mainly represented by the UML class diagram of the source code that includes classes and their relationships. Therefore, the design elements are classes, methods and relationships between classes. A relationship between two classes can be a generalization, an association, or a dependency.

The measures best support open source projects, where detailed information about each developer is not always available. Besides, most developers are volunteers from different countries. This consequently increases the difficulty of identifying the design contribution of developers. It is also not an easy task to identify developers who make the largest contributions to the structural design of the system.

The question that needs to be addressed is: how to measure developers’ contributions in the evolved structural design of software systems?

Measuring the contributions of developers in structural design provides many benefits to project managers. It leads to understand the nature of code changes activities of developers. For example, some developers do most structural design changes in one or two commits, while others make their changes in small chunks over longer period of time. It is important to project managers to identify developers who have made significant impacts on the structural design. In particular, they also need to know the contribution of a specific developer during a
specific period of time, or active developers in specific time duration.

Developers, who update class elements of the source code, contribute in building the structural design of the system. Measuring developers’ contributions can be achieved by defining new quantitative metrics based on analyzing the previous code changes activities of developers.

The paper proposes two historic-based quantity measures for developers’ contribution in the evolved structural design. The first measure reflects the added design elements by the developer, while second measure reflects deleted design elements. Code changes in software repositories are examined to calculate the contributions. The work is focused on measuring the contribution of developers to structural design of the source code which include classes, methods, and relationships. Minor code changes, such as adding a condition or renaming a variable, are not considered a design contribution.

The main research contributions that are addressed in this paper are:
1. Two historic-based measures for the contributions of developers towards evolved design.
2. A case study about applying the metrics on two open source projects.
3. A tool to automatically calculate design contributions for developers.

The paper is organized as follows. Section 2 explains the measured design elements. The proposed measures are described in Section 3. The developed tool that supports the measure is presented in Section 4. A detailed case study on two open source projects is discussed in Section 5. Related work is presented in Section 6 followed by conclusions and future work.

II. DESIGN CONTRIBUTIONS

Our concern is on measuring contributions of developers to the structural design of the source code. The structural design is mainly represented by classes and their relationships. The contents of classes include attributes and methods, while relationships include generalizations, associations and dependencies. The work presented in this paper is a continuation to our previous work that is presented in [1]. In [1], developers who updated design and their knowledge in the updated design are identified. It is shown that the identified design knowledge can help in handling high level change requests. In this paper, we defined measures for the contribution of developers who updated the structural design of the system.

The design contribution of a developer is measured by the total number of added or deleted design elements committed by the developer during specific time duration. In general, design contributions can be categorized into two categories; the addition of new design elements and the deletion of old design elements.

The benefits of measuring design contributions for developers can be summarized as follows:
- Identifying core developers who are important to the structural design of the system.
- Identifying the contribution of a developer during specific time duration.
- Understanding the nature of code changes activities for developers who updated the structural design.

The measured contributions include the following changing activities that directly affect the main elements of classes:
- Creating or deleting classes.
- Adding or deleting methods to/from classes.
- Adding or deleting relationships (generalizations, associations or dependencies) between classes.

These changes to the classes are extracted from code changes that are committed by developers. As a result, any developer committed at least one of the above changes has a design contribution to the structural design of the source code.

III. THE PROPOSED MEASURES

This section introduces and explains the proposed measures for the design contributions of software developers. Two historical-based measures are proposed to quantify the detailed design contribution of developers. The proposed measures are calculated from the historical code changes committed by developers.

Based on the definition of design contributions introduced in Section II, we propose two measures. The first measure is the total number of Added Design Elements (ADE) by the developer during specific time duration. The second one is the total number of Deleted Design Elements (DDE) by the developer during specific time duration. The following two equations show how the two measures are calculated for a specific developer $(d)$.

$$\text{ADE}(d) = w_1 \cdot \sum_{c=1}^{n} AC_c(d) + w_2 \cdot \sum_{c=1}^{n} AM_c(d) + w_3 \cdot \sum_{c=1}^{n} AR_c(d)$$

$$\text{DDE}(d) = w_1 \cdot \sum_{c=1}^{n} DC_c(d) + w_2 \cdot \sum_{c=1}^{n} DM_c(d) + w_3 \cdot \sum_{c=1}^{n} DR_c(d)$$

The first measure (ADE) sums the total number of added classes (AC), added methods (AM), and added relationships (AR) that are identified from all extracted commits C from 1 to n for a specific developer $d$. Meanwhile, the DDE is the summation of the total number of deleted classes (DC), deleted methods (DM), and deleted relationships (DR) for all extracted n commits of the developer.

One important factor in calculating the contribution is the weight of design change. For example, a developer who added three classes may be considered to have more
contributions to the design than the one who added three methods. Even if the ADE for both developers is three, adding a class may have more impact than adding a method. To include weights in ADE and DDE, each summation is multiplied by a weight value (w1, w2, or w3) such that:

$$\sum_{i=1}^{3} w_i = 1$$

These weights are determined by project managers to rank design contributions based on their types. For example, weight values could be set as (w1=0.6, w2=0.3, w3=0.1). These weight values give the highest design contributions to added/deleted classes then added/deleted methods followed by added/deleted relationships. In this case, the contribution of adding one class is counted as twice compared to adding one method and six times compared to adding a relationship.

Another option is to consider an equal weight to all design contributions regardless of their types and their impact on the whole design system. In this case, weights are set to be equal to 0.33+ or simply ignore the weight concept.

The process of finding design contributions for developer d in project P is calculated as follows:

1. A set of commits are extracted from the software repository of project P.
2. The set of commits Cn that are committed by developer d are identified from the extracted commits in Step 1.
3. All added/deleted classes, methods and relationships are identified from the set of commits Cn.
4. The total number of added design elements identified in Step 3 is reported as the ADE design contribution value for developer d.
5. The total number of deleted design elements identified in Step 3 is reported as the DDE design contribution value for developer d.

The proposed measures are based on counting the number of added and deleted design elements across all commits. All added and deleted classes, methods and relationships (generalization, association, or dependency) are counted. For a specific developer, all his commits for a specific period of time are extracted. Then, code changes in these commits are analyzed to identify commits that have design changes. In the next step, the number of added and deleted design elements per commit, with design changes, is counted. Finally, the sum of all added design elements is reported as ADE value and the sum of all deleted design elements is reported as DDE value.

For example, consider a developer who committed three commits during a month. In the first commit, he created a new class with four methods. In the second commit, he deleted an association relationship. In the third commit, the developer added another new method to some class. Therefore, his design contribution is six based on the ADE measure because he added six new design elements (one class + four methods + one method). Based on the DDE measure his contribution is one because he deleted only one design element (an association) from the design.

Combining ADE and DDE together helps in determining the total design contribution of a specific developer to design. Consequently, the summation of the two measures ADE and DDE gives the Total Design Contribution (TDC) of a developer d.

$$TDC(d) = ADE(d) + DDE(d)$$

TDC represents the total number of design changes that are committed by a certain developer regardless of the change type.

Another variation to TDC is the measuring of Actual Total Design Contribution (ATDC) for certain developer. ATDC is counted by excluding any design element that is added and then deleted by the same developer. This design element presents an intersected element between the ADE(d) and DDE(d). Consequently, this intersected element represents an overlap that is counted twice. Therefore, it has to be excluded once. This is because that design element is not part of the current status of the design anymore. For instance, suppose a developer has added one method and later he deleted it. This change is counted zero based in ATDC and it is counted two based on TDC. The summation of the two measures ADE and DDE excluding the Overlap Design Elements (ODE) gives the Actual Total Design Contribution (TDC) of a developer d.

$$TDC(d) = ADE(d) + DDE(d) - ODE(d)$$

Our focus in this paper is on measuring TDC for developers, which reflects the amount of their changes activities they applied.

Since ADE and DDE are numbers, it is also useful to show the total design contribution of a developer as a percentage of all other developers. The total design contribution of a specific developer is divided by the total number of design contributions committed by all developers. The result of the division represents the Percentage Total Design Contribution (PTDC) for the developer. The PTDC for developer d among n developers who involved in the development process during a specific period of time is calculated as shown in the following equation:

$$PTDC(d) = \left( \frac{TDC(d)}{\sum_{i=1}^{n} TDC_i} \right) * 100$$

IV. TOOL SUPPORT

A tool has been developed to automatically find the design contribution for developers. The tool supports both extracting commits, and measuring design contributions (ADE and DDE) based on specific weights determined by the user. The tool is named Contributor. Contributor finds the contributions values of a specific developer to design. Consequently, the summation of the two measures ADE and DDE gives the Total Design Contribution (TDC) of a developer d. Consequently, the summation of the two measures ADE and DDE gives the Total Design Contribution (TDC) of a developer d.
contribution is automatically calculated and shown to the user.

Contributor extracts commits from the determined repository and analyzes code changes of the extracted commits by using another tool called srcTracer [2, 3]. The srcTracer tool automatically analyzes C++ code changes to identify changes to design. The tool identifies added/deleted classes, methods, and relationships between classes.

The design contributions for all developers or a group of developers who are involved in the development process are calculated by the tool during predefined time duration. Also, developers can be ranked based on their design contributions (ADE, DDE or TDC). The tool is able to calculate the percentage total design contribution (PTDC) and actual total design contribution (ATDC) for a specific developer or a group of developers.

V. CASE STUDY

We applied the proposed contribution measures, by using the Contributor tool, on two C++ open source projects. The results are analyzed and discussed in this section. The two studied open source projects are:

1. The quantitative finance library QuantLib (www.quantlib.org)
2. The cross-platform GUI library wxWidgets (www.wxwidgets.org).

Subset of commits has been extracted from each of the two projects. The goal of the study is to investigate the following issues:

- What is the design contribution of a specific developer?
- Who are the developers with major design contributions?
- Do all developers vary in their design contributions?

The extracted commits from the two projects have been analyzed by the Contributor tool to calculate the design contributions for all developers who committed the extracted commits. Tables 1 and 2 show the top ten developers of wxWidgets and QuantLib with the design contribution values for each developer. The contribution is measured by ADE and DDE measures. The total design contribution (TDC) is shown for each developer in the last column which is used to rank developers. The weights are ignored in calculating ADE and DDE (i.e. all design changes have the same weight).

For example, in wxWidgets, developer VZ is ranked first based on his total design contribution. He updated the design by adding 3994 (class, method or relationship) and deleting 1729 (class, method, or relationship). The total number of added and deleted design elements by VZ is 5723 (3994+1729). Tables 1 and 2 also show that developers vary in their contributions to the evolved structural design.

It is observed that most structural design changes are applied by small number of developers. To investigate this observation, we calculated the cumulative design contribution of top five developers of wxWidgets and top three developers of QuantLib. Five developers of wxWidgets represent 20% (5/25) of all developers who contributed to the structural design. Also, the top three developers of QuantLib represent 20% (3/15) of developers with design contributions. The cumulative design contributions were calculated by summing the total design contribution (TDC) for each developer. Top developers mean developers with highest contributions values. The cumulative design contribution for the top developers of the two projects is shown in Table 3.

![Table 1. Design Contributions for Top Ten Developers of QuantLib](image1)

<table>
<thead>
<tr>
<th>Developer</th>
<th>ADE</th>
<th>DDE</th>
<th>TDC (ADE+DDE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lballabo</td>
<td>575</td>
<td>624</td>
<td>6382</td>
</tr>
<tr>
<td>Nando</td>
<td>254</td>
<td>993</td>
<td>3534</td>
</tr>
<tr>
<td>fdv1</td>
<td>128</td>
<td>59</td>
<td>1344</td>
</tr>
<tr>
<td>Giorfa</td>
<td>711</td>
<td>194</td>
<td>905</td>
</tr>
<tr>
<td>Klausspanderen</td>
<td>559</td>
<td>57</td>
<td>616</td>
</tr>
<tr>
<td>Markoshti</td>
<td>348</td>
<td>0</td>
<td>348</td>
</tr>
<tr>
<td>Chiforma</td>
<td>300</td>
<td>22</td>
<td>322</td>
</tr>
<tr>
<td>Kenzouzi</td>
<td>234</td>
<td>51</td>
<td>285</td>
</tr>
<tr>
<td>Coliminecco</td>
<td>182</td>
<td>48</td>
<td>230</td>
</tr>
<tr>
<td>Mariopucci</td>
<td>165</td>
<td>30</td>
<td>195</td>
</tr>
</tbody>
</table>

![Table 2. Design Contributions for Top Ten Developers of WXWidgets](image2)

<table>
<thead>
<tr>
<th>Developer</th>
<th>ADE</th>
<th>DDE</th>
<th>TDC (ADE+DDE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VZ</td>
<td>3994</td>
<td>1729</td>
<td>5723</td>
</tr>
<tr>
<td>JS</td>
<td>2478</td>
<td>155</td>
<td>2633</td>
</tr>
<tr>
<td>RR</td>
<td>1701</td>
<td>575</td>
<td>2276</td>
</tr>
<tr>
<td>SC</td>
<td>1088</td>
<td>390</td>
<td>1478</td>
</tr>
<tr>
<td>VS</td>
<td>861</td>
<td>427</td>
<td>1288</td>
</tr>
<tr>
<td>RD</td>
<td>1015</td>
<td>90</td>
<td>1105</td>
</tr>
<tr>
<td>ABX</td>
<td>669</td>
<td>336</td>
<td>1055</td>
</tr>
<tr>
<td>BJW</td>
<td>491</td>
<td>135</td>
<td>626</td>
</tr>
<tr>
<td>MW</td>
<td>270</td>
<td>108</td>
<td>378</td>
</tr>
<tr>
<td>KO</td>
<td>177</td>
<td>37</td>
<td>214</td>
</tr>
</tbody>
</table>

![Table 3. Cumulative Contributions for Top Five Developers of WXWidgets and Top Three Developers of QuantLib](image3)

<table>
<thead>
<tr>
<th>wxWidgets</th>
<th>TDC</th>
<th>QuantLib</th>
<th>TDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 5</td>
<td>13398 (76%)</td>
<td>Top 3 developers</td>
<td>11260 (77%)</td>
</tr>
<tr>
<td>Others</td>
<td>4242 (24%)</td>
<td>Others developers</td>
<td>3318 (23%)</td>
</tr>
</tbody>
</table>

We found that the top five developers of wxWidgets, contributed by adding and deleting 13398 design elements which represent 76% of all updated elements by all developers. The same observation is valid on QuantLib developers. Three developers only contributed about 77% of design changes in QuantLib project. These top developers may be considered as the core designers of the system because of their large contributions.
Fig. 1 visualizes the PTDC values of all developers based on the extracted commits. Developers with no design contributions are ignored. The PTDC of each developer is presented by a pie in the chart. The distribution of PTDC values for all developers is shown for both projects. The charts show that most developers have minor design contributions towards evolved structural design. For example, in QuantLib project shown in Fig. 1, only one developer contributed more than 25% of design changes. The same observation is applied for wxWidgets. Only two developers have contributed about 50% of all design changes. In summary, only small number of developers contributes the most changes activities towards evolved structural design. Project managers should give great attention to those developers due to their importance to the organization or to the development process.

VI. RELATED WORK

Ramil and Lehman [4] described some models that predict effort as a function of a suite of metrics of software evolution. A close work to ours is proposed by Gousios et al. [5]. They proposed a model for evaluating developer contribution. The model takes advantage of software development repositories to extract contribution indicators. The model assigns a weight to each action type by the developer. The approach measures many development activities but not contributions related to the structural design of source code. Di Penta and German [6] introduced a method to track the names of contributors, including those explicitly listed as copyright owners from licensing statements in source code file. Kpodjedo et al. [7] investigated the usefulness of elementary design evolution metrics to identify defective classes. The metrics include the numbers of added, deleted, and modified attributes, methods, and relations. The metrics are used to recommend a ranked list of classes likely to contain defects for a system.

Many studies as in [8, 9] showed that most of code changes are contributed by small number of developers. Bird et al. [10] mined e-mail archives to analyze the communication and co-ordination activities of the participants. In our work, we showed that most structural design changes are made by small number of developers.

Del Rosso [11] used collaborations and interactions between knowledge-intensive software developers to build a social network. Sowe et al. [12] presented a framework for investigating Free and Open Source Software (F/OSS) developers’ activities in both source code and mailing lists repositories. Kagdi et al. [13] presented an approach with a tool named xFinder. The approach recommends expert developers by mining version archives of a software system. The approach is based on the idea that the developers who contributed more in changing a specific part of source code in the past are likely to best assist in its current or future changes. The approach is used to recommend experts for files. The approach has been used in [14, 15] to recommend a ranked list of expert developers to assist in the implementation of software change requests.

Ben et al. [16] examined the contribution characteristics of developers in open source environment based on visual analysis, and presented approaches from three aspects-influencing factors, time characteristics and region characteristics. They found that the code which newcomers started to contribute with more people engaged in would lead to less contribution in some degree. They also found that there’s a relation between developers’ early and later period contribution.

Oosterman et al. [17] introduces a tool named EVOJAVA for extracting static software metrics from a Java source code repository. For each version of a program, EVOJAVA builds a comprehensive model of the semantic features described by Java code, and tracks the identity of these features as they evolve through sequential versions. This allows software metrics to be recorded over time. Eden and Mens [18] introduced the notion of evolution complexity and demonstrated how it can be used to measure and compare the flexibility of programming paradigms, architectural styles and design patterns. Canfora et al. [19] [20] investigated the relationship of source code complexity and disorganization. They used three factors which includes the number of contributors that modified the source code file. They found that entropy tends to increase with the number of file committers.

Our work differs in the type of the measured developer’s activities. We measure the contribution of developers based on their updates on the structural design of the system.
VII. CONCLUSIONS AND FUTURE WORK

Two measures have been proposed to measure the contribution of software developers in the evolved structural design of software systems. The presented measures consider only code changes that impact the corresponding UML class diagram representation of the source code. The measures are based on mining software repositories to analyze the historical code changes for developers. The number of updated design elements in the structural design is used for calculating the contributions. Measuring design contributions for developers provide useful information about the amount of contributions committed by developers to the structural design. The measures help project managers in locating developers who have major contributions to the structural design. The measures help project managers in locating developers who have major contributions to the structural design and empirically identify the amount of their contributions. The proposed contribution measures have been applied on two open source projects. Results show that developers vary in their design contributions. It is also shown that very small number of developers contributes in most of design changes.

The future work aims to introduce new metrics for developers that are based on different code activities. Possible measurements that are under investigation include the number of added/deleted packages and the amount of code refactorings. We are currently working on defining general metrics to measure the evolution of software systems.

REFERENCES


Maen Hammad is an Assistant Professor in Software Engineering Department at The Hashemite University, Jordan. He completed his Ph.D. in computer science at Kent State University, USA in 2010. He received his Master in computer science at Al-Yarmouk University—Jordan and his B.S. in computer science from The Hashemite University—Jordan. His research interest is Software Engineering with focus on software evolution and maintenance, program comprehension and mining software repositories.

Mustafa Hammad is an Assistant Professor at Information Technology department in Mu’tah University, Al Karak - Jordan. He received his PhD. in computer science from New Mexico State University, USA in 2010. He received his Masters degree in computer science from Al-Balqa Applied University, Jordan in 2005 and his B.Sc. in computer science from The Hashemite University, Jordan in 2002. His research interest is
Software Engineering with focus on static and dynamic analysis and software evolution.

Hani Bani-Salameh is an Assistant Professor in the Software Engineering Department at The Hashemite University since 2011. He holds a B.Sc. in Computer Science (1995), a M.Sc. in Computer Science (2007) from New Mexico State University (NMSU), and a Ph.D. in Computer Science (2011) from the University of Idaho (UI). His research interests include software engineering, computer supported cooperative work (CSCW), software development environments, collaborative software development in virtual environments, and social networking and social media. He studies social interactions in social networks and online environments, including Facebook and SourceForge.

Ebaa Fayyouni received the B.Sc. degree from The Hashemite University, Zarqa, Jordan, in 2000, the M.Sc. degree from the University of Jordan, Amman, Jordan, in 2002, and the Ph.D. degree from Carleton University, Ottawa, ON, Canada, in 2008. She has been with the Faculty of Prince Hussein Bin Abdalla II for Information Technology, The Hashemite University, since November 2008, where she is currently an Assistant Professor at the Computer Science Department. Her current research interests include statistical syntactical pattern recognition, micro-aggregation techniques, secure statistical databases, automata learning, applied algorithm, mobile application, data mining and e-learning. She got many awards during her academic life; one of them is Carleton University Medal on outstanding graduate work in 2008.
An Improved Face Recognition Technique Based on Modular Multi-directional Two-dimensional Principle Component Analysis Approach

Xiaoqing Dong
Department of Physics and Electronic Engineering, Hanshan Normal University, Chaozhou, 521041, China
Email: dxqzq110@163.com

Hongcai Chen
Department of Physics and Electronic Engineering, Hanshan Normal University, Chaozhou, 521041, China
Email: czhschc@126.com

Abstract—In this paper, a new method named modular multi-directional two-dimensional principle component analysis (M2D2DPCA) is proposed for face recognition. First, the original images are rotated at some predetermined angles so that we may extract features from the images in any direction. Then we divide the rotated images into smaller sub-images and apply 2DPCA approach to each of these sub-images. Finally we propose a fusion method named modular multi-directional 2DPCA (M2D2DPCA) to combine a bank of preliminary results in different directions. Compared with conventional 2DPCA based algorithms, the advantage of the proposed method is that it can extract significant features from the images in any direction and avoid the effects of varying illumination and facial expression. The results of the experiments on ORL and Yale datasets show that the proposed M2D2DPCA method can obtain a higher recognition rate than the conventional 2DPCA based methods.

Index Terms—face recognition, feature extraction, 2DPCA, MD2DPCA, M2D2DPCA, feature fusion

I. INTRODUCTION

Face recognition[1~4] is a hot topic in computer vision and pattern recognition. Yet, it is still a challenging task because the image of a face varies with facial expression, age, viewpoint, illumination conditions, noise etc. How to effectively extract informative features from training images is the key point of face recognition.

The methods based on principle component analysis (PCA)[5] are recognized as one of the most important feature extraction[6~8] instruments and still widely used in face image recognition. Sirovich and Kirby [9] first used PCA algorithm to represent the human face, while Turk and Pentland [10,11] proposed a famous "eigenface" method. The disadvantage of the traditional eigenface method based on PCA is that the 2D face image matrix must be previously transformed into a high-dimensional 1-D image vector. This transformation leads to some important issues such as small sample size problem and a high-dimensional image vectors space, where it is difficult to evaluate the covariance matrix accurately due to its large size [12].

To overcome this shortcoming, two-dimensional principle component analysis (2DPCA) [12] was proposed. The 2DPCA projects the image matrix onto the projection axis directly to extract the features from the image matrix, and the projected result is the so-called feature matrix[13]. 2DPCA is computationally more efficient than PCA because the total scatter matrix of 2DPCA has the lower dimensionality than that of PCA. Nevertheless, 2DPCA just extracts the information between the rows of the images and ignores the correlation between the columns. To alleviate this problem, two-directional two-dimensional PCA((2D)2PCA) was proposed by Zhang et al.[14]. (2D)2PCA shows that 2DPCA essentially works in the row direction of images and Zhang et al. proposed the alternative-2DPCA that works in the column direction of images. The alternative-2DPCA extracts feature in the column direction, which also can be viewed as the 2DPCA working in row direction. But, features extracted in one or two directions are always not sufficient for achieving high recognition rate. In Ref. [15], the authors first indicated that features extracted in different directions have different influences for accurate classification, and then proposed a method called multi-directional two-dimensional PCA (MD2DPCA). The MD2DPCA consists of two consecutive stages: First, using the directional 2DPCA(D2DPCA)[15] to extract features in different directions. In the second stage, using a matching score level fusion method to fuse several results of D2DPCA in different directions for face recognition.

A problem faced by all the methods which have been mentioned above is that they extract the most informative features all in a global manner. Consequently, some useful local information might be ignored. Also, it is difficult to avoid the influence of variations in illumination and facial expression when we extract features in a global manner. So, some researchers
propose modular based algorithms to solve this problem.

Modular PCA (MPCA) [16] divides images into smaller sub-images and uses PCA to extract features on each sub-image. MPCA considers each sub-image as a sample and projects all the sub-images onto a single projection matrix. Thus, variations in facial expression or illumination will affect only some of the sub-images. It has been shown that 2DPCA is a special case of BPCA [17]. Also, modular 2DPCA and block-wise two-directional 2DPCA employ a block-wise approach in feature extraction for reducing the computational complexity and obtain a higher recognition rate accuracy under the conditions of varying facial expression, illumination and pose [18,19]. But these modular based methods just extract features only in one or two directions, and some useful information in other directions might be lost.

In this paper, we propose a new algorithm called modular multi-directional two-dimensional PCA (MD2DPCA), which is an extension of the MD2DPCA method. In the MD2DPCA method, first, the original images are rotated in some predetermined directions (degrees). Then, we divide the rotated images into smaller sub-images and apply 2DPCA approach to each of the sub-images. Finally, we develop a fusion method to fuse the recognition results in different directions. The proposed method not only extracts significant features in any direction, but also extracts many more informative local features. Further more, variations in facial expression and illumination may affect only some of the sub-images. Hence, we expect the proposed method to have a better recognition rate. In the paper, we used two face databases to evaluate the performance of the MD2DPCA, 2DPCA, (2D)2PCA, M2DPCA, (M2D)2PCA and MD2DPCA.

The rest of this paper is organized as follows: Section II describes the D2DPCA and MD2DPCA method. Section III explains the proposed method modular MD2DPCA (MD2DPCA). The experiments on two public face datasets are given to compare the proposed method with several relevant methods in section IV. Finally, a conclusion is drawn in section V.

II. MULTI-DIRECTIONAL 2DPCA

A. 2DPCA

Consider an m by n random image matrix A. Let $X \in \mathbb{R}^{m \times d}$ be a projecting matrix with orthonormal columns. Projecting A onto X by the linear transformation $Y = AX$ yields an m by d matrix $Y$. $Y$ is the feature matrix and X is the projection matrix. The idea of 2DPCA is to find the optimal projection matrix $X_{opt}$.

Suppose that there are N training image samples in total and $A_k$ ($k = 1,2,\ldots,N$) represents the kth image of training samples. $\bar{A}$ is the mean of all training samples.

$$\bar{A} = \frac{1}{N} \sum_{k=1}^{N} A_k. \quad (1)$$

The total scatter matrix of the training images is defined as

$$G = \frac{1}{N} \sum_{k=1}^{N} (A_k - \bar{A})^T (A_k - \bar{A}) \quad (2)$$

Then, calculating the eigenvalues and eigenvectors of G, and the eigenvector of G corresponding to the largest eigenvalue is the best projection axis [20]. In general, it is not enough to have only one optimal projection axis. We usually need to select a set of projection axes, $X_1, X_2, \ldots, X_d$ [15]. In fact, these optimal projection axes are the orthonormal eigenvectors of G corresponding to the first d largest eigenvalues, and the $X_{opt} = [X_1, X_2, \ldots, X_d]$ is the optimal projection matrix.

B. Directional 2DPCA

In Ref [15], the authors indicated that features extracted only in one or two directions are not enough, and then proposed the directional 2DPCA (D2DPCA) that can extract features from the sample image in any direction. The D2DPCA is implemented by performing the 2DPCA on the rotated sample images which are obtained by using original image multiplying the rotation matrix $R(\alpha)$ in Euclidean space [15]. The rotation matrix $R(\alpha)$ is as follow:

$$R(\alpha) = \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \quad (3)$$

D2DPCA rotates sample images in a certain angle $\alpha$ and denotes the rotated sample images as $r_\alpha(A_1), r_\alpha(A_2), \ldots, r_\alpha(A_N)$. Since the rotated images will not be matrix form, they fill their four corners with 0 pixel values [15]. Fig.1 shows the image rotated with $\alpha = \frac{\pi}{6}$.

![Fig.1](a) The original image, (b) the image rotated with $\alpha = \frac{\pi}{6}$.

Similarly to 2DPCA, D2DPCA projects the rotated images $r_\alpha(A_1), r_\alpha(A_2), \ldots, r_\alpha(A_N)$ onto the projection matrix $X_{opt}$ by linear transformation as follow:

$$Y = r_\alpha (A_k) X \quad (k = 1,2,\ldots,N) \quad (4)$$

Where Y is the feature matrix of the rotated image. The idea of D2DPCA is to find the optimal projection matrix $X_{opt}^{D2DPCA}$.

Let $\bar{r_\alpha(A_k)}$ denote the mean of all the rotated training images and the total scatter matrix of the rotated image is defined as:

$$G_{rot} = \frac{1}{N} \sum_{k=1}^{N} (\bar{r_\alpha(A_k)} - \bar{A})^T (\bar{r_\alpha(A_k)} - \bar{A}) \quad (5)$$

The optimal projection matrix $X_{opt}^{D2DPCA}$ can be found using the eigenvalue decomposition of $G_{rot}$.
$$G_{D2DPCA} = \frac{1}{N} \sum_{i=1}^{N} (r_{A_i}(A_i) - \bar{r}_{A_i}(A_i))^T (r_{A_i}(A_i) - \bar{r}_{A_i}(A_i)) \quad (5)$$

Then solving for eigenvalues and eigenvectors of the total scatter matrix $G_{D2DPCA}$ and selecting $d'$ orthonormal eigenvectors as projection axes, i.e. $X_{1, D2DPCA}, X_{2, D2DPCA}, ..., X_{d', D2DPCA}$, corresponding to the $d'$ largest eigenvalues. The $X_{d', D2DPCA} = [X_{1, D2DPCA}, X_{2, D2DPCA}, ..., X_{d', D2DPCA}]$ is the optimal projection matrix of D2DPCA. As we can see from above, the 2DPCA and the alternative-2DPCA are two special cases of D2DPCA for $\alpha = 0, \pi / 2$ respectively[15].

C. Multi-directional 2DPCA

In Ref [15], the authors have proposed an effective fusion method named multi-directional 2DPCA(MD2DPCA) to fuse the features extracted by the D2DPCA in different directions for classification. The flowchart of MD2DPCA is shown in Fig.2 and its process[15] is implemented as follows:

Step1. Obtaining the projection matrix. It rotates all the training images with certain predetermined angles: $[0, \pi / 2, 2\pi / 2, ..., (l-1)\pi / l]$, where $l$ is the number of the feature extraction directions. Then apply 2DPCA to the rotated images to obtain the projection matrix. Now we have $l$ projection matrices.

Step2. Calculating the matching score of D2DPCA in one direction. Projecting all the rotated training and testing samples onto the corresponding projection matrices to obtain their feature matrices and then calculating the matching score between each testing sample and training sample[15] using:

$$s_{a_i, j} = \sum_{p=1}^{d'} \|Y_{i}^p - Y_{j}^p\| \quad (6)$$

Where $Y_{i}^p$ and $Y_{j}^p$ represent the $p$th feature vector of the feature matrix from sample $A_i$ and $A_j$, respectively. $\|Y_{i}^p - Y_{j}^p\|$ denotes the Euclidean distance between the two vector $Y_{i}^p$ and $Y_{j}^p$. Hence $s_{a_i, j}$ is Euclidean distance between the two feature matrices from sample $A_i$ and $A_j$ rotated with $a_i$ [15].

Step3. Adding up all the matching scores of D2DPCA in all directions by:

$$s_{i, j} = \sum_{a_i} s_{a_i, j} \quad (7)$$

Step4. Recognition using nearest neighbor classifier.

III. M$^2$D2DPCA: MODULAR MULTI-DIRECTIONAL 2DPCA

Under the conditions of varying facial expression and illumination, the methods (including PCA, 2DPCA, (2D)$^2$PCA, MD2DPCA) are not very effective since the features would vary considerably from the features of the face images with normal facial expression and illumination[16]. If the face images are divided into smaller sub-images, only some of the sub-images may vary and rest of the sub-images will remain the same. Therefore it is expected that the proposed method modular multi-directional 2DPCA($M^2$D2DPCA) not only extracts more informative features from different directions and local regions, but also has a higher recognition rate under the conditions of varying facial expression and illumination.

In this section, we present the details of the proposed $M^2$D2DPCA method. The flowchart of $M^2$D2DPCA is shown in Fig.3.

As shown in Fig.3, the process of the $M^2$D2DPCA is implemented by the following steps:

Step1. Rotating the sample images with the angles $[0, \pi / l, 2\pi / l, ..., (l-1)\pi / l]$, where $l$ is the number of the features extraction directions. Let image $A$ be an $m$ by $n$ image matrix and $A_i$ denotes the $ith$ sample image. The image matrix rotated with angle $\alpha$ is shown as follow:

$$r_{\alpha}(A) = \begin{bmatrix}
    r_{\alpha}(A)_{11} & r_{\alpha}(A)_{12} & \cdots & r_{\alpha}(A)_{1n} \\
    r_{\alpha}(A)_{21} & r_{\alpha}(A)_{22} & \cdots & r_{\alpha}(A)_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{\alpha}(A)_{m1} & r_{\alpha}(A)_{m2} & \cdots & r_{\alpha}(A)_{mn}
\end{bmatrix} \quad (8)$$

The new position of each pixel in the rotated image
matrix is obtained by using original position multiplication of the rotation matrix $R(\alpha)$ (Equation 3) in Euclidean space[15].

Step2. The rotated image matrix of $A_i$ is divided into $N = p \times q$ sub-images, where $p$ and $q$ are the numbers of the blocks in vertical and horizontal directions, as given in (9).

$$r_{\alpha}(A_i) = \begin{bmatrix}
    r_{\alpha}(B_{i1}) & r_{\alpha}(B_{i2}) & \cdots & r_{\alpha}(B_{i\alpha}) \\
    r_{\alpha}(B_{i2}) & r_{\alpha}(B_{i3}) & \cdots & r_{\alpha}(B_{i2\alpha}) \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{\alpha}(B_{i\alpha}) & r_{\alpha}(B_{i\alpha+1}) & \cdots & r_{\alpha}(B_{i\alpha\alpha})
\end{bmatrix}$$  \hspace{1cm} (9)

Where $r_{\alpha}(B_{ij})$ denotes each sub-image of the rotated image matrix. The width and height of each sub-image can be calculated as $m_i = m / p \quad \text{and} \quad n_i = n / p \quad \text{respectively}.$

It is expected that the image divided into smaller sub-images is helpful for extracting more local information and reducing the influence of varying facial expression and illumination. Nevertheless, the face images can not be divided into very small regions, because most of the global information of the face may be lost. The optimal block size ultimately depends on the characteristics of each particular database[16].

Step3. After dividing the images, performing D2DPCA on the sub-image, the size of each block is reduced.

The total scatter matrix of all the rotated sub-image is defined as:

$$G_{M^{d2DPCA}} = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} (r_{\alpha}(B_{ij}) - r_{\alpha}(B_{\alpha})) (r_{\alpha}(B_{ij}) - r_{\alpha}(B_{\alpha}))^T$$  \hspace{1cm} (10)

Where $M = Npq$ denotes the total number of all the training sub-images rotated with a certain angle. The average image $\overline{r}_{\alpha}(B_{ij})$ of all the training sub-images rotated with a certain angle is computed as:

$$\overline{r}_{\alpha}(B_{ij}) = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} r_{\alpha}(B_{ij})$$  \hspace{1cm} (11)

Next, we calculate the eigenvalues and eigenvectors of the total scatter matrix $G_{M^{d2DPCA}},$ and select $d'$ orthonormal eigenvectors as projection axes, i.e. $X_1^{M^{d2DPCA}}, X_2^{M^{d2DPCA}}, \ldots, X_{d'}^{M^{d2DPCA}}$ corresponding to the $d'$ largest eigenvalues. The $X^{M^{d2DPCA}} = [X_1^{M^{d2DPCA}}, X_2^{M^{d2DPCA}}, \ldots, X_{d'}^{M^{d2DPCA}}]$ is the optimal projection matrix of $M^{d2DPCA}.$

In this way, a feature matrix with the same number of blocks as before but with reduced size is obtained as given in (12):

$$Y_{i} = \begin{bmatrix}
    r_{\alpha}(B_{ij}) X_{1}^{M^{d2DPCA}} & r_{\alpha}(B_{ij}) X_{2}^{M^{d2DPCA}} & \cdots & r_{\alpha}(B_{ij}) X_{d'}^{M^{d2DPCA}} \\
    r_{\alpha}(B_{ij+1}) X_{1}^{M^{d2DPCA}} & r_{\alpha}(B_{ij+1}) X_{2}^{M^{d2DPCA}} & \cdots & r_{\alpha}(B_{ij+1}) X_{d'}^{M^{d2DPCA}} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{\alpha}(B_{ij\alpha}) X_{1}^{M^{d2DPCA}} & r_{\alpha}(B_{ij\alpha}) X_{2}^{M^{d2DPCA}} & \cdots & r_{\alpha}(B_{ij\alpha}) X_{d'}^{M^{d2DPCA}}
\end{bmatrix}$$  \hspace{1cm} (12)

Let a testing face image matrix $C$ rotated with the same angle and divided into the same number of blocks with the training sample. The rotated matrix $C$ is as follow:

$$r_{\alpha}(C) = \begin{bmatrix}
    r_{\alpha}(B_{j1}) & r_{\alpha}(B_{j2}) & \cdots & r_{\alpha}(B_{j\alpha}) \\
    r_{\alpha}(B_{j2}) & r_{\alpha}(B_{j3}) & \cdots & r_{\alpha}(B_{j2\alpha}) \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{\alpha}(B_{j\alpha}) & r_{\alpha}(B_{j\alpha+1}) & \cdots & r_{\alpha}(B_{j\alpha\alpha})
\end{bmatrix}$$  \hspace{1cm} (13)

Projecting $r_{\alpha}(C)$ onto the optimal projection matrix $X^{M^{d2DPCA}}$ by the linear transformation $Y = AX$ yields a feature matrix for the testing image as given in (14):

$$Y' = \begin{bmatrix}
    r_{\alpha}(B_{j1}) X_{1}^{M^{d2DPCA}} & r_{\alpha}(B_{j1}) X_{2}^{M^{d2DPCA}} & \cdots & r_{\alpha}(B_{j1}) X_{d'}^{M^{d2DPCA}} \\
    r_{\alpha}(B_{j2}) X_{1}^{M^{d2DPCA}} & r_{\alpha}(B_{j2}) X_{2}^{M^{d2DPCA}} & \cdots & r_{\alpha}(B_{j2}) X_{d'}^{M^{d2DPCA}} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{\alpha}(B_{j\alpha}) X_{1}^{M^{d2DPCA}} & r_{\alpha}(B_{j\alpha}) X_{2}^{M^{d2DPCA}} & \cdots & r_{\alpha}(B_{j\alpha}) X_{d'}^{M^{d2DPCA}}
\end{bmatrix}$$  \hspace{1cm} (14)

Step 4. After obtaining feature matrices of all the rotated training and testing images, we calculate the matching score between each testing sample and training sample, Similarly to MD2DPCA in section II.C.

$$s_{a,i,j} = \sum_{p=1}^{d'} \| Y_{p} - Y_{i,p} \|$$  \hspace{1cm} (15)

The $Y_{p}$ and $Y_{i,p}$ represent the $p$th feature vector of the feature matrix and $\| Y_{p} - Y_{i,p} \|$ denotes the Euclidean distance between $Y_{p}$ and $Y_{i,p}$. Then we calculate the final matching score of D2DPCA in all directions as following:

$$s_{i,j} = \sum_{a=1}^{A} s_{a,i,j}$$  \hspace{1cm} (16)

Step5. Classification using the nearest neighbor classifier.

IV. EXPERIMENTS AND ANALYSES

The performance of conventional algorithms (including 2DPCA, M2DPCA, (2D)$^{d}$PCA, M(2D)$^{d}$PCA and MD2DPCA) and the proposed method M$^{d}$2DPCA were evaluated with two image databases, ORL and Yale[21]. The ORL database was employed to examine the performance of mentioned algorithms for varying number of eigenvectors and sample size. The Yale database was used to test the performance of the system under the conditions of varying facial expression and illumination. We used MATLAB2010 for all of our experiments and performed on a PC with Pentium(R) Dual-Core E5800 3.20GHz CPU ,2GB memory and windows XP operating system.

A. Experiments on the ORL Database

The ORL database includes 400 gray-scale images of 40 individuals, each providing 10 different images. The images of each subject vary with facial expressions, facial details, scale and limited rotation. Moreover, there is also some tilting and rotation of the face of up to 20 degrees. All images are normalized to a resolution of 112 *92 pixels. Five sample images of one subject in the ORL
database are shown in Fig. 4. In the experiment, the first five images in each subject were used for training and the remaining images were used to test the recognition rates. Thus, the total number of training samples and testing samples were both 200. The D2DPCA can extract features in any direction and this paper proposed a framework to fuse multi-direction features. We tested the proposed method using three directions and five directions (denoted as 3d and 5d) in the experiments respectively. The face images were divided into 4*4 blocks for testing the performance of the modular based algorithms (including modular 2DPCA, modular (2D)^2PCA and the proposed method). Hence the size of each sub-image is 56*46, the total scatter matrix is 56*56. The feature matrixes of M^2DPCA, MD2DPCA and M'D2DPCA are all 112*4V (V denotes the number of eigenvectors). But for the 2DPCA, (2D)^2PCA and M(2D)^2PCA, the sizes of feature matrixes are 112^*V, V*V, 4V*4V, respectively. Fig. 5 shows the result of dividing a face image into 4*4 blocks.

In the experiment, for the proposed method, we first rotated the training samples and divided them into smaller sub-images according to (8) and (9). Then we computed the feature matrix of each training sample as given in (12). Similar to training sample, we calculated the feature matrix of the testing sample as given in (14). After obtaining feature matrixes of the rotated training and testing images, we calculated the matching score between each testing sample and training sample according to (15) and (16). Finally, we classified the test sample by using the nearest neighbor classifier.

In the first experiment, we tested the performance of the proposed method and the related methods for varying number of eigenvectors (V from 1 to 20). However, the proposed M^2D2DPCA method using five directions also obtains the highest recognition rate of 97%.

Then, we conducted the next experiment at V=10, i.e. eigenvectors corresponding to the ten maximum eigenvalues of the covariance matrix. This experiment was used to examine the performance of the related methods with varying sample size and varying V would have the same effect on the related algorithms as shown in Fig. 6. Thus, we only considered the first ten eigenvectors for the experiment. As shown in Fig. 7, the recognition rates are increasing in all methods as the increasing of S (S denotes samples size) and the recognition rates grow fast when S<5, but grow slowly when S>5. It can also be observed from Fig. 7 that the proposed method (M'D2D2DPCA using three and five directions) has the higher recognition rates than that of the other methods at most of S and the highest recognition rate of 97.3% was obtained by the proposed method at S=9.

From the results we note that the proposed M'D2D2DPCA method always has a slightly better recognition rates than the traditional methods with varying number of eigenvectors and sample size, since it not only extracts many more significant information from different directions, but also extracts local information from smaller regions, while the minor and insignificant variations are eliminated.

Furthermore, we varied M (number of sub-images) from 2 to 64 to observe the effect of varying M on recognition rate. The number of eigenvectors is 10 and sample size is 5. Here, we only compared the proposed method in different number of directions. We can observe from Fig. 8 that the recognition rates are increasing as the increasing number of sub-images when M<16, because
many more significant local information has been extracted from smaller regions and variations affect only some of the sub-images. But the recognition rates fall fast when M>16, because the global information of the face may be lost when the image is divided into very small regions. Also, it can be observed from Fig.8 that the recognition rates of the method in five directions are slightly better than the other numbers of directions, since it can extract much more significant information from the more directions.

\[ \text{Recognition rate}(\%) \]

\[
\begin{array}{|c|c|}
\hline
\text{MMD2DPCA (3D)} & 85.50 \text{ (Variations in illumination)} \\
\text{MMD2DPCA (4D)} & 81.55 \text{ (Variations in illumination)} \\
\text{MMD2DPCA (5D)} & 86.21 \text{ (Variations in illumination)} \\
\text{MMD2DPCA (5D)} & 88.12 \text{ (Variations in facial expression)} \\
\text{MMD2DPCA (5D)} & 90.55 \text{ (Variations in facial expression)} \\
\text{MMD2DPCA (5D)} & 92.80 \text{ (Variations in facial expression)} \\
\text{MMD2DPCA (5D)} & 95.16 \text{ (Variations in facial expression)} \\
\text{MMD2DPCA (5D)} & 97.50 \text{ (Variations in facial expression)} \\
\text{MMD2DPCA (5D)} & 99.80 \text{ (Variations in facial expression)} \\
\hline
\end{array}
\]

Fig.8 Recognition rates of the proposed methods with varying number of sub-images.

**B. Experiments on the Yale Database**

The Yale database has 165 images of 15 adults (each person has 11 images). The face images are with variations in facial expression (including normal, sad, happy and surprised expressions) and variations in illumination. There are also images with and without glasses. Each image was manually cropped and resized to 100×100 pixels in this experiment. Figs.9 shows the images of a person from the Yale database.

(a) Normal (b) Variations in illumination (c) Variations in facial expression

Fig.9 Images of one person in the Yale database. (a) normal image, (b) images with variations in illumination, and (c) images with variations in facial expression.

In order to compare the performance of related methods for facial expression and illumination variations, we designed two sub-experiments: the normal face image of each person is chosen as training sample for both two sub-experiments. Three face images with variations in illumination of each person are chosen as the testing samples for sub-experiment 1 and seven face images with variations in facial expression of each person are chosen as the testing samples for sub-experiment 2. Thus the total number of training samples for both two sub-experiments is 15. The total number of testing samples in sub-experiment 1 and 2 are 45 and 105, respectively. In these two sub-experiments, we used the case of the 4×4 blocks for the modular based methods (including M2DPCA, M(2D)2PCA and M(2D)2DPCA). Hence the size of each sub-image is 25×25, the total scatter matrix is 25×25. The number of eigenvectors V varies from 1 to 20. The feature matrix of M2DPCA, MD2DPCA and M(2D)2DPCA are all 100×4V. But for the 2DPCA, (2D)2PCA and M(2D)2PCA, the size of the feature matrices are 100×V, 4V×V, 4V×4V, respectively.

In these two sub-experiments, for the proposed method, we conducted the experiments similarly to the experiment mentioned in section IV.A. We first rotated the training samples and divided them into smaller sub-images according to (8) and (9). Then we calculated the feature matrix of each training sample and testing sample as given in (12) and (14), respectively. After obtaining feature matrixes of the rotated training and testing sample, we calculated the matching score between each testing sample and training sample according to (15) and (16). Finally, we classified the test sample by using the nearest neighbor classifier.

Tab.I shows the top recognition rates under the conditions of varying illumination and facial expression for all methods. It can be observed from Tab.I that modular based methods effectively reduce the influence of variations in illumination and improve the recognition performance. The proposed M(2D)2DPCA method using five directions obtains the best recognition rate of 86.21%.

Under the conditions of varying facial expression, compared with 2DPCA, (2D)2PCA and MD2DPCA, modular based methods also have better recognition rates, because the influence of variations in facial expression is reduced by dividing the image into smaller regions. In all the related methods, the performance of the proposed method using five directions is the most excellent and obtains the best recognition rate of 98.12%.
proposed method always has the best recognition performance. Since the face images were divided into smaller regions, the variations in illumination and facial expression would be more representative of the local information of the face, so the modular based methods can avoid the effects of varying illumination and facial expression. However, because of extracting information from many more directions, the proposed M²D2DPCA method can obtain much more significant information from not only the horizontal and vertical directions. These are helpful for obtaining higher recognition rates as observed in the experimental results. 2DPCA, (2D)²PCA and MD2DPCA methods were not very effective when the face images have variations in illumination and facial expression, because they consider the global information of each face image and under these conditions the eigenvectors of the test image and the training image with normal illumination and facial expression may have a big difference, so it is difficult to recognize them correctly.

V. CONCLUSION
In this paper, a modular multi-directional
2DPCA(M²D2DPCA) method, which is an extension of the multi-directional 2DPCA(MD2DPCA) method for face recognition has been proposed. The advantages of the method can be briefly expressed as follows: First, the proposed method can extract significant information from the images in any direction, and its implementation is very simple. Second, since the face images are divided into smaller regions, the variations in illumination and facial expression will be more representative of the local information of the face, hence the method can reduce the effects of varying illumination and facial expression. The experimental results on ORL and Yale database show that the proposed M²D2DPCA method is more powerful for classifying human face images than the conventional 2DPCA based methods.

ACKNOWLEDGMENTS
This work is supported by 2013 colleges and universities in Guangdong Province, the quality of teaching and teaching reform project—Electrical Engineering and Automation comprehensive reform pilot project (Guangdong teach high letter [2013] No. 113, Guangdong Finance Education [2013] No. 322). Finally, the authors would like to thank the anonymous reviewers for their constructive advice.

REFERENCE

Xiaoping Dong received his B.E. degree in Electronic Information Science and Technology from Zhaoqing University, China in June 2005 and his M.S. degree in signal and information processing from Guangdong University of Technology, China in June 2008. His current research interest includes automatic control technology and machine vision.
Hongcai Chen received his B.E. degree in electronic engineering from Northeast Normal University, China in June 1990 and his M.S. degree in machine design from Harbin Engineering University, China in June 2000. His current research interest includes electronic engineering and embedded system design. He is now an associate professor in Department of Physics and Electronic Engineering, Hanshan Normal University.
Designing and Implementation of an Online System for Electronic Contract Negotiation Based on Electronic Signature

Juntao Gu
Department of Information Management, Management School, University of Shanghai for Science & Technology, Shanghai, China
E-mail: stick90328@vip.qq.com

Xiaodong Zhu
Department of Information Management, Management School, University of Shanghai for Science & Technology, Shanghai, China
E-mail: zhuxd@usst.edu.cn

Abstract—The contract is the bridge to transaction, as well as the core of business practice. Accompanying with the development of electronic commerce, contract negotiation under more complicated environment in advantage of internet is becoming increasingly urgent. This paper briefly introduces the status of electronic signature and electronic contract in China, then presents a reference architecture for an online system for electronic contract negotiation and signature in detail. This architecture is systematically designed on the basis of electronic signature technology to realize online contract signing, electronic contract storage on the trusted third party and online or offline verifications. All above ensures electronic contract's confidentiality, integrity and non-repudiation, makes a breakthrough for electronic signature in business field, and realizes the electronization in the whole process of business negotiation.

Index Terms—e-commerce, e-contract, e-signature, online negotiation

I. INTRODUCTION

With the rapid development of computer technology, the rapid expansion of web services and the fast integration of global logistics, e-commerce, as one of important strategic emerging industries, has become the vanguard of reform and opening up to play a significant role in the process of innovation and transformational development in China. In 2013, promoted by government, enterprises and consumers, e-commerce in China kept a smooth trend of rapid development. According to statistics, its trading volume topped more than 10 trillion yuan, a 27.8% increase compared with last year, which almost equaled to 18.6% of Chinese GDP in 2014[1]. Though more and more procedures of commercial activities being executed electronically, as an essential procedure of commercial activities, especially in B2B activities, contract negotiation and signature as an essential procedure developed relatively slowly on its way to electronization. The advanced stage of e-commerce is the realization of electronic processing in the whole business process. If the electronic contract negotiation could not be researched and used effectively, it would definitely restrict the high-speed development of e-commerce. Hence, this paper presents an online system for electronic contract negotiation based on electronic signature technology.

In recent years, a lot of papers focused on modeling of the online contract negotiation and signature. In 1997, Y. Lei et al detailed a comparison of various workflow metamodels [2]. But there was no work on conceptually modeling e-contracts or a methodology for transforming an e-contract into workflows. In 2004, talking Entity Relationship Diagram as a framework, P. R. Krishna et al described an implementation model of an e-contract which bridges between the XML contract document and web services, but only based on workflow [3]. In 2005, started from the angle of web services, T. Bui et al put up with the online negotiation platform structure model and negotiation process [4]. In 2006, taking SOLACE as a standard, O. Abass et al proposed a universal and multi-issues negotiation frame to satisfy the need of electronic negotiation system [5-6]. In 2007, T. Kwok et al designed a web-based and e-mail driven electronic contract management system supporting both internal and intra-enterprises workflows, and was used in several IBM pilot programs [7]. In 2008, S. Angelov et al presented a reference architecture of three levels for the development of e-contracting systems, which is aimed at facilitating the design of logical views of concrete e-contracting systems [8]. In the same year, combined with the technology of cloud computing, T. Kwok et al presented the first of a kind multi-tenancy Software as a Service (SaaS) e-contract management application, which may be more suitable for the small and medium businesses and in particular the very small businesses[9]. In 2009, X. Yu et al defined a lifecycle of e-contract and recounted the establishment process of e-contract. This system help to protect businessmen's right and interests, and facilitate governmental monitoring of the e-commerce market [10]. In 2010, from the perspective of reducing the manpower,
material and financial resources, W. Xin et al put forward another platform of electronic contract based on electronic signature technology to meet the needs of ever-developing business model and the operations of supply chain [11]. In the same year, S. Liu et al also proposed solutions for electronic contact’s supervision and notarization, which could be a breakthrough point for E-commerce supervision, related economic data analysis and related public service expansion, thus promoting the development of online trading in term of safety and normativity [12]. In 2012, J. Yang et al explained the status of current business negotiation in China, analyzed the difficulties in the popularizing of applying the electronic signature into commercial activities, and then raised the pioneer Network Platform for Electronic Contract Making as well as its applications [13]. In 2013, based on the automatic synthesis of fair non-repudiation protocols, K. Chatterjee et al showed how to specify the objectives of the participating agents, then explained clearly that assure-guarantee synthesis can be computed efficiently as the secure equilibrium solution of three-player graph games[14].

This paper presents an architecture for an online system for electronic contract negotiation and signature. It also presents electronic contract’s storage in the trusted third party (TTP) and two different verification ways, online and offline. This would realize the application and popularization of electronic signature in business contract field.

II. STATUS OF E-CONTRACT, E-SIGNATURE AND ONLINE NEGOTIATION

Nowadays, e-commerce in China showed that internet is largely popularized and online sales scale is increasing explosively. However, there are also several problems for example, invasion of intellectual property right, counterfeit and shoddy products, malicious falsehood and other kinds of transgressions. Inadequate supervision may be one of the main reasons causing these situations, so the constraint for transactions is not enough. However, the goal of electronic contract, as the same as the traditional contract, is to constraint both parties in trading activities like production and business operation by the legal effectiveness in itself. Once an e-contract has been signed, the rights of both parties are under the protection of the law. Meanwhile, e-commerce trading environment will be greatly improved.

A. Electronic Contract

According to Chinese ‘Electronic Signature Law’, electronic contract can be defined as an agreement among both or multi litigants to establish, modify and terminate the civil rights and duties in property through electronic information net and by using electronic format.

According to statistics of year 2011 in China, there were around 11 million medium and small-sized enterprises, and 36 thousand large and medium-sized enterprises. Their average online selling rate was about 29.33%, and average online procurement rate was above 28.91% [15]. Supposing every medium and small-sized company had 10 contract negotiations per year, there would have 1.1 billion contract negotiations totally. If one fifth of them adopt online trading, there would be 22 million contract negotiations per year. Compared with traditional paper contract, electronic contract could decrease negotiation cost for enterprises, such as business travel expenses, express fee and time cost, etc. Meanwhile, the electronic form of e-contract is also easy to store and manage.

On the other hand, Chinese new ‘Contract Law’ expanded traditional contract to data message format. The No.11 item of new contract law says ‘written form of contract is a form where contract papers, letters and data message (including telegram, telex, fax, electronic data exchange and email) can present its content in a visible way[16]. It means valid electronic contract will be empowered the same legal effectiveness as traditional paper contract. Once contract disputes appear, electronic contract and paper contract share the same reliable legal evidence [17-18].

B. Electronic Signature


On the other hand, from the perspective of law, in many countries, including the United States, the European Union, India, China, Brazil and Australia, electronic signatures (when recognized under the law of each jurisdiction) have the same legal consequences as the more traditional forms of executing of documents.

In traditional commerce, contracting parties should sign or stamp in the paper. This kind of signature and stamp has several functions: first, showing the source of the document; second, demonstrating that signatories have confirmed the paper’s content; third, constructing an evidence that signatories are responsible for the correctness and integration of the content of the paper [21]. And in No. 14 item of ‘Electronic Signature Law’ stipulates that ‘reliable electronic signature has equivalent legal effectiveness as written signature and stamp’. Hence, electronic signature can realize the confirmation of traditional paper signatories and the verification of confidentiality, integrity and non-repudiation.

So far, most studies of electronic signature focused on the encryption technology. As a fundamental technology, data encryption is the corner stone of the security of all the communication. Scholars around the world offered symmetric and asymmetric encryption methods to ensure the security of data encryption. However, there was no breakthrough in the application of electronic signature. So far, the application of electronic signature in China is
only limited to certain parts of electronic e-governance, such as foreign trade declaration. In other fields, especially in production and business fields, it has still not found a right breakthrough point. Several reasons why electronic signature is still distant to our people could be concluded as follows:

First of all, lack of awareness. People even have no impression on the function of electronic signature; second, no breakthrough point is selected in its application; last but not least, technologies of electronic signature has not been well developed.

All in all, in terms of current situation, the breakthrough point of electronic signature application should be the procedures of online business negotiation and contract signing where electronic signature can be conducted frequently and exert its utility to maximum. The following research raises the online system for electronic contract negotiation based on electronic signature technology to a commercial environment.

C. Online Negotiation

Negotiation, such as commercial contract negotiation and international dispute negotiation, is a process where each party involved in negotiating tries to gain an advantage for themselves by the end of the process [22]. Compared with traditional negotiation that costs lots of social resources, productivities and opportunities for innovation, online negotiation, which can greatly improve negotiation efficiency, is more in line with the requirements of low-carbon life. On the other hand, in recent years, with the rapid development of e-business, globalization of e-commerce emerges inevitably. Additionally, modern companies have trended to be scalization, conglomeration as well as internationalization, which all contribute to an acute exploitation of a better negotiation manner with characteristics of economy, efficiency, effectiveness and security.

However, in terms of current situation, the research and popularize of online negotiation, as a key process of e-business transaction, still processes slowly. No mature electronic contract negotiation system is available nowadays. Although text, audio and video exchanging technologies have been sophisticated and applied in online chatting software, such as MSN, ICQ and Skype, nothing to do with negotiations for ceremonial electronic contracts.

III. SYSTEM ARCHITECTURE OF SYSTEM

The trade standard ‘The Process Specification of Online Setting Electronic Contracts (SB/T 11009-2013)’, which is issued by the Ministry of Commerce of the PRC, entered into force in December 1, 2013[23]. The article in this standard stipulates that only through the system established by the third party, the fairness of negotiation process and the effectiveness of the signed electronic contract can be protected by law.

Hence, according to this trade standard, this paper presents such a system based on the independent third party platform of electronic contract negotiation and signature, separating the management and verification of electronic signature and the storage verification function of signed contract from platform. System architecture of the whole system is illustrated as Figure 1.

This system architecture sets electronic signature server and electronic contract storage server in the Certificate Authority center and electronic contract storage center that are independent from the platform of electronic contract negotiation and signature, which has three advantages as follows:

1. Confidentiality: Put electronic contract storage center in national authority organizations makes it less easy to steal and destroy.
2. Independency: Once certain part of this system has problems, it can disconnect connection with other server immediately, thus minimize lost.
3. Conditionality: Expand from the independent third party to realize balance among several third parties and ensure both parties’ interests in the negotiation.

IV. FRAME AND FUNCTION DESIGN

In the light of the basic functions that electronic contract negotiation system ought to possess, the trade standard SB/T11009-2013 puts forward seven requirements as follows[23]:

1. Allow users to input contract templates, and automatically generate the contract text compliance with the requirements of law in user’s format requirements.
2. Allow for online negotiation and contract revision.
3. Able to realize a variety ways of multimedia communication, such as text, video, audio, mobile.
4. Able to realize real-time transmission and backup of the information of contract and negotiation in the process. Meanwhile, the backup data can be traced back and replicate.
5. Apply electronic signature or other technologies to verify the confidentiality, integrity and availability (CIA triad) of the contract.
6. Allow for storage online and third party at the same time.
7. Be in line with the technology, equipment and management systems related to national security and confidentiality standards.

For meeting these above requirements, the main functions of the system frame are designed as six major parts: user management, contract management, contract negotiation, contract preview, online signature and online
contract storage, as shown in Figure 2.

(1) **User management:** mostly including system user’s registration and its information storage and modification;

(2) **Contract management:** the classification and re-edition for user’s contract template;

(3) **Contract negotiation:** a core function of the system. After the negotiation initiator inputs the contract draft, the negotiation part begins. Both parties modifies every item on the contract till they reach an agreement that they have no dissent for each item in the contract.

(4) **Subsidiary Chatting:** mostly including various real-time multimedia communication methods, such as text, video and audio. Its main purposes include:
   a) Promote final contract signing via both negotiation parties’ conversation;
   b) Confirm each other’s identity through audio and video chatting;
   c) Save the communication record as an effective evidence.

(5) **Online signature:** after several rounds of confirmation by both negotiation parties, they can sign this contract online by using the USBKEY that registered in local Certificate Authority center. Once signed successfully, this electronic contract will share the same legal effectiveness with paper contract.

(6) **Contract storage:** mainly including storage in local places, e-negotiation database and e-contract database.
   a) **Local storage:** mostly backing up locally for singed contract, unfinished contract and contract templates, which makes it convenient for enterprises or individuals to check and record.
   b) **Storage on e-negotiation database:** store contract templates and unfinished contracts mostly. Many enterprises have long-term relationship of supply and demand, so contents of contract are similar each time. Store contract templates in negotiation server can decrease negotiation time and increase negotiation efficiency. On the other hand, some contracts may be too complicated to reach a consensus in one or two days, so the unfinished versions must be saved till the end of negotiation.

(c) **Storage on e-contract database:** store signed contract only. It is also called contract third party storage, which is convenient for users to review, and regarded as a reliable legal statement when disputes appear between both negotiation parties.

V. **DETAILED PROCESS DESIGN**

A. **Application for Digital Signature**

Traditional contract in paper is valid only if it is signed or stamped. It is the same with electronic contract. In network cryptography, a certificate authority (CA) is an entity that issues digital certificates, which certifies the ownership of a public key by the named subject of the certificate. This allows others (relying parties) to rely upon signatures or assertions made by the private key that corresponds to the public key that is certified. A Certificate Authority Center is an authoritative, reliable and equitable third party that is trusted by both the subject (owner) of the certificate and the party relying upon the certificate, and it is responsible for issuing digital certificates, authenticating digital certificates and managing issued digital certificates. In that sense, Figure 3 illustrates an implementation that one corporate representative applies for digital signature Certificates. After the corporate representative submitted corporate’s related information, the CA center would verify the related information. Once confirmed, CA administrators will adopt PKI technologies to issue a digital signature certificate and save the encrypted certificate into USB card that used for digital signature, and then save the information of digital certificate to E-signature database.

<table>
<thead>
<tr>
<th>Corporate Representative</th>
<th>CA Administrator</th>
<th>RA Server</th>
<th>CA Server</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Corporate RepresentativeApply for Digital Signature Certificates</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. UML activity chart for corporate representative applying for digital signature in USB card stage

The main function of the CA Administrator, RA Server and CA Server is clarified as follows:

- **CA Administrator:** assist corporate representatives in accomplishing the certificate application, and manage issued certificates.

- **Registered Authorization (RA) Server:** extract the required fields from the verified information of certificate application/update/invalidation, then submit to the CA server in a given format and, at
the same time, receive and record the returned results.

- **CA Server**: firstly, generate the public and private key pairs by means of software or hardware; secondly, receive the requirements of certificate application/update/invalidation from RA service; thirdly, regularly issue certificate revocation list (CRL).

### B. Electronic Contract Negotiation and Signature

This section is divided into three parts: negotiation invitation, contract negotiation and online signature.

#### A) Negotiation invitation

Before this stage, the contract negotiation parties made a negotiation appointment firstly. By searching through the authentication database, the initiator confirms partner’s information by each other and sends negotiation request. If the partner is online, they can confirm each other’s identity and then prepare for negotiation; if the partner is offline, other methods, such as SMS and e-mail, could be used to notice the partner and wait to enter the negotiation stage together.

#### B) Contract negotiation

After the stage of negotiation invitation, both negotiation parties begin the contract negotiation stage. Firstly, one party of the contract edits an existed or new contract template, generates contract draft and enters negotiation page; at the same time, the other party waits for the contract draft’s generation and then enters negotiation page automatically. After checking contract content, both can edit contract items if necessary. To prevent the possibility that both parties are editing the same content at the same time and the inconvenience it brings, the system adopts simplex model which means only one party can edit content at one time, and only after one party finishes the edition, the other party can start editing. To better and safer reach an agreement, the subsidiary chatting function is added in this contract negotiation stage.

If one party has confirmed contract items, the contract content cannot be edited anymore. Once the contract has been confirmed by both parties, they have to preview the contract and confirm it one more time to browse the entire contract, and then both enter online signature stage. At this point, the system prompts that this contract will be saved as contract template for future usage. If the negotiation pauses or either party has disagreements towards the contract, it can be re-edited, and negotiation server will save details till the final confirmation of the contract.

#### C) Online signature

Once both parties have confirmed the contract, they have to finish online signature and empower the contract legal effectiveness. Both parties use the USB card that represents their enterprises identities to sign the contract online. After the digital certificates are verified through digital signature server, the online signature is done. This platform sends the signed contract to e-contract storage database, and will generate the contract in PDF format and unique QR code for both parties to download and store.

The whole process of electronic contract negotiation and signature is showed in Figure 4.

![Figure 4. UML activity chart for negotiation and signature stage](image-url)
signed, the signer’s certificate is embedded in the PDF file. Therefore verification of identity of both contract parties can be realized by verifying the digital signature in the PDF format contract offline.

B) Contract online verification

If users still have doubt on the authenticity, integrity and effectiveness of contract, users can verify contracts online, as shown in Figure 5. Once the application is made, users submit the contract document they need to verify and corresponding QR code to the system administrator and then verification can be done by comparing with contract backup copy in the e-contract storage database immediately. If they are well matched, this authentication is successful. If not matched, the contract may be modified after electronic signature.

<table>
<thead>
<tr>
<th>Contract Online Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User</strong></td>
</tr>
<tr>
<td>Apply for the contract online verification</td>
</tr>
<tr>
<td>Upload the PDF file of contract and the corresponding QR code</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Figure 5. UML Activity Chart for Contract Online Verification

VI. INNOVATIONS OF SYSTEM

A. Satisfy Enterprises’ Urgent Need of Remote Business Negotiation

According to the study of Borough et al, negotiation activities take 20% of manager’s working time in organizations [25]. However, in those real negotiations, negotiators can hardly reach satisfactory results. The ineffectiveness decreases useful resources in society, productivity and opportunities of creativity, and increases social conflicts and self-destruction.

With the prosperous development of domestic and international E-commerce, enterprises has the tendency of scale expansion and collectivization, the frequency different kinds of remote business negotiation will increase, so enterprises’ need for remote business negotiation becomes more urgent Online negotiation and contract signature system is not limited to geography restrictions and can meet enterprises’ urgent for remote business negotiation. Besides, this system can effectively reduce negotiation cost, including business trip expenses, express fee and time cost, etc. [26–27].

B. Realize Electronization in the Whole Process of Online Business Negotiation

In this research, based on the ‘Electronic Signature Law’ and existing mature electronic signature technology, this system realizes the electronization of business negotiation and contract signature and the formation of the entire online business negotiation platform setting the functions of supplier/customer’s connection, negotiation invitation, contract signature and signed contracts tracking management in one.

C. Realize the Expansion and Application of Electronic Signature in Business Field

The research provides users a whole new concept of business negotiation through the creative application of electronic signature on contract making, realizing the equal legal effect between paper contracts and electronic ones. Meanwhile, this research presents an electronic contract application mode to meet the legal requirements and designs the application process and framework with the electronic signature technology and independent third party electronic contract preservation to fill the blank in the expansion and application of electronic signature in domestic and foreign business fields.

D. Develop a Standard Structure of Electronic Contract Templates to Adapt the Form of Online Negotiation Independently

It seems that no standard structure of electronic contract templates is widely used in online business negotiations all over the world currently. In this research, XML electronic contract markup language and XML documents are used for automatically generating the contract templates through paste the text from Microsoft word or excel, where pictures are also acceptable, according to the requirements of users. Meanwhile, this research also benefits for the normalized construction in digital business market.

E. Achieve the Preservation of Electronic Evidence on an Independent Third Party

This system achieves online preservation of electronic evidence including electronic contracts and documents. Meanwhile, the use of a VPN dedicated channel makes sure of safe delivery of electronic contracts, and the comprehensive utilization of the technologies of time stamps, network attached storage (NAS), Storage Area Network (SAN) and third-party data security and disaster recovery to meet the data security requirements of the ‘Electronic Signature Law’.
VII. CONCLUSIONS

In this paper, we focused on the development environment of electronic commerce and the current situation of e-contract, e-signature and online negotiation, especially in China. According to the Chinese latest trade standard ‘The Process Specification of Online Setting Electronic Contracts’, then we proposed that e-contract should involve electronic signature as its primary element and an independent third party should be introduced into the negotiation and signature of e-contracts. Based on these views, we broke the ice and raised a whole new system framework for the online system for electronic contract negotiation based on electronic signature technology, where we defined a lifecycle of e-contract negotiation and two different ways of verification, online and offline. Finally, we cited the innovations of this system in list.

Contracts are required in most business transactions in companies of all sizes as they constitute the binding relationships between a company and its suppliers, business partners, or customers. Likewise, e-contract is a sensitive point in e-commerce area, which involves many parties’ rights and interests. Especially on the view of management, we need a mature and practical solution to regulate and accelerate the development of e-contract. This paper gives a good point of penetration to achieve above goals, but we still need more study on related area to improve our system, including its security, practicability and operability. On the other hand, there are still many blanks in the law of electronic commerce field, and people lack of enough trust on the application of electronic signature. Hence, it still has a long way to go to achieve massive substantive expansion and application of electronic signature in business field.

ACKNOWLEDGMENT

This paper is supported by Ph.D. Programs Foundation of the Ministry of Education of China (Project Number: 20121120120004) and Innovation Fund Project for Graduate Student of Shanghai (Project Number: JWCXSL1302).

REFERENCES


Juntao Gu was born in Jiangsu, China, on Mar. 1990. He received the BS degree from Department of Computer Engineering, Jiangsu University of Technology in 2012, where he was awarded the National Scholarship. Since then, he has been studying in the Department of Information Management in the University of Shanghai for Science & Technology, Shanghai, China. Now he is also a student fellow in the China Computer Federation (CCF). His current research interests include e-commerce, e-signature, e-contract and software modeling.

Xiaodong Zhu was born in Anhui, China, on 1981. He received the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, in 2009. Now he is an associate professor in the Department of Information Management, Management School, University of Shanghai for Science & Technology, Shanghai, China. He is also a senior member of the China Computer Federation (CCF). His current research interests include data mining, e-commerce, knowledge engineering and software engineering.
Lexical-semantic SLVM for XML Document Classification

Jun Long
School of Information Science and Engineering, Central South University, Changsha 410075, China
Email: jlong@csu.edu.cn

Luda Wang
School of Information Science and Engineering, Central South University, Changsha 410075, China
Xiangnan University, Chenzhou 423000, China
Email: wang_luda@csu.edu.cn

Zude Li, Zuping Zhang, Huiling Li and Guihu Zhao
School of Information Science and Engineering, Central South University, Changsha 410075, China
Email: zli@csu.edu.cn, zpzhang@csu.edu.cn, lihuiling2013@csu.edu.cn

Abstract—Structured link vector model (SLVM) and its improved version depend on statistical term measures to implement XML document representation. As a result, they ignore the lexical semantics of terms and its mutual information, leading to text classification errors. This paper proposed a XML document representation method, WordNet-based lexical-semantic SLVM, to solve the problem. Using WordNet, this method constructed a data structure for characterizing lexical semantic contents of XML document, and adjusted EM modeling to disambiguate word stems. Then, synset matrix of lexical semantic contents was built in the lexical-semantic feature space for XML document representation, and lexical semantic relations were marked on it to construct the feature matrix in lexical-semantic SLVM. On categorized dataset of Wikipedia XML, using NWKNN classification algorithm, the experimental results show that the feature matrix of our method performs F1 measure better than original SLVM and frequent sub-tree SLVM based on TF-IDF.

Index Terms—Semi-structured document, SLVM, Lexical semantics, Classification, Feature matrix

I. INTRODUCTION

In order to record semi-structured information, lots of document standards, such as HTML, BibTex and SGML/XML, are recommended by many main international standards organizations. The structural flexibility of these semi-structured documents offers an approach to representing information in specified structure. Contrasting conventional plain documents, semi-structured documents represent their syntactic structure via the use of document structural elements marked by user-specified tags, and the associated schema specified in Schema format [1]. Extensible Markup Language (XML) is a semi-structured document format with strong support via Unicode for different human languages. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures,[6] for example in web services.

Considering the structure property, classification methods for semi-structured document analysis have to consider the information embedded in both the element tags and their associated contents. Recently, the Structured Link Vector Model (SLVM) [2] was proposed for XML documents analysis. Especially, the extended version of SLVM uses the closed frequent sub-trees as structural units for content extraction from the XML document. For XML document representation, existing models depend on statistical term measures for feature extraction. However, in the information retrieval field, statistical term measures causes XML document analysis to perform on the level of term string basically, and neglect lexical semantic contents in the structured elements.

Semantic approach is an effectively used technology for document analysis. It can capture the semantic features of words under analysis, and based on that, characterizes and classifies the document. Close relationship between the syntax and the lexical semantics of words have attracted considerable interest in both linguistics and computational linguistics. For XML document classification, the design and implementation of lexical-semantic SLVM take account of the lexical semantics particularly. Unlike present models, using WordNet [3], our model developed a new term measure which can characterize lexical semantic contents and relations, and provides a practical method for XML document representation which can handle the impact of synonyms and other relations. Theoretical analysis and relevant experiments are carried out to verify the effectiveness of this model.
II. AN OVERVIEW ON ORIGINAL SLVM AND FREQUENT SUB-TREE SLVM HELPFUL HINTS

Structured Link Vector Model (SLVM) was proposed by Jianwu Yang [2], which forms basis of our work, was proposed for representing XML documents. It was extended from the conventional vector space model (VSM) by incorporating document structures (represented as term-by-element matrices), referencing links (extracted based on IDREF attributes), as well as element similarity (represented as an element similarity matrix). On the other hand, based on original SLVM, an extended VSM utilize the closed frequent sub-trees as structural units for content extraction from the XML document [1], which are called frequent sub-tree SLVM in this context.

A. XML Document Representations

SLVM represents a XML document \(doc_x\) using a document feature matrix \(\Delta_x \in \mathbb{R}^{n \times m}\), given as [2]

\[
\Delta_x = \left[ \begin{array}{c}
\Delta_{x(1)} \\
\Delta_{x(2)} \\
\vdots \\
\Delta_{x(m)}
\end{array} \right],
\]

where \(m\) is the number of distinct XML document elements, \(\Delta_{x(i)} \in \mathbb{R}^n\) is the TFIDF feature vector representing the \(i\)th XML element \((e_i)\), given as

\[
\Delta_{x(i)} = TF(w_j,doc_x,e_i) \cdot IDF(w_j),
\]

for all \(j=1\) to \(n\), and \(TF(w_j,doc_x,e_i)\) is the frequency of the term \(w_j\) in the element \(e_i\) of \(doc_x\).

In frequent sub-tree SLVM, each document is represented as a matrix based on SLVM, where the selected closed frequent sub-trees are regarded as structural unit [1]. In order to deal with exception that a document do not include any the selected closed frequent sub-trees, it adds the vector of the document based on VSM into the matrix as a column.

B. Similarity Measures

The SLVM-based document similarity between two XML documents \(doc_x\) and \(doc_y\) is defined as [2]

\[
Sim(doc_x,doc_y) = \frac{1}{m \times m} \sum_{j=1}^{m} \sum_{i=1}^{m} M(e_i,j) \cdot \Delta_x(i) \cdot \Delta_y(j),
\]

where \(\Delta_x(i)\) or \(\Delta_y(j)\) is the normalized document feature matrix of documents \(doc_x\) or \(doc_y\), and \(M\) is a matrix of dimension \(m \times m\) and named as the element similarity matrix. The matrix \(M\) captures both the similarity between a pair of document structural elements as well as the contribution of the pair to the overall document similarity. To obtain an optimal \(M\) for a specific type of XML data, SLVM-based document similarity learn the matrix using pair-wise similar training data (unsupervised learning) in an iterative manner [4].

C. Analysis of Feature Extraction

In the above, these models for document representation are perceived as the mode using statistical term measures. As a sort of ontology methods [5], XML document representations based on statistical term measures ignore recognition of lexical semantic contents. It causes the document representation to lose the mutual information [6] of term meanings which comes from synonyms in different samples. Our comment on statistical term measures and XML document representation can be clarified by analyzing a small XML corpus Example 1.

Example 1

```
<!--Doc A--> 
<article> 
<title>human treasures trees.</title> </article>

<!--Doc B--> 
<article> 
<title>maple</title> 
<body>Men prize maple.</body> </article>
```

Figure 1. The SLVM feature matrices for Example 1

In Example 1, the two simple XML documents are viewed as two document samples, and these two documents comprise the small corpus. Evidently, the meanings of Doc A and Doc B are extremely equivalent. Thus, the correlation and semantic similarity between these two documents are considerable. But, SLVM and frequent sub-tree SLVM can not display the document similarity between Doc A and Doc B. Obviously, because document representation matrices shown in Fig.1 make each \(\Delta_{A(i)j} \cdot \Delta_{B(j)j} = 0\), so the \(Sim(doc_A,doc_B) = 0\), using the Eq. (2). Then, on behalf of statistical term
measures, the document representations on Example 1 did not perform well for semantic similarity.

III. PROPOSED PROGRAM

A. Preliminary Conception and Theoretical Analysis

For document analysis, document representations which depend on statistical term measures shall lose mutual information of term meanings. Besides, in different documents, term meanings are relevant to specific synonyms which are involved by lexical semantic contents. Thus, our model resorts to WordNet [3], a lexical database for English, for extracting lexical semantics. Then, the method of document representation will construct a lexical-semantic SLVM of XML document in order to define feature matrix for classification.

In WordNet, a form is represented by a string of ASCII characters, and a sense is represented by the set of (one or more) synonyms that have that sense [3]. Synonymy (syn same, onyma name) is a symmetric relation between word forms [3]. Synonymy is WordNet’s basic relation, because WordNet uses sets of synonyms (synsets) to represent word senses. Videlicet, shown as Fig 1, one word refers to several sets of synonyms (synsets).

Based on the above definition, involved semantic-factors can characterize the lexical semantic contents of Example 1, which shall accomplish feature extraction of lexical semantic contents. For instance, in Fig. 2, the words human and man belong to different document samples in Example 1, and the common semantic-factor synsetp (homo) that simultaneously describes the meanings of human and man can gain mutual information [6] between term meanings. Moreover, our document representation is able to capture the lexical semantic mutual information between samples which lies with a number of synonyms in different documents.

According to the statistical theory of communications, our conception needs further analysis for theoretical proof. The analysis first introduces some of the basic formulae of information theory [6, 7], which are used in our theoretical development of samples mutual information. Now, let $x_i$ and $y_j$ be two distinct terms (events) from finite samples (event spaces) $X$ and $Y$. Then, let $X$ or $Y$ be random variable representing distinct lexical semantic contents in samples $X$ or $Y$, which occur with certain probabilities. In reference to above definitions, mutual information between $X$ and $Y$, represents the reduction of uncertainty about either $X$ or $Y$ when the other is known. The mutual information between samples, $I(X;Y)$, is specially defined to be [7]

$$I(X;Y) = \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}.$$

In the statistical methods of SLVM, probability $P(x_i)$ or $P(y_j)$ is estimated by counting the number of observations (frequency) of $x_i$ or $y_j$ in sample $X$ or $Y$, and normalizing by $N$, the size of the corpus. Joint probability, $P(x_i, y_j)$, is estimated by counting the number of times (related frequency) that term $x_i$ equals (is related to) $y_j$ in the respective samples of themselves, and normalizing by $N$.

Taking the Example 1, according to SLVM feature matrices (shown in Fig. 1), between any term $x_i$ in Doc A and any term $y_j$ in Doc B, there is not any counting of times that $x_i$ equals $y_j$. As a result, in Example 1, the statistical term measures indicate $P(x_i, y_j) = 0$ so the samples mutual information $I(X;Y) = 0$. Thus, the analysis verifies that the statistical methods of feature extraction lose mutual information of term meanings.

On the other hand, for feature extraction of lexical semantic contents, our method uses several particular semantic-factors to describe the meaning of one word or term. In different samples, words can be related to other words which are described by same synsets. Then, lexical
In our work, XML documents are represented using the lexical-semantic SLVM. In this model, each XML document is represented as a document feature matrix in the lexical-semantic structured link vector space. For organizing the lexical-semantic SLVM, the procedures are as follows. First, (1) a data structure of semantic-factor information is composed for feature extraction of lexical semantic contents. Secondly, (2) the EM modeling is used to disambiguate word stems. Furthermore, (3) this work constructs the feature space of lexical-semantic SLVM and builds the synset matrix in the space to characterize lexical semantic contents of XML. Lastly, (4) to characterize lexical semantic relations, it marks each vector in synset matrix with weights of 5 semantic relations between synsets [3]. Thus, the feature matrix in lexical-semantic SLVM is constructed via marking semantic relations on synset matrix.

B. Lexical-semantic SLVM

In our work, XML documents are represented using the lexical-semantic SLVM. In this model, each XML document is represented as a document feature matrix in the lexical-semantic structured link vector space. For organizing the lexical-semantic SLVM, the procedures are as follows. First, (1) a data structure of semantic-factor information is composed for feature extraction of lexical semantic contents. Secondly, (2) the EM modeling is used to disambiguate word stems. Furthermore, (3) this work constructs the feature space of lexical-semantic SLVM and builds the synset matrix in the space to characterize lexical semantic contents of XML. Lastly, (4) to characterize lexical semantic relations, it marks each vector in synset matrix with weights of 5 semantic relations between synsets [3]. Thus, the feature matrix in lexical-semantic SLVM is constructed via marking semantic relations on synset matrix.

In Example 1, joint probability $P(x_i, y_j)$ is estimated by calculating the frequency of common semantic-elements that relate to lexical semantic contents or relations of $x_i$ and $y_j$, and modulo $N$. For instance, shown in Fig. 3, the words human and man are described by the common semantic-factor synset$_{(homo)}$. Actually, according to document representation matrices shown in Fig. 4, $P$(human, man) = $F$(synset$_{(homo)}$) mod $N > 0$. In addition, joint probability of semantic relation $P$(tree, maple) = $F$(Hyponymy(tree, maple)) mod $N > 0$. As a result, $I(X; Y) > 0$, so mutual information from lexical semantic contents and relations between Doc A and Doc B is positive. Thus, the analysis proves that the semantic-factors and feature extraction of lexical semantic contents can provide the probability-weighted amount of information (PWI) [7] between XML documents on the lexical semantic level.

B. Lexical-semantic SLVM

In our work, XML documents are represented using the lexical-semantic SLVM. In this model, each XML document is represented as a document feature matrix in the lexical-semantic structured link vector space. For organizing the lexical-semantic SLVM, the procedures are as follows. First, (1) a data structure of semantic-factor information is composed for feature extraction of lexical semantic contents. Secondly, (2) the EM modeling is used to disambiguate word stems. Furthermore, (3) this work constructs the feature space of lexical-semantic SLVM and builds the synset matrix in the space to characterize lexical semantic contents of XML. Lastly, (4) to characterize lexical semantic relations, it marks each vector in synset matrix with weights of 5 semantic relations between synsets [3]. Thus, the feature matrix in lexical-semantic SLVM is constructed via marking semantic relations on synset matrix.

In Example 1, joint probability $P(x_i, y_j)$ is estimated by calculating the frequency of common semantic-elements that relate to lexical semantic contents or relations of $x_i$ and $y_j$, and modulo $N$. For instance, shown in Fig. 3, the words human and man are described by the common semantic-factor synset$_{(homo)}$. Actually, according to document representation matrices shown in Fig. 4, $P$(human, man) = $F$(synset$_{(homo)}$) mod $N > 0$. In addition, joint probability of semantic relation $P$(tree, maple) = $F$(Hyponymy(tree, maple)) mod $N > 0$. As a result, $I(X; Y) > 0$, so mutual information from lexical semantic contents and relations between Doc A and Doc B is positive. Thus, the analysis proves that the semantic-factors and feature extraction of lexical semantic contents can provide the probability-weighted amount of information (PWI) [7] between XML documents on the lexical semantic level.

B. Lexical-semantic SLVM

In our work, XML documents are represented using the lexical-semantic SLVM. In this model, each XML document is represented as a document feature matrix in the lexical-semantic structured link vector space. For organizing the lexical-semantic SLVM, the procedures are as follows. First, (1) a data structure of semantic-factor information is composed for feature extraction of lexical semantic contents. Secondly, (2) the EM modeling is used to disambiguate word stems. Furthermore, (3) this work constructs the feature space of lexical-semantic SLVM and builds the synset matrix in the space to characterize lexical semantic contents of XML. Lastly, (4) to characterize lexical semantic relations, it marks each vector in synset matrix with weights of 5 semantic relations between synsets [3]. Thus, the feature matrix in lexical-semantic SLVM is constructed via marking semantic relations on synset matrix.
word refers to more than 1 word stems, the list of Semantic Member will expend the node of original word to register all word stems.

Meanwhile, the list of Semantic Members Frequency is shown in Fig. 5(b). It records the frequency of each original word one by one in their order of Semantic Member.

<table>
<thead>
<tr>
<th>Item</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synset ID</td>
<td>Identification of synonym set</td>
</tr>
<tr>
<td>Set of Synonym</td>
<td>Synonymy is WordNet’s basic relation. WordNet uses sets of synonyms (synsets) to represent word senses.[3]</td>
</tr>
<tr>
<td>Weight (Frequency)</td>
<td>Frequency of semantic-factor in an element (sum of Semantic Members Frequency)</td>
</tr>
<tr>
<td>Sample ID</td>
<td>Identification of semi-structured document sample</td>
</tr>
<tr>
<td>Element ID</td>
<td>Identification of structural element or unit in semi-structured document</td>
</tr>
<tr>
<td>Semantic Member</td>
<td>A linked list (shown in Fig. 3) which carries all Original Words of Terms referring to the semantic-factor and their Word Stem(s)</td>
</tr>
<tr>
<td>Semantic Members Frequency</td>
<td>A linked list (shown in Fig. 4) which carries frequency of each Original Words of Terms (that refer to the semantic-factor) one by one</td>
</tr>
</tbody>
</table>

Table I

DATA STRUCTURE OF SEMANTIC-FACTOR INFORMATION

(2) On the basis of data structure of semantic-factor information, Semantic Member needs to disambiguate word stems of original word. In case of an original word referring to more than 1 word stem in base form, semantic-factors must ensure that one original word refers to only 1 word stem. Then, in order to select only 1 word stem for an original word (shown in Fig. 6), we employ the Maximum Entropy Model [10].

ME modeling provides a framework for integrating information for classifying from many heterogeneous information sources [11]. In our model, we put an assumption that diversity [12] of Semantic Member implies the significance of the semantic-factor and the rationality of existing Semantic Members in a document.

Assume a set of original words $X$ and a set of its word stems $C$. The function $cl(x) : X \times C \rightarrow \mathbb{R}$ chooses the word stem $c$ with the highest conditional probability, which makes sure original word $x$ only refers to:

$$cl(x) = \arg \max_c p(c \mid x).$$

Each feature [11] of original word is calculated by a function that is associated to a specific word stem $c$, and it takes the form of Eq. (5), where $S_j^i$ is the number of Semantic Member in semantic-factor $i$, $P_j$ is the proportion of the Frequency of original word $j$ to Weight in semantic-factor $i$, and the $\hat{a}$ indicates Semantic Member diversity of semantic-factor $i$ in a document, in the form of Shannon-Wiener index [12].

The conditional probability $p(c \mid x)$ is defined by Eq. (6). The parameter of the semantic-factor $i$ [11], $\alpha_i$, is the Frequency of original word $x$ in semantic-factor $i$. $K$ is the number of semantic-factors which word stem $c$ refers to, and $Z(x)$ is a value to ensure that the sum of all conditional probabilities for this context is equal to 1.

$$f_i(x,c) = \frac{S_j^i}{\sum_j^P} P_j \log_2 P_j$$ if $x$ refers to $c$ and $c$ refers to semantic-factor $i$.  \quad (5)$$

$$p(c \mid x) = \frac{1}{\sum_{i=1}^K \alpha_i^f(x,c)} \quad (6)$$

Figure 6: 1:1 reference of original word

Above equations aim at finding the highest conditional probability $p(c \mid x)$, and using the function $cl(x)$ to ensure that original word $x$ refers to only 1 word stem (like Fig. 6). After semantic-factors characterizing lexical semantic contents of XML document preliminarily, the specified ME modeling is applied to implement disambiguation of word stems. Necessarily, the relevant items in the data structure of semantic-factor information shall be modified, such as the Semantic Member, the Frequency of original word, and the Weight. Furthermore, some relevant semantic-factors shall be eliminated.

(3) As for XML document dataset, all referred semantic-factors are fixed by disambiguation of word stems. Then, in feature space of lexical-semantic SLVM, a XML document $doc_c$ is preliminary represented using a synset matrix $\mathbf{A}_x \hat{1} R^m m$, defined as

$$\mathbf{A}_x = \left[ \frac{\mathbf{A}_x(1,1)}{\mathbf{A}_x(1,2)} \ldots \mathbf{A}_x(m) \right] , \quad (7)$$

$$\mathbf{A}_x(i) = \left[ \mathbf{A}_x(i,1) \mathbf{A}_x(i,2) \ldots \mathbf{A}_x(i,n) \right] , \quad (8)$$

where $m$ is the number of distinct XML document elements and units, $\mathbf{A}_x(i)$ is the synset vector representing the $i^{th}$ XML element or unit, given as $\mathbf{A}_x(i,j) = FS(s_j,doc_c,e_j)IDF(s_j)$ for all $j=1$ to $n$, and $FS(s_j,doc_c,e_j)$ is the frequency of the semantic-
factor \( s_j \) in \( e_j \), in which \( e_j \) is the \( i^{th} \) XML element or unit of \( \text{doc}_c \).

(4) In feature space of lexical-semantic SLVM, to characterize lexical semantic relations, our method marks Antonymy, Hyponymy, Meronymy, Troponomy and Entailment on each dimension of the synset matrix. The processing is formalized as

\[
\Delta \Delta_x(i, q) = \sum_{p=1}^{\lambda} R(p, q) \Delta \Delta_x(i, p), \quad (9)
\]

where \( p \) and \( q \) are 1 to \( n \), \( n \) is dimensional number of the synset vector of the \( i^{th} \) XML element, and \( \Delta \Delta_x(i, p) \) is value of the \( p^{th} \) synset vector element of the \( i^{th} \) XML element. \( \Delta \Delta_x(i, q) \) is semantic relation increment to the \( q^{th} \) dimensional value of the synset vector, and function \( R(p, q) \) denotes semantic relation coefficient for \( \Delta \Delta_x(i, q) \). Specifically, when the synset of \( q^{th} \) dimension is related to synset of \( p^{th} \) dimension via semantic relation such as Antonymy, Hyponymy, Meronymy, Troponomy or Entailment, the \( R(p, q) \) assignment is shown in equation (7). The assignments of \( R(p, q) \) reflect the semantic relations which are organized into synsets by WordNet.

As for XML documents, all synset dimensions carry the corresponding semantic feature values. Then, lexical-semantic SLVM represents a XML document using a feature matrix \( \Delta_x \) defined as

\[
\Delta_x = \Delta x(1), \Delta x(2), \ldots, \Delta x(m) \quad (11)
\]

where \( m \) is the number of distinct XML elements and units. \( \Delta_x(i, j) R_m \) is the lexical-semantic vector representing the \( i^{th} \) XML element, given as

\[
\Delta x(i, j) = \left( \Delta x(i, 1), \Delta x(i, 2), \ldots, \Delta x(i, n) \right), \quad (12)
\]

\[
\Delta x(i, j) = \Delta x(i, j) + \Delta \Delta_x(i, j) \quad (13)
\]

where \( n \) is the number of identical Synset ID of all semantic-factors in XML dataset.

IV. EXPERIMENT AND RESULT

A. The Experiment

In our work, experiments use three sorts of matrices to represent document sample: 1) document feature matrix based on original SLVM, 2) document feature matrix based on frequent sub-tree SLVM, 3) the lexical-semantic matrix in the lexical-semantic feature space based on lexical-semantic SLVM.

In the XML document classification, the dataset consisting of 20 categories is composed of XML documents of the Wikipedia XML, the training set is composed of 5000 XML documents, and the test set is composed of 3000 XML documents.

In the experiments, to tackle unbalanced text dataset, we select an optimized KNN classification, the NWKNN (Neighbor-Weighted K-Nearest Neighbor) algorithm defined to be Eq. (14) [13]. As for NWKNN, each XML document \( d \) is considered to be a feature matrix based on original SLVM, frequent sub-tree SLVM, or lexical-semantic SLVM.

\[
\text{score}(\text{doc}, c) = \text{Weight} \cdot \text{Sim}(\text{doc}, \text{doc}_c) \delta(\text{doc}_j, c), (14)
\]

In the process of Eq. (8), this algorithm uses the SLVM-based document similarity between feature matrices of \( \text{doc} \) and \( \text{doc}_c \) to calculate the \( \text{Sim}(\text{doc}, \text{doc}_c) \). Besides, according to experience of NWKNN algorithm [13], the parameter of \( \text{Weight} \), Exponent [13], is equal to 3.5.

B. The Result

To evaluate the classification systems, we use the F1 measure [14]. Then, we can observe the effect of different kinds of data on a classification system [14]. For ease of comparison, we summarize the F1 scores over the different categories using the macro-averages and micro-averages of F1 score, and display mean average precision.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Macro-F1</th>
<th>Micro-F1</th>
<th>Mean Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original SLVM [1]</td>
<td>0.241853</td>
<td>0.295418</td>
<td>0.486702</td>
</tr>
<tr>
<td>Frequent sub-tree SLVM [1]</td>
<td>0.479748</td>
<td>0.518635</td>
<td>0.701861</td>
</tr>
<tr>
<td>Lexical-semantic SLVM</td>
<td>0.492090</td>
<td>0.521977</td>
<td>0.727203</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In the work, a data structure of semantic-factor information is constructed to record relevant information of each semantic-factor in element of XML document. It can characterize lexical semantic contents and be adapted for disambiguation of word stems. Furthermore, in the the feature space, lexical semantic relations are marked on the synset matrix. Using the NWKNN algorithm, lexical-semantic SLVM achieve better performance of
classification than document feature matrix built by original SLVM and frequent sub-tree SLVM which stand for the typical statistical method of feature extraction.

Our future research includes using more current algorithms based on the lexical-semantic matrix for XML document analysis, and developing a method for analyzing WSDL document on the basis of semantic-factor.

ACKNOWLEDGMENT

This work was supported by National High Technology Research and Development Program (863 Program) of China (No. 2012AA011205), National Natural Science Foundation of China (No. 61272150, 61472450, 61379109, 61103034), Research Fund for the Doctoral Program of Higher Education of China (No. 20120162110077), and Excellent Youth Foundation of Hunan Scientific Committee, China (No. 11JJ1012).

REFERENCES


Jun Long was born in Anhui Province, China in 1972. He is a professor and Ph. D. supervisor of Central South University of Information Science and Engineering. His research interests include distributed computation, software architecture and trusted computing et al.

Luda Wang was born in Jinan, Shandong Province, China in 1981. He received the M.S. degree in Computer Application Technology from Hunan University in 2009, and he is currently a Ph. D. candidate of Central South University of Computer Science and Technology.

He has been a faculty member of Computer Science at Xiangnan University, Chenzhou, China, since 2005, where he is currently a lecturer. His research interests include AI, data analysis and information retrieval.

Zude Li is an Assistant Professor at Central South University (China). He obtained his Ph.D. degree in 2010 from The University of Western Ontario (Canada). His research interests are in the fields of software architecture, evolution and quality. He collaborates with IBM Canada and is appointed as a Software Engineering expert in SKANE SOFT.
Novel Metrics for Bug Triage

V. Akila
Dept. of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, India
Email: akila@pec.edu

Dr. G. Zayaraz
Professor, Dept. of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, India
Email: gzayaraz@pec.edu

Abstract—Bug triaging is a vital part of issue management systems. Bug triaging deals with assigning a developer the task of an incoming bug. This activity is error prone and time consuming if done manually. There is a need for automated support to accelerate this process. The current automated bug triaging systems exploit the text contents of the bug and the tossing relations among the developers. The automated bug triaging systems estimate the optimal path between the first assignee of the bug and the bug resolver using the tossing relations. The metrics used for assessing the efficiency of bug triaging systems that are based on tossing relations is Mean number of Steps To Resolve (MSTR). This metric quantifies the number of steps reduced by the predicted path compared to the original path. It does not capture how far the retrieved path is in alignment with the actual path. MSTR does reveal the information regarding the extent to which the order of the developers in the retrieved path is in line with that of the original path. In addition, there are no indicators for measuring the strength of the retrieved path. In this paper, we propose two metrics (i) Path Similarity Metric which quantifies path alignment based on pair wise path alignment and (ii) Path Alignment Indicator that measures the effectiveness of the retrieved path based on degree centrality. The effectiveness of the two proposed metrics is validated using bug reports extracted from the Eclipse project.

Index Terms—Open Source Software, Bug Triaging, Mean Steps To Resolve

I. INTRODUCTION

Software Maintenance contributes to about 50% of the costs incurred in a software project [1]. Issue management is an important cog in Software maintenance. Issue management pertains to solving of bugs/defects/issues that arise after the deployment of the software system. This solving of bugs comprises of (i) assessing the issue, (ii) assigning the issue to the correct developer for solution and (iii) verifying the solution. Software Maintenance of Open Source systems brings has its own set of peculiarities. Assigning a developer to an issue in Open Source systems parlance is Bug Triaging. Here the cost incurred does not literally translate to currency. It is more about effort and time incurred in a project. The triaging of a bug is a non trivial task. This task when done manually is error prone and time consuming. Automated support is essential to perform triaging. The several methods that provide automated support bank on the textual content of the bug report and the previous tossing relations that exist among the developers. Usually machine learning methods are used to exploit the textual content. The first assignee for an incoming bug may always not be the one to solve the bug. The bugs get transferred among developers due to various reasons before getting solved. The reasons may be because of wrong assignment, inability to solve the bug etc. The fact is all bug tosses may not be detrimental [2]. Some tosses are necessary to fix the several fields like component, severity in the bug report. The tossing relations are captured in a Bug Toss Graph. The Bug Toss Graphs used in the literature are based on Markov model. The transition probabilities are calculated based on the number of tosses on that link. Shortest path algorithms are employed to find the path from the assignee to the developer. The set of developers are identified such that the transition probabilities are maximized. This leads to a set of developers who have collaborated frequently in earlier assignments. The task of bug triaging is not only to identify a correct developer, but also to retrieve a set of developers who may collaborate on an incoming bug.

At present, the performance of the bug triaging systems that are based on Bug Toss Graphs -are evaluated by the metric, MSTR. This metric quantifies how many number of hops in the retrieved path have been reduced by the bug triaging system. - The comparison of number of hops alone will not indicate the effectiveness of the bug triaging systems. There is need of some indicator which informs about how far the retrieved path is in alignment with the actual path. The information will show the extent to which the developers are in the retrieved path in the same order and position as that of the original path. Now a days many techniques have been borrowed from other diverse fields. To this end, we propose that pairwise path alignment techniques available in bio- informatics field will serve as the appropriate indicator for path alignment in bug triaging systems. This paper presents a path alignment metric based on pair wise path alignment and a path alignment indicator that serves as a goodness measure for the retrieved path based on degree centrality.
II. RELATED WORK

The related work is examined along two directions. The first one throws light on the metrics used in Resolution Networks. The later one investigates the various pair wise path alignment techniques and their applications in different fields.

A. Resolution Network

P. Battacharya et al., [1][5] has proposed a bug triage system which is based on multi-featured bug toss graphs using incremental learning framework. The Bug Toss Graph here is based on goal-oriented path model which captures only the tossing relations with the final resolver. The metrics used to evaluate the system are prediction accuracy and reduction in toss length. Gaeul Jeong et al., [3] introduced the idea of toss graphs for Open Source systems. The Bug Toss Graphs are modeled after goal-oriented path model. The metric used for evaluation is prediction accuracy. Liguo Chen et al., [4] has advocated for a bug triaging system which combines the support vector machine technique with Bug Tossing Graphs. Weight based breadth first search algorithm is used to find the path between the assignee and developer. The performance of the system is compared with the existing work based on mean length of tossing paths. Sung hun Kim et al., [6] introduces a crash graph method which captures the crashes reported present in a bucket. The crash graph is used to provide a high level view of the crashes reported as bug. The evaluation parameter used is graph similarity.

Qihong Shao et al., [7] models the passing of tickets in an enterprise network as a ticket resolution graph based on Markov Model. The variable order multiple active search algorithm is used to determine the shortest path to a resolver. The performance metric used is Mean Steps To Resolve. Gengxin Miao et al., [8][9] presents a unified generative model, the optimized network model to capture the life cycle of a ticket. A probabilistic algorithm is used to make the ticket routing recommendations. The parameter used for performance evaluation is Mean number of Steps To Resolve. Gengxin Miao et al., [10] has generated a collaborative network based on software bugs and resolving consumer problems. A routing model that maps the task driven information routing has been developed. The evaluation measure proposed is Mean number of Steps To Resolve. Tao Zhang et al., [11] proposes a hybrid bug triage algorithm using a probability model and smoothed unigram model. The probability model also uses bug reopened count to fix the probability of a candidate developer. The experience model uses another new factor called as activity factor.

The evaluation is done using mean reciprocal rank and F measure. Shuo Chang et al., [12] focuses on routing a question to set of potential collaborators who may collaborate to answer a question in community question answer system. They present that when a question warrants more number of comments then the lasting value of the thread is more. A model based on compatibility, expertise and availability is presented. Baichuan Li et al., [13] proposes a framework for question routing in CQA. The framework passes questions to users who are likely to answer questions quickly. The framework consists of four parts: (i) performance profiling, (ii) expertise estimation, (iii) availability estimation and (iv) answerer ranking. The performance metric used is mean reciprocal ranking.

B. Sequence Alignment

Comparison of protein structures is an important part of bioinformatics.[14] There are two major types of aligning protein structures: (i) pair wise alignment and (ii) multiple sequence alignment. In pair wise sequence alignment, this is further classified as local and global alignment.[14][15]. In global alignment, the correspondence of the entire sequence is evaluated.

C. Protein Matching

Kevin W DeRonneet al., [16] investigates the process of Pareto optimality for pair wise alignment. A dynamic programming based heuristic method by combining different profile scoring function is presented. The combination of different scoring function is treated as a multi objective optimization problem. Jayendra Gnanaskandan Venkateswaran et al., [17] has developed a triplet based iterative alignment algorithm for computing the number of transformations that will maximize the number of aligned protein sequences. Sudha Sadasivam et al., [18] proposes an efficient method to sequence alignment. The algorithm exploits the computational parallelism of hadoop grids to improve accuracy and speed.

D. Other Applications


Adesina Simon Sodiya et al., [24] uses a Cross – semi global algorithm to improve the efficiency of Masquerading attack. Nan Li et al., [25] presents a new algorithm that is composed of a local alignment algorithm-GASBSLA (Generation of Attack Signatures Based on SequenceLocal Alignment) and a multi-sequence alignment algorithm-TGMSA (Tri-stage Gradual Multi-Sequence Alignment) for generation of attack signatures. Cyril Labbé [26] et al., presents a method to calculate the intertextual distance to detect duplicates in scientific publications. Hyoungshick Kim et al., [27] proposes a method for detecting code clones, software plagiarism, code theft, and polymorphic malware using sequence alignment algorithms.
Birthmark sequences extracted from malicious programs with a hybrid approach based on "non-consecutive insertion" and "highest frequency deletion" were proved to be effective.

E. Inference from the Survey

There is a significant amount of commonality in techniques for expert identification in ticket resolution in enterprise systems, bug triaging in Open Source systems and question routing in community question answering systems. The problem that is to be solved in all of the three domains is identification of appropriate expert. The underlying graph has been modeled based on Markov Models. Routing algorithms are employed to identify the optimal route from the source resolver. Routing algorithms basically try to do information based routing. The semi-automated systems that assist in identification of experts are validated against the number of hops reduced by the semi-automated system. The target of the automated system is to identify the set of optimal resolvers who can collaborate in partial ordering to solve the problem. While the metric - Mean number of Steps To Resolve (MSTR) encodes the performance of the system, it does not indicate anything about the extent to which the predicted route is in sync with the actual path in the test set. Route or path similarity in terms of ordering and presence or absence of developers is needed to be evaluated for the retrieved set of developers. In order to improve the automated bug triaging system, the research contribution presented in the paper is

- Path Similarity Metric based on pair wise sequence alignment.
- Path Alignment Indicator based on degree centrality

III. RESEARCH CONTRIBUTION

In Open Source software development, the developers are loosely coupled, the contribution from each developer is strictly voluntary. There is no centralized control in the development of the software system. Added to this, there is the complication of a developer becoming inactive with time and the expertise area of a developer evolving. When a software developed using Open Source development model develops bugs, these bugs are reported using issue tracking tools. In such a scenario it is the job of the triager to assign the correct developer for the solving the bug. The problem described is akin to the problem of expert retrieval.

The bug tossing relationships are retrieved from the activity data in the issue tracking system. The Bug Toss Graph is evolved from the bug tossing relations. The task at hand is to route a bug to the correct developer in the best route. Best route here means that passing through developers who can contribute positively towards the resolution of the bug. Thus, any automated system in the realm of bug triaging provides information based routing. The overview of the proposed system is given in Figure 1.

The bug reports from the bug repository are partitioned into training data set and test data set. Using the training data set, the automated bug triage system retrieves the optimal path for the incoming bug reports. The retrieved path between the assignee and the resolver is compared against the original path extracted from the test data set by pair wise path similarity. This module compares the two paths based on the number of edit operations needed to transform one path to another. This number of edit operations is quantified as Path Similarity Metric between the retrieved path and the original path. An exploratory analysis of Levenshtein Similarity, Smith Waterman Similarity and Jaro Wrinkler similarity was performed. The edit operations are used to identify the nature of developers who needs to be inserted, deleted or substituted. Combining this information with the degree centrality of each developer, the Path Alignment Indicator is derived. The formal problem definition is given in the next section.

![Flow Diagram of Path Alignment Indicator](image)

Figure 1. Flow Diagram of Path Alignment Indicator

A. Problem Formulation

Markov models are useful in capturing the transfer dependencies that prevails in the bug tossing relations. This model is used to predict future bug tossing relations. Long sequences of bug tossing has in them a few incorrect transfers. Any learning system must capture the predominant tosses. The bug tossing relations can be captured as a graph \( G = (V,E) \). Here \( V = \{v_1,v_2,\ldots,v_n\} \) is the set of states which represents the developer. \( E = \{e_1,e_2,\ldots,e_n\} \) represents the set of edges where each edge represents the previous bug tossing relation among the developer. The transition probabilities among the developers capture the local decisions made by the developer. The transition probability of an edge or link from \( v_i \) to \( v_k \) is \( P(v_i|v_k) \).

\[
P(v_i|v_k) = \begin{cases} \frac{N(v_i,v_k)}{N(v_k)} & \text{if } N(v_i) > 0 \\ 0 & \text{otherwise} \end{cases} \quad [7]
\]
Here, \( N(v_i,v_k) \) is the number of transfers from \( v_i \) to \( v_k \). \( N(v) \) is the total number of transfers from \( v_i \). An automated Bug Triage System uses this model with a shortest path algorithm to find the shortest path between the assignee and the developer. This retrieved path needs to be validated in terms of number of steps reduced as well as similarity with that of the actual path. The Path Alignment Indicator quantifies the alignment between the retrieved path and the actual path.

Definition 1: ‘V’ is the set of nodes of developers in the graph. A retrieved path ‘R’ is an array of nodes from ‘V’ that are contiguous and occupy a unique position.

Definition 2: Node alignment is based on the pairs of nodes. It is based on
- Match – if nodes matches
- Mismatch – if a insertion, deletion or substitution is needed

B. Path Alignment Indicator

Based on the given definitions, the Path Alignment Indicator (PAI) is given as

\[
PAI = f(I,D,S)
\]

I - Insertion
D - Deletion
S - Substitution

The cost of the insertions, deletions and substitutions is done based on the social network analysis. In particular, the node centrality measure is used. The degree of a particular node is the total number of incoming edges and outgoing edges from a particular node. This measure reveals how well connected a node is. In bug triaging parlance, this reveals about the number of bugs that has been tossed to a developer and how many bugs has been tossed by the developer. Based on the degree centrality of the node, the cost of insertion, deletion and substitution is calculated as given below.

\[
I = \sum_{i=0}^{n} (deg_{in} - Wd(N_i))
\]

\[
D = \sum_{i=0}^{m} (Wd(N_i) - deg_{out})
\]

\[
S = \sum_{i=0}^{k} ((deg_{in} - Wd(N_i)) + (Wd(ND_i) - deg_{out}))
\]

\( Wd(N) \) is the sum of total weight of all the edges to a node \( N \)

\( n \) is the total number of insertions.
\( m \) is the total number of deletions.
\( k \) is the total number of substitutions.

\( deg_{in} \) - Average weighted degree of the Bug Toss Graph

The procedure for Path Alignment Indicator (PAI) is depicted in Figure 2. The input is the developer from the retrieved path who is misaligned from that of the original path and the type of operation. This input is received from the Levenshtien path similarity module. If the degree of the developer to be inserted is greater than the threshold degree then it is assumed that a good developer has been missed by the automated bug triage system. If the degree of the developer to be deleted from the retrieved path is less than the threshold degree then a below average developer has been included by the bug triage system. If the operation is substitution then the degrees of the developer to be substituted are compared.
The result of the exploratory analysis using Levenshtein similarity, Smith Waterman similarity and Jaro Wrinkler similarity for the 30% test data set is depicted in Figure 4.

The relation between the precision and recall values for the 20% data set experiment is shown in Figure 5. The correlation coefficient $R^2$ indicates the positive correlation between the precision and recall values.

The results of the exploratory analysis using Levenshtein similarity, Smith Waterman similarity and Jaro Wrinkler similarity for the 20% test data set is depicted in Figure 6.

The correlation coefficients of the results obtained are presented in Table 1. From the results, it is inferred that Levenshtein similarity offers the best correlation coefficient value with respect to precision. Thus, Levenshtein similarity is employed to assess the alignment between the retrieved path and the original path.

The relation between the number of hops and the PAI is illustrated in Figure 7. From the figure it can be understood that the number of hops increases the alignment strength.

V. CONCLUSION

Bug triaging is an important activity of software maintenance in Open Source software systems. Automated support during bug triage improves the overall bug management by retrieving the optimal path from the first assignee to the final resolver. These automated bug triage systems are evaluated with Mean Steps To Resolve (MSTR) parameter. This parameter encodes the number of steps reduced while solving the bug. Bug resolution can be viewed as a collaborative activity among multiple members of the Open Source software community. While reducing the number of steps to resolution is vital, it is also necessary to measure the alignment between the retrieved path and the original path. Two metrics – Path
Similarity Metric and Path Alignment Metric are proposed in this paper. The Path Similarity Metric uses the pairwise path alignment method borrowed from bio-informatics. Path Alignment Indicator quantifies the strength of the retrieved path using degree.

REFERENCES
[25] Li, Nan, Xia, Chunhe, Yang, Yi and Wang, Haquan, ”An Algorithm for Generation of Attack Signatures Based on Sequences Alignment”, JSEA 1, no. 1 : 76-82, 2008

V. Akila is working as Assistant Professor at Department of Computer Science and Engineering at Pondicherry Engineering College from 2007. Her areas of interest includes Mining Software Repositories, Complex Event Processing and Swarm Intelligence.

G. Zayaraz is working as Professor at Department of Computer Science and Engineering at Pondicherry Engineering College. His areas of interest includes Software Architecture, Mining Software Repositories, and Network Security.
Implementing Six Sigma to Reduce Online Petitions on the E-petition System in Taiwan Power Company

Pei–Shih Chen
General Education Center, Central Police University, Taoyuan County, 33304, Taiwan, R.O.C

Chang–Hsien Hsu
Department of Business Administration, Asia University, Taiwan, R.O.C
pci@asia.edu.tw

Shih–Yun Hsu
Department of Leisure and Recreation Management, Asia University, Taichung, 41354, Taiwan, R.O.C

Chun–Ming Yang
Department of Management Sciences, Tamkang University, Taiwan, R.O.C

Abstract—This study presents a case study demonstrating the application of Six Sigma (DMAIC) to successfully reduce the number of online petitions on the e-petition system in Taiwan Power Company (TPC). First, in the define (D) stage, a total of 1,292 petitions for the year 2010 in TPC are received and divided into 9 categories. In the measure (M) stage, the category that received the most petition letters is “line removal”, which accounts for 29.3% of the total petitions. The analyze (A) stage indicates that TPC did not provide relevant application procedures for “line removal” service. In the improve (I) stage, this study presents three approaches to speed up the “line removal” application process. Finally, a well-organized standard of procedure and workflow control charts has been developed in order to monitor the overall application and construction processes in the control (C) stage. After implementing Six Sigma (DMAIC), the number of petitions for “line removal” decreased by 66% from 379 in 2010 to 129 in 2011.

Index Terms—Six Sigma; DMAIC; Power Company; Online Petition; E-Petition System

I. INTRODUCTION

In the 21st century, the product itself or the ability to sell is no longer the key to a successful business, but rather meeting customer needs, decreasing customer complaints, and increasing customer satisfaction. According to Singh and Pandya [1], when customers experience dissatisfaction with a product or service, they will take the initiative in seeking compensation. Several studies have also pointed out that a business might be rejected by the customer due to its low service quality [2]. Saraei and Amini [3] present that a business with high service quality and dealing tactfully with customer complaints will increase its customer satisfaction while retaining customer loyalty.

In general, handling customer complaints properly is an act to reclaim customer satisfaction when a service provider makes mistakes. A mishandled complaint may often result in the loss of customers. Corporate image can also be easily damaged by dissatisfied customers or their negative word-of-month.

In recent years, due to the advances in information technology and emphasis on customer relationship management, more and more application systems are being established on the Internet for providing an easy and convenient way to complete transactions, services and applications online between customers and businesses [4]. The E-service mailbox, as a suite of electronic services (e-services) application, has been adopted in many organizations, from large organizations to smaller companies, even in Taiwan. The e-service mailbox provides several benefits in terms of being cost effective, providing fast services, and return on investments [5]. Although, the e-service mailbox is considered as a well-known way of handling customer complaints, the businesses should seek to decrease the number of customer petitions as well as to increase the level of customer satisfaction and market competition. Therefore, businesses need a significant and systematic approach to improve business service quality and handle customer complaints.

Taiwan Power Company (TPC), founded in 1946, is a government-run company in Taiwan. TPC is a vertically integrated power company; it provides services which include power generation and power supply to residential and commercial users. Driven by advances in information technology, TPC offers more diversified services by having set up an online petition system with e-service mailbox (service@taipower.com.tw) in 1996. It not only provides information about TPC to the general public, but
also acts as a platform for customers to express their opinions and comments regarding TPC’s services; the more petitions TPC receives, the more dissatisfied customers TPC has.

The Six Sigma DMAIC (Define, Measure, Analyze, Improve, and Control) method is often described as an approach for solving problems and improving quality. The Six Sigma has been widely embraced by many Fortune 500 consumer products manufacturers with multiple business groups in the U.S. [6]. Many case studies also show that the Six Sigma approach is very popular in solving problems in the real world [7]. van Iwaarden et al. [8] present the Six Sigma process as a value-added creation through quality improvement. Cheng [9] concludes that Six Sigma could be inferred from both definitions and could be seen as a philosophy or strategy used to improve the efficiency and effectiveness of business processes to meet or exceed customers’ needs and expectations. Easton and Rosenzweig [10] propose that the Six Sigma approach is a useful tool to discover a number of key factors associated with project team success and failure in practice. Shafer and Moellerm [11], by investigating a sample of 84 Six Sigma firms, indicate that adopting Six Sigma may positively impact organizational corporate performance. Therefore, the aim of this study is to apply the Six Sigma DMAIC approach to improve service quality and increase the competitiveness of TPC.

The five stages of the implementation are described as follows. First, in the “define” stage, petition letters in the year of 2010 are collected from the online service mailbox. Of all the petition letters, TPC receives the most in the “line removal” category. In the “measure” stage, a pie chart is used to map the proportions of different petition categories. The petition category of “line removal” accounts for the largest proportion. In the “analyze” stage, a cause and effect diagram is used to analyze the causes of the core problem. Two supervisors, who have worked in TPC for over ten years, and two scholars brainstorm the causes of the petitions in the “line removal” category regarding the nature of petitioners, persons in charge and TPC’s system. In the “improve” stage, potential solutions are developed, evaluated and implemented. A two-way communication platform is recommended to provide better service. Finally, in the “control” stage, a knowledge-based system and in-service training are recommended in order to monitor the overall operation through a line chart, to illustrate the before-and-after differences. Through the steps of DMAIC, the number of petition letters in “line removal” category decreases and customer satisfaction increases.

The remainder of the paper is organized as follows. The next section presents a literature review of the development of the Six Sigma approach (DMAIC). A case study is described in Section 3. Section 4 describes the DMAIC framework used in this study. In the final section the conclusions are presented.

II. SIX SIGMA APPROACH (DMAIC)

The Six Sigma approach (Define-Measure-Analyze-Improve-Control, DMAIC), a business management strategy, originally was developed in 1986 by Motorola. The so-called Six Sigma is a quality measure for the technical point of view for each process, product, or service in terms of the probability of occurrences of defects per million being less than 3.4 times (i.e. 3.4 DPMO). Banuelas et al. [12] point out that if a business wants to implement the Six Sigma DMAIC approach efficiently, support from higher authorities which includes staffing and funding are needed. The Six Sigma approach works best in solving problems in the manufacturing sectors, particularly in reducing costs, improving yield and increasing production, in order to emphasize catering to customer needs. Canato et al. [13] indicate that the Six Sigma procedure is representative of continuity forces for manufacturers. Pillai et al. [14] indicate that Six Sigma is used in many manufacturing industries to achieve dramatic improvements in their operations, maintenance, engineering and business processes.

Currently, with the positive effects of Six Sigma on the manufacturing sectors, important issues of the non-manufacturing sector have been gradually noted. Aksoy and Dincmen [15] show increased quality and profitability through Six Sigma on process improvements concerning the issues of product value-added, low-cost products and services, by using statistical tools within a structured roadmap. Jou et al. [16] use Six Sigma to evaluate and improve the performance of new product development procedures along with performance matrix, factor analysis and theory of constraints. Moosa and Sajid [17] posit Six Sigma as a useful tool at management level to deal with complex organizational problems that need simultaneous extensive analysis of data, confirmation of results and validation of long-term actual benefits. Gutierrez et al. [7] propose the Six Sigma approach to measure teamwork and process management, as well as absorptive capacity in the organization. Kumar et al. [18] find that the Six Sigma has a significant impact on the bottom-line and working culture of an organization in their attempt to manage changes. Prisani et al. [19] put forward the Six Sigma methodology to focus on reducing and ultimately eliminating defects in business processes. Chang et al. [20] utilize the Six Sigma managerial improvement method to obtain a strategic solution for applying information systems integration as a way to improve production planning planning procedure, upgrade customer satisfaction and increase the performance of the production planning procedure. Hsu et al. [21] point out that using the Six Sigma control method with the DMAIC framework could curtail the power supply over time in order to enhance the power supply efficiency in Taiwan Power Company (TPC). In accordance with the above reasons, the Six Sigma approach is a very practical management method.
III. CASE STUDY

Taiwan Power Company (TPC), a vertically integrated power company, has provided services of power generation and power supply to residential and commercial users since May 1, 1946. TPC is the only company in the Republic of China that offers a power supply to residential and commercial users. All private-owned power plants generate their electrical power and then sell it to TPC. At the end of the year 2009, TPC had a total capacity of 4,025 megawatts (MW), including 3,206 MW stored in TPC’s own facility and 819 MW stored in private-owned power plants. Electricity is generated mainly by hydro power, thermal power, nuclear power and other renewable energy. As for the electrical power grid, there are 579 substations at all levels with 340,006 kilometers of transmission and distribution grid connections, which provides electrical power for 23 million residents in Taiwan, Penghu, Kinmen and Matsu [22].

In general, if customers have any questions about TPC’s services, they can file a petition for requesting a response. There are three ways to file a petition: (1) written petition through mail, fax, or public opinion reports; (2) oral petition through in-person visit, or phone, and (3) online petition system through service mailbox emails. Recently, e-mail is considered the fastest and the most convenient form of petition. Hence, in order to fully understand its customer needs and to increase service quality, TPC set up an online petition system with e-service mailbox (service@taipower.com.tw) on July 3, 1996. If customers have any kinds of questions regarding TPC’s services, they can email their requests to this mailbox [23].

Chen et al. [24] confirm that there is a positive correlation between the level of convenience and the number of petitions. Owing to the convenience and low-cost of filing an online petition, people tend to overuse this option, which causes a greater workload and higher costs for processing. When a customer complains, it means he/she is not satisfied with the product or the service. When the level of customer satisfaction is low, it may influence customer loyalty. Therefore, this study adopts the Six Sigma DMAIC approach to improve the service quality of the category that receives the most petitions. A managerial system is developed to control the situation and then to decrease the number of petitions, as well as to increase the level of customer satisfaction.

IV. DMAIC FRAMEWORK

A. Stages of Define and Measure

This study explores ways to reduce TPC’s number of complaints and email petitions by using the Six Sigma DMAIC approach. In TPC’s e-service mailbox, there are 10 petition categories: power usage applications, billing, power distribution quality, line removal, compensation, service attitude, operational suggestions, questions regarding TPC regulations, illegal power usage, and others. The category of “others” means the general public can indicate any complaints that are not listed in the previous 9 categories. Owing to the broad nature of issues in this category, the contents are not taken into account in this study.

In the define (D) stage, TPC collected a total of 1,292 petition letters from the e-service mailbox during the year of 2010, divided into nine categories (see Figure 1).

![Figure 1. Number of cases in nine petition categories in Taiwan Power Company in 2010](image)

This study analyzed these nine petition categories as follows:

- The first category is power usage application, which includes petitions about time allocation of power usage, location and range. There are 122 cases in this category.
- The second category is billing, which includes complaints about abnormality in power consumption and sudden increase in payments; 304 cases belong to this category.
- The third petition category is power distribution quality, which includes complaints about voltage surge or outage situations; 298 cases are in this category.
- The fourth petition category is line removal, which includes issues and complaints regarding safety, transportation and construction due to line removal; 379 cases are in this category.
- The fifth category is compensation. Customers claim compensation for damages to their electrical equipment caused by unstable power supply; 27 cases are included in this category.
- The sixth category is service attitude. This category contains customers’ complaints about employees’ attitudes, manners and tone of voice; 113 cases belong to this category.
- The seventh category is operational suggestions, which includes any proposals or suggestions about the management and planning of TPC; 32 cases are counted in this category.
- The eighth category includes any questions or doubts regarding TPC regulations. There are only 5 cases in this category.
- The last category is illegal power usage. In this category, the general public feels discontent regarding illegal power usage and payment recovery; 12 cases are in this category.
The define stage shows that the fourth petition category, “line removal”, which includes a total of 379 cases, is the first in line for improvement. Next, in the measure (M) stage, as shown in Figure 2, a pie chart is used to illustrate the proportion for each petition category.

Figure 2. Percentages of cases in nine petition categories in TPC in 2010

Figure 2 shows that the number of petition cases in “line removal”, “billing” and “power distribution quality” account for over 75% of the total petitions, while the six remaining categories only represent less than 25% of the total petitions. The category that receives the most petition letters is “line removal”, which accounts for 29.3% of the total petitions, followed by “billing” and “power distribution quality”, which account for 23.5% and 23.1%, respectively. The six remaining categories in order of percentage are “power usage application”, “service attitude”, “operational suggestion”, “compensation”, “illegal power usage” and “questions regarding TPC regulations”.

B. Analyze, Improve and Control Stages

In recent years, TPC wants to process their customers’ requests in a faster and more efficient manner. Thus, TPC specially set up an e-service mailbox system. The general public can easily use this e-service mailbox system to apply for power usage or to make any other requests online. However, when an excessive number of petitions occur, it may result in a much longer processing time for responding to these petitions. Therefore, TPC urges finding a way to provide a more prompt service, increase the level of customer satisfaction and reduce the number of complaints.

As described in the measure (M) stage, the petition category of “line removal” is the one that needs immediate attention. A cause and effect diagram is used to analyze the category of “line removal” in the analyze (A) stage. In the “line removal” category, most petitioners complain about their safety, route changes, or construction building quality.

Two scholars and two supervisors, who have worked at TPC for over 10 years, brainstormed the main reasons why “line removal” receives many petitions; they find that “petitioners”, “employees” and “the Company’s system framework” are the three main reasons. “Petitioners” usually file complaints because TPC does not provide specific application procedures or sufficient public notification prior to construction. Petitions about “employees” are usually complaints about inexperienced employees, i.e. that they are not sufficiently service-minded. “Company system framework” includes complicated regulations and problems so that the cases are sent to the wrong department. A cause and effect diagram is shown in Figure 3:

Figure 3. Cause and Effect diagram for “line removal” petition category

Figure 3 shows the cause and effect relationships among the three main reasons why “line removal” receives many petitions. Below, the diagram is explained in greater detail.

1. Petitioners
   1) Insufficient public notification
      Insufficient public notification about “line removal” construction caused inconvenience to the people in the nearby area. TPC did not provide enough information or clear notification about the “line removal” construction to the neighborhood prior to the construction, causing great inconvenience to the neighborhood around the construction site. Neighbors often complain about the traffic congestion around the neighborhood and power outage caused by “line removal” construction.
   2) Lack of specific application procedures
      TPC did not provide a specific application procedure online for “line removal”. When the general public applies for “line removal” service, the submitted documents are often rejected due to insufficient information or ambiguous description of the construction site.

2. Employees
   1) Inexperience
      Employees dealing with customers have limited knowledge about their work, so they cannot solve the problems effectively and efficiently. This situation often leads to delays in processing and later creates potentially dissatisfied customers.
   2) Not service-minded enough
      If employees did not take the initiatives in assisting customers’ problems during the application procedures, more customers might complain about being dissatisfied with the service.

3. Company system framework
   1) Complicated regulations
      Many companies have strict regulations that seem complicated and overly-lengthy. For example, if the previous step is not done during the processing procedure, it could not advance to the next level to be processed.
This often results in inflexibility and affects the overall progress, and thus causes more complications.

2) Mis-assigned cases
   Head office sometimes assigned “line removal” projects to the wrong department. It often causes processing and time delays in each unit. Along with the complicated procedures and high degree of manual intervention, more and more complaints accumulate over time.

The above analyze (A) stage indicates that TPC did not provide relevant application procedures for “line removal” service, and the general public is not familiar with the instructions for applying for such service. Moreover, TPC often performs government-assigned “line removal” improvement projects but do not provide enough information to the general public about the improvement projects prior to the construction work. Therefore, in the improve (I) stage, this study presents three approaches to speed up the “line removal” application process, reduce the numbers of complaints and improve service quality in TPC as follows:

1. To establish a two-way communication platform to solve petitioners’ problems immediately;
2. To design a specific “line removal” application procedure and provide complete instructions, especially emphasizing the most frequently asked questions and solutions. Samples of completed application procedures should also be provided.
3. When performing any projects, TPC should take the initiative to notify the neighborhood with detailed information, such as the time period of construction, the construction site and contact information prior to commencement.

At the end of improve (I) stage, evaluations should be based on the Six Sigma standards. If the results fail to meet the expectations, the improve stage should be revised until the problems are fixed.

In the control (C) stage, petition letters in 2011 from TPC’s e-service mailbox are collected for the purpose of monitoring the improvement of TPC’s service quality. As shown in Figure 4, after implementing the Six Sigma DMAIC approach, the number of cases in the “line removal” petition category decreases from 379 in 2010 to 129 in 2011, with a reduction rate of 66%, while the total number of TPC e-service mailbox petitions decreases from 1,292 in 2010 to 481 in 2011, with a reduction rate of 62.77%. In other words, implementing Six Sigma successfully improves TPC’s service quality in the “line removal” category.

This indicates that there was significant improvement in the TPC e-service mailbox petition. If the improvement approach is well-implemented, a systematic knowledge-based system and in-service training programs should be developed. A well-organized standard of procedure and workflow control charts should be developed in order to monitor the overall application and construction processes in the control (C) stage. After a year of observation and improvement implementation, this study found that TPC did not provide the standardized application procedure to the general public for applying for “line removal” service. As a result, applicants often did not know which proper documents were needed. All kinds of cluttering documents have caused an overall delay with discrepancies in construction sites.

For these reasons, this study established a standardized application procedure for “line removal” service, including three stages of pre-construction, construction and post-construction. Descriptions are as follows:

1. Before the construction, the applicants should fill out their names, contact information, construction sites (preferable with photos) and date of application.
2. During the construction, TPC or the contractors should confirm the construction site and time with the applicants. After this two-way communication, applicants should confirm all the information and sign on the application form. This way, there will not be problems about any discrepancies in regard to the construction sites. Applicants, therefore, will not have any dissatisfaction regarding the “line removal” process.

After the construction, TPC inspectors should be sent to investigate how applicants feel about the “line removal” service. This way, TPC would be more aware of the problems in the process and try to solve them immediately. TPC service quality will therefore be improved.

V. CONCLUSIONS

The study demonstrates how the Six Sigma DMAIC approach is adopted to reduce the number of petitions for TPC’s online petition system (e-service mailbox system). 1,292 petition letters are collected from TPC e-service mailbox system in 2010, and then divided into 9 petition categories; 379 cases are in the “line removal” petition category, which is the highest among all the categories. Thus, “line removal” petition category is the first priority for improvement.

Based on the cause and effect diagram analysis, the main reasons for “line removal” petitions are “petitioners”, “employees” and “the Company’s system framework”. TPC should develop a two-way communication system. A complete “line removal” application manual and a sample of an application procedure with proper attached documents should be included. Also, a week prior to the construction, a sign with information of construction date and location along with person-in-charge’s contact information should be provided to notify the neighborhood.
After implementing the Six Sigma DMAIC approach, the number of cases in the "line removal" petition category decreases from 379 in 2010 to 129 in 2011, with a reduction rate of 66%, while the total number of TPC e-service mailbox petitions decreases from 1,292 in 2010 to 481 in 2011, with a reduction rate of 62.77%. This shows significant improvement. In order to have better control, a standardized application procedure for "line removal" service was developed, including pre-construction, during construction and post-construction.

REFERENCES


Pei-Shih Chen is an Assistant professor in the General Education Center at Central Police University in Taiwan. Her research interests are in the field of quality management, professional English and learning style. She has published in TELKOMNIKA.

Chang-Hsien Hsu is an Associate Professor in the Department of Business Administration at Asia University in Taiwan. He earned his PhD in Graduate Institute of Management Science from Tamkang University. His research interests are in the field of process capability index and quality management. He has published in International Journal of Production Research, International Journal of Information and Management Sciences, Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, International Journal of Advanced Manufacturing Technology, and Quality & Quantity.

Shih-Yun Hsu is an Associate Professor in the Department of Leisure and Recreation Management at Asia University in Taiwan. Her research interests are in the field of tourism marketing and consumer behavior.
Chun-Ming Yang is a PhD Candidate in the Department of Management Sciences at Tamkang University in Taiwan. He earned his MS in Department of Business Administration from Asia University. His research interests are in the field of process capability index and decision analysis. He has published in International Journal of Production Research, Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, and International Journal of Information and Management Sciences.
Energy Constrained Target K-coverage Algorithm in Heterogeneous Wireless Sensor Networks

Yong-sheng Wang
Lanzhou Jiaotong University, Lanzhou, China
wys_lz@126.com

Abstract—This paper presents a novel energy constrained target K-coverage algorithm, and the proposed algorithm is suitable to be exploited in heterogeneous wireless sensor networks. Particularly, the network activity in heterogeneous wireless sensor networks in this paper is organized in rounds, and each round is constructed by initial step and information sensing step. Furthermore, to prolong the network lifetime in target K-coverage process, the remaining energy and the sensing ability of each sensor are calculated in advance. Afterwards, the sensing unit set is defined to record sensing attributes that can be only covered by the target sensors. Furthermore, the attribute set contains several sensing attributes, and these attributes can cover the targets utilizing the specific sensing unit. In order to guarantee the K-coverage constraints, for each sensing attribute, we set a function to test whether the sensing attribute is covered. If the sensing attribute is covered by at least K sensors, the function return true, and then the decisions which represent if a sensor should be turned on or not are broadcast to each sensor’s one-hop neighbors. Finally, a series of experiments are conducted to make performance evaluation. In these experiments, all targets and wireless sensors are randomly allocated in the sensing field, and six types of experimental settings with different number wireless sensor are utilized. Experimental results show that, our proposed algorithm perform better than EF method and KTC_MNL method, and the performance of our algorithm is close to the optimal method.

Index Terms—Heterogeneous wireless sensor networks, Target K-coverage, sensing unit, sensing attribute, Integer linear programming

I. INTRODUCTION

Wireless sensor network is made up of a large number of resource-limited tiny devices, named sensors. A typical wireless sensor includes CPU, storage, battery power, and communication bandwidth. These sensors can sense task-specific environmental phenomenon, perform in-network processing on the sensed, and communicate wirelessly to other sensors or to a data gathering node, usually through multi-hop communications. Wireless sensor network can be utilized for a variety of applications dealing with monitoring, control, and surveillance[1]. Particularly, Wireless sensor networks have attracted a lot of attention recently due to the plenty of the applications and their connections with the physical world[2].

WSNs can be classified into two types, “homogeneous” and “heterogeneous”, in terms of the capacities of sensor resources such as energy supply, memory space, processing power and communication bandwidth. A homogeneous WSN is composed of the sensors with similar or identical resource capacities. A heterogeneous WSN (HWSN) is made up of sensors with different levels of resource capacities. To balance sensor resource capacities in a HWSN, its sensor are often organized into a number of clusters. A sensor with a high level of resource capacity normally acts as a cluster head to provide a data aggregation point for its cluster members served by sensors with lower resource capacities [3].

In the research field of wireless sensor networks, coverage problems are fairly fundamental and crucial. The target coverage problem is finding an optimal scheduling for sensors such that the lifetime to detect every target can be as long as possible. However, the target coverage problem has been proved to be NP-complete[4]. Most of previous research work only considers one or two factors exclusively and thus fails to prolong the lifetime to near the optimum.

Furthermore, a high level of redundancy in wireless sensor networks can benefit for enlarging the network lifetime, as well as for improving the resilience of coverage and network connectivity. The target coverage problem can be classified into two kinds: 1) field coverage, where the overall field is covered by sensors and no coverage hole can be tolerated at any time, and 2) target coverage, where each target is monitored continuously by at least one sensor[5]. In this paper, we concentrate on the target coverage problem of wireless sensor networks.

On the other hand, energy efficiency is an important issue in wireless sensor networks since sensors are battery powered. Therefore, reducing power consumption and prolonging network lifetime are the primary challenges in the design of wireless sensor networks. In this paper, we focus on the target coverage problem with the objective of maximizing network lifetime and reduce energy constrained[6].

To enhance the system performance, the K-coverage problem is presented. K-coverage means that every target should be covered by at least k sensors, and a property
known as k-coverage, where the maximum value of k is called the coverage degree. The limited battery power of sensors and the difficulty of replacing and recharging batteries in hostile environments require that the sensors be deployed with high density in order to extend the network lifetime. Moreover, to cope with the problem of faulty sensors due to low battery power and achieve high data accuracy, redundant K-coverage of the same region is necessary[3].

Based on the above analysis, in this paper, we study on how to design an efficient target K-coverage algorithm in heterogeneous wireless sensor networks with the energy conserving policy. The main innovations of this paper lie in that we organize the network activity in heterogeneous wireless sensor networks in round mode, and we propose a function to test if a specific sensing attribute is covered.

II. RELATED WORKS

The target coverage problem of wireless sensor networks is often studied together with the network lifetime in existing research works.

In the field coverage category, Pazand presented an algorithm named MDS based on minimum dominating set without using GPS/location information. It finds many subsets of sensors, each of which guarantees a high level of field coverage, with the aim to extend the overall network lifetime[7].

Similarly, Shen et al. present a novel density control algorithm called ESSC(Enhanced Sponsored Sector Coverage). ESSC can make full use of the Sponsored Sector Coverage concept. This method can prolong the network lifetime and provide a high degree of field coverage. Particularly, in this work, the authors presented ESSC algorithm takes note of the deficiency of the SSC algorithm and modifies the model of SSC to minimize the active node number based on the Sponsored Sector Coverage[8].

Gu et al. studied on the partial target coverage problem in wireless sensor networks under a novel coverage model. The authors showed that the partial target coverage problem can be optimally solved in polynomial time. First, they built a linear programming formulation considering the total time that a sensor spends on covering targets[9].

To prolong the network lifetime of wireless sensor network with multiple type sensors, the sensor need to monitor deployed targets. Hence, in paper [10], Mostafaei proposed an efficient scheduling method based on learning automata. In the proposed method, each node is equipped with a learning automaton to help the node selecting its proper state (active or sleep), at any given time.

Mini S. et al. aimed to examine phase-transition behavior of maximum lifetime target coverage problem in wireless sensor networks domain. This paper showed that as they vary the uniform sensing range of sensors, the number of hard instances of the problem sharply rises around a critical value independent of other parameters. To facilitate the hardness analysis, they devised a new and efficient algorithm to solve target coverage problem; this algorithm has the distinct ability of distinguishing hard problem instances[11].

Zorbas et al. described the problem of the minimum sampling quality, where an event must be sufficiently detected by the maximum possible amount of time. Since the probability of detecting a single target using randomly deployed static nodes is quite low, we present a localized algorithm based on mobile nodes. The proposed algorithm sacrifices a part of the energy of the sensors by moving them to a new location in order to satisfy the desired detection accuracy. It divides the monitoring process in rounds to extend the network lifetime, while it ensures connectivity with the base station. Furthermore, this paper propose two redeployment schemes that enhance the performance of the proposed approach by balancing the number of sensors between densely covered areas and areas that are poorly covered[12].

Montoya et al. proposes a MWSN architecture with an initial random distribution in a specific work area, and a centralized management to perform autonomous decision making about the movement and connectivity of the sensors. The work area presents mobile targets with interesting events. Particularly, these events must be covered by the mobile sensors, and thus, send the collected information through the network to any base station available[13].

Medagliani et al. addresses the problem of engineering energy-efficient target detection applications, using unattended Wireless Sensor Networks (WSNs) with random node deployment and partial coverage, for long-lasting surveillance of areas of interest. As battery energy depletion is a crucial issue, an effective approach consists in switching on and off, according to proper duty cycles, sensing and communication modules of wireless sensors [14].

Deng et al. discuss the energy-efficient area coverage problem considering boundary effects in a new perspective, that is, transforming the area coverage problem to the target coverage problem and then achieving full area coverage by covering all the targets in the converted target coverage problem[15].

However, in many cases, wireless sensors can be easily damaged or the sensed data is noised. Therefore, it should be ensured that each target in the sensing field must be covered by at least k sensors, such that even if some sensor fails or has no energy, the region covered by it still can be covered by other wireless sensors. It can be seen that target K-coverage algorithm can enhance the security and accuracy of the wireless sensor networks.

Zhao et al. concentrate on the problem of scheduling sensor activities to maximize network lifetime while maintaining both discrete K-target coverage and network connectivity. In K-target coverage, it is required that each target should be simultaneously observed by at least K sensors. The data generated by the sensors will be transmitted to the sink node via single or multiple hop communications[16].

Yun et al. study on the deployment patterns to achieve full coverage and k-connectivity (k <= 6) under different ratios of the sensor communication range (denoted by
R(c)) to the sensing range (denoted by R(s)) for homogeneous wireless sensor networks (WSNs). They propose new patterns for 3- and 5-connectivity[17].

Ammari et al. propose the Reuleaux tetrahedron model to characterize k-coverage of a 3D field and investigate the corresponding minimum sensor spatial density. They prove that a 3D field is guaranteed to be k-covered if any Reuleaux tetrahedron region of the field contains at least k sensors. Furthermore, they compute the connectivity of 3D k-covered WSNs as well. [18].

Li et al. investigate the Sensor Scheduling for k-Coverage problem and find that solving this problem requires to efficiently schedule the sensors, such that the monitored area can be k-covered throughout the whole network lifetime with the purpose of maximizing network lifetime. The proposed problem is NP-hard and the authors propose two heuristic algorithms under different scenarios[19].

Yang et al. analyze the critical conditions for connected-k-coverage using the percolation theorem and demonstrate their effectiveness using simulation results. Connected-k-coverage has been recognized as an effective notion for prolonging the sensor network lifetime. [20].

Different from the above methods, the novelty of the proposed method lies in that 1) Each round of the wireless sensor network is made up of initial step and information sensing step, and 2) To guarantee the K-coverage constraints, a decision function is presented for each sensing step. To prove that a 3D field is guaranteed to be k-covered if any Reuleaux tetrahedron region of the field contains at least k sensors. Furthermore, they compute the connectivity of 3D k-covered WSNs as well. [18].

In this paper, the network activity in heterogeneous wireless sensor networks is organized in rounds. This design means that each sensor runs this algorithm at the interval of a round time unit. Each round consists of two steps: 1) initial step and 2) information sensing step. In the first phase, all the wireless sensors should decide which sensing unit should be turned on in the second step. To make all the sensors’ activities can be synchronized, all the sensors are assumed to have a clock with a uniform starting point.

Firstly, the formal description of the energy constrained target K-coverage problem[21,22] in heterogeneous wireless sensor networks is given through the parameters definition.

Definition 1 (Energy constrained target K-coverage problem in heterogeneous wireless sensor networks): Given an integer K, and the condition K > 0 is satisfied. The proposed problem aims to find a subset of sensors C ⊆ V, and then the following conditions are satisfied:

i) Each sensor node in the set V could be covered by at least K different wireless sensors in set C.

ii) The number of wireless sensors in set C is minimized.

iii) The wireless sensors in set C are connected with each other.

Furthermore, the proposed target K-coverage problem can be converted to an integer programming as follows. Supposing there are n wireless sensors( N1, N2, ⋯, Nn), and αij denotes the coefficients which represent the coverage relationship between different wireless sensors. Particularly, αij can be computed as follows.

\[ α_{ij} = \begin{cases} 1, & \text{if the node } N_j \text{ can be sensed by } N_i \\ 0, & \text{otherwise} \end{cases} \] (1)

Next, the variables \( x_j \) (j ∈ {1, 2, ⋯, n}) represent boolean values.

\[ x_j = \begin{cases} 1, & \text{if the node } N_j \text{ can be selected in set } C_j \\ 0, & \text{otherwise} \end{cases} \] (2)

Based on the above analysis, the energy constrained target K-coverage problem can be represented as follows.

\[
\text{Minimize } \sum_{i=1}^{n} x_j \\
\text{S.T. } \sum_{j=1}^{n} α_{ij} x_j ≥ k, i \in \{1, 2, ⋯, n\} \\
x_j \in \{0, 1\}, i \in \{1, 2, ⋯, n\} \tag{3} 
\]

We can see that the condition \( \sum_{j=1}^{n} α_{ij} x_j ≥ k \) can ensure that each wireless sensor in the set V can be covered by at least K sensors in the set C.

Assuming that there is a wireless sensor work with N sensors having the ability to provide sensing coverage to a set of m target in the given region. Particularly, the sensors in the sensing range are randomly located. Supposing that N denotes the set of all the wireless sensors and T represents the set of all the targets. Before we provide the target K-coverage algorithm, some important definitions should be given in advance.

Definition 2 (Sensor utility): Sensor utility of sensor \( N_j \in (U(N_j)) \) means the total number of the target points \( t ∈ T_j \), and the points are located in the range of the sensor \( N_j \).

Definition 3 (Coverage): Target coverage of a given...
target \( t(C_t(t)) \) refers to the total number of sensors covering the target \( t \).

**Definition 4 (Cover utility):** Cover utility of the cover \( C_s(t) \) is defined as follows.

\[
U(C) = \sum_{N_s \subseteq C} U(N_s) \tag{4}
\]

**Definition 5 (Set Coverage):** Coverage of a set \( T \) for the target points which are related to the wireless sensors set is defined as follows.

\[
C_s(T) = \sum_{i=1}^{N_s} C_s(T) \tag{5}
\]

To enhance the network lifetime in target K-coverage process, the sensor’s energy should be utilized with high efficiency. Therefore, the remaining energy and the sensing ability of a given sensor should be calculated accurately. To calculate sensor priority, we should obtain its sensing ability in advance. The value of sensing ability for the sensor \( N_j \) is obtained by the following equation can be computed as followings.

\[
A_i = \frac{1}{\sum_{m=1}^{k} r_i} \tag{6}
\]

In the next section, the sensing attribute set that can only be covered by \( N_j \) is represented as \( d_j \). Hence, the sensor priority of sensor \( N_j \) can be calculated as follows.

\[
P_j = \frac{1}{2} (A_i + \gamma_i) \times \left[ (A_i - 1 - \gamma_i) \times \tau + 2E_i - 2 \right] + A_i \tag{7}
\]

where the parameter \( \gamma_i \) is used to modify the influence of sensing ability on sensor priority, and \( \gamma_i \) is computed through Eq.8.

\[
\gamma_i = \left[ \frac{E_i - 2}{\tau} \right] \tag{8}
\]

Based on the above analysis and related definitions, the target K-coverage algorithm is illustrated as follows.

**Algorithm1:** Energy conserving based k-coverage algorithm for wireless sensor networks

**Input:** Number of sensors, the sensing range, number of targets, number of sensing attribute, \( K \)

**Output:** Determining that which sensor should be turned on to assure that all the targets could be covered by at least \( K \) sensors.

1) Computing the sensing ability \( (K) \) of all sensors;

2) Computing the energy of all sensors:

\[
E_i = E_i - \sum_{j=1}^{k} E(d_i) \tag{9}
\]

3) Integrating \( E_i \) and \( A_i \) to obtain the sensing priority \( P_j \) by Eq.7

4) Setting the decision time in which the neighbor sensors are listened continuously.

5) Define \( \mu_p \) as the sensing unit set, that can be enabled by the sensor \( N_p \), to cover the sensing attributes, and the sensing attributes can only be covered by \( N_p \).

6) Define \( v_p \) as the attribute set containing the element \( \beta_{jk} \), and \( \beta_{jk} \) represents the sensing attribute covering the target \( t_j \) with the sensing unit \( \sigma_k \).

7) Swapping the information of \( \mu_p \) and \( v_p \) with neighbors of each sensor.

8) For all sensing attribute \( \beta_{jk} \) belonged to \( v_p \)

9) For the sensors belonged to one of the neighbors of \( N_p \) and they can cover target \( t_j \) utilizing \( \beta_{jk} \)

10) If \( \beta_{jk} \in \mu_p \) and the remaining energy of sensor \( N_i \) is larger than \( N_p \)

11) \( v_p = v_p - \{ \beta_{jk} \} \)

12) End if

13) End for

14) End for

15) For each sensing attribute \( \beta_{jk} \) belonged to \( v_p \)

16) Defined the function \( \text{is\_cover}(\beta_{jk}) \) to show whether the sensing attribute \( \beta_{jk} \) is covered.

17) If \( \text{is\_cover}(\beta_{jk}) = \text{false} \) then

18) Turn on the sensing units which can cover the attributes \( \beta_{jk} \);

19) \( N(\beta_{jk}) = N(\beta_{jk}) + 1 \)

20) If \( N(\beta_{jk}) = K \) then

21) \( \text{is\_cover}(\beta_{jk}) = \text{true} \);

22) End if

23) broadcast all the above decisions to each sensor’s one-hop neighbors

24) End if

25) End for

26) End if

© 2014 ACADEMY PUBLISHER
A. Experimental Settings

Before analyzing the experimental results, the experimental settings should be illustrated in detail. We select OMNET++\cite{25} as the simulation tool in this experiment. OMNET++ refers to the Modular discrete event simulator implemented in C++. Furthermore, OMNET++ provides a powerful GUI library for animation and tracing and debugging support. Particularly, the topologies of the wireless sensor network used in this experiment are shown in Fig.2, and routing algorithm used in this paper is presented in paper \cite{26}.

Parameters of experimental settings are explained in advance (shown in Table.1). In the proposed experimental settings, the numbers of wireless sensors and targets are limited for each simulation setting. Furthermore, the number and types of sensing units on sensors are constrained for each simulation as well, and the number and types of sensing units on each wireless sensor are randomly chosen. Meanwhile, the number and types of sensing attributes which are needed to be sensed at each target are chosen in a random mode. Particularly, all targets and wireless sensors are randomly allocated in the sensing field, and six types of experimental settings with different number wireless sensors are utilized. As is shown in Fig.2 (a)-(f), six types of sensors installation mode in the sensing field is described. To be short, the allocation of sensors and sensing targets are not allowed to be changed in the simulation process. The sensing range of each sensing unit is fixed to be 60 meters. Moreover, the communication range of each sensor is equal to two times of the sensing range distance. To make the experimental results more objective, all experiments are executed 15 times, and the final results are obtained by averaging all the 15 experiments.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensing range</td>
<td>60m</td>
</tr>
<tr>
<td>Area of sensing field</td>
<td>400m × 400m</td>
</tr>
<tr>
<td>Communication range</td>
<td>120m</td>
</tr>
<tr>
<td>Number of wireless sensors</td>
<td>{50, 100, 150, 200, 250, 300, 350}</td>
</tr>
<tr>
<td>Number of sensing attributes</td>
<td>{2, 3, 4, 5, 6, 7, 8, 9}</td>
</tr>
</tbody>
</table>

In table.2, the value of energy conserving for each working state of wireless sensors is given, and three sensor working states are considered: 1) active state, 2) transmitting state and 3) receiving state.

<table>
<thead>
<tr>
<th>Description of sensor working state</th>
<th>Energy consumption(mA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active state</td>
<td>11.25</td>
</tr>
<tr>
<td>Transmitting state</td>
<td>12.07</td>
</tr>
<tr>
<td>Receiving state</td>
<td>7.18</td>
</tr>
</tbody>
</table>

Afterwards, relationship between the sensing unit and its energy consumption should be explained (shown in Table.3). In this table, each sensing unit in our
experimental settings and its energy consumption is presented.

<table>
<thead>
<tr>
<th>The sensing unit id</th>
<th>Energy consumption (J/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1</td>
</tr>
<tr>
<td>$u_2$</td>
<td>2</td>
</tr>
<tr>
<td>$u_3$</td>
<td>3</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$u_l$</td>
<td>$l$</td>
</tr>
</tbody>
</table>

B. Experimental results and related analysis

To make the experimental analysis more objective, three methods are compared with the proposed algorithm. In paper [23], the wireless sensor networks’ target coverage problem is formulated as integer linear programming (ILP) to obtain the optimal results. It is well known that solving ILP is an NP-complete problem. Due to the limited energy and computing ability, it is not possible to utilize ILP solution to solve the target K-coverage problem in heterogeneous wireless sensor networks. Further, the ILP solution is implemented by the optimization toolbox in Matlab.

Firstly, we test the energy efficiency for each method by the standard metric “Network lifetime” with the number of targets and number of wireless sensors changing. Considering the ILP solution is the optimal approach for energy constrained wireless sensors target K-coverage, and there are no control and computation overheads in ILP solution. We design two experimental schemes to make performance evaluation, that is, 1) scheme 1: all the methods are compared with the ILP solution to show the efficiency without the control and computation overheads, where the ILP solution is implemented by matlab tool box optimization library, and 2) Scheme 2: the above methods are evaluated considering the control and computation overheads.
Fig. 3-Fig. 8 shows the performance of our proposed algorithm, EF method and KTC_MNL method in a comprehensive mode which integrates the numbers of wireless sensors and targets to test their influence on the network lifetime. Particularly, the unit of the network lifetime we used is “Round”. In this experiment, the number of wireless sensors is set from 50 to 350 with a step of 50, the number of targets is varied from 10 to 50 with a step of 10, and the number of attributes is fixed at 4. It can be seen from Fig. 3-Fig. 8 that all the methods under scheme 1 performs better than scheme 2. The reasons lie in that all in scheme 1 we do not consider the control and computation overheads in target converge process. However, in scheme 2, these factors are all considered in target coverage to enhance energy consumption.

On the other hand, we find that our proposed algorithm performs better than EF and KTC_MNL.

From Fig. 9, we can see that the network lifetime can be prolonged with the number of sensors increasing for all the methods. Because the targets covered by more sensors can enhance the network lifetime and reduce the energy consumption as well. As the ILP is an optimal solution, it can be regarded as the performance evaluation metric. Comparing with the ILP solution, the performance of network lifetime of EF method, KTC_MNL and our proposed algorithm decreasing 36.85%, 17.25% and 12.78% respectively.

Fig. 10 gives the experimental results of network lifetime with number of targets increasing. The conclusions can be drawn from the data in Fig. 10 that for all the methods, network lifetime reducing with the number of targets increasing. The reasons lie in that if the number of targets increases, additional energy consumption of sensors are increasing as well. Regarding the ILP solution as baseline, the performance of network lifetime of EF method, KTC_MNL and our proposed algorithm decreasing 40.45%, 24.88% and 15.25% respectively.

As this paper aims at the heterogeneous wireless sensor networks, energy consumption should be tested with different number of sensing attributes. In Fig. 11, network lifetime evaluating with number of attributes changing.
enhancing is demonstrated. When the number sensing attributes increasing, network lifetime decreases obviously for all methods, because more sensing attributes would increase the computation of wireless sensor networks. Based on the ILP performance, the performance EF method, KTC_MNL and our proposed algorithm decreasing 37.47%, 28.62% and 19.93% respectively.

V. CONCLUSIONS

In this paper, we propose energy constrained target K-coverage algorithm for heterogeneous wireless sensor networks. Firstly, the remaining energy and the sensing ability of each sensor are computed. Secondly, the sensing unit set and the attribute set are defined. Thirdly, a function is designed to test if the sensing attribute is covered. Fourthly, if the sensing attribute is covered by at least K sensors, the function return true, and then the finally decisions are transmitted to each sensor’s one-hop neighbors. Finally, experimental results show the effectiveness of our proposed algorithm.

Although this paper proposed an effective constrained target K-coverage algorithm, there are some aspects to be improved and deeply studied. In the future, the sensor with more than one sensing unit should be considered, and the sensors with different sensing ranges should also be taken into account. Particularly, other simulation platform and routing algorithm will be exploited in the experiment.

REFERENCES

[24] Vu Chinh T, Gao Shan, Deshmukh Wiwek P, Li Yingshu, Distributed energy-efficient scheduling approach for k-
coverage in wireless sensor networks, In proceedings of IEEE Military Communications Conference, 2006, pp.1-7


Research on Deep Web Query Interface Clustering Based on Hadoop

Baohua Qiang1, Rui Zhang2, Yufeng Wang3, Qian He1, Wei Li1 and Sai Wang1
1 Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology, Guilin 541000, China
2 North China University of Water Resources and Electric Power, Zhengzhou 450045, China
3 The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050000, China

Abstract—How to cluster different query interfaces effectively is one of the most core issues when generating integrated query interface on Deep Web integration domain. However, with the rapid development of Internet technology, the number of Deep Web query interface shows an explosive growth trend. For this reason, the traditional stand-alone Deep Web query interface clustering approaches encounter bottlenecks in terms of time complexity and space complexity. After further study of the Hadoop distributed platforms and Map Reduce programming model, a Deep Web query interface clustering algorithm based on Hadoop platform is designed and implemented, in which the Vector Space Model (VSM) and Latent Semantic Analysis (LSA) are employed to represent “Query Interfaces-Attributes” relationships. The experimental results show that the proposed algorithm has better scalability and speedup ratio by using Hadoop architecture.

Index Terms—Hadoop, Map Reduce, Deep Web, LSA, Query Interface Clustering

I. INTRODUCTION

According to the depth of the information, the Web can be divided into “Surface Web” and “Deep Web”. With the rapid development of Internet technology, the information contained on the Web, especially on the Deep Web, is showing an explosive growth trend. As Bright Planet speculated in year 2000 that the entire Internet contains 40 to 90 thousands of Deep Web pages, the information capacity of which is about 7500T [1]. MetaQuery had made more accurate statistics about the whole internet Deep Web pages in the year 2004; the results show that there were some 450 thousands of Deep Web databases [2]. It turns out that the amount of the pages had increased by nearly 9 times after only 4 years. Compared to that contained on the “Surface Web”, the information on the “Deep Web” has 5 characteristics below: (1) it could not be obtained by traditional search engines; (2) users acquire the information by filling out a form; (3) the information has a higher quality and a larger quantity; (4) the domain characteristics is more obvious and ; (5) most of the information has a free access. The explosive growth of information contained on Deep Web as well as the great value strongly attracts the attention of academia and business community.

As the Deep Web information portal, how the Deep Web query interface can be clustered effectively is one of the core issues need to be addressed while generating the integrated query interface [3]. At present, researches on Deep Web mainly focus on query interface integration algorithm based on single machine [4-7]. The proposed algorithms can effectively match the related query interfaces among a few Deep Web sites. But facing with the huge amounts of emerging Deep Web databases, the present approaches encounter great challenges in terms of time complexity and space complexity. So it is absolutely necessary and meaningful to study how to use distributed platforms to analyse the massive information on the Deep Web.

In view of the massive characteristic of Deep Web query interfaces, an effective approach is to introduce the parallel processing technology and design a rational and efficient parallel clustering algorithm [8]. Hadoop as a software framework which is able to make a distributed processing on massive data has been widely used. It has high reliability, scalability, efficiency and high fault tolerance [9]. On the basis of further study of the Hadoop platform, we designs and implements Deep Web query interface clustering algorithm on Hadoop platform, and before clustering, we employ constructed domain ontology and latent semantic analysis to make semantic expansions. Thus, the effectiveness of the query interfaces clustering is further improved. Besides, we have verified the correctness and effectiveness of the parallel algorithm design from the recall ratio and precision ratio in contrast to the results on single machine experiment. The results also show that the proposed parallel algorithm has good scalability and speedup ratio.

The main body of this paper is organized as follows. Firstly, we introduce the Hadoop architecture for Deep Web query interface clustering in Section II. And Section III describes the approach of parallelizable clustering algorithm based on VSM and LSA. Then the detailed Deep Web query interface clustering algorithm based on Hadoop is presented in section IV. The experiments and conclusion are given in Section V and Section VI, respectively.
II. HADOOP PLATFORM ARCHITECTURE FOR DEEP WEB QUERY INTERFACE CLUSTERING

Taking Hadoop Distributed File System (HDFS) and MapReduce as the core, Hadoop provides users with distributed infrastructure which is transparent in system bottom layer [10]. With high fault tolerance, high scalability of HDFS, users can deploy Hadoop on cheap hardware to form a distributed system. MapReduce allows users to develop parallel applications without knowing the details of the distributed system bottom layer. Thus, users can easily build their own distributed platforms and finish the processing of massive data using the computing and storage capability of the cluster.

HDFS uses Master/Slave structured distributed file system and an HDFS cluster consists of a NameNode and several DataNodes. NameNode as the primary server, manages the file system namespace and client access to file operations; DataNode manages the stored data. HDFS allows users to store the data in the form of documents. Internally, the file is divided into several data blocks and they are stored in a group of DataNodes. NameNode performs file system namespace operations, such as open, close, rename a file or a directory, it is also responsible for the mapping from data blocks to specific DataNode. The task of DataNode is processing the read and write requests of the system clients and handling create, delete, and copy to data blocks under the unified coordination of NameNode. Fig. 1 shows the architecture of HDFS.

MapReduce is a parallel programming model that enables software developers to write distributed parallel programs easily. In the Hadoop architecture, MapReduce is a software framework that is easy to use, the principle of which is: use an input <key, value> collection to produce an output of <key, value> collection; Specifically, MapReduce framework consists of two stages: Map and Reduce. In Map stage, MapReduce divides the input data of the task into fixed-size split, each split is then further broken into a number of key values<key1, value1>. After that, Hadoop creates a Map task for each split to execute user-defined Map functions, takes the corresponding split in <key1, value1> as input, and then calculates and generates an intermediate <key2, value2> collection. MapReduce collects all the value collections that have the same key, forms <key2, list(value2)>, and then divides the meta group into several groups according to the range of the key, corresponding to different Reduce tasks. In the Reduce stage, Reducer integrates the received data from different Mappers together and sorts them according to the key value, then calls the user-defined reduce function and processes the input <key2, list(value2)> to obtain the key value <key3, value3> and then output to HDFS. Fig. 2 shows the process of MapReduce data processing.

To sum up, the distributed storage used by Hadoop platform can improve the read and write speed and expand the storage capacity; using MapReduce programming model to integrate data on HDFS will ensure the efficiency of data analysis and processing. In view of the rapid growth of the information contained on the Deep Web, if we want to store and manage these useful data and information efficiently and then make further analysis, Hadoop platform is undoubtedly an excellent choice.

III. DETERMINE THE LATENT SEMANTIC RELATIONSHIPS BASED ON VSM AND LSA

A. The Vector Space Model of Deep Web Query Interface

The first step in clustering Deep Web query interface is to convert the Deep Web query interface set to Vector Space Model (VSM) [11]. Assume that we get N Deep Web query interface expressed as \( F = \{f_1, f_2, \ldots, f_n\} \) and consider it as the column index of VSM model, with \( A = \{a_1, a_2, \ldots, a_m\} \) representing all attributes obtained from F and consider it as the row index. So that we get a “Query Interfaces-Attributes” matrix C:

\[
C = (c_{ij})_{m \times n}
\]

Each row of this matrix represents a single attribute and each column stands for a single query interface, the element indicates the number of attributes occurred in query interface. TF-IDF weight will be selected to evaluate the importance of attributes, its basic thought is: if an attribute in a query interface appears a lot, it will also appears much in another similar query interface, and vice versa. Weight is calculated as follows:

\[
w_{ij} = TF_{ij} \times log(N/DF_{i})
\]

Where \( TF_{ij} \) represents the numbers of times attribute \( a_i \) occurs in query interface \( f_j \); \( N \) stands for the total number of query interfaces; \( DF_i \) signifies the total number of attribute \( a_i \) appears in the N query interfaces. While calculating the distance between cluster objects, the general way is to use Euclidean distance, but considering the existing difference about the number of query interface in different areas, we choose the Cosine Similarity, and it is calculated as follows:
\[
\text{sim}(i_1, i_2) = \frac{\overrightarrow{V(i_1)} \cdot \overrightarrow{V(i_2)}}{\|\overrightarrow{V(i_1)}\| \cdot \|\overrightarrow{V(i_2)}\|}
\] (3)

At this point, the vector space model of Deep Web query interface is set up.

B. Determine the Latent Semantic Relationships by LSA

VSM can be used to compute the similarities of Deep Web query interfaces by evaluating keywords matching literally. But it is difficult to determine their latent semantic relationships. Latent semantic analysis (LSA) is an effective indexing and searching approach [12], which can be employed to discern the latent semantic similarities among Deep Web query interfaces through constructing latent semantic model.

The core issue of LSA is how to project high-dimensional “Query Interfaces-Attributes” matrix to lower dimensional latent semantic space by low rank approximation effectively [3]. Singular value decomposition (SVD) is the mathematical basis of LSA. Let \( C \) be the \( m \times n \) “Query Interfaces-Attributes” matrix, it can be represented as formula 2.1 by SVD:

\[
C = U\Sigma V^T
\] (4)

\( U \) and \( V \) stand for orthogonal matrix of \( m \times n \), and their columns are orthogonal feature vector of \( C^T \), respectively. \( \Sigma \) is \( m \times n \) matrix, \( \Sigma \) needs to be explained specially:

1. The eigenvalues of \( C^T_m C_m \) are \( \lambda_1, \lambda_2, \ldots, \lambda_r \);
2. \( \forall i \in [1, r] \), there exists \( \delta_i = \sqrt{\lambda_i} \) and \( \lambda_i \geq \lambda_{i+1} \), \( \Sigma_m \) meets \( \Sigma = \delta \) and other elements of matrix is 0. \( \Sigma \) is also called singular value of \( C_m \).

In LSA, noise data can be removed by low rank approximation [4]. Low rank approximation is defined as follows. Suppose \( C \) is a matrix of \( m \times n \), its rank is \( r \), and \( C_k \) is a matrix of \( m \times n \) with rank \( K \) and \( r \geq k \). Let \( X = C - C_k \). As formula 5 is the smallest one, we call \( C_k \) is the low rank approximation matrix of \( C \) when \( k \) is much smaller than \( r \).

\[
\| X \|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2}
\] (5)

SVD is an effective means to solve the problem of low rank approximation. We can firstly obtain \( \Sigma_k \) by reserving the \( k \) biggest singular values and setting other \( r-k \) singular values as 0 of \( \Sigma_m \), then calculate \( C_k = U\Sigma_k V^T \) according to formula 4, finally we can obtain the approximation \( C_k \) of \( C \). Theorem is shown as follow:

\[
\min_{\|C - \tilde{C}\|_F} \|C - \tilde{C}\|_F = \sum_{i=1}^{r-k} \delta_i^2
\] (6)

It can be demonstrated that the above process will produce a matrix \( C_k \) with rank \( k \), and its norm has the minimum error. By VSM and LSA, the latent semantic similarities among Deep Web query interfaces can be obtained effectively.

IV. DEEP WEB QUERY INTERFACE CLUSTERING ALGORITHM BASED ON HADOOP

A. Parallelizable Clustering Algorithms

K-Means and K-medoids clustering are the most commonly used partition-based algorithms in clustering domain [13]. The latter is more “robust” than the former when there were noise and outliers. Unlike mean value, the K-medoids clustering is not sensitive to outliers or other extreme values. However, it is not suitable for distributed scenarios. In contrast, the similarity calculation between each node and the center point in K-Means algorithm is independent, and the center point calculation is done in one cluster. In terms of the code implementation, a new center point calculation could be finished in one reduce function; therefore it is very suitable for parallelization transformation. Sequential execution procedure of the K-Means algorithm is given below firstly, and its parallel design will be given in subsequently.

\[
\begin{array}{c}
\text{Input:} \\
k:\text{the number of the Clustering;}
S: \text{Deep Web Query Interfaces Set}
\end{array}
\]

\[
\begin{array}{c}
\text{Output:} \\
K \text{ clusters}
\end{array}
\]

\[
\begin{array}{c}
\text{Steps:} \\
(1) \text{Represent “Query Interfaces-Attributes” matrix with VSM and find the latent semantic relationships by LSA;}
(2) \text{Randomly select k query interfaces from S as the initial center point;}
(3) \text{Repeat}
(4) \text{Classify each query interface into the most similar cluster of center point;}
(5) \text{Calculate the mean value of each cluster as the new center point;}
(6) \text{Until no change occurs}
\end{array}
\]

B. Map Function Design

TextInputFormat is the default input method of Hadoop, each split is separately a map input, each row of data will generate a record and each record is represented as a form of <key, value> which can be accepted by map function, and key represents record byte offset in current split, the type of which is LongWritable, value stands for the content of each row, the type of which is Text. As in this experiment, value is represented by the column string of the vector space model matrix of the query interface.

The setup function is executed prior to map in Mapper, and it is executed only once in the Mapper life cycle. Therefore, we can do some initialization operation in the function. The role of setup function in this algorithm is initializing the center point of each cluster, and then storing it into centerList. The pseudo-code of map function is given below:

© 2014 ACADEMY PUBLISHER
map(LongWritable key, Text value){
    /* Parse the value to node object */
    Node node = parse(value);
    /* Calculate the center point that each node belongs to */
    node.setCenter(centerList);
    /* Take the serial number of cluster that nodes belong to as key, the string representation of current node as value */
    context.write(new IntWritable(node.center.id),
    new Text(node.toString()));
}

C. Combine Function Design

Combine process is a part of Mapper and executed after map function. Normally, it can effectively reduce the number of intermediate results; thereby reduce network traffic during data transmission. If designed properly, this process can significantly enhance the execution efficiency of the program. The pseudo-code of the Combine is as follows:

combine(IntWritable key, Iterator<Text> values){
    /*Parse "values" to Node, record the number of node in values set, then use "count" to save the node number belonging to cluster key in current split */
    int count = 0;
    float[] vector;
    while(values.hasNext()){
        Node node = parse(value.next());
        /*Accumulate the component of each node and prepare for the new center point that reduce function will calculate */
        vector = plus(vector, node.vector);
        count++;
    }
    /*Splice "vector" and "count" into string value */
    Text value1 = toString(count) + "#" + toString(vector);
    /* Output key and value */
    context.write(new IntWritable(key),
    new Text (value1));
}

D. Reduce Function Design

The parameter that reduce function receives is <IntWritable key, Iterator<Text> values>, in which key is the serial cluster number, values are the string representation of all the node component values in cluster. Reduce function is similar to combine function and pseudo-code is as follows:

reduce(IntWritable key, Iterator<Text> values){
    /*Parse values to Node and use count to record the number of node in values set, define newCount to store the number of each cluster. Use vector to store the component of center point */
    int count = 0;
    int newCount = parseInt(value);
    float[] vector;
    while(values.hasNext()){
        Node node = parse(value.next());
        vector = plus(vector, node.vector);
        count++;
    }
    /* Update newCount */
    newCount += count;
    /* Update the component of center point */
    vector /= newCount;
    /* value1 is the string representation of vector */
    Text value1 = toString(count) + "#" + toString(vector);
    /* Output key and value */
    context.write(new IntWritable(key),
    new Text (value1));
}

After a round of , we get newCenters, then compute the next round until convergence.

V. EXPERIMENTS

A. Selection of the Experiment Data

The data in this experiment comes from the Web integration resource library of UIUC which contains the query interface in many fields and store in the form of XML. 221 query interfaces in 4 domains i.e. airfares, automobiles, books and musicRecords are selected in the experiment. In order to simulate the situation of big data, we have made a proportional copy of the original data to expand the size. The experiment data is shown in Table I:

<table>
<thead>
<tr>
<th>Data Groups</th>
<th>Data Size</th>
<th>Number of Query Interfaces</th>
<th>Copy Multiples</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1G</td>
<td>210613</td>
<td>953</td>
</tr>
<tr>
<td>B</td>
<td>2G</td>
<td>421005</td>
<td>1095</td>
</tr>
<tr>
<td>C</td>
<td>3G</td>
<td>631397</td>
<td>2857</td>
</tr>
<tr>
<td>D</td>
<td>4G</td>
<td>841789</td>
<td>3809</td>
</tr>
<tr>
<td>E</td>
<td>5G</td>
<td>1052181</td>
<td>4761</td>
</tr>
</tbody>
</table>

B. Experiment Data Preprocessing

The irregular definition of query interface attributes has brought too much noise, because of that, some part of query interfaces are lack of enough semantic information. In the light of the characteristics of this experiment, four steps are needed to process the data:

(1) Remove the stop words, but retain the words that have domain meanings, for example in the airfares domain “to”, “from” etc.

(2) Stemming reduction and morphology normalization. Revert the different state of attribute words, singular and plural forms to the stemming of the word.

(3) Semantic expansion of query interfaces. Build domain ontology for each of these domains. Take the aviation domain for example, we can build ontology. Fig. 3 shows the hierarchy diagram of aviation domain ontology:
If an attribute of query interface appears in the ontology, the related attributes of the entire path will be added to the attribute set of the query interface and make semantic expansion.

(4) The latent semantic analysis to query interface is employed to find the latent semantic relations among Deep Web query interfaces and improve the similarity of query interface belonging to the same domain.

C. Experiment Environment

The experiment environment is composed of 4 HP desktops of the same model, configured as follows:

<table>
<thead>
<tr>
<th>EXPERIMENT ENVIRONMENT CONFIGURATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
</tr>
<tr>
<td>RAM</td>
</tr>
<tr>
<td>OS</td>
</tr>
<tr>
<td>Hadoop Version</td>
</tr>
</tbody>
</table>

D. The Recall and Precision

The results of Deep Web query interface are unordered collections, so we choose precision ratio $P$ and recall ratio $R$ as criteria to evaluate our proposed algorithm. The formula of $R$ and $P$ is given below:

\[
R = \frac{tp}{tp + fn}
\]

\[
P = \frac{tp}{tp + fp}
\]

Where $tp$ (true cases) represents the correct assignment to the corresponding cluster cases; $fp$ (pseudo-positive cases) indicates the wrong assignment cases; $fn$ (pseudo-negative cases) represents the cases that is assigned to the cluster but is not retrieved. The specific values of precision $P$ and recall ratio $R$ is shown in Fig. 4:

Since this experiment expands the size of the data by replicating the raw data, 5 groups of Deep Web query interface data: A, B, C, D, E are respectively tested, each group of data has been run respectively on the cluster of 2, 3, 4 machines. Fixed initial centers, recall ratio and precision ratio remains unchanged. We also get the same $P$ and $R$ value in the case of single machine. Besides, the experiment results that the value of $P$ and $R$ are greater than 90% are also very encouraging. So we can say that the map/reduce distributed algorithm of Deep Web query interface clustering is reliable and significant in the era of big data.

E. Cluster Scalability

Run the data of A, B, C, D, E group on the cluster formed by different number of nodes and compare the run-time. The results are shown in Fig. 5:

From the Figure above we can see that when process the same number of query interface, if the number of nodes in the cluster increases, time consuming significantly reduces; the larger the data size, the faster the run-time speed decreases. Therefore, while dealing with the large-scale data, we can improve the process capability of the system by increasing the number of nodes, which reflects the good scalability of the system.

F. Cluster Speedup Ratio

Speedup ratio [15] is the time-consuming ratio of the same task running on a single-processor system and parallel processor system. It is used to measure the performance and effects in parallel system or program parallelization. Speedup ratio is calculated as follows:

\[
Sp = \frac{Sp_{single}}{Sp_{parallel}}
\]

Where $Sp$ represents Speedup ratio, denotes the running time in a single processor, denotes the running time in a $p$-processors parallel system. The experiment results are shown in Fig. 6:

As can be seen from Fig. 6, the speedup ratio of the algorithm is close to linear speedup ratio, with the increment of the data scale, the speedup ratio of the distributed system tend to stabilize. This fully demonstrates the advantages of handling with big data on Hadoop platform. In the view of algorithm design, we introduce the combine function between map and reduce...
which can effectively merge locally, reduce the network data transmission between different nodes in cluster, and greatly reduce the unnecessary time consuming. What’s more, due to the reasonable design of data structure, the extra system time consuming is also reduced correspondingly.

VI. CONCLUSIONS

In light of the bottleneck that Deep Web query interface clustering meets in handling with massive data on traditional single machine, we designed and implemented the Deep Web query interfaces clustering algorithm based on Hadoop. By experiment on different scale of data sets and different nodes of cluster, the results show that our proposed algorithm has excellent scalability and speedup ratio. However, there is still room for improvement in the implementation details of the algorithm and the platform configurations, for example: how to compress the data to reduce the pressure of network bandwidth; how to set a more reasonable number of reducer, etc. Therefore, the next step we will focus on the Hadoop platform and algorithm design, and further tap the potential of cluster computing.

ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (grant 61163057, 61201250, 61363029, 61462020), Guangxi Nature Science Foundation (grant 2012jjAAG0063), Open Fund of Guangxi Key Laboratory of Trusted Software (kx201308). The authors would also like to express their gratitude to the anonymous reviewers for providing helpful suggestions.

REFERENCES


Baohua Qiang was born in 1972. He received the B.E. degree and M.A. degree from Southwest University in 1996 and 2002 respectively and the Ph.D. degree from ChongQing University in 2005. He went to University of Illinois as a visiting scholar from May to August in 2007. He had finished his postdoctoral research at South China University of Technology from 2007 to 2009. Now he works at Guilin University of Electronic Technology. His major research interests are Web information processing, intelligent search, massive data processing and network information integration.

Rui Zhang was born in 1980. She received her Master degree at information institute in Southwest University at 2006 and now is studying for her doctorate. Her research interests are Software Engineering, Data Mining and Big Data.
Call for Papers and Special Issues

Aims and Scope.

Journal of Software (JSW, ISSN 1796-217X) is a scholarly peer-reviewed international scientific journal focusing on theories, methods, and applications in software. It provides a high profile, leading edge forum for academic researchers, industrial professionals, engineers, consultants, managers, educators and policy makers working in the field to contribute and disseminate innovative new work on software.

We are interested in well-defined theoretical results and empirical studies that have potential impact on the construction, analysis, or management of software. The scope of this Journal ranges from the mechanisms through the development of principles to the application of those principles to specific environments. JSW invites original, previously unpublished, research, survey and tutorial papers, plus case studies and short research notes, on both applied and theoretical aspects of software. Topics of interest include, but are not restricted to:

- Software Requirements Engineering, Architectures and Design, Development and Maintenance, Project Management,
- Software Testing, Diagnosis, and Validation, Software Analysis, Assessment, and Evaluation, Theory and Formal Methods
- Design and Analysis of Algorithms, Human-Computer Interaction, Software Processes and Workflows
- Reverse Engineering and Software Maintenance, Aspect-Orientation and Feature Interaction, Object-Oriented Technology
- Component-Based Software Engineering, Computer-Supported Cooperative Work, Agent-Based Software Systems, Middleware Techniques
- AI and Knowledge Based Software Engineering, Empirical Software Engineering and Metrics
- Software Security, Safety and Reliability, Distribution and Parallelism, Databases
- Software Economics, Policy and Ethics, Tools and Development Environments, Programming Languages and Software Engineering
- Mobile and Ubiquitous Computing, Embedded and Real-time Software, Database, Data Mining, and Data Warehousing
- Internet and Information Systems Development, Web-Based Tools, Systems, and Environments, State-Of-The-Art Survey

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the “Call for Papers” to be included on the Journal’s Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. “Special Issue: Selected Best Papers of XYZ Conference”.
- Sending us a formal “Letter of Intent” for the Special Issue.
- Creating a “Call for Papers” for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees.
- Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jsw/.