FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Katri Säily

FEEDBACK AGGREGATION IN CROWD FEEDBACK SYSTEMS

Bachelor's Thesis
Degree Programme in Computer Science and Engineering
February 2021

Säily K. (2021) Feedback aggregation in crowd feedback systems. University of Oulu, Degree Programme in Computer Science and Engineering. Bachelor's Thesis, 28 p.

ABSTRACT

The purpose of this literature review is to research the way different crowd feedback systems aggregate and visualize their data for the user. First the concept of crowdsourcing for design purposes is introduced as well as four different crowd feedback systems, which are Voyant, CrowdCrit, Decipher and Paragon. Crowdsourcing means giving a task for a crowd of people to perform, usually online. Crowdsourcing is often used when there is a need for a large amount of responses because of its low cost compared to other methods. Crowd feedback systems use crowdsourcing to achieve their goal that is collecting feedback from a crowd.

For a crowd feedback system to provide value into the design process, they should not only collect feedback but also convey the collected data to the designer in an informative but also easily understandable manner. This requires that the system provides support for non-experts for them to give feedback in a professional manner. The results of this thesis give an insight into how crowd feedback systems differ from each other.

The results showed that different crowd feedback systems collect and present their feedback in very different ways. Voyant and CrowdCrit both visualize feedback using visual markers and stacked bar charts, but Voyant also uses word clouds for this purpose. Decipher shows whether the feedback is considered negative or positive and what the feedback provider had to say about the design. Paragon presents collected feedback with the help of examples that the feedback provider has chosen to help describe their feelings about the design. Voyant and CrowdCrit were eventually considered to be the most visually pleasing of these four crowd feedback systems. The way Voyant aggregated its feedback was seen more versatile but CrowdCrit collected feedback in a way that provided more useful feedback from non-experts.

Keywords: crowdsourcing, crowd feedback systems, feedback visualization

Säily K. (2021) Palautteen koostaminen joukkoistamisen palautejärjestelmissä. Oulun yliopisto, tietotekniikan tutkinto-ohjelma. Kandidaatintyö, 28 s.

TIIVISTELMÄ

Tämän kirjallisuuskatsauksen tarkoitus on tutkia, millä tavoin erilaiset joukkoistavat palautejärjestelmät koostavat ja visualisoivat keräämänsä palautteen käyttäjälle. Ensin esitellään joukkoistamisen rooli suunnittelussa ja sen myötä myös neljä palautejärjestelmää, jotka ovat Voyant, CrowdCrit, Decipher ja Paragon. Joukkoistamisella tarkoitetaan tehtävien antamista joukolle ihmisiä, yleensä verkossa. Joukkoistamista käytetään usein, kun tarvitaan iso määrä palautetta, johtuen sen käytön edullisuudesta verrattuna muihin metodeihin. Joukkoistamisen palautejärjestelmät hyödyntävät joukkoistamista saavuttaakseen tavoitteensa, joka on palautetteen kerääminen joukolta ihmisiä.

Jotta palautejärjestelmä voisi tuoda lisäarvoa suunnitteluprosessiin, täytyy sen palautteen keräämisen lisäksi myös esittää saatu data käyttäjälleen informatiivisessa, mutta myös helposti ymmärrettävässä muodossa. Tämä vaatii, että palautejärjestelmä tukee jollain tavalla ei-asiantuntijoita, jotta he voisivat antaa palautetta asiantuntevalla tavalla. Tämän kandidaatintyön tulokset antavat käsityksen siitä, miten joukkoistamisen palautejärjestelmät eroavat toisistaan.

Tulokset osoittivat, että eri joukkoistamisen palautejärjestelmät keräävät ja esittävät keräämänsä palautteen hyvin eri tavoilla. Voyant ja CrowdCrit visualisoivat palautteen visuaalisten markkereiden ja pinottujen pylväsdiagrammien avulla, mutta Voyant käyttää myös sanapilviä tähän tarkoitukseen. Decipher ilmoittaa, onko palaute nähty positiivisena, negatiivisena vai neutraalina ja mitä mieltä palautteen antaja on ollut designista. Paragon esittää keräämänsä palautteen esimerkkikuvien avulla, jotka palautteenantaja on valinnut kuvaamaan tuntemuksiaan. Lopulta Voyant ja CrowdCrit nähtiin visuaalisesti miellittävimpinä näistä neljästä palautejärjestelmästä. Voyantin tapa koostaa palaute koettiin monipuolisempana, mutta CrowdCrit keräsi palautetta tavalla, joka tuotti hyödyllisempää palautetta ei-asiantuntijoilta.

Avainsanat: joukkoistaminen, palautejärjestelmät, palautteen visualisointi

TABLE OF CONTENTS

FOREWORD

I would like to thank Jonas Oppenlaender for supervising my work and providing me with an interesting subject for this thesis. Working on this made me think a lot about my interests in the field of Computer Science and this felt very suitable for my future goals.

Oulu, 9.2.2021

Katri Säily

ABBREVIATIONS

ACM	Association for Computer Machinery
MTurk	Amazon Mechanical Turk

1. INTRODUCTION

Crowdsourcing has brought us means to collect opinions and views from a group of people by using tools such as Amazon Mechanical Turk (MTurk), where workers can choose which tasks they want to contribute in. This information can further be used for example by entrepreneurs, software developers or data scientist. Crowd feedback systems make it easier for today's designers to receive feedback on how to improve their design according to the opinions of a crowd. There are some studies on crowd feedback systems but very few concerning the actual feedback aggregation and none that really compare how different systems end up aggregating and visualizing the feedback. This Bachelor's Thesis will focus on the way crowd feedback systems collect feedback, aggregate it and how this information is visually expressed to the designer.

For clarity, this paper will refer to people who review the design and give feedback as 'feedback providers' and the people who have submitted the designs and need the feedback as 'designers'.

2. METHOD

This Bachelor's Thesis was conducted as a literature review and the goal was to find different crowd feedback systems and compare the way they aggregate and visualize feedback. To limit the amount of search results, the main source material for this research was obtained from ACM Digital Library. The focus on choosing which studies to include was on how they describe the feedback systems functionalities and whether it shows how feedback is aggregated and visualized.

The first search with the following search string:

crowdsourcing AND (feedback OR critique) AND design,

came up with 3 970 results. Next, a filter was added to only show research articles and this narrowed the search down to 3 155 results. It was noticed that based on the abstract and title, the first result, CrowdCrit [5], already seemed like a good start, so that paper was saved to be looked at later. To achieve a smaller amount of search results, some keywords were added to the previous search so the results would reveal actual feedback systems:

crowdsourcing AND (feedback OR critique) AND design AND "feedback system".

This search gave 51 results, which was manageable. When looking through the abstracts and titles of these articles, it was noted that many of these papers focused on software rather than design so "NOT software" was added to the previous search criteria, reducing the results to 23. A look through these articles and especially the figures in them revealed which ones portray the feedback aggregation interface and concentrate on the design aspect. Two systems, Voyant [4] and Critiki [11], stood out but a closer read-through revealed that the study concerning Critiki did not portray in which form the results are visualized for the designer and this paper was therefore not included in this thesis.

ACM Digital Library has a 'suggested articles' function, which first revealed that there was further research made of both CrowdCrit and Voyant so those articles [6, 9, 12] were also added as supporting sources. A search using the names of these systems also provided the same articles and confirmed that these systems are very suitable for this thesis. Also, all these papers had one common author, Brian P. Bailey, so a search was made with the following search string, attempting to narrow the search to systems that visualize feedback:

feedback AND design AND visualiz*.

After adding a filter for the author to be Brian Patrick Bailey, this search provided 24 results. Some papers were concerning the crowd feedback systems already known [5, 12], but also some new papers [10, 13], were found with this search. Other results from this search did not fit the criteria of this thesis based on their abstracts and titles. Decipher [10] ended up being the only one fitting the criteria, since the other paper [13] focused more on the effect of reflection in the design process and did not portray well enough how the feedback system works.

Initially Voyant, CrowdCrit and Decipher were chosen as the main systems to look at but to broaden the scope, a search was made to find another system to study. It was noted that another author had participated in the studies of Voyant [4] and CrowdCrit [5, 9] as well as the system that got cut earlier [13]. One of the previous searches yielded 3 970 search results:

crowdsourcing AND design AND (feedback OR critique),

but by adding a filter for the author to be Steven Dow, the results were cut down to 30. This search revealed another potential system, Paragon [14], which turned out to be

suitable for this literature review and it also provided a different perspective on what type of feedback crowd feedback systems can collect, since it is not always in the form of written text.

In the end it was decided that these four systems, CrowdCrit [5, 9], Voyant [4, 6, 12], Decipher [10] and Paragon [14] would be the focus points of this thesis.

3. RELATED WORK

3.1. Crowdsourcing

To understand how crowd feedback systems work, one must first grasp the concept of crowdsourcing. Crowdsourcing was first introduced to the public in an article by Jeff Howe (2006) [1] to describe the combination of ‘crowd’ and ‘outsourcing’ [2]. This term describes how people can utilize the knowledge and competence of other people practically for free or for a small cost, still cheaper than hiring an expert to do the job. For bigger tasks, the reward is larger because it takes up more of the person’s time. Smaller tasks pay less but they can usually be done quite quickly and without putting too much thought into them. This type of task could for example be to recognize an object from a picture. This information could then be used as data to train artificial intelligence.

There are and has been many definitions for the term ‘crowdsourcing’ and the meaning has evolved from when it was first introduced to public. Also, the definitions of crowdsourcing differ based on the context in which the term is used [24]. One definition, that is considering the current role of social media and mobile devices, is presented in an article by Jan H. Kietzmann (2016):

“The use of IT to outsource any organizational function to a strategically defined population of human and non-human actors in the form of an open call.” [3]

Li et al. (2013) considered the fact that usually the crowd consists of non-experts and provided the following definition:

“Crowdsourcing has recently emerged as a powerful alternative. It outsources tasks to a distributed group of people (usually called workers) who might be inexperienced on these tasks.” [23]

Due to the generalization of the Internet in the recent years, crowdsourcing has grown even bigger than before [2]. One can utilize crowdsourcing in basically everything, including software design, funding, health care and many other fields. Internet has made it very easy to reach people around the world and crowdsourcing is making use of just that.

3.2. Crowdsourcing for design purposes

Today a lot of designing is done with the help of a computer instead of in physical form. All the modern apps, websites and programs we use daily would not exist if there was nobody to design them. There have been some studies on how crowdsourcing can provide value to the design process [16, 17] and how feedback can help designers iterate their designs to achieve better results [13, 18]. Crowdsourcing provides the perfect opportunity for designers to gain large amounts of feedback from random people of all ages and ethnicities, but it can also limit the quality of feedback received [4] due to designers needing quite detailed and well-reasoned critique for them to be able to improve their designs [5]. If a designer receives feedback such as ‘I do not like

this design' or 'The colours are not good', it does not help the designer to improve because they would not know what to change or why does the feedback provider feel this way. It would always be important to hear why people feel the way they do and not all crowd feedback systems are able to provide that information. Feedback is also hard to interpret correctly when it comes from behind a screen instead of face-to-face interaction where one could see a person's expressions and ask them to elaborate on their feelings.

Designers have varying opinions on what type of feedback is wanted from non-experts, but mainly it can be divided into four categories [6]: visual hierarchy such as layout and balance, first impressions, understanding the emphasis and message of the design and technical assessment based on design guidelines, for example appropriateness, simplicity [5], contrast and alignment [6].

3.3. Crowd feedback systems

For a non-expert to give feedback based on the principles mentioned in the previous chapter, they would most likely require some guidance from the system that they are using to give feedback, and therefore crowd feedback systems were created. The benefits of using a crowd feedback system instead of collecting feedback individually are that the designer can receive larger amounts of feedback with less money and organizational investment needed [4, 16]. The designer should consider what type of crowd feedback system to use based on what type of feedback they wish to receive.

3.3.1. Voyant

Voyant is a Web-based crowd feedback system focused on providing designers with feedback from regular people who are not experts in the field of design [6]. Compared to other crowd feedback systems, it is one of the most cited systems in the ACM Digital Library with 83 citations.

To use Voyant, the designer uploads and submits their graphic design in the system through the Web application and decides who they want to receive feedback from according to age, gender and location [4]. From there, Voyant will set up five types of feedback: *Elements*, *First Notice, Impressions*, *Guidelines* and *Goals* [4, 6]. To make it easier for the non-expert crowd to give useful feedback, the process is divided into smaller tasks and submitted to MTurk [7], where workers with 95% or higher task approval rate can participate in these microtasks [4, 6]. These workers will from now on be referred to as 'feedback providers'.

The process of giving feedback is divided into two phases, which are *description* and *interpretation* [4, 6]. This means that the feedback provider first describes what type of *Elements* they see in the design and mark the parts of the design where they see these elements using visual markers. During this phase, the system will urge the feedback provider to make a minimum of two markings of the colors, objects, shapes and activities they observe in the design [6]. In the interpretation stage the feedback provider must decipher the design according to the four types of feedback categories left, which are *First Notice, Impressions*, *Guidelines* and *Goals.* For different categories, the way feedback is given varies. For *First Notice* the feedback provider is presented with elements found in the design, in a random order, and they select which element they spot the earliest. In the *Impressions* tab the feedback provider provides

the system with words they first subconsciously link to certain parts of the design and try to justify their reaction. In the *Guidelines* as well as *Goals* category the feedback provider will consider one of the design guidelines at a time and give the design a score between -3 and +3 to indicate how well they feel the guideline is met, marks that area, and justifies the rating in their own words. In the end of the review, each feedback provider's feedback is aggregated, and a visual summary of each category is shown to the designer [4, 6].

3.3.2. CrowdCrit

Like Voyant, CrowdCrit is also Web-based, implemented with Python and JavaScript, and collects its feedback from non-experts. CrowdCrit can be used with any kind of crowd but it is mostly designed to gain its feedback with the assistance of MTurk [5, 7]. The system uses techniques like *scaffolding* to help the feedback providers give better critique on the designs. Scaffolding is defined as follows:

"When a teacher provides significant support to a student to help them learn new things." [8]

The designers of CrowdCrit have reviewed systems such as Voyant, Five Second Test and Feedback Army and tried to upgrade from them by focusing on giving feedback in a more professional manner. CrowdCrit's emphasis is on the language and process of giving criticism and the system tries its best to guide non-experts into giving as detailed critique as possible [9]. The feedback process is divided into seven key principles of design: *Readability, Layout, Balance, Simplicity, Emphasis, Consistency* and *Appropriateness* [5].

To begin, feedback providers fill out a questionnaire regarding their level of expertise and previous experience in the field of design so that they can be categorized accordingly [5]. Feedback providers can then focus on each of the previously introduced principles individually by choosing a category and going through a total of 70 pre-authored statements that include both positive and negative options [5]. Feedback providers can also use visual markers and provide a short explanation to help them make grounds for their feedback [5, 9]. Not all statements have to be answered so the feedback providers can freely choose which categories they want to give feedback on. From there, the feedback will be aggregated accordingly and later shown to the designer.

3.3.3. Decipher

Decipher is also a Web-based system built with JavaScript, HTML/CSS, JQuery and Python Django framework [10]. Unlike the other crowd feedback systems presented previously, Decipher's goal is not to collect feedback but to aggregate feedback that the designer has previously received in some form, analyse it and present it in a more visually pleasant and understandable way [10]. The feedback can be collected in any form, for example from a colleague or using an online questionnaire. In the study conducted by Yen et al. [10] the feedback was written by graphic designers recruited on Upwork [25]. This system was especially designed for less-experienced designers who do not yet have the ability to effectively gather critical information from the feedback that is received. With the help of Decipher, they should quickly pick up on important information and what type of changes to make.

The way Decipher works is that the designer will import their existing feedback into the system, where Decipher will help the designer sort feedback by letting them mark phrases or words from the feedback as positive, neutral or negative. Designer can also state whether they need clarification of the feedback, want to fix the problem in question, keep it in mind for future or if they just disagree with it. From there, if preferred, the designer can also choose to group the feedback according to who has given it, for example client or external user and if they are an expert or a non-expert in the field of design. [10]

3.3.4. Paragon

Paragon differs from the previously introduced systems in the way that it bases its operations on providing a non-expert crowd with examples that they can filter and search through to give better feedback. The examples are collected from other designers and from the Internet. The system itself is Web-based and built with Node.js. [14]

The system starts by introducing the feedback provider with the design and its description, as well as a gallery where they can search and filter for example designs. These examples combined with rubrics formed by experts provide a way for the non-expert crowd to give more specific feedback and show the designer what their design is lacking for example. The 8 rubrics made based on a study by Yuan et al. [15] are: "1. Need to consider audience 2. Provide better visual focus 3. Too much information 4. Create a more sensible layout 5. Use complementary visuals and text 6. Needs a clear visual hierarchy 7. Thoughtfully choose the typeface and colors 8. Other" [14]. The feedback providers choose a statement and review the design based on it. If they have some thoughts related to the statement, they can browse the gallery of examples and choose an example of another design that they feel is visually better in that aspect. After this they can explain why they chose that certain example. [14]

3.4. Other types of feedback

All the previously introduced systems collect their feedback in mainly text-form but that is not the only type of feedback there is. In some cases, other types of feedback are desired.

CrowdUI [19] is a crowd feedback system focused on web design. This system allows the feedback provider to mould the web page to their own liking by moving and modifying elements on the page or removing them all together. The feedback is then aggregated into heat maps, which show the critical points in the design where most feedback providers have made changes. It did however seem that this system was more suitable for looking into the layout or usability of the website [19] and not for example assessing the key principles for design.

Apparition [20] is a real-time crowdsourcing system designed to make prototyping easier and faster for designers and design teams. Starting with a blank canvas, designers can add low-fidelity elements and functionalities into their design by sketching or talking. Apparition conveys the user interface elements it recognizes into higher quality versions and the elements that are left are created by paid crowd workers from the sketches provided. This is then repeated to achieve a high-fidelity prototype. This helps the designers iterate their designs and convey their thoughts to the other

team members in the form of a functioning prototype. Even though this system provides quick results, it does not allow the crowd workers to comment on the design and the actual feedback is expected to be received and looked at among the design team.

Zipt [22] is a crowdsourcing system built to see how a user interface works in the eyes of a feedback provider. Designers come up with tasks they would like the feedback provider to perform in an app and Zipt monitors the way feedback providers operate during these tasks. The system then aggregates the information it has collected and shows the metrics and visualized versions of the data to the designer. This type of feedback proved to be valuable in the field of mobile design and helped the designers find for example usability problems in some apps.

For certain purposes, such as testing the usability of an interface, feedback in text-form would not provide much value since the designer cannot see how the user interacts with the interface. By using crowd feedback systems specific for the design category, designers can reach better results. Even though written critique is the most common type of feedback, in some cases it does not serve the purpose well enough and complimentary methods are needed.

4. RESULTS

The results show how feedback was aggregated and visualized in the crowd feedback systems in question. The differences between the systems will be discussed in the next section.

4.1. Voyant

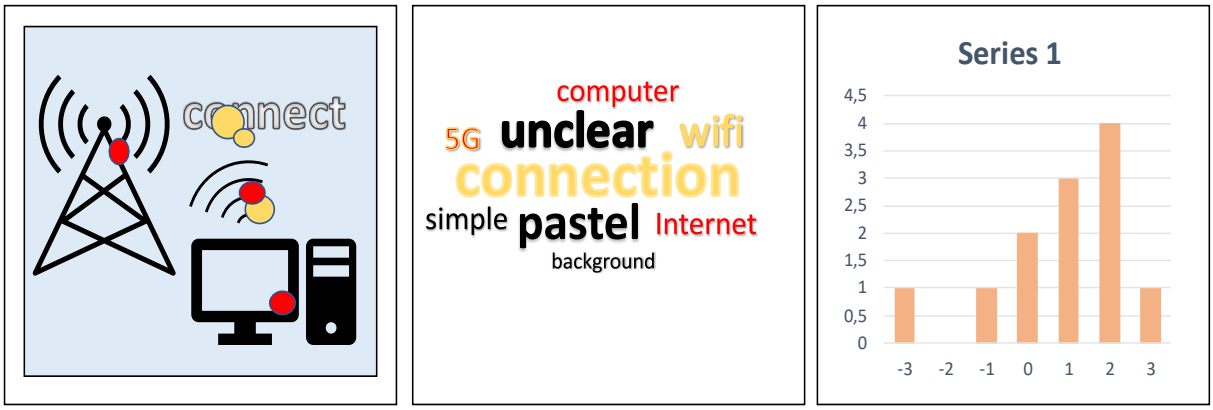
Voyant allows the designer to look at the feedback from two different angles: as an overview representation or just see the design's markers [6]. When looking at the feedback on the markers the aggregation and visualization of the feedback depends on the category that is viewed.

The *Elements* section shows a list of elements that the crowd has noticed in the design according to categories, which can be different based on what categories have been set for a design in advance [12]. From there, the designer can click on a certain word to see the visual markers that the feedback provider has placed on the design indicating where in the design they felt like these elements can be seen.

In the *First Notice* tab the designer can see how the visual hierarchy of the design is perceived by feedback providers [12]. This list of words can be organized either as an ascending list based on how many feedback providers saw the element in question first or based on category similarly as in the *Elements* section. Also, visual marking works the same way in both first two sections.

The third tab, *Impressions,* differs from the previous ones. Here the designer can see a word cloud where the largest number of first impressions documented by feedback providers will be presented as a bigger word in the cloud [12]. From there the designer can once again select a word to see the visual markers as well as what justification, if any, has the feedback provider given for the impression. The designer can also click on the markers to see what words were associated with a certain part of the design.

The fourth and fifth tabs, *Goals* and *Guidelines*, have the same type of aggregation and visualization. These sections use a bar chart to convey the points given by feedback providers on a seven-point scale on how the design meets the goal or guideline in question. The options on the scale vary from strongly disagree (-3) to strongly agree (+3). In addition, the feedback providers also had the option to place visual markers, which the designer can later look at to see what in the design created these ratings among feedback providers [12].



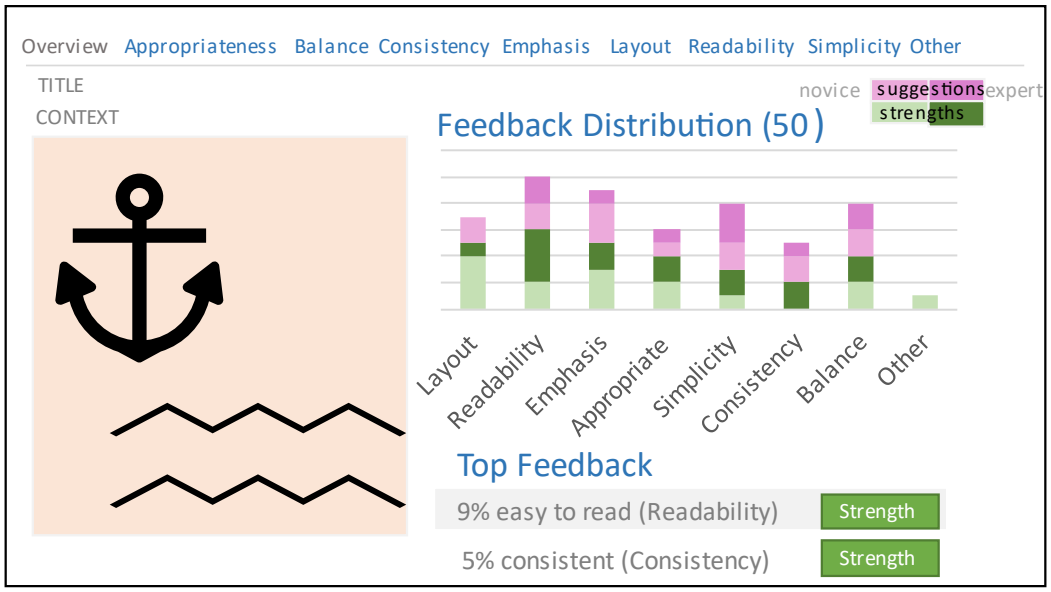
(a) Visual markers (b) Word cloud (c) Bar chart

Figure 1. A sketch made by Katri Säily presenting different styles of feedback aggregation [6, Figure 3]. (a) An example of how the visual markers are presented to the designer. (b) The designer can see the word cloud and if some words are marked with visual markers, they appear in a different color. (c) An example of the bar chart where feedback providers have rated the design from a scale from -3 to +3.

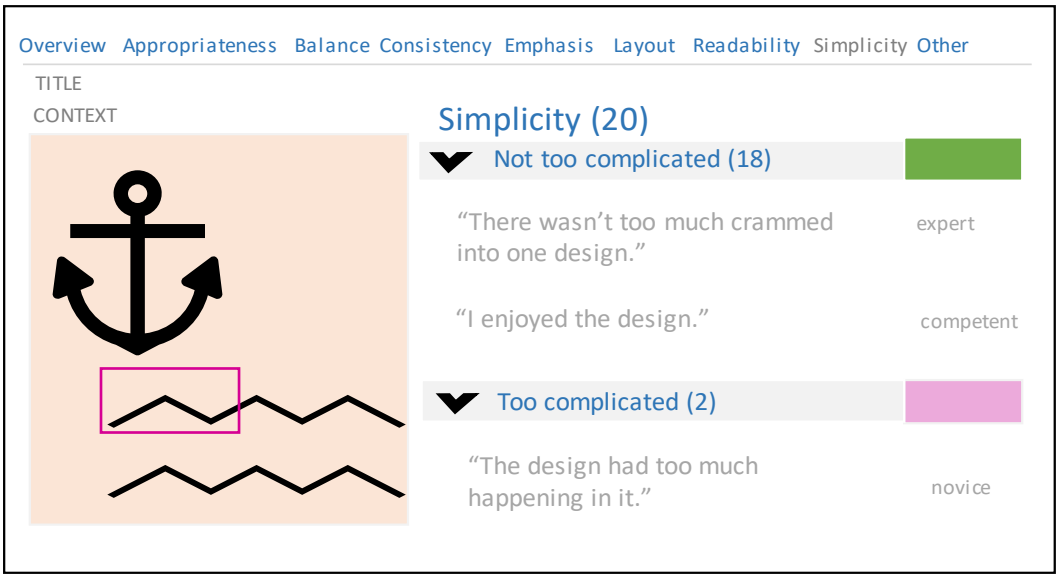
4.2. CrowdCrit

In CrowdCrit the aggregated feedback is divided into two main sections: *Overview* and *Design Principle View* [5]. The first one being the general overview of feedback showcasing all main categories in the form of a stacked bar chart and latter being a more detailed view of all the individual feedback and comments in each of the seven design principles. When looking at the feedback, colors provide a lot of information for the designer, green being positive feedback and purple being something to improve on [9]. Also, the darker the color, the more expertise the provider of the feedback has had [5, 9]. Besides the stacked bar chart, the *Overview* section holds a 'Top Feedback' section, that shows in which category there was most feedback but also the feedback statements that were selected most frequently [9].

In the *Design Principle View* the designer can see a more detailed take on the feedback based on the individual design principles [9]. For every statement provided in the questionnaire, the feedback provider's comments will be showed with the darkness of the text representing their expertise and visual markers in case the feedback provider has used them to support their feedback. Due to the large amounts of information, the interface will reveal the information gradually, so that going through the feedback would not become too overwhelming [5].



a) Overview



b) Simplicity principle is selected

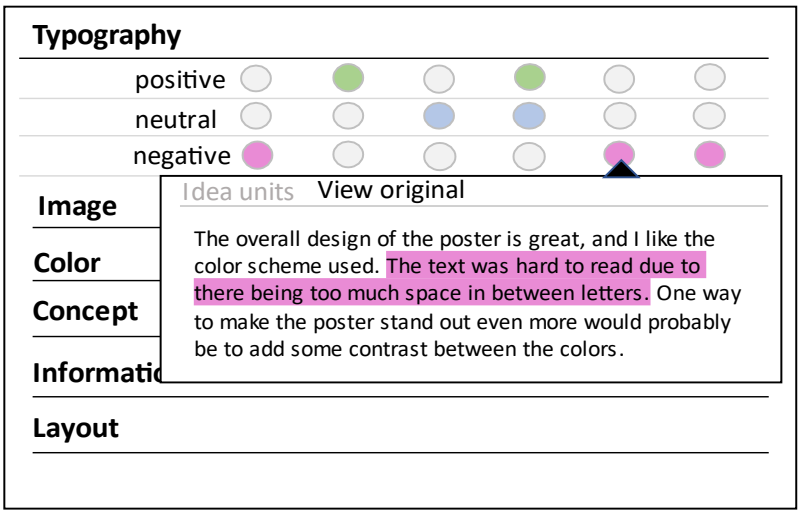
Figure 2. Feedback aggregation in CrowdCrit, a sketch made by Katri Säily based on: [5, Figure 4]. (a) The Overview tab where the designer can see how the feedback is distributed, darker colors meaning that the feedback provider's expertise was greater and green being positive and purple negative feedback. (b) One of the principles, Simplicity, is selected and the designer can see exactly what was said about the design and they could also see if the feedback providers have placed any visual markers.

4.3. Decipher

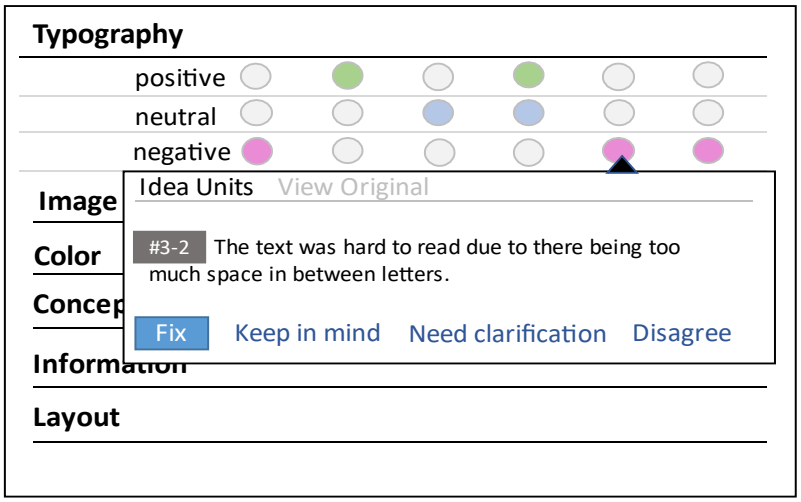
The feedback in Decipher is aggregated into positive (green), neutral (blue) and negative (pink) groups and also in sections based on the part of the design that the feedback is considering: typography, image, color, concept and information, but the designer could have also added categories of their own [10]. On rows one can find

feedback given by all the feedback providers and feedback given by one provider is placed in one column. If there is no feedback on some topic from the feedback provider, it will remain grey. If the feedback on one section is considered partly positive and partly negative, the circle would be half-pink and half-green. When looking through the feedback the designer can then filter it based on, for example, keywords that appear in the original texts and are marked as 'keep in mind'. If they are looking for feedback connected to a certain topic, they can see it by clicking on a plus icon and a pop-up will show up. [10]

Clicking on a 'See original' button will present the feedback in the certain section in its whole context. This way the designers can easily see which actions to take in the future and in which part of the design these actions are needed. By using Decipher the designer can quite quickly see the overall reaction to their design without reading the full feedback [10].



b) View original



a) View when hovering over a selection

Figure 3. Feedback aggregation in Decipher, a sketch made by Katri Säily based on: [10, Figure 3]. (a) Shows the view for the designer when they hover over a certain selection, in this case feedback that is marked as negative and seen connected to

Typography category. Designer has chosen that they would regard this feedback as fixable. (b) Presents the ‘View original’ selection, where the designer can see the selected feedback in its whole context.

4.4. Paragon

Paragon presents the collected feedback to the designer by simply showing what each feedback provider has written, which rubrics the statement is connected to as well as what example they have chosen and why [14]. Compared to other systems it does not aggregate feedback in a very effective way even though it collects feedback in an informative manner. This means that Paragon is not very suitable for large amounts of feedback, since the designer would have to go through them individually.

It was found that examples helped the feedback providers convey their thoughts more easily and gave them further ideas for complementary feedback. According to the studies conducted by Kang et al. [14], giving feedback by using examples was preferred by many non-experts and feedback acquired this way provided more accurate, lengthy and helpful critique compared to feedback provided in textual form. Some other studies [21] also support this statement. On the downside, this method does take more time and some feedback providers considered it to be too much work and hard to master. In the studies the best method was found to be that feedback providers choose a rubric, write their first critique and then search for an example giving grounds for their feedback and explaining the choice. [14]

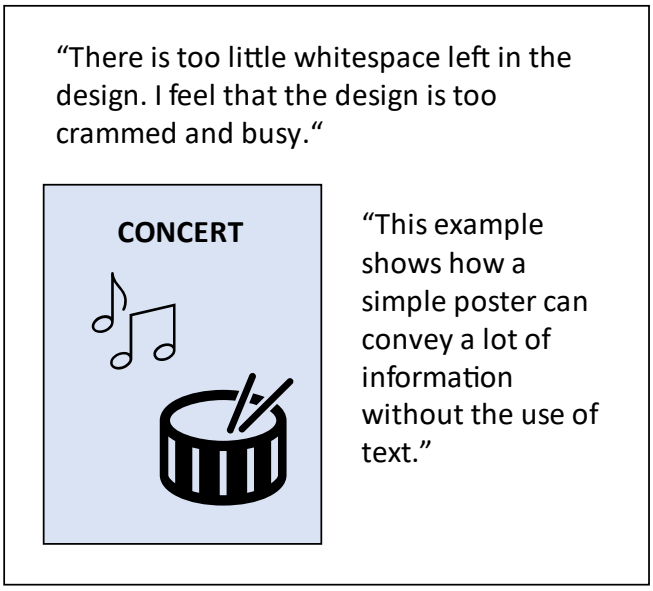


Figure 4. Feedback aggregation in Paragon, a sketch made by Katri Säily based on: [14, Figure 10]. Paragon shows the initial feedback given with the help of rubrics on top and the feedback made with the help of an example below the initial feedback.

4.5. Summary

This summary shows the most important attributes of each crowd feedback system researched in this thesis. The table compares how different systems collect their feedback, who they collect it from and how they aggregate and visualize the collected feedback.

Table 1. A summarizing table of all four crowd feedback systems introduced in this thesis.

Feedback system / **Attribute**	**Voyant**	**CrodCrit**	**Decipher**	**Paragon**
Who provides feedback?	Non-experts	Non-experts	Anyone	Anyone, but mostly non-experts
How is feedback collected?	With MTurk, workers with >=95% approval rate can participate	Mostly with MTurk	System uses previously collected (existing) feedback	In the system using a gallery containing examples
Technique used for collecting feedback	Microtasks consisting of two phases: interpretation and description	Scaffolding, 70 pre-authored statements in each category	Marking feedback according to its attributes: positive, neutral, negative, provider, Fix, Keep in mind, need clarification and disagree	Rubrics
What type of categories are used?	Elements, First Notice, Impressions, Guidelines and Goals	Readability, Layout, Balance, Simplicity, Emphasis, Consistency and Appropriateness	Up to the designer but usually Typography, Image, Color, Concept and Information	Need to consider audience, Provide better visual focus, Too much information, Create a more sensible layout, Use complementary visuals and text, Needs a clear visual hierarchy, Thoughtfully choose the typeface and colours, Other
Can the feedback provider use visual markers?	Yes	Yes	No	No

How is feedback presented visually?	Visual markers, word clouds and stacked bar charts	Visual markers, stacked bar charts	All received feedback is on a row and feedback given by one provider on a column, colors state whether feedback is negative, positive/neutral	The provided example image and feedback together
Other attributes	Feedback aggregation and visualization is different based on the category	Colour of feedback shows the expertise of the feedback provider	Designers can add their own categories for feedback	None

5. DISCUSSION

The study by Luther et al. (2015) [5] compared Voyant and CrowdCrit briefly. Voyant was referred to as a system that can handle large quantities of feedback, but it was also mentioned that Voyant mainly focuses on people's impressions of the design, and not as much on the language or process of giving feedback as CrowdCrit. After reviewing the feedback aggregation in both systems, I would agree with this statement. The disadvantage of CrowdCrit as well as Paragon, however, is that it only presents the feedback in a textual form rather than making a proper visual representation for the designer. This can be a major disadvantage compared to the other systems since designers are usually very visual people and they would most likely prefer visually pleasant presentations of information.

In Decipher's case it was found that presenting the feedback in a simple and interactive form helped especially young designers make sense of the feedback and find the problems in their designs, compared to them just reading the feedback from different sources and trying to aggregate it by themselves [10]. Although, it must be considered that usually the designers would have to annotate the feedback themselves and in the study in question [10] it was done for them, so there would have to be more research made on whether designers can do that effectively themselves. Also, in the study conducted by Yen, Kim and Bailey (2020) [10] it was stated that designers' reactions might be different if they were to receive feedback of their own designs instead of someone else's because in that case negative feedback might feel more personal.

The main benefits of Decipher compared to Voyant, CrowdCrit or Paragon are that it shows what part of the critiques are considered for example positive, and that the designer can very easily filter out feedback that is not needed at the time and focus on critique relevant at the time. One of the improvements that was suggested for Voyant was to make filtering available so the designers could filter it according to age or gender, for example [6]. Decipher does offer this function, Paragon does not and CrowdCrit only on the aspect of whether the feedback providers are experts or novices.

Voyant on the other hand is very versatile in the way that it aggregates feedback based on category. It uses word clouds, visual markers and bar charts, whereas CrowdCrit and Decipher use the same type of visualization for every category and Paragon does not even use categories. Voyant, CrowdCrit and Paragon focus more on helping feedback providers give feedback based on the actual design principles whereas Decipher's goal is just to sort the existing written feedback into a more visually pleasing form, and therefore it does not give tools for non-experts to give relevant feedback. Decipher basically expects the feedback provider to be competent in giving design critique, which is not a realistic expectation when talking about collecting the feedback using crowdsourcing. Also, unlike Voyant and CrowdCrit, Decipher does not make it possible to use visual markers whereas in the other systems it was considered by designers to be one of the biggest advantages [5, 6]. Paragon on the other hand uses example designs as reference, which was very helpful for the feedback provider and designer [14]. Annotation tools were also found to help designers get a grasp of how the feedback providers see their designs and what elements they link with these perceptions,[6] but the tools also added some perspective on feedback that would otherwise be very vague [5].

After reviewing all these crowd feedback systems, I feel that my own preference would be to use a system like Voyant because of the versatile visualization of the

collected feedback. This means that the feedback would be pleasant to go through and not feel too clustered or busy. Then again, if there was a great need for professional feedback, I would opt to use CrowdCrit due to its ability to help feedback providers give more expert-like critique. Also, I feel that visual markers would make it easier to see where the problems in the design are.

From the designer's perspective crowd feedback systems seemed to be helpful in all cases [5, 6, 10, 14] and were praised for providing quick results for a low cost [5]. It was also mentioned that the crowd was able to provide a wide range of feedback from many different perspectives and reaching this many people would not necessarily be possible without these kinds of systems [5]. In the future it might be necessary to do a case study where these crowd feedback systems are compared from the designers' perspectives. This thesis mostly contains assumptions about which of these systems designers would prefer and why, so it would be beneficial to see which system is the best for different situations. It would also be interesting to see if the same design is submitted to each of these systems, how the results would differ from each other due to the different methods of collecting feedback in each system. From there designers could share their impressions of the interfaces and how they compare to each other.

6. CONCLUSION

Designing is never easy, and one might be blinded to their own design after staring at it for too long. Therefore, an outsider's perspective is often desired to see how other people perceive the design. Crowdsourcing gives us access to the opinions of a large crowd, although they are usually considered as novices in the field of design. Crowd feedback systems try to steer these people into giving professional feedback so they can be more helpful in the design process. In this study it was found that crowd feedback systems each focus on different things and therefore provide differing results. Decipher focuses on aggregating and visualizing existing feedback by categories and user demographics, so people who use it should have previously collected feedback available. CrowdCrit's aim is to make use of scaffolding to provide expert-like feedback in a simple and easily readable form. Voyant collects feedback and illustrates it in the form of word clouds, bar charts and visual markers. Paragon uses examples to help feedback providers give more informative feedback so the designers can get a better understanding of their thoughts. All these systems were found helpful in the studies by designers, which means they all can provide value to the design process.

REFERENCES

[1] Howe, J. (2006). The Rise of Crowdsourcing. *Wired*, 1-4. Retrieved from https://www.wired.com/2006/06/crowds/

[2] Schenk, E., & Guittard, C. (2009). Crowdsourcing: What can be Outsourced to the Crowd, and Why ? Retrieved from https://www.researchgate.net/publication/40270166_Crowdsourcing_What_can_be_Outsourced_to_the_Crowd_and_Why

[3] Kietzmann, J. H. (2016). Crowdsourcing: A revised definition and introduction to new research. *Business Horizons*. Retrieved from https://beedie.sfu.ca/sms/admin/_DocLibrary/_ic/512c2e31791539b761580701430ec95e.pdf

[4] Anbang Xu, A., Rao, H., Dow, S. P., & Bailey, B. P. (2015). A Classroom Study of Using Crowd Feedback in the Iterative Design Process. *CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1637-1648). BC, Vancouver, Canada: Association for Computing Machinery. DOI: https://doi.org/10.1145/2675133.2675140

[5] Luther, K., Tolentino, J-L., Wu, W., Pavel, A., Bailey, B. P., Agrawala, M., . . . Dow, S. (2015). Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. *CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 473-485). Vancouver, Canada: Association for Computing Machinery. DOI: https://doi.org/10.1145/2675133.2675283

[6] Xu, A., Huang, S.-W., & Bailey, B. P. (2014). Voyant: generating structured feedback on visual designs using a crowd of non-experts. *CSCW '14: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1433–1444). Maryland, Baltimore, USA: Association for Computing Machinery. DOI: https://doi.org/10.1145/2531602.2531604

[7] Amazon's Mechanical Turk, https://www.mturk.com/

[8] Beed, P. L., Hawkins, E. M., & Roller, C. M. (1991). Moving Learners toward Independence: The Power of Scaffolded Instruction. *The Reading Teacher, Vol 44, No. 9*, 648-655. Retrieved from https://www.jstor.org/stable/20200767

[9] Luther, K., Pavel, A., Wu, W., Tolentino, J.-l., Dow, S., Hartmann, B., & Agrawala, M. (2014). CrowdCrit: Crowdsourcing and Aggregating Visual Design Critique. *CSCW Companion '14: Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 21-24). Maryland, Baltimore, USA: Association for Computing Machinery. DOI: https://doi.org/10.1145/2556420.2556788

[10] Yen, Y.-C. G., Kim, J. O., & Bailey, B. P. (2020). Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). Honolulu, USA: Association for Computing Machinery. DOI: https://doi.org/10.1145/3313831.3376380

[11] Greenberg, M. D., Easterday, M. W., & Gerber, E. M. (2015). Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. *C&C '15: Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition* (pp. 235-244). Glasgow, United Kingdom: Association for Computing Machinery. DOI: https://doi.org/10.1145/2757226.2757249

[12] Xu, A., & Bailey, B. P. (2014). A System for Receiving Crowd Feedback on Visual Designs. *CSCW Companion '14: Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 37-40). Maryland, Baltimore, USA: Association for Computing Machinery. DOI: https://doi.org/10.1145/2556420.2556791

[13] Yen, Y.-C. G., Dow, S. P., Gerber, E. & Bailey, B. P. (2017). Listen to Others, Listen to Yourself: Combining Feedback Review and Reflection to Improve Iterative Design. *C&C '17: Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition* (pp. 158-170). Singapore, Singapore: Association for Computing Machinery. DOI: https://doi.org/10.1145/3059454.3059468

[14] Kang, H. B., Amoako, G., Sengupta, N. & Dow, S. (2018). Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). Montreal QC, Canada. Association for Computing Machinery. DOI: https://doi.org/10.1145/3173574.3174180

[15] Yu, L. & Nickerson, J. V. (2011). Cooks or Cobblers?: Crowd Creativity Through Combination. *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing* (pp. 1393-1402). BC, Vancouver, Canada. Association for Computing Machinery. DOI: https://doi.org/10.1145/1978942.1979147

[16] Yu, L., Kittur, A. & Kraut, R. E. (2016). Encouraging "Outside- the- box" Thinking in Crowd Innovation Through Identifying Domains of Expertise. *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1214-1222). California, San Francisco, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/2818048.2820025

[17] Andolina, S., Schneider, H., Chan, J. Klouche, K., Jacucci, G. & Dow, S. (2017). Crowdboard: Augmenting In-Person Idea Generation with Real-Time Crowds. *C&C '17: Proceedings of the 2017 ACM SIGCHI Conference*

on Creativity and Cognition (pp. 106-118). Singapore, Singapore. Association for Computing Machinery. DOI: https://doi.org/10.1145/3059454.3059477

[18] Dow, S., Glassco, A., Kass, J., Schwarz, M., Schwartz, D. L. & Klemmer, S. R. (2010). Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction,* Volume 17, Issue 4, Article No 18. Association for Computing Machinery. DOI: https://doi.org/10.1145/1879831.1879836

[19] Oppenlaender, J., Tiropanis, T. & Hosio, S. (2020). CrowdUI: Supporting Web Design with the Crowd. *Proceedings of the ACM on Human-Computer Interaction,* Article No. 76. Association for Computing Machinery. DOI: https://doi.org/10.1145/3394978

[20] Lasecki, W. S., Kim, J., Rafter, N., Sen, O., Bigham, J. P. & Bernstein, M. S. (2015). Apparition: Crowdsourced User Interfaces that Come to Life as You Sketch Them. *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1925-1934). Seoul, Republic of Korea. Association for Computing Machinery. DOI: https://doi.org/10.1145/2702123.2702565

[21] Robb, D. A., Padilla, S., Kalkreuter, B. & Chantler, M. J. (2015). Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It? *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1355-1364). Seoul, Republic of Korea. Association for Computing Machinery. DOI: https://doi.org/10.1145/2702123.2702470

[22] Deka, B., Huang, Z., Franzen, C., Nichols, J., Li, Y. & Kumar, R. (2017). ZIPT: Zero-Integration Performance Testing of Mobile App Designs. *UIST '17: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (pp. 717-736). QC, Québec City, Canada. Association for Computing Machinery. DOI: https://doi.org/10.1145/3126594.3126647

[23] Li, H., Yu, B., & Zhou, D. (2013). Error rate analysis of labeling by crowdsourcing. *ICML Workshop: Machine Learning Meets Crowdsourcing.* Atalanta, Georgia, USA. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.724.1773&rep=rep1&type=pdf

[24] Hosseini, M., Shahri, A., Phalp, K., Taylor, J. & Ali, R. (2015). Crowdsourcing: A taxonomy and systematic mapping study. *Computer*

Science Review, (pp. 43-69) Volume 17. DOI: https://doi.org/10.1016/j.cosrev.2015.05.001

[25] Upwork. (2019). https://www.upwork.com/