Universal Access in the Information Society (2018) 17:635–647
https://doi.org/10.1007/s10209-018-0610-z

LONG PAPER

Automated comprehensive evaluation approach for user interface satisfaction based on concurrent think-aloud method

Weipeng Chen[1] · Tao Lin[1] · Li Chen[1] · Peisa Yuan[1]

Published online: 31 January 2018

Abstract

The concurrent think-aloud protocol (CTA) is an effective method for collecting abundant product comments related to user satisfaction during the execution of evaluation tasks. However, manual analysis of these audio comments is time-consuming and labor-intensive. This paper aims to propose an approach for automated comprehensive evaluation of user interface (UI) satisfaction. It takes advantage of text mining and sentiment analysis (SA) techniques instead of manual analysis in order to assess user comments collected by the CTA. Based on the results of the SA, the proposed approach makes use of the analytic hierarchy process (AHP) method to evaluate the overall satisfaction and support developers for UI design improvements. In order to enhance the objectivity of evaluation, a sentiment matrix originating from text mining and SA on user comments is used to replace the criteria and the relative weights of the AHP method which were previously defined by experts. A comparison between the questionnaire survey method and the proposed approach in the empirical study suggested that the latter can efficiently evaluate UI satisfaction with high accuracy and provide designers abundant and specific information directly related to defects in design. It is argued that the proposed approach could be used as an automated framework for handling any type of comments.

Keywords Satisfaction evaluation · Concurrent think-aloud · Text mining · Sentiment analysis · Analytic hierarchy process

1 Introduction

Usability refers to the degree to which a specific product can be used by specific users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specified context of use [1]. Satisfaction is an objective measure of the needs and expectations that reflect the subjective desires of the user, which is an important index in evaluating product usability. Moreover, user interface (UI) improvement is one of the main targets of usability engineering and overall product usability can be prompted by enhancing the user satisfaction of UI [2]. It is, therefore, important to apply an effective evaluation method in order to accurately measure the user satisfaction of UI.

Whitefield [3] proposed a framework based on the views of users and computers, with real or representational presence, to classify evaluation methods into four classes which included analytic methods, specialist reports, user reports, and observational methods. The analytic methods are applied in early development as suitable modelling techniques without real users and computers, which produce results that are usually of uncertain validity and reliability. In the specialist report methods, the views of experts are integrated instead of those of real users when assessing real computers. For the Usability Evaluation (UE), expert interview, heuristic evaluation, and cognitive walkthrough are the three typical methods [4–6]. On the contrary, questionnaire survey is the widely used method of user evaluation because of its simplicity and feasibility, which involves real users and representational computers with less objectivity [2]. Furthermore, it requires questionnaire designers to think about how to effectively draft the questionnaire contents in order to get the user comments they want to know. In comparison with the other three classes, observational methods have more powerful capabilities in obtaining abundant detailed and accurate data for quantitative analysis due to direct interactions between real users and computers. These methods, however, have some integrating, interpreting, and analyzing difficulties.

✉ Tao Lin
richardcwp@163.com

[1] College of Computer Science, Sichuan University, Chengdu, Sichuan, China

User testing and the Think-Aloud protocol (TA) are representative methods of user reports in usability tests. In the former, users will be asked to implement a series of specific tasks, which do not seem to be a better way of understanding the mind of users because the user testing method focuses on task implementations rather than feelings. The TA, on the other hand, originates from cognitive psychology and is an ideal evaluation method in revealing user satisfaction by making use of interviewees continuously verbalizing their instantaneous thoughts [7].

In addition, Nielsen [8] points out that *Thinking aloud may be the most unique and valuable method in usability engineering*. Moreover, it has been demonstrated as an effective and accurate method in collecting more information per datum and determining usability problems [9–11]. Based on the timing of verbalization during or after the task, TA can be divided into concurrent think-aloud protocol (CTA) and Retrospective Think-Aloud protocol (RTA). The segments reported by CTA are far more than the ones reported by RTA [12]. However, manual analysis on the data collected using TA still requires greater amounts of both time and labor, which is why automated analysis is necessary.

Based on the four dimensions of automated UE as defined by Ivory [12], the proposed approach applies usability test as the method class, CTA as the method type, text mining and sentiment analysis (SA) as automation types, and interface usage as the effort level in the processing of the comprehensive evaluation for UI satisfaction. It mainly includes three stages: UI assessing, comments analysis, and comprehensive satisfaction evaluation. In the first stage, interviewees will evaluate the product UI with CTA and the audio records of the CTA will be transformed into text corpus. In the second stage, text mining and SA techniques will be applied to the text corpus for further analysis. In the last stage, based on the analytic hierarchy process (AHP) analysis, global and specific dissatisfaction ranks of UI elements are computed for design improvements. There are three advantages in the proposed approach. First, user comments collected by CTA are abundant and effective for further analysis. Second, automated analysis techniques with high-efficiency are used to replace manual analysis. Finally, the proposed approach more objectively selects criteria and balances their weights in the process of AHP since they are based on the actual results of SA instead of subjective judgments made by experts.

An empirical case, with the proposed automated approach and a manual analysis of the questionnaire survey, was applied to verify and validate our approach. Results showed that both methods can present the overall satisfaction of UI. Moreover, the results indicated that the proposed approach had higher rates of accuracy and the dissatisfaction ranks of the UI element from the proposed approach, along with their corresponding specific user comments which were easily queried from the database, were effectively demonstrated to improve UI defects.

2 Related work

2.1 UI assessment using CTA

First introduced into usability studies by Clayton Lewis [13], TA has the ability of creating oral reports that reflect instantaneous cognitive processes stimulated by external influence. It can be further divided into CTA and RTA, and the major difference between the two being the timing of verbalization that CTA occurs during task; RTA comes up after task. In particular, CTA can avoid deviations caused by retrospection and introspection when the information is extracted from long-term memories of human [3].

In usability evaluation using CTA, Ericsson and Simon [14] utilized the short term memory characteristics of human beings to acquire useful and reliable information from the continuous communication between the interviewee and the experimenter with the use of protocol analysis. However, Boren and Ramey [15] argued that the previous research might create a mismatch between theory and practice. They then proposed a new approach, which was based on observation and empirical studies, to obtain a higher degree of task completion with less rate of loss from verbal interactions. Furthermore, their theory placed an emphasis on maintaining confirmations and feedback during communication by keeping the *backward channel* between interviewees and experimenters when executing the task. On the basis of their contribution, Hertzum [16] placed the concept of *relaxation* into a cognitive process and enhanced the participation of interviewees especially when in trouble. In some cases, CTA has been used as a usability test method to ask interviewees to share their cognitive experiences [9]. Moreover, CTA can acquire higher accuracy of test results while working with pre-coaching conditions [17]. In conclusion, a relaxed environment and pre-training will prompt interviewees to fulfill the CTA test and breezily speak out their experiences. Combining the above advantages together, interviewees, who have had pre-training, can process more accurate evaluations using CTA and their comments can then be used as information bases for further analysis.

However, most of these user comments are analyzed manually and recent studies have shown that the time required for the manual analysis of user comments is about 100 times greater than the evaluation time during the TA test. It is therefore necessary to establish an automated mechanism that can replace manual analysis on these user comments for time and labour saving.

2.2 Comments analysis

In general, customers who are very satisfied or dissatisfied with products would likely add their extra positive or negative comments in the typical bimodal distribution, of which the average rate of the products is either extremely high or low [18]. However, it is difficult to read and summarize usability information from the huge numbers of user comments [19] by manual. Text mining is an effective method for analyzing user comments by deriving high-quality information from texts typically with the use of the SA technique.

Most of the pioneer text mining studies focused on the summarization of text documents with similar information. Text mining was later adopted to analyze text contents in more detail and for more complex purposes. For instance, Hedegaard [20] has used text mining to extract and classify online reviews into different dimensions of usability or user experience. Although such text categorizations in coarse granularity can identify and classify usability information from user comments, they somehow miss the importance of analyzing the sentiment polarity of the information. A further progress in text mining was then made by Hu and Liu [21] on capturing product features and identifying sentiment orientations of opinions from user comments, which are divided into three categories—positive, neutral and negative [22]. In fact, opinions contain not only sentiment polarities but also emotional intensities. Thus, Daekook Kang [23] classified sentiment expressions into 5 scales namely the very positive, positive, neutral, negative and very negative with corresponding sentiment values of 2, 1, 0, − 1 and − 2, respectively. Moreover, he compiled two dictionaries, of product attributes and sentiment words, for SA processing. However, he neglected that negative words and adverbs of degree within sentences are also important factors in computing sentiment scores. In addition, online reviews, analyzed by either Wu or Daekook Kang [22, 23], usually lack detailed information about the specific situations of product use or measurements, which are not easy to have a precise evaluation of against comments of TA.

Aside from the extraction and analysis of product attributes and verbal sentiment expressions, negative word and adverbs of degree, which are also an important part of sentiment emotional expression, should be taken into account in both text mining and SA. Furthermore, in reference to the templates with Natural Language Processing (NLP) applied by Chetan Arora [24], patterns, according to traditional Chinese grammar, can be used in text mining and SA for finer granularity.

2.3 AHP comprehensive evaluation

AHP is a multi-criteria decision making (MCDM) method for resolving conflict management issues by organizing and analyzing complex decisions based on mathematics and psychology [25]. It has the ability to decompose the overall evaluation into a hierarchy of more easily comprehensive sub-assessments including fundamental components and interdependencies with a large degree of flexibility [26]. In other words, the basic principle of AHP is to study complex problems as a whole system by analyzing a number of elements and dividing the system into levels associated with each element. Therefore, the construction of a hierarchy model, priority analysis, and consistency verification are involved in the use of AHP. In addition, element rank results of each level, computed by AHP, have been used to help in decision making and selecting the correct ways to resolve the problem. Based on the evaluation on both tangible and intangible attributes or characteristics of UI with AHP, Matsuda [27] selected the most suitable UI for each user in the UI design. In the meantime, Delice [28] also took advantage of AHP to analyze the website usability and helping developers make decisions.

In general, the methodology of AHP is articulated in the following four steps: First of all, the problem or evaluation goal is analyzed into a hierarchical structure by setting the overall evaluation target of the problem at the top level, multiple criteria decomposed from the top element in the middle level, and decision alternatives at the bottom level. The second step is to construct a decision matrix based on pair comparisons of criteria or alternatives. Then, the third step is to compute the weight of the criteria or alternatives. The final step is calculate the global score and check the consistency of the decision matrix which should have a consistency ratio less than 0.1 [29].

3 Proposed evaluation approach

The automated framework of the proposed evaluation approach, as depicted in Fig. 1, mainly contains three phases: UI assessing, user comments analysis, and comprehensive satisfaction evaluation. In the first phase, interviewees evaluate the product UI with CTA after dictionary compilation, then the audio comments are transformed into text corpus. After extracting a 5-tuple m_k (UI object, attribute, sentiment word, adverb of degree and privative) from each sentence or phrase using text mining, the SA technique is utilized in the second phase to construct sentiment vectors based on the 5-tuples. Here the 5-tuple m_k will be extracted and constructed as in Eq. 1 if at least one UI object, attribute, and sentiment word are found in the same sentence or phrase.

$$m_k = (O_k, A_k, S_k, \mathrm{AD}_k, \mathrm{NA}_k) \tag{1}$$

where m_k is the k-th 5-tuple found in a sentence or phrase; O_k is the UI object of m_k, $O_k \in \{E_i | i = 1 \ldots 13\}$; A_k is the related attribute of O_k in m_k, $A_k \in \{F_j | j = 1 \ldots 10\}$, E_i and F_j are listed in Tables 2 and 3; AD_k and NA_k are the adverb

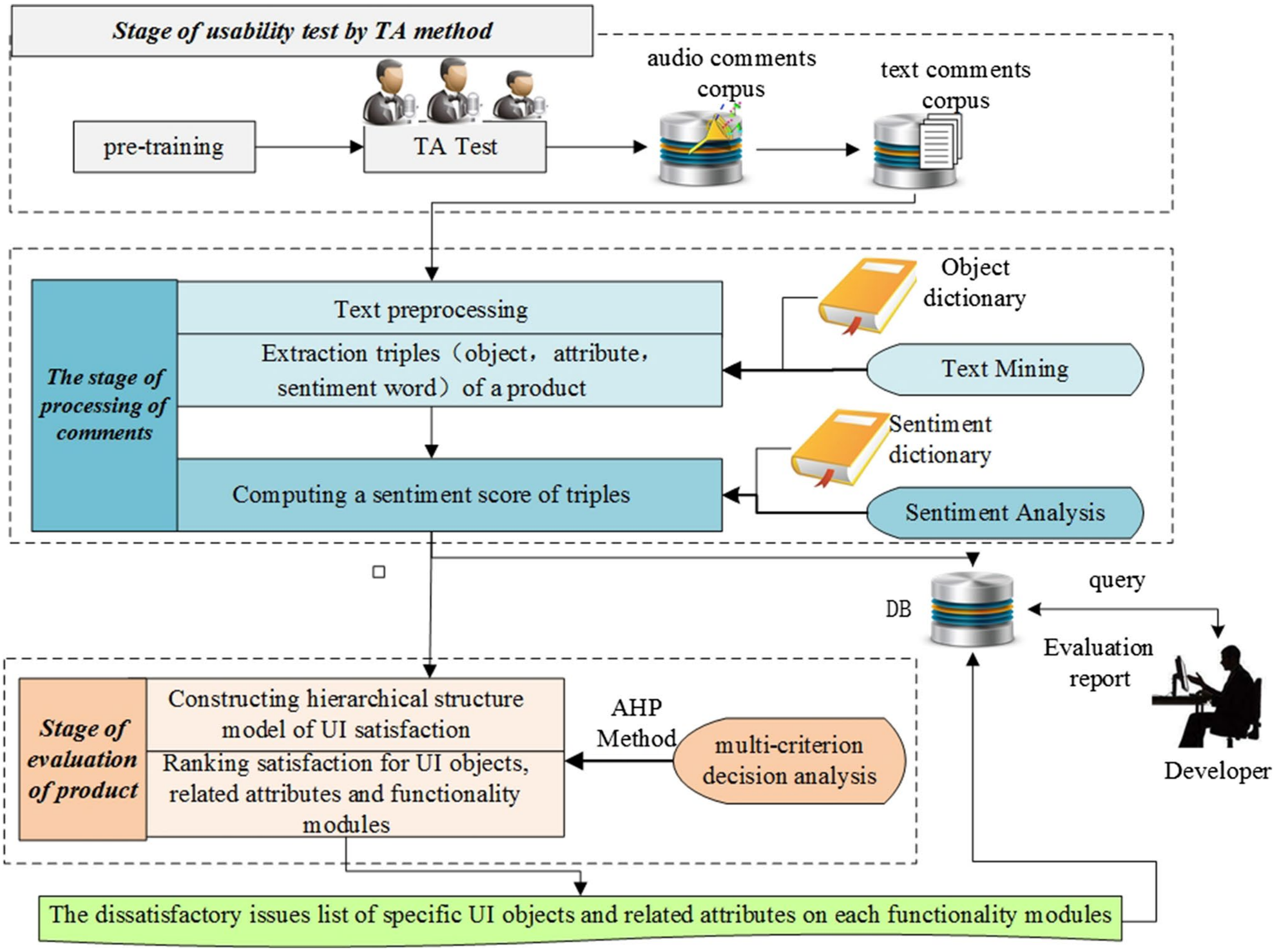
Stage of usability test by TA method
pre-training
TA Test
audio comments corpus
text comments corpus
The stage of processing of comments
Text preprocessing
Extraction triples（object，attribute，sentiment word）of a product
Computing a sentiment score of triples
Object dictionary
Text Mining
Sentiment dictionary
Sentiment Analysis
DB
query
Evaluation report
Developer
Stage of evaluation of product
Constructing hierarchical structure model of UI satisfaction
Ranking satisfaction for UI objects, related attributes and functionality modules
AHP Method
multi-criterion decision analysis
The dissatisfactory issues list of specific UI objects and related attributes on each functionality modules

Fig. 1 Framework of the proposed approach

of degree and negative word in m_k. Here O_k, A_k and S_k are essential but AD_k and NA_k are optional for m_k.

In the final phase, the AHP method is adopted to evaluate user satisfaction of different objects and their related attributes in all functional modules of the product based on the results of the SA. Developers can then have a fast sentiment orientation tracking of UI defects and improve them according to the dissatisfaction ranks from the AHP evaluation results.

3.1 Dictionary construction

According to Chinese grammar rules, the sentiment polarity of a sentence or phrase relies on its sentiment word and negative word in and the emotional intensity is related to its adverb of degree. Therefore, three dictionaries for sentiment words, negative words, and adverbs of degree were compiled for text mining and SA. Here, sentiment words (i.e. nice or bad), negative words (i.e. not), and adverbs of degree (i.e. very) were quoted from those released by Hownet[1] in 2007. According to the classifications of sentiment intensities in Hownet, we defined the emotional intensities of adverbs using three emotion coefficients which are: 3 (i.e. most), 2 (i.e. very), and 0.5 (i.e. less). For similarity, negative words such as *not* or *nor* were compiled into a negative dictionary. The overall information for each of the dictionaries is shown in Table 1.

To match keywords of UI widgets and their properties when text mining, synonyms about UI objects and their attributes are compiled into an object synonym dictionary and an attribute synonym dictionary. There are many widgets with different properties in any UI. Some of them are invisible while others (such as the carousel widget) have modern effects. In general, users will focus on the widgets in sight and pay more attention on those with eye-catching properties. As an exploratory research, the proposed approach only selected some static UI components and their visible

[1] http://www.keenage.com/html/e_index.html

Table 1 Dictionaries for comments analysis

Dictionary	Number	Description
Emotion dictionary	Positive: 4372 Negative: 4565	sentiment words and their polarities
Negative dictionary	11	Words with negative sentiment
Degree adverb dictionary	218	Words enhancing or weakening emotionalintensity and their emotion coefficients

Table 2 Terms of selected UI objects, attributes and their Chinese synonyms

Objects	Description	Chinese-Synonym
textBox (E1)	Input characters in designer-determined length	(text box), (input box)
checkBox (E2)	Offer a limited set of no-mutuality exclusive visible choices	(multi select), (multi), (check box), (multi key), (multiple), (check button), (multi button), (check button),(multi switch), (multi switch)
listBox (E3)	Offer an array of options, values or properties for single or multi-selection	(list), (list box), (pull-down list)
comboBox (E4)	Select an option from the list or enter text in the text box	(combo box), (combo box), (dropdown list)
radioButton (E5)	Allow only single choice and toggle off when clicking on another one in the set	(radio), (radio button), (radio button)
commandButton (E6)	Press it to proceed actions like submit,cancel	(press button), (press key)
scrollBar (E7)	Can be dragged (back/forth or upward /downward) with cursor, or click the array key to control the thumb's position	(scroll bar), (horizontal scroll bar), (vertical scroll bar)
icon (E8)	Be a sign and represent a significant degree of cognitive complexity	(icon), (picture), (image)
dialogBox (E9)	Support less common task and provide sets of related functionality in a well-defined container	(dialog box), (pop-up box)
menu (E11)	Afford access to execute or the selected menu option leading to a dialogue	(menu)
label (E12)	To display texts which do not need to be edited	(label), (caption)
Tab (E13)	To group functions and select it to keep track of different windows	(multipage), (tab)

attributes, whose defects are easily discovered and described by users, to simplify the evaluation task of the interviewees. From the point of view of the developers, widgets definitions have been described in user interface description languages (UIDLs), such as UIML, XIML, UsiXML. However, most of the interviewees we invited tested the UI closed to the view of actual users. Based on the UI components introduced by Heim's interaction design [30] instead of the ones defined in UIDLs and the property definitions described in the CSS,[2] the selected UI objects and related attributes are listed in Tables 2 and 3.

3.2 UI Accessing

At the initial phase of UI accessing, interviewees with basic knowledge about HCI participate in training about UI design and CTA. First, designers introduce the basic UI widgets and their conspicuous properties to interviewees with a brief PPT presentation as shown in Table 2. Second, a small game is played, wherein interviewees are praised once they select the correct UI widgets and properties, in order to arouse the interests of interviewees and strengthen their impressions. The game scores of the interviewees can serve as the basis to assess their masteries of related UI terms. Finally, the interviewees are guided to describe product UI defects by using the terms learned from the PPT and those they practiced in the game during CTA task. The final training goal is to teach the way interviewees describe their experiences in specific speech patterns.

[2] https://www.w3.org/TR/CSS/

Table 3 Terms of attributes and their Chinese synonyms

Attributes	Description	Chinese-synonym
name (F1)	An identifier is an unambiguous name for a resource	(name), (designation), (monicker)
dimension (F2)	For the properties, i.e. width, min-width, max-width, height, min-height and max-height	(size), (length), (width), (height), (size)
Positioning (F3)	For setting the properties, i.e. positioning, z-index, top, right, bottom and left...	(position), (stack), (top), (bottom),(right), (left)
color (F4)	For the properties, i.e. color and opacity...	(color), (hue), (color and lustre)
layout (F5)	For the properties, i.e. display, float, clear, visibility, clip, overflow, overflow-x and overflow-y...	(layout), (arrangement), (distribution), (drift), (visible), (transparent), (align), (appearance), (expression), (aspect), (show), (expression)
font (F6)	For the properties, i.e. font, font-style, font-variant, font-weight, font-size, font-family and font-stretch	(font), (font), (typeface), (character)
background (F7)	For the background properties of a component, i.e. background color, image, position, size...	(background), (background), (background color), (background size), (background size), (background position)
border (F8)	For the border properties of a component, i.e. border width, style,...	(frame), (border), (boundary), (limit), (round), (edge)
margin or padding (F9)	For the properties, i.e. margin, padding, margin/padding,top,margin/padding right, margin/padding bottom,margin/padding left	(margin), (padding), (blank), (margin), (margin), (space), (distance)
test (F10)	For the test properties, i.e. indent, align, transform, decoration, shadow, letter-spacing, word-spacing, word-wrapand line-height...	(test), (content), (interval), (line feed), (uppercase), (lowcase), (capital and small letter), (row height), (array pitch), (number of words), (shadow), (number of characters)

On the other hand, experimenters should guide interviewees when they experience difficulties using the product or when some clarifications are needed so as to help them proceed to the next operations. Through this guided CTA test, more audio comments about user experience with the UI can be collected with less labor consumed and more time saved. In general, audio records collected using CTA cannot be directly analyzed with ease in an automated way and would need monitoring by analysts. On the contrary, automated analysis of text corpus is well equipped with the techniques of text mining and SA. Therefore, CTA comments can also be used for automated analysis in the case of high transforming accuracy rate from audio to test. In November 2016, iFLYTEK Co. Ltd. pushed forward their Automated Speech Recognition (ASR)[3] technique innovation, which is capable of transforming Chinese audio into Chinese text at a high accuracy rate (up to 97%) in quiet environment settings. Furthermore, interviewees have been very familiar with the UI keywords after training, so that UI words can be spoken out clearly. Sentiment words, negatives and adverbs of degree in sentences are only a small part of the total user comments. Hence, the error rate of recognition for the keywords can be less than 3% when all audio records are transformed into text comments, which can be directly applied with text mining and SA for automated analysis.

3.3 Comments Analysis

With the techniques of text mining and SA, the proposed approach extracts common UI objects and attributes from text comments of different functionality modules to assess and compare their respective user satisfaction.

All comments are collected together into a comments set grouped using specific functionality modules. Therefore, comments sets $D = \{d_1, d_2, \dots, d_r\}$ are created as pre-processing corpus, where d_r is the comment set about the r-th functionality module. The main idea of comment analysis is to initially match the keywords of UI objects and related attributes. Once UI keywords are matched in a sentence, the proposed approach will try to match sentiment words based on the relationship of word co-occurrence within the same sentence or phrase [31]. As a result, automated extraction will first be processed on words of each d_r which have been parsed with Parts-Of-Speech (POS) tags. Nouns, adjectives, adverbs and negative words, within a sentence or phrase of d_r, will be mined and distinguished if they match the words in the dictionaries, which will be used to construct the 5-tuple m_k.

For each m_k, its sentiment vector SOA_k related to specific object and attribute can be deduced as a 3-tuple SOA_k in Eq. (2), and the sentiment score SC_k of SOA_k can be calculated as in Eq. (3):

$$\mathrm{SOA}_k = \{O_k, A_k, \mathrm{SC}_k\} \tag{2}$$

[3] http://www.voicecloud.cn/

$$\mathrm{SC}_k = SW_k * \mathrm{SADV}_k * \mathrm{SNEG}_k \tag{3}$$

where SOA_k is the sentiment vector of m_k; SW_k is the polarity value of sentiment word S_k which can be 1 or − 1; SNEG_k and SADV_k are the sentiment coefficients of NA_k and AD_k; SNEG_k will be set as − 1 if NA_k founded in m_k. If AD_k is found in m_k, the coefficient of SADV_k can be retrieved from the dictionary for the emotional intensity of AD_k, which is set as 0.5, 2 or 3.

Thus, the sentiment score of a specific object and its related attribute can be calculated using Eq. (4):

$$O_iA_j = \sum_{k=1}^{N} \mathrm{SC}_k \tag{4}$$

where O_iA_j is the aggregate sentiment score for O_i and attribute A_j in all m_k. So the sentiment matrix (SM_r) of d_r based on each O_i and A_j can be constructed as Eq. (5):

$$\mathrm{SM}_r = \begin{bmatrix} O_1A_1 & O_1A_2 & \cdots & O_1A_{j-1} & O_1A_j \\ O_2A_1 & O_2A_2 & \cdots & O_2A_{j-1} & O_2A_j \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ O_{i-1}A_1 & O_{i-1}A_2 & \cdots & O_{i-1}A_{j-1} & O_{i-1}A_j \\ O_iA_1 & O_iA_2 & \cdots & O_iA_{j-1} & O_iA_j \end{bmatrix} \tag{5}$$

The aggregate sentiment matrix (SM) of comments set D for all functionality modules can be computed as summarized in Eq. (6):

$$\mathrm{SM} = \sum_{r=1}^{N} \mathrm{SM}_r \tag{6}$$

Obviously SM_r is a non-rectangular sparse matrix because not all O_i and their related A_j have been assessed by the interviewees. As already discussed, those UI objects and their related attributes, which are mostly dissatisfaction, will impact the satisfaction of interviewees the most. So, a sentiment threshold (γ) is defined to filter the O_iA_j with value over than γ and set O_iA_j as 0 in SM.

When text mining is processed, the sentences without UI keywords (UI object and attribute) will be skipped. On the other hand, it is impossible that all sentiment words are already stored in the sentiment dictionary before the CTA test. The resolution used in the proposed approach is to calculate the sentiment polarity of adjective in the current sentence if UI keywords have been matched but there are no sentiment words in the sentence. By using the method proposed by Ge and Li [32], sentiment orientation (SW_k) of the rest of the Chinese adjective words in this sentence can be identified by semantic similarity computation based on Hownet, which computes sememe (sentiment word) similarity by employing the sememe tree's depth and density information and combining the description structure of relation sememe and relation symbol with different weights, is given in Eq. (7):

$$\mathrm{weight}(k) = \frac{2 \times (\mathrm{depth} - k)}{\mathrm{depth} \times (\mathrm{depth} + 1)} \tag{7}$$

where weight (k) is the function to calculate the weight of the k-th edge and depth is the height of current sememe in the hierarchical tree. The density information of the sememe can be calculated as Eq. (8):

$$f(w) = \frac{connNodes}{totalNodes} \tag{8}$$

where *connNodes* is the connected node counter of the current sememe (including current sememe itself) and *totalNodes* is the total counter of the sememe tree. Here we used the Least Common Node (LCN) to represent the least common parent node of two sememes in tree. Therefore the similarity of these two sememes can be deduced as shown in Eq. (9)

$$\begin{aligned} simSememe(a,b) = {} & \tfrac{1}{2} \times \tfrac{\alpha}{\alpha + \sum_{i=1}^{n} \mathrm{weight}(\mathrm{level}(i))} \\ & + \tfrac{1}{2} \times \tfrac{2 \times \log f(LCN)}{\log f(a) + \log f(b)} \end{aligned} \tag{9}$$

where level(i) is the function to compute the level of the i-th edge in the sememe tree and α is 1.6 [32]. So the polarity of the new adjective, which is not included in the sentiment database, should be aligned to the one of the sentiment words which is of maximum similarity with Eq. (9), and its value is equal to or over the similarity threshold β ($\beta = 0.5$). Thus, the polarity of this adjective can be determined and the sentiment score (SC_k) can be calculated using Eq. (3).

If the identified sentiment orientation is neutral, i.e. the maximum similarity of adjective is less than β, the SC_k of this sentence will be zero. In the meantime, the identified Chinese adjectives with positive or negative polarity will be viewed as a new sentiment word added into the sentiment dictionary for automated update in order to improve the matching efficiency.

3.4 Comprehensive evaluation of satisfaction using AHP

In our approach, UI satisfaction is the final evaluation goal of AHP, which depends on the satisfaction of all the widgets (i.e. commandbutton, textbox) on the UI. In addition, the user satisfaction of all these widgets relies on the feeling of their properties (i.e. color, size); thus, the satisfaction related to the properties of widgets (O_iA_j) can be used as criteria in the middle level of the hierarchy structure model. In the end, actual user satisfaction of each module (SM_r), based on

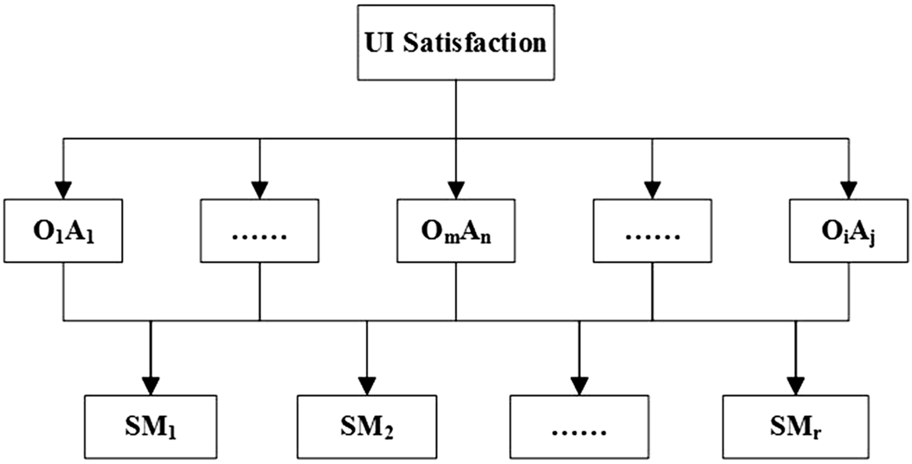

Fig. 2 Hierarchical structure mode of UI satisfaction

the satisfaction of their corresponding widgets and related properties, are placed on the bottom level as alternatives for qualitative and quantitative analyses. In brief, the AHP hierarchy structure for UI satisfaction is depicted in Fig. 2.

There are many UI criteria that affect user satisfaction and it is unnecessary to have a comprehensive evaluation taking into account all of these factors. In most cases, the UI criteria that users dislike or are mostly dissatisfied with will have a strong impact on overall user satisfaction during assessment [23]. It is therefore a key point in satisfaction evaluation to select the major criteria which users are most concerned about. In general, criteria selection and weight balancing in the traditional AHP method are decided by experts based on their subjective experiences and judgments, which would relatively risk the objectivity of the evaluation process. Thus, the criteria selection and weight balancing are determined according to their actual user satisfaction (O_iA_j) instead of empirical judgments from experts in order to improve the objectivity of AHP assessment.

During the CTA test, the interviewees were asked to speak out their experiences when using the product and their comments were mostly related to negative words which make most O_iA_j less than 0. In other words, the lower the value of O_iA_j is, the more dissatisfaction the interviewees felt and the more attention the interviewees paid to attribute A_j of object O_i. Since identifying usability problems is the main characteristic of CTA, most user comments are related to negative feedback on product defects. Thus, this study focuses on analyzing negative comments to assess user dissatisfaction. Owing to this assessing purpose, a sentiment threshold (γ) is defined to filter the sentiment vectors where sentiment scores are over γ to skip lower dissatisfaction.

Therefore, the criteria set C and its related importance set S are defined as follows: Let $C = \{c_1, c_2, c_k, \dots, c_n\}$ and $S = \{s_1, s_2, s_k, \dots, s_n\}$, where c_k is the k-th element of sentiment matrix SM and its related sentiment score s_k is less than γ. So, the relative importance a_{ij} between c_i and c_j can be defined as $a_{ij} = s_i/s_j$. The pairwise comparison matrix A, constructed by the relative importance a_{ij}, is given as Eq. (10):

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{10}$$

The satisfaction rank of each criterion can be deduced from the largest eigenvector of A. Its consistency radio (CR) must be less than 0.1, otherwise matrix A needs to be adjusted to avoid inconsistency of comparisons among its elements.

4 Empirical Case Study

To verify and validate the proposed approach, an empirical case on the student information management system for the Southwestern University of Finance and Economics (SWUFE), which at the time of writing was still under development, was conducted to evaluate their satisfaction with the proposed approach for the UI and with a questionnaire survey. Four major management modules of the software which include *message*, *event*, *vacation* and *notification* were selected for assessment.

4.1 Satisfaction test using CTA

In the CTA test phase, ten computer science postgraduate students were invited as interviewees to evaluate the UIs of the management modules after some training on UI design. Every interviewee had processed a 1-min CTA test on each module and described his/her experience by using keywords in the dictionaries as much as they could. In addition, with reference to the idea of using templates with NLP [21], two patterns, according to the traditional Chinese grammar, were

Table 4 Test comments collected through TA method

Module names	Total words
Message	3809
Event	2727
Vacation	3007
Notification	2422

used in text mining and SA for finer granularity. Therefore, interviewees were suggested to describe their experiences in the following two patterns as:

1. ⋯ + **UI object** + ⋯ + **attribute** + ⋯ + [negative] + ⋯ + [adverb of degree]+ ⋯ + [negative] + ⋯ + **sentiment word** + ⋯ + [negative] +⋯ + [adverb of degree] + ⋯ + [negative] + ⋯
2. ⋯ + [negative] + ⋯ + [adverb of degree] + ⋯+ [privative] + ⋯ + **sentiment word** + ⋯ + [negative] + ⋯+ [adverb of degree] + ⋯ + [negative] + ⋯ + **UI object** + ⋯ + **attribute** +⋯

On the above Chinese speech patterns, UI object, attribute and sentiment word in bold are the essential while the rest are optional. In particular, if the UI object and its related attribute had been found in a sentence though without any sentiment word, the polarity of the adjective word in this sentiment will be identified using the method of semantic similarity computation based on Hownet [32]. If the polarity of the adjective word is not neutral, this adjective word would be added into the sentiment dictionary as a new sentiment word. As the result, the current sentence or phrase would match one of the above patterns. Besides, the proposed approach can also be applied to other languages (i.e. German or English), of which dictionaries should be constructed for UI objects, attributes, sentiment words and so on. The speech patterns should also be adapted to the language used by the interviewees.

Following the CTA test, 10-min audio comments were collected from each functionality module and then a total of 40 min of audio comments were transformed into text corpus using ASR. The text transformation results of each module are summarized in Table 4:

4.2 Text Comments Analysis

Using Stanford parser,[4] POS tags were assigned to all words based on the contexts, which were utilized to locate different types of words in text comments. Then, the tagged words which matched with keywords in the dictionaries, would be constructed together into a 5-tuple which would be inserted into database in the meantime. After text mining, 78 5-tuples were extracted from the module *message*, 57 from the module *event*, 60 from the module *vacation*,and 54 5-tuples were extracted from the module *notification*. Then, the proposed approach queried the sentiment vectors from the database by filtering the scores greater than the sentiment threshold γ ($\gamma = -4$), which represent vectors with less dissatisfaction. The reason for using γ is that only the major dissatisfied sentiment vectors were expected to be analyzed and the rest of the vectors with minor dissatisfaction were skipped. From the total 249 extracted 5-tuple sets, there were 70 valid summarized sentiment vectors less than γ grouped using UI objects (E_i) and their related attributes (F_j). For each valid sentiment vector, its involved interviewee, module, object, attribute, sentiment score, and original sentence were stored into a database for searching and further statistical analysis. Because the aggregate sentiment scores of vectors were all negative, the signs of the scores were flipped for convenient calculation. In addition, the elements in the matrix without sentiment scores were set as 0. The sentiment matrix is presented in Table 5.

Table 5 Summarized Sentiment Matrix

	E_1	E_6	E_8	E_{19}	E_{10}
F_1	0	10	0	0	0
F_2	14	0	0	7	0
F_3	4	16	11	0	11
F_4	0	27	0	5	14
F_5	16	45	0	0	7
F_6	10.5	0	20	0	0
F_7	0	21	0	0	9
F_{10}	6	0	0	0	0
SUM	50.5	119	33	12	30

4.3 Satisfaction Comprehensive Evaluation

As shown in Table 5, there were 17 valid sentiment vectors with aggregate scores larger than 4 ($-\gamma$), which could be used as criteria for judgment. Since psychologists think that it is inappropriate for AHP when the counter of criteria in a same level exceeds 9, the criteria level was divided into two sub-levels: the UI objects level and their related attributes level. Thus, the hierarchical structure model of UI satisfaction is depicted in Fig. 3.

Using the square root method, the eigenvector (W) of each level was calculated as presented in Table 6, and their *CRs* were all less than 0.1, which satisfied the consistency ratio checking.

From the eigenvector (W) of level B, the dissatisfaction rank order of UI objects was *commandButton* > *textBox* > *icon* > *hyperLink* > *dialogBox*, which easily implies that users were mostly dissatisfied with the *commandButton*,

[4] http://nlp.stanford.edu/software/segmenter.shtml

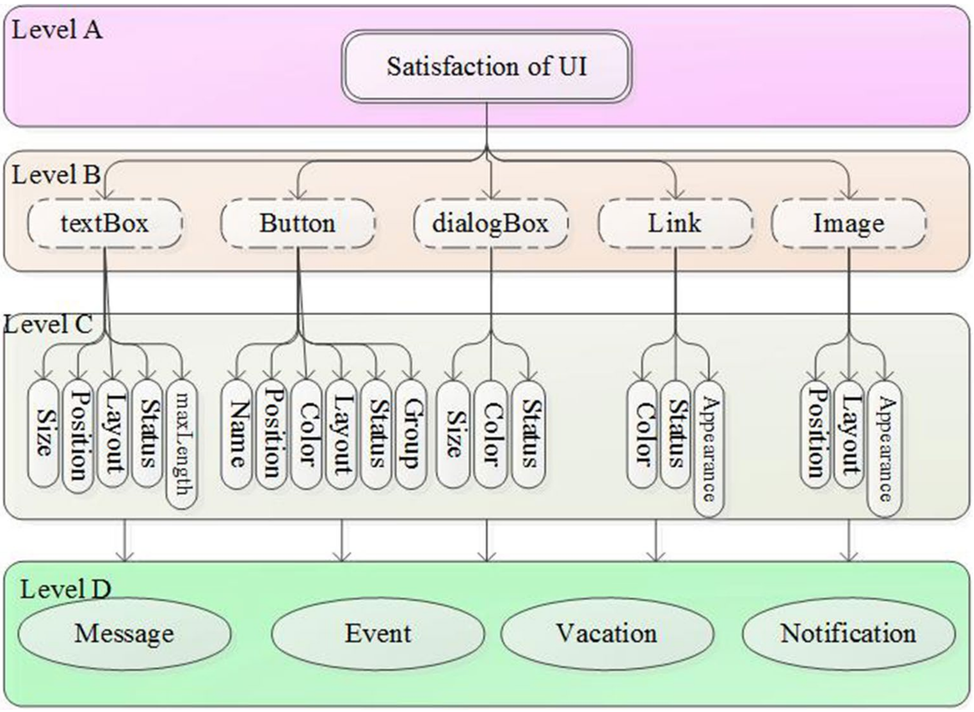

Fig. 3 Hierarchical structure mode of empirical case study

Table 6 Elements sets and their total sequencing of each level

Level	Criteria set	Eigen vector (W)
B	E_1, E_6, E_8, E_9, E_{10}	$(0.1278, 0.0495, 0.4907, 0.2082, 0.1237)^T$
C	(E_1, F_2), (E_1, F_3), (E_1, F_5), (E_1, F_6), (E_1, F10), (E_6, F_1), (E_6 , F_3), (E_6, F_4), (E_6, F_5), (E_6, F_7), (E8, F_3), (E_8, F_6), (E_9, F_2), (E_9, F_4), (E_{10}, F_4), (E_{10}, F_5), (E_{10}, F_7)	$(0.0577, 0.0165, 0.066, 0.0433, 0.0247, 0.0412,0.066, 0.1113, 0.1856, 0.0866, 0.0453, 0.0825, 0.0289, 0.0206, 0.0577, 0.0289, 0.0371)^T$
D	Message, Event, Vacation, Notification	$(0.3515, 0.3826, 0.1675, 0.0984)^T$

followed by *textBox*. From the maximum eigenvector of level C, the dissatisfaction rank order in Table 6 for specific pairs of UI objects and their related attributes was (*commandButton, layout*) > (*commandButton, color*) > (*commandButton, background*) > (*textBox, layout*) > (*commandButton, positioning*), (*textBox, dimension*) > (*hyperLink, color*), (*icon, positioning*) > (*textBox, font*) > (*commandButton, name*) > (*hyperLink, background*) > (*dialogBox, dimension*), (*hyperLink, layout*) > (*textBox, test*) > (*dialogBox, color*) > (*textBox, layout*). Thus, developers could go over their designs and improve the defects of specific UI pairs with high dissatisfaction, such as *commandButtons* and their *layouts*, in an assembly line work to increase the efficiency of design reworking.

In a similar way, the dissatisfaction rank order of level D for functionality modules was *vacation* > *event* > *message* > *notification*, and developers could arrange their efforts to improve UI design in this order. Based on these three dissatisfaction rank orders of all levels in the hierarchical structure, a list for specific dissatisfaction of UI objects and their related attributes querying from the database group by 4 modules was displayed in the format: (**module**, (**object, attribute**): *user comment*). Here are some specific dissatisfaction issues on different functionality modules listed for developers

- *Message, (Button, Layout)* In addition, the layouts of two commandButton for sending and posting picture are not good.
- *Event, (Radio-Button, layout)* Moreover, there is no layout format on radioButton.
- *Vacation, (Button, Color)* Command button for submitting vacation request does not align with the other command buttons at the same level.
- *Notification, (Button, Layout)* I see the background of quick response button with improper color.

With the use of the SQL technique, any comments were easily queried by fields of involved object, attribute, frequency, interviewee or score beyond the threshold from the

Table 7 Matching rate of the top dissatisfaction attributes from questionnaire and proposed approach

Module	top3 (%)	top4 (%)	top5 (%)
Message	66.7	75	80
Event	33.3	50	60
Vacation	33.3	75	80
Notification	0	25	60

Table 8 Top3 dissatisfaction attributes from the questionnaire

Module name	top3 dissatisfaction UI objects and related attributes
Message	(E_2, DF_5), (E_9, F_{11}), (E_1, F_5)
Event	(E_2, F_2), (E_8, F_{11}), (E_9, F_{11})
Vacation	(E_1, F_5), (E_2, F_5), (E_3, F_5)
Notification	(E_{13}, F_8), (E_1, F_5), (E_3, F_6)

Table 9 Top3 dissatisfaction attributes from the proposed approach

Module	top3 dissatisfaction UI objects and related attributes
Message	(E_6, F_5), (E_{10}, F_4), (E_8, F_3)
Event	(E_6, F_5), (E_6, F_3), (E_6, F_4)
Vacation	(E_6, F_4), (E_8, F_5), (E_6, F_7)
Notification	(E_1, F_2), (E_6, F_4), (E_6, F_7)

database, which was helpful to designers in the improvement of their UI design.

To verify the effectiveness of the proposed approach, a questionnaire survey was used to evaluate the same functionality modules. An additional 22 computer postgraduates and UI designers were invited to give scores for all UI objects (O_k) and their related attributes (A_k) using the evaluation scale used by Kang and Park [23] as follows: 2: excellent; 1: good; 0: neutral; − 1: bad; and − 2: terrible.

Based on the design principle of consistency, UI components should be shown in a similar design style and bring about similar experiences to users. Thus, the dissatisfaction objects and related attributes from the questionnaire and proposed approach should both correspond to this rule. By comparing the sentiment vectors of the questionnaire survey with that of the proposed approach, the UI attributes with most dissatisfaction in these two sentiment vector sets were similar. In particular, the 5 attributes with most dissatisfaction in the two sets were almost the same. Table 7 shows the matching rates of most 3, 4 and 5 dissatisfaction attributes, demonstrating that both methods could achieve similar overall user impressions.

Table 10 Performance comparison between proposed approach and questionnaire survey

Module	Our method (%)	Questionnaire survey
Message	74.0	33.9
Event	73.5	28.2
Vacation	78.0	28.2
Notification	72.4	25.2
Summary	74.5	28.9

However, the objects and their related attributes, with most dissatisfaction, were entirely different between Tables 8 and 9. With the exception of the alternatives of testers, the difference of thinking methods in the two approaches was the primary cause leading to such inconsistent results. Normally, survey questionnaires require testers to fill or select their answers by retrospection from long term memory, which would make them easily miss some detailed information (i.e. some specific UI objects and their attributes). The proposed approach with CTA, however, can record instant and detailed user experiences, which are suitable for more detailed and specific evaluation compared to survey questinnaire.

To validate the above two methods, three usability experts and three senior UI designers respectively checked the dissatisfaction user comments from the proposed approach and the dissatisfaction options from the questionnaire survey. From the dissatisfaction items of the two methods, they respectively identified them corresponding to the defects in the UI design that need to be improved. Therefore, the evaluation accuracy is the fraction of dissatisfaction items correctly identified as design defects among all dissatisfaction items. The comparative performances of the two different methods are shown in Table 10.

The comparison shows that the accuracy of the proposed approach was more than twice that of the questionnaire survey. The empirical case, aims to identify the feelings of specific widgets and their specific properties and then evaluates their satisfaction. To find out the root cause leading to these differences, a discussion with interviewees and testers was made. Uncovering two reasons for these diversities. First, from the view of physiology, information retrieved from short-term memory is more accurate than that retrieved from long-term memory in the human brain. The testers who took part in the questionnaire survey said that it was difficult to clearly recall the feeling of specific widgets and their specific properties on specific modules after completing the test task. For example, it is hard for testers to give scores to the layouts of text boxes on the module *Message* after completing the task. In contrast, interviewees responded that they were relaxed when asked to simultaneously describe their feelings when executing the task. Second, from a psychological point of view, testers were hedged with rules and regulations of the questionnaire survey which easily made psychological inversions in them. In the drafted survey, testers were asked too many similar questions regarding the feelings of some widgets and their properties, which were very tedious and tiresome. Testers needed to laboriously recall their impressions for each question one by one and later felt bored and filled in their answers halfheartedly. Interviewees in the CTA task, on the other hand, felt more freedom, fun, and relaxation.

In particular, the accuracy of UI objects and their related attributes, with sentiment scores of less than γ in the proposed approach, were higher than 90%. Moreover, the more dissatisfaction with UI objects and related attributes users felt, the more likely it was to indentify defects in the UI. Thus, the dissatisfaction ranks computed using the AHP method, which took advantage of actual user data to select criteria and determine their related weights, were really helpful for UI designers in the improvement of their design. Moreover, they could rapidly and efficiently locate and look up the corresponding specific user comments of the objects and attributes from the database of the proposed approach and meet customer demands more efficiently.

5 Conclusions

In this study, an automated comprehensive evaluation approach was applied for the user satisfaction of UI. The techniques of text mining and SA were used to analyze the user comments collected using the CTA. Based on the results of SA, the AHP method was adopted for the evaluation of user's satisfaction. The evaluation comparisons of the proposed approach and that of questionnaire survey showed that the proposed approach was more valid and had higher performance than the questionnaire in finding user dissatisfaction which mostly correspond to defects in the UI design. We think that our approach could be used as an automated framework for handling any type of comments. However, there are still some limitations that should be addressed. First, concentrations of interviewees are easily affected by rigid speech patterns. Second, the words of dictionaries, especially the words of sentiment dictionary based on Hownet, are still incomplete and limited. Third, it costs great effort to learn a new sentiment word whenever the proposed approach tries to process a semantic similarity computation. Fourth, it is often hard to have a one-to-one correspondence between actual UI objects, related attributes, and the specific words described in user comments. In future studies, enhanced text mining with the NLP should be used for untying the speech constraint of interviewees in order to make users pay more attention on sharing their experiences. Furthermore, sentiment word dictionary can be updated by pre-processing semantic similarity computation in online comments for learning additional new sentiment words. In the meantime, key UI words of design documents user manuals, and so on can be utilized as other sources to expand the terms for the dictionaries of UI objects and attributes. Recently, Eye tracking has been applied to improve usability test moderation using triggered TA protocols [33]. It may be revolutionary for further studies to set up accurate connections between the UI elements in which interviewees have concentrated and the feeling they described to enable interviewees to conveniently speak out their experiences without any speech patterns. To extend automated UI evaluation, the approach will propose another novel way in a future study, by performing a static analysis on source code against usability and accessibility guidelines of website [34].

Acknowledgements We would like to thank the anonymous interviewees and reviewers for their detailed and insightful comments. This research work has been partially supported by the Supporting Programs of Science and Technology Department of SiChuan Province of China Grants, Nos. 2013GZX0138, 2014GZ0063.

References

1. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: CHI 00 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 345–352 (2000)
2. Calleros, J.M.G., Garca, J.G., Vanderdonckt, J.: Advance human machine interface automatic evaluation. Univers. Access Inf. Soc. **12**(4), 387–401 (2013)
3. Whitefield, A., Wilson, F., Dowell, F.: A framework for human factors evaluation. Behav. Inf. Technol. **10**(1), 65–79 (1991)
4. Davids, M.R., Chikte, U.M., Halperin, M.L.: Effect of improving the usability of an e-learning resource: a randomized trial. Adv. Physiol. Educ. **38**(2), 155–60 (2014)
5. Pinelle, D., Wong, N., Stach, T.: Heuristic evaluation for games: usability principles for video game design. In: CHI 13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1453–1462 (2008)
6. Grigoreanu, V., Mohanna, M.: Informal cognitive walkthroughs (ICW): paring down and pairing up for an agile world. In: CHI 13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3093–3096 (2013)
7. Schoonewille, H.H., Heijstek, W., Chaudron, M.R.V.: A cognitive perspective on developer comprehension of software design documentation. In: SIGDOC 11 Proceedings of the 29th ACM International Conference on Design of Communication, PP. 211–218 (2011)
8. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco (1995)
9. Cooke, L.: Assessing concurrent think-aloud protocol as a usability test method: a technical communication approach. IEEE Trans. Prof. Commun. **53**(3), 202–215 (2010)
10. Peute, L.W.P., Keizer, N.F.D., Jaspers, M.W.M.: The value of retrospective and concurrent think aloud in formative usability testing of a physician data query tool. J. Biomed. Inf. **55**, 1–10 (2015)

11. Stefano, F., Borsci, S., Stamerra, G.: Web usability evaluation with screen reader users: implementation of the partial concurrent thinking aloud technique. Cognit. Process. **11**(3), 263–272 (2010)
12. Ivory, M.Y., Hearst, M.A.: State of the art in automating usability evaluation of user interfaces. ACM Comput. Surv. **33**(4), 470–516 (2001)
13. Lewis, C.H.: Using the thinking-aloud method in cognitive interface design. Research report RC-9265, IBM (1982)
14. Ericsson, K.A., Simon, H.A.: Protocol analysis: verbal reports as data. J. Market. Res. **23**(3), 381–423 (1986)
15. Boren, T., Ramey, J.: Thinking aloud: reconciling theory and practice. IEEE Trans. Prof. Commun. **43**(3), 261–278 (2000)
16. Hertzum, M., Hansen, K.D., Andersen, H.H.K.: Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? Behav. Inf. Technol. **28**(2), 165–181 (2009)
17. Olmsted-Hawala, E.L., Murphy, E.D., Hawala, S., Ashenfelter, K.T.: Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: Proceedings of 28th International Conference on Human Factors in Computing Systems, PP. 2381–2390 (2010)
18. Anderson, E.W.: Customer satisfaction and word of mouth. J. Serv. Res. **1**(1), 5–17 (1998)
19. Hu, N., Pavlou, P.A., Zhang, J.: Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online word-of-mouth communication. In: EC 06 Proceedings of the 7th ACM Conference on Electronic Commerce, PP. 324–330 (2006)
20. Hedegaard, S., Simonsen, J.G.: Extracting usability and user experience information from online user reviews. In: CHI'13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2089–2098 (2013)
21. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD 04 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
22. Wu, M., Wang, L., Li, M., Long, H.: An approach of product usability evaluation based on Web mining in feature fatigue analysis. Comput. Ind. Eng. **75**(1), 230–238 (2014)
23. Kang, D., Park, Y.: Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach. Expert Syst. Appl. **41**(4), 1041–1050 (2014)
24. Arora, C., Sabetzadeh, M., Briand, L., Zimmer, F.: Automated checking of conformance to requirements templates using natural language processing. IEEE Trans. Softw. Eng. **41**(10), 944–968 (2015)
25. Saaty, T.L.: Decision making with the analytic hierarchy process. Int. J. Serv. Sci. **1**(1), 83–98 (2008)
26. Saaty, T.L.: Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process, Revista de la Real Academia de Ciencias Exactas, Fsicas y Naturales. Serie A. Matemticas, Vol. 102, No. 2, September 2008, pp. 251–318 (2008)
27. Matsuda, M., Uesugi, Y., Nonaka, T., Hase, T.: Design method of UI of AV remote controller based on AHP. In: Proceedings of 26th IEEE International Conference on Consumer Electronics, pp. 1–2 (2008)
28. Delice, E.K., Güngör, Z.: The usability analysis with heuristic evaluation and analytic hierarchy process. Int. J. Ind. Ergon. **39**(6), 934–939 (2009)
29. Albayrak, E., Erensal, Y.C.: Using analytic hierarchy process (AHP) to improve human performance: an application of multiple criteria decision making problem. J. Intell. Manuf. **15**(4), 491–503 (2004)
30. Heim, S.: The Resonant Interface: HCI Foundations for Interaction Design. Addison-Wesley Longman Publishing Co., Inc., Boston (2007)
31. MA, Y.H., Wang, Y.C., SU, G.Y., Zhang, Y.M.: A novel Chinese text subject extraction method based on character co-occurrence. J. Comput. Res. Dev. **40**(6), 874–878 (2003)
32. GE, B., Li, F., Guo, S., Tang, D.: Word s semantic similarity computation method based on Hownet. Appl. Res. Comput. **27**(9), 3329–3333 (2010)
33. Freeman, B.: Triggered think-aloud protocol: using eye tracking to improve usability test moderation, In: CHI 11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1171–1174 (2011)
34. Vanderdonckt, J., Beirekdar, A.: Automated web evaluation by guideline review. J. Web Eng. Arch. **4**(2), 102–117 (2005)