

INHALTSVERZEICHNIS

Teil I

THESIS

EINLEITUNG

1.1 MOTIVATION

Die Qualität von Software für Menschen verständlich und greifbar zu machen, die nicht täglich mit Quellcode arbeiten, stellt nach wie vor eine Herausforderung dar. Aber auch erfahrene Entwickler*innen stehen regelmäßig vor der Aufgabe, sich schnell einen Überblick über ein bestehendes Softwaresystem zu verschaffen: Wo liegen die problematischen Stellen? Wo lohnt sich ein tieferer Blick? Und wie lassen sich Verbesserungspotenziale effizient identifizieren?

Um solche Fragen beantworten zu können, sind geeignete Visualisierungen essenziell. Sie helfen, komplexe Strukturen zu abstrahieren, Muster zu erkennen und technische Schulden sichtbar zu machen – ohne dass zunächst jede Zeile Code gelesen werden muss. Während einfache Diagramme oder Dashboards nützliche Einstiegspunkte bieten, zeigen immersive Ansätze, insbesondere dreidimensionale Darstellungen, ein deutlich höheres Potenzial, Software auf eine intuitivere und erfahrbare Weise zugänglich zu machen [MFM03; WLo7; Rei95; KM00].

Insbesondere 3D-Metaphern, wie die Stadt-Metapher [WLo7], haben sich in Forschung und Praxis als äußerst wirkungsvoll erwiesen (QUELLEN). Durch die Übertragung von Softwarestrukturen auf städtische Elemente – Gebäude, Blöcke, Straßen – entsteht ein räumliches Abbild, das verschiedene Code-Metriken kombinieren kann. Die Beliebtheit dieser Methode beruht nicht zuletzt auf der hohen Informationsdichte und der intuitiven Lesbarkeit räumlicher Strukturen. Im Vergleich dazu wirken alternative Metaphern, etwa waldartige Darstellungen [Atz+21], häufig unübersichtlich und verlieren bei wachsender Komplexität schnell an Aussagekraft.

Gleichzeitig zeigt sich, dass auch die Stadt-Metapher nicht frei von Schwächen ist. Gerade bei sehr großen Codebasen stößt die Übersichtlichkeit an ihre Grenzen oder wichtige Details gehen verloren [LFo8]. Viele dieser Visualisierungen beruhen auf Treemap-Layouts – einer etablierten Technik zur Darstellung hierarchischer Strukturen. Doch gerade in Bezug auf Lesbarkeit, Informationsdichte und Benutzerfreundlichkeit stellt sich zunehmend die Frage, ob der klassische Treemap-Ansatz nicht in Bezug auf 3D-Softwarevisualisierungen verbessert werden kann.

Außerdem ist offen, ob es alternative Layout-Ansätze gibt, die eine noch klarere, verständlichere oder flexiblere Darstellung ermöglichen würden – sowohl in 2D als auch in 3D. In der Praxis ist dieser Aspekt bisher wenig beleuchtet worden, obwohl er entscheidend für die Nützlichkeit solcher Visualisierungen ist.

1.2 GRUNDLAGEN

[ht]

1.2.1 *Software-Qualitätsmetriken*

Software-Qualitätsmetriken sind ein wichtiges Werkzeug zur Analyse der Qualität von Software. Sie ermöglichen es, verschiedene Aspekte der Software quantitativ zu bewerten. Software-Metriken werden oft sehr ähnlich definiert:

“Eine Softwaremetrik, oder kurz Metrik, ist eine (meist mathematische) Funktion, die eine Eigenschaft von Software in einen Zahlenwert, auch Maßzahl genannt, abbildet. Hierdurch werden formale Vergleichs- und Bewertungsmöglichkeiten geschaffen.”[WPo4]

“Softwaremetrik ist ein quantitatives Maß, das verwendet wird, um die Eigenschaften eines Softwareproduktes oder des Softwareentwicklungsprozesses zu bewerten.”[Sof]

“Eine Softwarequalitätsmetrik ist eine Funktion, die eine Software-Einheit in einen Zahlenwert abbildet, welcher als Erfüllungsgrad einer Qualitätseigenschaft der Software-Einheit interpretierbar ist.”[EH98]

Alle definitionen ist zu entnehmen, dass Sie nur einen Speziellen Eigenschaft von Software messen. Das ist auch eine der wichtigsten Learnings hier und auch eine Kritik an Software-Qualitätsmetriken: Sie messen nur bestimmte Aspekte der Software und können keine vollumfängliche Aussage über die Qualität der Software treffen.

Den definitionen am anfang ist zu entnehmen, dass Software-Qualitätsmetriken also Kennzahlen sind, die zur Bewertung der Qualität von Software verwendet verwendet werden und ein wichtiges Werkzeug zur Analyse und Verbesserung sind, obwohl sie nur in der lage sind bestimmte Aspekte der Software zu messen.

“Eine Metrik alleine kann nie eine vollständige Aussage über die Qualität der Software treffen. Metriken sollten immer in Kombination bewertet werden. Die Güte und Reife von Software kann nur in Kombination mit statischer Code Analyse, Code Reviews und funktionalen Tests final beurteilt werden.”[Sch19]

Es ist nicht möglich eine einzige Metrik zu finden, die eine vollumfängliche aussage über die Qualität einer Software trifft - was ist schon gute Software? Wann ist software gut? SEblst wenn man als Ziel von software definieren würde, dass sie fehlerfrei ist, wäre es nicht möglich eine einzige Metrik zu finden, die dies beschreibt. Wann ist eine software fehlerfrei? Ist sie fehlerfrei, wenn sie keine Bugs hat? Wenn sie keine Bugs hat, aber trotzdem nicht benutzbar ist, ist sie dann fehlerfrei? Also ist es nicht möglich eine Metrik zu identifizieren, die eine vollumfängliche aussage über ein programm trifft. es ist also notwenig verschiedene metriken im zusammenspiel miteinander zu betrachten, um die aussagekraft von metriken zu erhöhen.

Dieser Eigenschaft macht es notwendig verschiedene Metriken im Zusammenspiel zu betrachten und sich einen Überblick über die Indizien zu verschaffen, die auf eine gute oder schlechte Qualität hinweisen. Die Besonderheit bei Software-Qualitätsmetriken im Vergleich zu anderen Metriken mag sein, dass die Daten hierarchisch strukturiert sind. Meist wird dies auf die Ordner- und Datei-Struktur abgebildet. So werden die Metriken in der Regel auf der Ebene der einzelnen Dateien ermittelt und dann auf die Ordner-Ebene aggregiert, wodurch eine hierarchische Baumstruktur entsteht.

Probleme mit Software-Metriken treten vor allem dann auf, wenn initial gar nicht klar ist, welches Ziel eine Software hat [HF94]. Metriken müssen immer so definiert und gemessen werden, dass sie die Qualität der Software in Bezug auf das definierte Ziel messen. Wenn zum Beispiel das Ziel der Software ist, möglichst schnell zu sein, ist vielleicht das Arbeiten mit einer Komplexitätsmetrik sinnlos oder sogar kontraproduktiv, da dadurch die Software zwar einfacher, aber nicht schneller wird. Zudem muss auch beachtet werden, dass das Erfassen selbst und die Orientierung an Metriken Zeit einnimmt [HF94].

Außerdem wird die Kritik ausgeübt, dass Metriken oft nicht wissenschaftlich fundiert sind und die wahre Kosten-Nutzen-Bilanz von Metriken gar nicht klar ist, man sollte sich nicht zu sehr an Metriken orientieren und seine Entscheidungen darauf basieren, wenn nicht wirklich klar ist, wie viel es zu dem eigentlichen Ziel der Software beiträgt [VK17].

Zudem muss klar sein, dass sobald Metriken definiert sind, die nicht exakt das Software-Ziel verfolgen, es passieren kann, dass die Software-Entwickler versuchen, diese Metriken zu optimieren, anstatt das eigentliche Ziel der Software zu verfolgen. Wodurch zwar die Metriken optimiert werden, aber die Software nicht besser wird. Es dürfen also keine unrealistischen Erwartungen an Software-Metriken gestellt werden. Es muss klar sein, dass Software-Qualitätsmetriken nur Indizien für die Qualität der Software liefern und niemals, Prozesse wie Code-Reviews oder Tests ersetzen können. Außerdem muss klar sein, wie man Metriken interpretiert. Zum Beispiel kann eine Lines-Of-Code-Metrik, obwohl sie immernoch weit verbreitet ist, keine Aussage über die geleistete Arbeit eines Entwicklers oder die Komplexität / Funktionalität eines Programms sein. Allein der Unterschied von Programmiersprache zu Programmiersprache macht dies undurchsichtig und Code-Qualität wird hier auch in keiner Weise gemessen - Stichwort duplizierter Code - das selbe gilt übrigens über die Anzahl an Commits [Wor].

Einfache Metriken können keine komplexen Qualitätsprobleme aufdecken [VK17].

Oft sind Metriken auch stark von der Sprache oder dem individuellen Programmierstil abhängig [VK17].

Die ISO/IEC 25010 Norm [Iso] beschreibt, wie Software-Attribute und -Qualitäten miteinander in Beziehung stehen. Diese Norm zerlegt die Software-Qualität in acht Qualitätsmerkmale, wie zum Beispiel funktionale Eignung. Dies ist der Blick von außen auf die Software. Zum Beispiel, wie oft stürzt die Software ab, wie gut ist die Benutzeroberfläche, wie gut ist die Performance

und so weiter. Diese Merkmale sind wichtig, um die Software aus der Sicht des Benutzers zu bewerten.

“Qualität wird als Abwesenheit von Fehlern im Systemverhalten verstanden, die Software verhält sich demnach so, wie der Benutzer es erwartet” [Wit18, S. 1]

Aber was sind die Gründe für diese Fehler? Gibt es eventuell eine Möglichkeit SoftwareQualität auch von innen her zu messen? In den Siebziger Jahren entstand die Notwendigkeit Programm-Code selbst zu bewerten und dadurch die Qualität von Software zu messen [Lud]. Es entstanden erste Metriken wie die McCabe Komplexität. Diese Innere Qualität- oder im folgenden einfach Code Qualität genannt interessiert besonders Entwickler und Auftraggeber und Stakeholder von Software, damit die Software langfristig wartbar, fehlerfrei und erweiterbar bleibt. In dieser Arbeit wird ausschließlich von Code-Qualitätsmetriken gesprochen, wenn es um Software-Metriken geht.

Software-Qualitätsmetriken werden meist auf Software-Einheiten angewendet. Also z.B. auf Dateien des Quellcodes, auf Module oder auf Klassen. Diese Software-Einheiten sind in der Regel hierarchisch strukturiert. Auch die im folgenden vorgestellten Ansätze zur Visualisierung von hierarchischen Daten sind nicht speziell für Software-Qualitätsmetriken gedacht, sondern stellen allgemeine Ansätze dar. Ich behaupte, dass die Struktur von Software Gemeinsamkeiten hat, z.B. ein Ordner hat nicht mehr als 20 Kinder, was interessante Annahmen darstellt, die eventuell in die Visualisierungs Algorithmen verbessernd einfließen können.

1.2.2 Software-Visualisierung

1.2.3 3D-Software Visualisierung

Warum 3D? Ist es nicht ausreichend Software-Qualität einfach als Zahlenwerte darzustellen oder zumindest in 2D darzustellen? Wie in Abschnitt ?? aufgezeigt, wollen wir Software Qualität auch für nicht-Software-Entwickler greifbar machen, einfache Werte sind dafür nicht ausreichend.

Despite the proven usefulness of 2D visualizations, they do not allow the viewer to be immersed in a visualization, and the feeling is that we are looking at things from outside". 3D visualizations on the other hand provide the potential to create such an immersive experience... [WLo7, S. 1]

2D Visualisierung bietet zwar bestimmt einen guten Überblick, aber macht die Software nicht wirklich greifbar.

Die Idee Software in 3D darzustellen ist nicht neu. Schon 1995 stellte Steven P. Reiss einen Ansatz vor, der es ermöglichte Software in 3D darzustellen [Rei95]. Die meisten dieser ersten Ansätze verfolgten das Ziel Software für Entwickler greifbar zu machen, die Struktur aufzuzeigen und einen Highlevel Überblick über eine Software zu geben, was speziell hilfreich ist, wenn Entwickler ein System neu kennenlernen müssen [YM98]. Die-

se Ansätze sind in der Regel nicht für die Visualisierung von Software-Qualitätsmetriken gedacht, sondern zielen darauf ab, die Struktur und den Aufbau der Software zu verdeutlichen (siehe Abbildung ??).

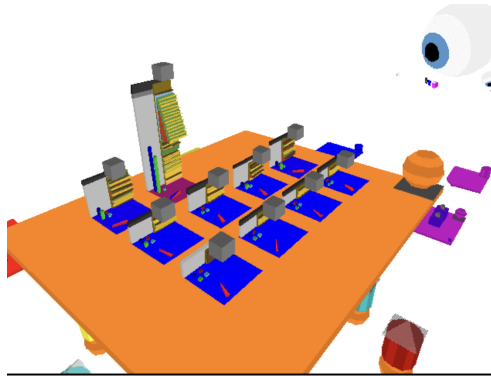


Abbildung 1.1: Beispiel für eine 3D-Visualisierung von Young und Munro [YM98, S. 6]

Trotzdem können wir auch für unser Ziel der Software-Qualitätsmetrik-Visualisierung Kriterien ableiten, die für eine gute Visualisierung wichtig sind [YM98]:

- **Darstellung:** Der wichtigste Aspekt ist die Darstellung der Software. Die Visualisierung sollte die Struktur und den Aufbau der Software verdeutlichen. Die Frage ist also, wie wird die Software dargestellt?
 - Informationsgehalt: Die Visualisierung sollte so viele Informationen wie möglich enthalten.
 - Niedrige visuelle Komplexität: Als Gegenspieler zum Informationsgehalt steht die visuelle Komplexität. Die Visualisierung sollte so einfach wie möglich gehalten werden, um den Betrachter nicht zu überfordern.
 - Skalierbarkeit: Die Visualisierung sollte auch bei großen Software-Systemen noch gut lesbar sein. Dies ist besonders wichtig, da wir hier über große Software-Systeme sprechen. Die Autoren von *Visualising Software in virtual reality* [YM98] sagen zudem, dass Mechanismen Nötig sind, um Komplexität und Informationsgehalt zu steuern und je nach Software-System anpassen zu können.
 - Stabilität gegenüber Änderungen: Die Visualisierung sollte stabil gegenüber Änderungen in der Software sein. Das bedeutet, dass die Visualisierung sich nur so sehr wie nötig ändert, wenn sich die Software ändert, um eine Versions konsistente Vergleichbarkeit zu ermöglichen und bereits mit der Visualisierung vertraute Nutzer nicht zu überfordern.
 - Gute Visuelle Metaphern: Die Visualisierung sollte gute visuelle Metaphern verwenden, um bereits bekannte Konzepte zu verwenden, um die Software verständlicher zu machen.

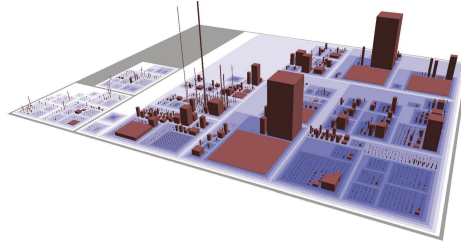


Abbildung 1.2: Beispiel für eine original CodeCity-Visualisierung [WLo7, S. 2]

- **Abstraktion:** Das Ziel von Visualisierung muss sein, unwichtige Details auszublenden und ein verständliches Modell der Software zu erstellen.
- **Navigation:** Da wir häufig über große Software-Systeme sprechen, kann es schnell passieren dass Nutzer in der Visualisierung verloren gehen. Es muss also möglich sein, sich gut zurecht zu finden und intuitiv zu wissen, wo was ist, um so ein Gefühl für die Software zu erhalten.
- **Korrelation mit dem Code:** Die Visualisierung sollte eine gute Korrelation mit dem Code haben. Wenn man die Visualisierung sieht, soll man diese auch mit dem Code in Verbindung bringen können. Es sollte also möglich sein, die Visualisierung mit dem Code zu verknüpfen und so ein besseres Verständnis für die Software zu bekommen.
- **Automatisierung:** Die Visualisierung sollte automatisiert werden können - ein Punkt der trivialerweise gegeben ist, da wir hier über Algorithmen sprechen, die keine manuelle Eingabe benötigen.

1.2.3.1 CodeCity

Der Cornerstone von 3D Software-Qualitätsvisualisierung ist das Konzept von CodeCity [WLo7]. Das Paper von Richard Wetzel und Michele Lanza verbindet einige Aspekte, die zu diesem Zeitpunkt neu waren. Sie reden zwar nicht konkret von Software-Qualitäts-Visualisierung, nutzen aber trotzdem Qualitätsmetriken auf Klassenebene, um die richtige Granularität von Software Visualisierung zu finden. Herkömmliche Paper waren oft noch auf niedrigeren Ebenen z.B. Marcus Adrian et al. auf Ausdrucksebene [MFMo3]. Zudem nutzen sie, wie von Young und Munro [YM98] gefordert, eine gute visuelle Metapher, um die Software darzustellen. Sie nutzen das Konzept einer Stadt, um die Software darzustellen. Dabei wird jede Klasse als Gebäude dargestellt, dabei hat jedes Artefakt (Klassen, Pakete und Ordner) verschiedene Attribute: Die Dimension, die Position, die Farbe, die Farbsättigung und die Transparenz. Ein Beispiel für eine solche Visualisierung ist in Abbildung ?? dargestellt.

Das Layout der Stadt lässt sich dabei herunterbrechen auf ein 2D-Layout Problem. Um dieses Problem zu lösen, implementieren sie eine Abwandlung von Treemap-Algorithmen (Was Treemaps sind wird in Abschnitt ?? erklärt). Auffällig bei Ihrer Implementierung ist, dass es viele ungenutz-

te leere flächen ohne gebäude gibt und gebäude anhand ihrer Größe sortiert platziert werden (groß weiter unten links, klein weiter oben rechts). Leider ist der algorithmus nicht frei zugänglich, sodass in dieser Arbeit zum vergleich eine eigene Implementierung des Algorithmus verwendet wird, wodurch natürlich auch die Ergebnisse, von der original implementierung abweichen werden (außer wenn anders gekennzeichnet).

1.2.4 Treemap-Layouts

Eine Treemap visualisiert einen Baum, indem jedem Knoten ein Rechteck mit der Fläche A zugewiesen wird, proportional zu seinem zugewiesenen Wert (z.B. Datenmenge oder Marktwert). Nicht-Blatt-Knoten werden dabei üblicherweise durch Rahmen (Container-Rectangles) gekennzeichnet, um die Gruppierung der Kinder zu zeigen. [BHVW00] Die Rechtecke aller Blätter füllen die Fläche des Wurzelrechtecks vollständig aus. Mathematisch entspricht die Eingabedatenstruktur einem gewichteten Baum, bei dem jede Blatteinheit eine numerische Größe hat. Die Fläche eines Eltern-Rechtecks entspricht der Summe der Flächen (Werte) seiner Kinder.

Die konkrete Idee hierarchische Daten in form von Treemaps darzustellen wurde erstmals 1991 von Shneiderman und Johnson [JS91] vorgestellt. Sie stellten fest, dass die Darstellung von hierarchischen Daten in Form von Bäumen in der Regel nicht sehr anschaulich ist. Sie entwickelten eine Methode, um diese Daten in Form von Rechtecken darzustellen, die die Fläche der Knoten proportional zu ihrem Wert darstellen. Diese Methode wurde als "Treemap" bezeichnet. Als Ziele dieser Visualisierung formulierten sie unter anderem diese Aspekte:

- **Effiziente Nutzung des Platzes:** Generell soll es darum gehen möglichst viele Informationen auf einem kleinen Raum darzustellen.
- **Verständlichkeit:** Die Visualisierung soll so gestaltet sein, dass sie für den Betrachter leicht verständlich ist. Es soll möglich sein schnell und mit nur niedrigem kognitiven Aufwand die dargestellten Informationen zu erfassen.
- **Ästhetik:** Die Visualisierung soll ansprechend gestaltet sein.

Zuvor bestehende Ansätze zur Visualisierung von hierarchischen Daten waren in der Regel nicht sehr anschaulich, besonders, wenn es um große Datenmengen ging. Listen, Baumdiagramme (siehe Abbildung ??) oder andere Darstellungen (auch bekannt als Node oder Link-Diagramme) sind nicht in der Lage alle diese Aspekte zu erfüllen. Bei einem typischen Baumdiagramm zum Beispiel werden teilweise mehr als die Hälfte der Fläche für Hintergrund genutzt [JS91, S. 3] außerdem ist es schwer, außer der Struktur der Daten auch die Metriken darzustellen. Sie kritisieren auch die Darstellung von hierarchischen Daten in Form von Venn-Diagrammen (siehe Abbildung ??): "The space required between regions would certainly preclude this Venn diagram representation from serious consideration for larger structures." [JS91,

S. 5] Es ist zwar möglich durch die Größe der Kreise eine Metrik darzustellen, es sei aber nicht möglich eine große Anzahl an Knoten sinnvoll darzustellen.

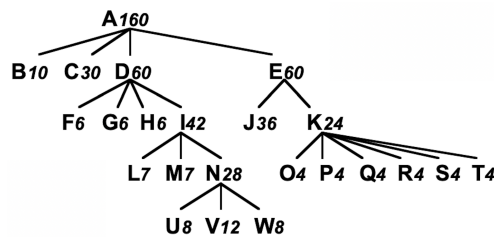


Abbildung 1.3: Beispiel für ein Baumdiagramm

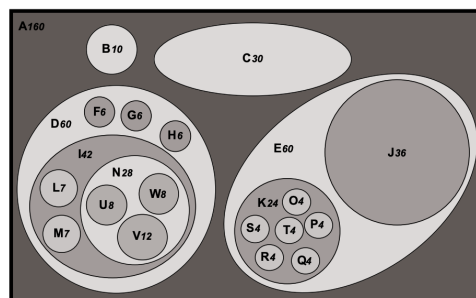


Abbildung 1.4: Beispiel für ein Venn-Diagramm

“Using boxes instead of ovals and a bin-packing algorithm could partially solve this space problem. But bin-packing is an NP-complete problem and does not preserve order.” [JS91, S. 5] Sie stellen fest, dass es theoretisch eine dem Venn Diagramm ähnliche Lösung gibt, die allerdings NP-Hard ist.

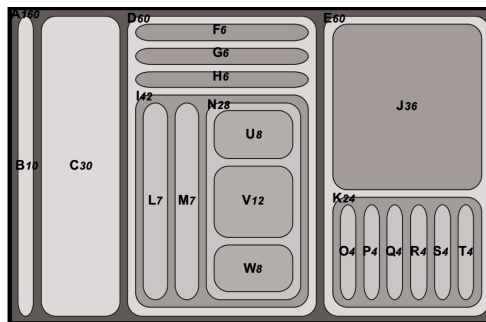


Abbildung 1.5: Beispiel für ein Boxed Venn-Diagramm

Shneiderman und Johnson schlagen zur Lösung dieser Schwierigkeiten ihren Treemap ansatz vor. Sie legen vier Eigenschaften fest, die bei der Erstellung der Treemaps gewährleistet werden:

- Wenn ein Knoten 1 ein Vorfahre von Knoten 2 ist, dann ist der Bereich von Knoten 1 vollständig enthalten in dem Bereich von Knoten 2.
- Die Bereiche von zwei Knoten schneiden sich, wenn ein Knoten ein Vorfahre des anderen ist.

- Knoten belegen eine Fläche, die streng proportional zu ihrem Gewicht ist.
- Das Gewicht eines Knotens ist größer oder gleich der Summe der Gewichte seiner Kinder.

Sie stellen auch einen Algorithmus vor, der diese Eigenschaften erfüllen soll. Ein von diesem Algorithmus erzeugtes Layout ist in Abbildung ?? dargestellt.



Abbildung 1.6: Beispiel für eine Treemap

Der Algorithmus unterteilt den Raum abwechselnd vertikal und horizontal, je nach Größe der Knoten. Der Algorithmus arbeitet sich rekursiv von der Wurzel bis zu den Blättern herunter und hat eine Laufzeit von $O(n)$, wobei n die Anzahl der Knoten ist. Im nächsten Abschnitt wird ein ähnlicher Algorithmus im Detail beschrieben, was zum besseren Verständnis dieser Art von Algorithmen helfen wird.

Obwohl es sich hier um ein renommiertes und viel Zitiertes Paper handelt, machen die Autoren einen entscheidenden Fehler, der in manchen Fällen sogar dazu führen kann, dass Rechtecke komplett verschwinden. Die Autoren betrachten dieses Problem in ihrem Paper leider nicht. Der Abstand zwischen den Knoten wird nämlich dadurch erzeugt, dass die Rechtecke diesen Abstand links, rechts, oben und unten abgezogen bekommen. Dadurch ist die dargestellte Fläche der Rechtecke nicht mehr proportional zu den Werten, die sie darstellen sollen. Eigenschaft 3 wird also verletzt. Dies ist besonders problematisch, wenn die Rechtecke sehr klein bzw. sehr langgezogen sind. In diesem Fall kann es dazu kommen, dass die Rechtecke so klein werden, dass sie nicht mehr dargestellt werden können. Dies stellt in der Praxis ein riesiges Problem dar, speziell wenn das Problem dieser Arbeit also 3D im Kopf behalten wird. Es kann ja theoretisch vorkommen, dass als Flächenmetrik die lines of code verwendet wird und dort ein file, weil er wenige lines hat sehr klein wird und deswegen aufgrund der margins nicht angezeigt wird. Wenn jetzt aber die metrik für die Höhenberechnung die prozentuale test abdeckung der lines ist und der file nicht getestet ist, dann würde ein potentiell großes problem, was auch eigentlich direkt in Auge springen sollte, nicht angezeigt werden.

Ein weiteres Problem bei diesem Algorithmus ist auch generell, dass unter Umständen die Rechtecke sehr langgezogen werden können, was unter Um-

ständen auch Kriterium 3 der Ästhetik verletzt. Es gibt viele Ideen dieses Problem anzugehen. Eine Möglichkeit, den Squarify Algorithmus zu verwenden, wird im nächsten Abschnitt vorgestellt.

Zuletzt noch eine Klarstellung im Bezug auf die Erweiterung von 2D treemaps um eine Dimension. In der Literatur kommt es hier öfter zu Unklarheit. Eigentlich sind 3D treemaps so ähnlich wie 2D treemaps, nur dass sie anstatt ein Rechteck zu unterteilen. Würfel bzw. Quader unterteilen, wobei dann nicht die Fläche sondern das Volumen betrachtet wird. In dieser Arbeit geht es aber nicht darum, sondern um sogenannte 2.5D Treemaps, wobei das 2D Treemap layout nach oben extrudiert wird, um eine dritte Dimension zu erhalten. Wenn in dieser Arbeit also von 3D Darstellung die Rede ist, dann ist damit im Kontext von Treemaps eigentlich eine 2.5D Treemap gemeint.

1.2.4.1 *Squarify-Algorithmus*

Der Squarify-Algorithmus ist ein Layout-Algorithmus für Treemaps, der darauf abzielt, die Fläche der Rechtecke so ausgewogen wie möglich zu gestalten. Bedeutet, dass die Rechtecke möglichst quadratisch sind. Die ursprüngliche Form des Algorithmus wurde im Jahr 2000 von Bruls et al. [BHVW00] vorgestellt. Sie stellten fest: "another problem of standard treemaps [is] the emergence of thin, elongated rectangles" [BHVW00, S. 1]. Wenn Rechtecke nicht mehr so langgezogen sind, Es ist einfacher auf Rechtecke zu zeigen, diese wahrzunehmen, sie zu vergleichen und ihre Größe einzuschätzen.

Der Algorithmus arbeitet rekursiv und teilt die Fläche in Rechtecke auf, wobei er versucht, die Seitenverhältnisse der Rechtecke so nah wie möglich an 1 zu halten. Sie stellen einen rekursiven Ansatz vor, bei dem, wie auch bei den meisten anderen Treemap-Algorithmen, die Rechtecke von oben nach unten (also vom Wurzelknoten bis zu den Blattknoten) aufgeteilt werden.

Im Folgenden wird der Algorithmus beschrieben, da er eine wichtige Grundlage für das Verständnis des Problems darstellt und außerdem eine gute Grundlage zum Verständnis der anderen Algorithmen bietet, da viele Algorithmen ähnliche Ideen verwenden.

Der Algorithmus wird anhand eines Beispiels aus dem originalen Squarify-Paper [BHVW00, S. 5] erläutert. Wir werden den Algorithmus jedoch anders erklären als im Paper, da wir uns näher an der Implementierung orientieren, wie sie in der bekannten d3-Bibliothek [D3t] umgesetzt ist. Es sollen Rechtecke mit den Größen 6, 6, 4, 3, 2, 2, 1 in ein 6 mal 4 Rechteck einsortiert werden.

Der Algorithmus arbeitet immer in Reihen, die er versucht zu füllen und dabei die Rechtecke möglichst quadratisch zu halten. Das erste Rechteck ist breiter als lang. (In dieser Arbeit werden wir, anders als in den herkömmlichen Papern zu Layout Algorithmen, das Wort breit als x-Koordinate und das Wort lang als y-Koordinate nutzen - das hat den Hintergrund, dass das Wort hoch im drei-dimensionalen meist für die z-Komponente genutzt wird und es sonst zu Verwirrungen kommen könnte) Da das Rechteck in das wir einfügen breiter als lang ist, werden wir eine imaginäre horizontale Reihe so lange mit Rechtecken befüllen, bis ein Threshold erreicht ist. Das erste

Rechteck mit gröÙe 6 fügen wir also in Schritt 1 in diese Reihe ein. Das Seitenverhältnis dieses Rechtecks beträgt 8 zu 3 (das Rechteck ist 1.5 Einheiten breit und 4 Einheiten lang). Das zweite Rechteck mit gröÙe 6 fügen wir in Schritt 2 in diese horizontale Reihe ein, dabei wird die Reihe entsprechend breiter. Das Seitenverhältnis des jetzt eingefügten Rechtecks beträgt 3 zu 2 (das Rechteck ist 3 Einheiten breit und 2 Einheiten lang). Jetzt kommt das nächste Rechteck mit gröÙe 4. Dieses Rechteck hat ein Seitenverhältnis von 4 zu 1 (das Rechteck ist 4 Einheiten breit und 1 Einheit lang). Das hinzufügen dieses Rechtecks führt nun aber dazu, dass das schlechteste Seitenverhältnis der Reihe von 3 zu 2 auf 4 zu 1 ansteigt. Deshalb wird diese Reihe als abgeschlossen angesehen und die nächste Reihe wird begonnen - Schritt 4.

Dieser Schritt des suchens des schlechtesten Seitenverhältnisses lässt sich von der rechenkomplexität her gut optimieren, sodass der Ratio berechnet werden kann, ohne dass die Reihe wirklich mit Rechtecken gefüllt werden muss. Anstatt für jedes Rechteck $\max(w/l, l/w)$ zu berechnen, ziehen wir folgende vereinfachung heran.

$$\frac{w_i}{l_i} = \frac{w_i \cdot l_i \cdot w^2}{l_i \cdot l_i \cdot w^2} \quad (1.1)$$

$$= \frac{w_i \cdot l_i \cdot w^2}{l_i^2 \cdot \left(\sum_{j=0}^n w_j\right)^2} \quad (1.2)$$

$$= \frac{w_i \cdot l_i \cdot w^2}{\left(l_i \cdot \sum_{j=0}^n w_j\right)^2} \quad (1.3)$$

$$= \frac{w_i \cdot l_i \cdot w^2}{\left(\sum_{j=0}^n l_i \cdot w_j\right)^2} \quad (1.4)$$

$$= \frac{w_i \cdot l_i \cdot w^2}{\left(\sum_{j=0}^n l_j \cdot w_j\right)^2} \quad \text{da } \forall i, j \in \{0, \dots, n\}, l_i = l_j \quad (1.5)$$

$$= \frac{V_i \cdot w^2}{sV^2} \quad (1.6)$$

Analog dazu gilt das gleich auch für $\frac{l_i}{w_i} = \frac{sV^2}{V_i \cdot w^2}$. Da wir nur an dem maximalen Wert beider Ausdrücke interessiert sind und die Länge (l_i) aller Rechtecke in der Reihe gleich ist, reicht es den Wert für das größte und das kleinste Rechteck zu berechnen und davon den maximalen Wert zu nehmen.

w ist für den gesamten Zeitraum des füllens einer Reihe konstant und muss daher nur einmal berechnet werden. sV wird mit jedem Rechteck aktualisiert.

Das Ziel ist es das Verhältnis der Seitenlängen gleich zu halten. Im ursprünglichen Paper von Bruls et al. [BHVWoo] wird darauf noch nicht so eingegangen, aber viele implementierungen z.b. die von d3.js [D3t] ermögli-

chen es, das Verhältnis nicht nur an den Wert 1 anzunähern, sondern auch an andere Werte, zum Beispiel den goldenen Schnitt [Lu+17].

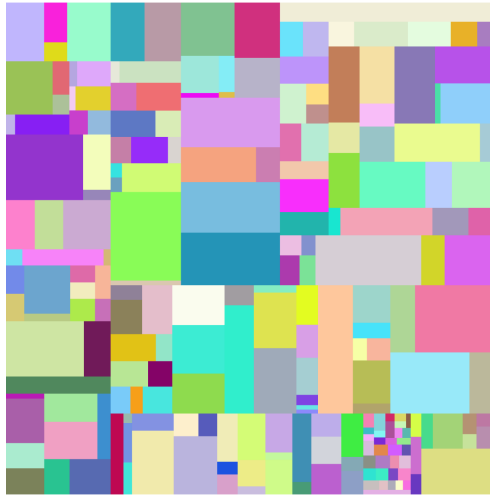


Abbildung 1.7: Beispiel für ein Squarify-Layout mit Annäherung an quadratische Rechtecke (Durchschnittliches Seitenverhältnis 1.42)



Abbildung 1.8: Beispiel für ein Squarify-Layout mit Annäherung an den Wert 5 (Durchschnittliches Seitenverhältnis 2.79)

1.3 PROBLEMSTELLUNG

Die fundamentale und schwerste Frage, bei den Stadt-Analogien ist, wie das Layout der Stadt aussehen soll. In Abschnitt ?? wurde das erste Layout vorgestellt, welches die Stadt-Analogie verfolgt. Die verbindende Komponente dieser Arbeit ist, dass die 3D-Visualisierung am Ende darüber entsteht, dass ein 2D-Layout generiert wird, welches dann durch eine weitere Metrik ins 3 dimensionale extrudiert wird. Und Anschließend weitere Metriken als Einfärbung genutzt werden.

Die übergeordnete Frage dieser Arbeit, die sich aber gar nicht wirklich beantworten lässt, ist:

Welches Layout ist am besten geeignet, um Code-Qualitätsmetriken zu visualisieren? Gibt es bessere Layouts (als das von CodeCity), die eine bessere Übersichtlichkeit und Verständlichkeit bieten?

Daraus leiten wir zwei konkrete Fragen ab, die in dieser Arbeit beantwortet werden sollen:

1. Wie lassen sich die Treemap-Layout-Algorithmen für das konkrete Problem dieser Arbeit verbessern? Welche Vor- und Nachteile haben diese Anpassungen?
2. Lassen sich über Recherche zu verwandten Arbeiten und bestehenden Tools Layouts finden, die sich besser eignen, um Code-Qualitätsmetriken zu visualisieren?
 - a) Beispiel: Ist das Sunburst Layout besser geeignet als das "beste" Treemap-Layout, um Code-Qualitätsmetriken zu visualisieren?
 - b) Beispiel: Ist das Stack Layout besser geeignet als das "beste" Treemap-Layout, um Code-Qualitätsmetriken zu visualisieren?

Laut Marcus Adrian et al. gibt es 5 Dimensionen die man beachten muss, wenn es um Software Visualisierung geht:

- Tasks - why is the visualization needed?
- Audience - who will use the visualization?
- Target - what is the data source to represent?
- Representation - how to represent it?
- Medium - where to represent the visualization? [MFMo3, S. 2]

Wir beantworten diese Fragen, um das Ziel dieser Arbeit zu begründen. **Warum ist die Visualisierung von Codequalitätsmetriken wichtig?** Die Visualisierung von Codequalitätsmetriken ist wichtig, um die Qualität von Softwareprojekten zu bewerten und zu verstehen, aber auch um einen schnellen Überblick über die Codebasis zu geben und einen Einstieg in vertiefende Codeanalysen zu ermöglichen. Eine effektive Visualisierung kann helfen, schnell Hotspots im Code zu identifizieren, die möglicherweise verbessert werden müssen, und somit die Wartbarkeit und Qualität des Codes zu erhöhen. Außerdem soll die Visualisierung ermöglichen verschiedene Metriken in Verbindung zu setzen, um so eine höhere Aussagekraft über den Code, den die einzelne Betrachtung jeder Metrik nicht bieten kann. Der wichtigste Punkt ist, das subjektive *Greifbar* machend der CodeQualität.

Wer wird die Visualisierung nutzen? Die Visualisierung ist vorallem an Personen gerichtet, die sich nicht mit der Codebasis auskennen. Das können Entscheidungsträger sein, die keine ahnung von software entwicklung haben, das können aber auch entwickler sein, die sich neu in ein Projekt einarbeiten müssen, um die qualität einer software zu erhöhen.

Was ist die Datenquelle? Die Datenquelle sind hierarschiche Codequalitätsmetriken, die aus dem Quellcode eines Softwareprojekts extrahiert werden. Dabei wird jeder Knoten in dieser Hierarchie als "Node" bezeichnet. Jede Node hat folgendes Schema: `""json "node": { "name": string, "children": List[Node] | "value": number, ""`

Wo soll die Visualisierung dargestellt werden? Die Visualisierung soll digital auf herkömmlichen Bildschirmen dargestellt werden. Speziell wird in dieser Arbeit beispielhaft eine darstellung in einem Webbrowser angestrebt und die Algorithmen in Typescript implementiert. Natürlich können aber alle Ergebnisse auch in anderen Programmiersprachen und Umgebungen umgesetzt werden.

Wie soll die Visualisierung dargestellt werden? Im GRunde soll eine Visualisierung in Anlehnung an den in Abschnitt ?? beschriebenen Stadt-Metapher ansatz verfolgt werden. - Speziell soll es in dieser Arbeit um das Layout der Knoten gehen, aber im hinterkopf soll die stadtmeter bleiben und immer als grundlage für die bewrtung des 2d layouts dienen. Wie in der CodeCity arbeit beschrieben, soll es möglich sein Metriken in form von Fläche, Höhe und farbe (ob jetzt nur farbe oder durchsichtigkeit wie bei dem codacity paper oder sogar textur von knoten - wird hier ignoriert). Das heißt also, dass das 2D layout in gewisser weise eingeschränkt wird, zB. wenn Farbe als visulaisrung von struktur verwendet werden soll oder wie in [BHVWoo] schattierung.

Speziell soll die Visualisierung am Ende optimiert auf folgende Aspekte sein, die sich aus den im Abschnitt ?? beschriebenen Grundlegenden Aspekten von Softwarevisualisierung, leicht angepasst and das Problem dieser Arbeit, ableiten lassen:

- **Informationsgehalt und Effiziente Nutzung des Platzes:** Die Visualisierung sollte so viele Informationen wie möglich auf so wenig Platz wie möglich darstellen.
- **Niedrige visuelle Komplexität und Verständlichkeit:** Als Gegenspieler zum Informationsgehalt steht die visuelle Komplexität. Die Visualisierung sollte so einfach und verständlich wie möglich gehalten werden, um den Betrachter nicht zu überfordern.
- **Skalierbarkeit:** Die Visualisierung sollte auch bei großen Software-Systemen noch gut lesbar sein. Dies ist besonders wichtig, da wir hier über große Software-Systeme sprechen.
- **Korrelation mit dem Code:** Die Visualisierung sollte eine gute Korrelation mit dem Code haben. Wenn man die Visualisierung sieht, soll man diese auch mit dem Code in Verbindung bringen können. Es sollte also

möglich sein, die Visualisierung mit dem Code zu verknüpfen und so ein besseres Verständnis für die Software zu bekommen.

- **Zweitrangig ist Stabilität gegenüber Änderungen:** Die Visualisierung sollte stabil gegenüber Änderungen in der Software sein, damit der Qualitätszustand der Software einfacher über die Zeit verfolgt werden kann.

1.3.1 *Das Treemap Problem*

In Abschnitt ?? wurde aufgezeigt, dass bereits der initiale Algorithmus von Johnson und Shneiderman [JS91] ein fundamentales Problem aufweist, wenn Treemaps mit Abständen zwischen Knoten dargestellt werden sollen. Zudem ist bereits das Problem, welches von Johnson und Shneiderman adressiert wird NP-Hard [BHVW00, S. 3] und das Problem mit Abständen macht es nicht einfacher.

- Da der Abstand von der Fläche der Knoten abgezogen wird, ist die dargestellte Fläche nicht mehr proportional zum Wert des Knotens.
- Durch das Abziehen der Abstände kann es passieren, dass Knoten verschwinden, wenn entweder die Länge oder die Breite der Knoten kleiner oder gleich dem Abstand ist.

In den Abbildungen ??, ?? und ?? sind Treemap Layouts abgebildet, die mit dem Squarify Algorithmus nach Abschnitt ?? generiert wurden. Der Visualisierten Metriken sind händisch erstellt, um das Problem zu verdeutlichen (siehe Anhang HIER REF EINRÜGEN). In Bild ?? sind alle Knoten sichtbar und die Fläche der Knoten ist proportional zu den Werten der Knoten. In Bild ?? sind immernoch alle Knoten sichtbar, aber die Fläche der Knoten ist nicht mehr proportional zu den Werten der Knoten. Zum Beispiel hat der Große Knoten mit Wert 3000 eine Fläche von ca. 2600, während der kleine Knoten oben links mit Wert 30 nur eine Fläche von ca. 5 hat. In Bild ?? mit Abstand 10 sind dann schon einige Knoten (zum Beispiel der Eben genannte Knoten oben links) nicht mehr sichtbar, da die Breite der Knoten kleiner als der Abstand ist.

Es ist nicht trivial dieses Problem zu lösen, da es die grundlegende Annahme der Treemap Algorithmen verletzt, dass die Fläche aller Knoten bekannt ist, bevor die Knoten platziert werden. Bevor die Knoten platziert werden, ist nicht klar, wie die Fläche der Knoten aussieht, das heißt, es ist auch nicht klar, wie viel Platz für die Abstände zwischen den Knoten benötigt wird. Dies wird klar wenn man sich die Abbildung ?? anschaut. Dort sieht man, dass die Fläche, die für Knoten in ihren Eltern benötigt wird, größer ist als die Fläche, die für die Knoten selbst benötigt wird und diese benötigte Fläche stark vom Layout der Knoten selbst abhängt. Somit ist auch unklar, wie groß die Fläche aller Elternknoten sind.

Das Problem ist jetzt aber, dass wenn die Fläche der Knoten nicht bekannt ist auch das Layout der Knoten nicht berechnet werden kann, da die Fläche



Abbildung 1.9: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 0

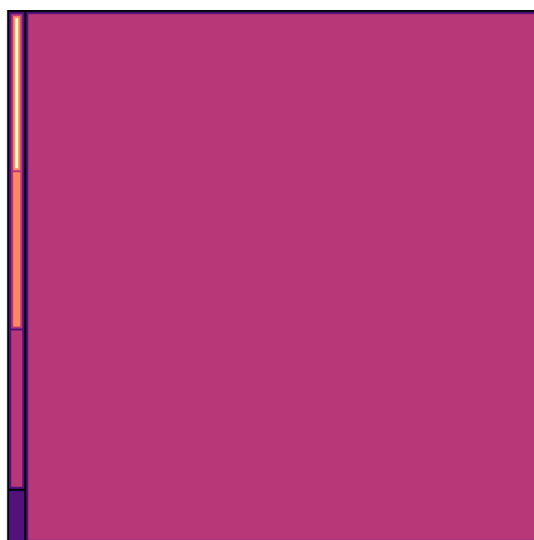


Abbildung 1.10: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 5

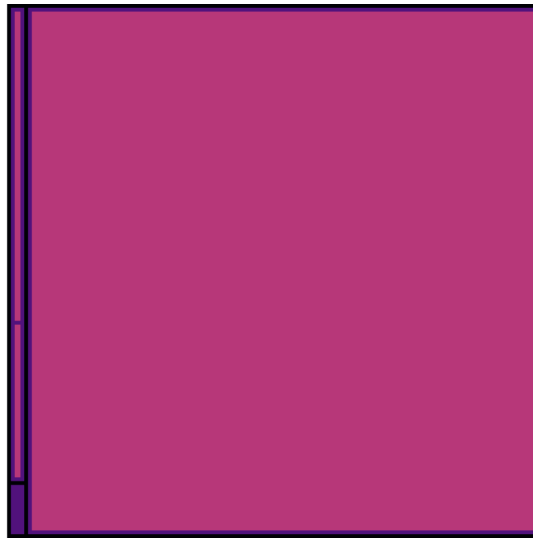


Abbildung 1.11: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 10

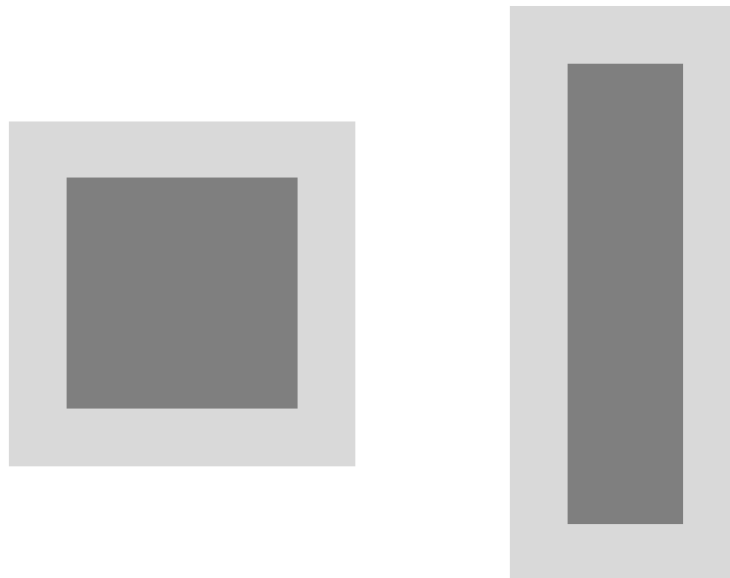


Abbildung 1.12: Abbildung eines zweier Rechtecke mit der Fläche 16 in dunkel grau und in hellgrau der Abstand von 1 um die Flächen herum. Das Linke Rechteck (4x4) mit Abstand nimmt eine Fläche von 25 (5x5) ein. Das rechte Rechteck (2x8) mit Abstand nimmt eine Fläche von 40 (4x10) ein.

der Knoten für das Layout benötigt wird. Hier ergibt sich also ein Zirkelschluss: Die Fläche ist nicht klar, ohne das Layout und das Layout kann nicht berechnet werden, ohne die Fläche zu kennen.

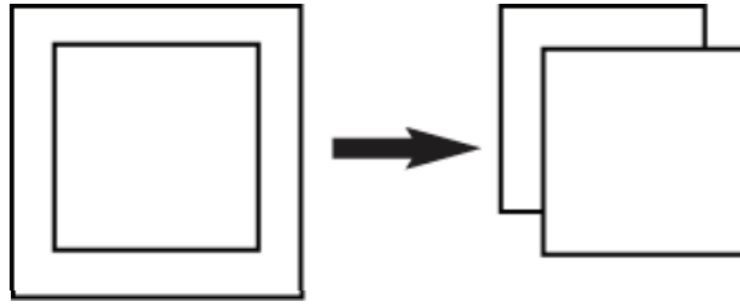


Abbildung 1.13: Links beispielhaft der Nested treemaps Ansatz, bei dem der Kindknoten einfach in dem Elternknoten gezeichnet wird. Rechts der Cascaded Treemap Ansatz, bei dem der Kindknoten als rechteck leicht versetzt nach rechts unten *über* dem Elternknoten gezeichnet wird. Abbildung aus [LFo8, S. 3].

1.4 VERWANDTE ARBEITEN

In diesem Abschnitt werden verwandte Arbeiten vorgestellt, die auch das Layoutproblem betrachten. Außerdem will ich auch konkrete Tools anschauen, die Metriken visualisieren und deren Ansätze betrachten.

1.4.1 *layouts*

So etwas ähnliches auch sagen: A number of other hierarchy visualization techniques have been developed [18, 23, 29, 32], including space-filling visualizations like step trees [6], Voronoi treemaps [2] and generalized treemaps [34]. Although relevant to hierarchy visualization, we pursue contributions that are sufficiently distinct from such work that we do not dwell on extensive comparisons. [LFo8]

1.4.1.1 *Treemap layouts*

Es gibt bisher noch zu wenige ARbeiten, die sich mit dieser Frage beschäftigen. (Oder eigentlich gibt es nur eine die wirklich in die richtung geht - mehr dann bei verwandte arbeiten)

[LFo8]: Hao Lü and James Fogarty stellten in ihrem Paper *Cascaded Treemaps: Examining the Visibility and Stability of Structure in Treemaps*[LFo8] fest: “an important limitation of treemaps is the difficulty of discerning the structure of a hierarchy”[LFo8, S. 1] Das stellt im Grunde das Problem von Treemaps dar, welches auch algorithmisch nicht einfach gelöst werden kann (siehe Abschnitt ??). Die Idee ist anders als bei Nested Ansätzen die Kindknoten nicht einfach in den Elternknoten zu zeichnen, sondern sie leicht versetzt *über* dem Elternknoten zu zeichnen (siehe Abbildung ??). Dadurch soll weniger Platz verloren gehen und es entsteht ein leichter 3D-Effekt.

Sie stellen in ihrem Paper auch fest, dass manche Knoten verschwinden können, da der Platz der für Beschriftung und abstände benötigt wird, beim



Abbildung 1.14: Beispiel für eine Cushion Treemap [VWW99, S. 4]

Layoutschritt nicht berücksichtigt werden kann. Sie stellen einen Zwei Schrit-tigen ansatz vor, der im ersten schritt mit den squarify algorithmus [BHVW00] das layout erstellt. Im zweiten schritt wird dann die GröÙe, aber nicht die Platzierung der Knoten angepasst, indem der Abstand und platz für Beschriftung berücksichtigt wird. Das Problem von verschwindenden Knoten wird dadurch nicht komplett gelöst, aber Knoten verschwinden nur noch, wenn der platz für die Beschriftung und den Abstand größer ist als der zur Verfügung stehende platz. In dem Paper wird leider nicht auf das Problem eingegangen, welches wir in Abschnitt ?? betrachtet haben, dass die Fläche, die durch Abstände eingenommen wird nicht exakt berechnet werden kann. Außerdem wird nicht beleuchtet, wie sehr sich die relation von KnotengröÙe und Knotenwert ändert. In diesem Paper werden wir auf beide Probleme in der tiefe eingehen und die Auswirkungen auf die darstellung untersuchen.

Die Autoren des ursprungs squarify Algorithmus [BHVW00] stellen in einem anderen Paper [VWW99] eine Idee vor, um Struktur ohne Änderung des Layouts darzustellen und zwar mit Schatten (siehe Abbildung ??). Dabei bekommt jeder Knoten, egal ob Eltern- oder Kindknoten, einen Innenrenschatten, wodurch die Struktur der Knoten sichtbar wird.

Peter Demian und Renate Fruchter stellen die Idee vor dass dickere outlines je höher der Knoten ist auch die Struktur verdeutlichen können.

Nicholas Kong et al schlagen vor verschiedene Umriss dicken zu nutzen um die Struktur der Knoten darzustellen.[KHA10] (siehe Abbildung ??).

Scheibel et al. stellen in ihrem Paper *Survey of treemap layout algorithms*[SLD20] eine Übersicht über die verschiedenen Treemap Layout Algorithmen vor. Sie unterscheiden zwischen den verschiedenen Ansätzen.

Sie kategorisieren die verschiedenen Ansätze in 4 Kategorien: Art der Aufteilung, zusätzliche Attribute, Layout Form und Referenzraum Dimension. Uns interessiert hier besonders das zusätzliche Attribut der "Wertetünd die



Abbildung 1.15: Beispiel für eine Treemap mit verschiedenen dicken outlines, die die Struktur verdeutlichen sollen. Abbildung aus [KHA10, S. 1]

Dimension 2D (wie in Abschnitt ?? beschrieben). Arten der Aufteilung sind: Packing und splitting. Packing ist dabei die Idee, die Knoten so zu packen, dass sie möglichst wenig Platz verbrauchen und splitting ist die Idee Knoten in kleinere Knoten zu unterteilen. Alle in den Grundlagen (Abschnitt ??) vorgestellten Algorithmen sind also splitting Algorithmen. Sie stellen vier Layout Formen vor: Kreisförmig, rechteckig, konvex und nicht konvex. Alle in den Grundlagen vorgestellten Algorithmen sind rechteckige Layouts. Von den untersuchten 81 Algorithmen sind 54 rechteckige Layouts und 58 splitting Algorithmen. Der in dieser Arbeit speziell untersuchte Treemap Algorithmus passt also genau in die Kategorie der meist verwendeten Ansätze.

2.1 STADT LAYOUT ANPASSUNGEN

In diesem Abschnitt geht es nicht um die Algorithmen, die zur erzeugung der Layouts verwendet werden, sondern es geht um die Layouts selbst. Hier wird immer konkret begründet, warum ich mich für bestimmte layouts entschieden habe und eventuell auch warum ich mich gegen andere entschieden habe. Dann werden diese Layouts - auf das in dieser Arbeit behandelte konkrete Problem - angepasst, damit sie dann am Ende evaluiert werden können.

CodeCity: Optionen: Eins: Die erste Änderung ist, dass die Knoten nicht mehr quadratisch sind, um den Platz besser auszunutzen und weniger freie ungenutzte Flächen zu haben. Zwei: Abstände zwischen Knoten könnten auf top level höher sein und nach unten hin immer kleiner werden, um den Platz besser auszunutzen und Probleme mit den Algorithmen zu verkleinern. Drei: Ordner könnten ein Label bekommen, dass die Fläche in eine dimension vergrößert. Anonsten wird das schwer zb. mit schwebenden Labels, da wird alles unübersichtlich. Das vielleicht auch nur auf den Top Level Ordnern?

Ganz Andere Idee: Street maps??? also order werden zu straßen und die Knoten zu Häusern

Ganz Andere Idee: Kreis förmig anordnen? Es gibt ja auch runde städte (oder bar chart)

Ganz Andere Idee: Circle packing

2.2 ERWEITERUNG DES SQUARIFY ALGORITHMUS

In diesem Abschnitt wird der Squarify Algorithmus [BHVWoo], wie er in Abschnitt ?? beschrieben wurde, auf verschiedene Weisen erweitert und angepasst, um das Layoutproblem, wie es in Abschnitt ?? beschrieben wurde, anzugehen.

2.2.1 *Approximative Fläche*

Die Grundlegende Idee dieser Erweiterung ist es, die Fläche der Knoten plus die Benötigte Fläche für die Abstände vor berechnung des Layouts zu approximieren.

Dafür brauche ich erstmal einen guten Algorithmus der das Layout gut macht, dann kann ich mit KI die Fläche lernen. Ist die frage ob das wirklich so gut funktionieren kann.

Problem man kann sich sehr gut beispiele konstuieren, bei denen das nichth funktionieren wird. Man kann das natürlich mit skalierung wieder lösen, aber das ist natürlich nicht optimal.

2.2.2 *Zweifache Berechnung*

Die Grundlegende Idee dieser Erweiterung ist es, dass sich die Fläche der Knoten mit Abstand durch das Layout und die Fläche der Knoten ohne Abstand approximieren lässt. Die Idee ist es also einen ersten Durchlauf zu machen, bei dem das Layout ohne Abstand berechnet wird. Dann werden die Größen der Knoten entsprechend dem Layout angepasst, sodass die Größe der Knoten nun auch den Abstand berücksichtigt. Anschließend wird ein zweiter Durchlauf mit diesen angepassten Größen durchgeführt, um das finale Layout zu berechnen.

Bevor wir uns die Details und Ergebnisse dieser Erweiterung anschauen, wollen wir vorweg nehmen, dass diese Erweiterung natürlich nicht optimal funktionieren kann und das auch klar ist, da sich die die Änderung der Größe der Knoten natürlich auch das Layout im Zweiten Durchlauf ändern wird, wodurch die Größen der Knoten wieder nicht korrekt sind. Was ja überhaupt erst das Grundlegende Problem ist (siehe Abschnictt ??). Allerdings ist es ein erster Schritt sich dem Problem zu nähern und zu schauen, ob es sich lohnt in diese Richtung weiter zu forschen.

Der Grundlegende Algorithmus bleibt also (fast) gleich, nur dass zwischen dem ersten und dem zweiten Durchlauf ein zusätzlicher Schritt *Größenanpassung* eingefügt wird. Die Einzige änderung die vorgenommen werden muss ist, dass Knoten nur mit dem definierten Abstand zwischen dem Elternknoten platziert werden können und generell die Fläche des Elternknotens um den Abstand verkleinert wird. Außerdem ist es nötig nach dem zweiten Durchlauf die Knoten, deren größenwert ja nun den abstand beinhalten, zu verkleinern, um auch den abstand zwischen den Geschwistern herzustellen. Es ist zu erkennen, dass dadurch der Abstand sowohl zwi-

schen Geschwistern als auch zu den Elternknoten den doppelten wert des definierten Abstands hat, dieses Problem ignorieren wir hier, da man es trivialerweise lösen könnte, indem man immer nur die hälfte des Abstands zwischen Geschwistern und Elternknoten abzieht, was wir hier der Einfachheit halber nicht tun. – ODER VIELLEICHT HIER IN DER THESIS DOCH? DANN KÖNNTE ICH MIR DIESEN ABSCHNITT SPAREN, AUCH WENN ES IN DER IMPLEMENTIERUNG AM ENDE ANDERS IST

Wir stellen verschiedene Ansätze vor, was sowohl die Größenanpassung als auch die Anpassung der Knoten nach dem zweiten Durchlauf angeht. Die Algorithmen funktionieren allerdings alle nach ähnlichem Prinzip: Es wird zunächst für jeden Knoten die Fläche die der Abstand in diesem Layout benötigen würde addiert, indem die Fläche aus der neuen Länge (alte Länge + 2 mal den Abstand) und der neuen Breite (alte Breite + 2 mal den Abstand) berechnet wird. Zusätzlich wird für Elternknoten die Flächenvergrößerung aller Kinderknoten addiert. An dieser Stelle ist allerdings nicht klar, wie sich die Flächenvergrößerung der Kinderknoten auf die Fläche der Elternknoten auswirkt, da diese Änderung selbst von der Anordnung der Kinderknoten abhängt. Wir testen verschiedene Ansätze, um die Flächenänderung der Elternknoten in Abhängigkeit zu der Flächenänderung der Kinderknoten zu approximieren.

Nach dem zweiten Durchlauf wird nun die Fläche der Knoten so reduziert, dass sowohl der Abstand zwischen Geschwistern als auch der Abstand zu den Elternknoten den gewünschten Wert hat. Dies ist straight forward und wird hier nicht weiter erläutert. Anzumerken ist aber das dieser Schritt speziell abhängt von der Art der Größenanpassung, die im ersten Schritt durchgeführt wurde.

2.2.2.1 Einfache Größenanpassung

Dies ist die einfachste naive Version der Größenanpassung, der Algorithmus zeigt aber gut die zuvor beschriebenen Probleme auf. Die Fläche der Knoten wird um den Abstand in beiden Richtungen vergrößert. Zusätzlich wird die Fläche der Elternknoten um die Fläche der Kinderknoten vergrößert.



Abbildung 2.1: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 0 und der einfachen Größenanpassung auf der händisch erstellen map (siehe Anhang).

Algorithm 1 Einfache Größenanpassung

```

1: function INCREASEVALUESIMPLE(node: SquarifyNode, margin: number)
2:   childrenValueIncrease  $\leftarrow$  0
3:   if node.children then
4:     for child in node.children do
5:       childrenValueIncrease += increaseValuesSimple(child, margin)
6:     end for
7:   end if
8:   valueIncrease  $\leftarrow$  width * margin * 2 + length * margin * 2 + margin *
   margin * 4 + childrenValueIncrease
9:   node.value += valueIncrease
10:  return valueIncrease
11: end function

```

Das Problem des Algorithmus ist am besten an einem Beispiel zu verdeutlichen. Wir visualisieren den ZWITEN AnHANG - auch wieder eine händisch erstellte Map, um das Problem zu verdeutlichen. In Abbildung ?? ist das Layout mit einem Abstand von 0 zu sehen.

In Abbildung ?? ist das Layout mit einem Abstand von 1 zu sehen. Es ist zu erkennen, dass die Knoten auf der linken Seite des Layouts sich außerhalb ihrer Elternknoten erstrecken. Warum passiert das? Knoten 10 ist im zweiten Layout-Schritt deutlich schmaler als im ersten Layout-Schritt. Dadurch wird die Fläche die durch den Abstand eingenommen wird größer als angenom-

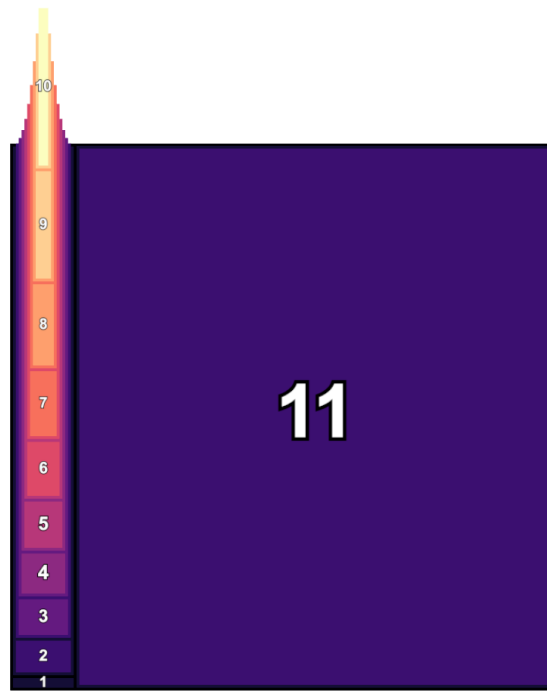


Abbildung 2.2: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 1 und der einfachen Größenanpassung auf der händisch erstellen map (siehe Anhang).

men, weshalb die Fläche des Knotens nach abzug des Abstands im letzten Schritt kleiner ist, als gewünscht. Dementsprechend ist auch die Fläche der Knoten unten links größer als gewünscht, da das Layout der Knoten im zweiten Layout-Schritt quadratischer wird. Knoten 5, der Knoten, der am Quadratischsten ist, wird also am größten erscheinen. Obwohl beide den selben wert haben ist Knoten 5 ca. 1,2 mal größer als Knoten 10) In diesem Beispiel erscheint der unterschied kaum merklich, aber es gibt ihn trotzdem und in anderen fällen kann dieser Unterschied merklich werden. Viel signifikanter ist aber der Effekt, dass Elternknoten ebenfalls immer schmaler werden, wodurch die fläche, die der innerere abstand einnimmt, ebenfalls größer wird und dass sogar immer mehr von ebene zu ebene, wenn man runter geht. Dadurch wird die Fläche für die Kindknoten immer kleiner, was dazu führt, dass Knoten teilweise über ihre Elternknoten hinauswachsen.

Dieser Effekt kann einfach behoben werden, wie wir in Abschnitt ?? sehen werden.

2.2.2.2 Relative Größenanpassung

Bei der berechnung zuvor wurde einfach der margin fläche eines Elternknotens berechnet, auf basis der alten Fläche, ohne die Flächenänderung durch die Kinderknoten zu berücksichtigen. dadurch wird die resultierende Fläche der Elternknoten zu klein sein, da die Fläche der Kinderknoten nicht berücksichtigt wird. Die Idee bei dieser Erweiterung ist es die Änderung der Kinder schon vor dem Hinzufügen der Abstandsfläche zu berücksichtigen.

Dafür muss die relative Flächenänderung durch die Kindknoten berechnet werden. und damit dann die Seitenlängen anpassen und dann die margins hinzufügen, um die neue Fläche zu erhalten. Der Algorithmus wird im folgenden als Pseudocode dargestellt (siehe Algorithmus ??).

Algorithm 2 Relative Größenanpassung

```

1: function INCREASEVALUES(node: SquarifyNode, margin: number)
2:   childrenValueIncrease  $\leftarrow$  0
3:   if node.children then
4:     for child in node.children do
5:       childrenValueIncrease += increaseValues(child, margin)
6:     end for
7:   end if
8:   ratioChildrenValueIncrease  $\leftarrow$  (node.value + childrenValueIncrease)
   / node.value
9:   valueIncrease  $\leftarrow$  Math.sqrt(ratioChildrenValueIncrease) * width *
   margin * 2 + Math.sqrt(ratioChildrenValueIncrease) * length * margin
   * 2 + margin * margin * 4 + childrenValueIncrease
10:  node.value += valueIncrease
11:  return valueIncrease
12: end function

```

Problem: Der Algorithmus in dieser Form zeigt einige Probleme auf. Die Flächenvergrößerung der Kindknoten sagt nichts darüber aus, in welche Richtung sich die Fläche ändert. Es wird davon ausgegangen, dass sich die Fläche gleichmäßig in beide Richtungen ändert (siehe Zeile 13 und 14 in Algorithmus ZEILEN ANPASSEN ??). es kommt also zu ähnlichen Problem wie davor. es können knoten sowohl zu viel platz als auch zu wenig Platz bekommen, jenachdem ob Sie im zweiten schritt quadratischer oder schmaler werden. Siehe Abbildung ?? für ein Beispiel.

2.2.3 *Scaling der Knoten*

Wenn die Fläche innerhalb von Elternknoten immer kleiner wird, kann es passieren, dass Knoten über ihre Elternknoten hinauswachsen, wie in Abbildung ?? zu sehen ist. Es kann genauso passieren, dass die Fläche innerhalb der Knoten größer wird, wie in Abbildung ?? auf der linken seite zu erkennen ist. Dieser Effekt kann trivialer weise behoben werden, indem der zweite Layout-Schritt angepasst wird, sodass die Knoten immer auf die Fläche des Elternknotens skaliert werden. Vor jeden Squarify-Schritt wird dafür die wirklich zur verfügung stehende Fläche des Elternknotens berechnet und die Kindknoten entsprechend dieser Änderung skaliert, sodass sie genau in die Fläche des Elternknotens passen.

Der Nachteil dieser Methode ist in Abbildung ?? zu erkennen. Die Knoten werden dadurch natürlich nicht mehr proportional zu ihren Werten sein.

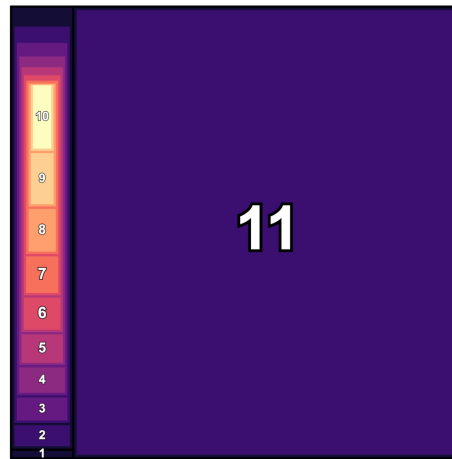


Abbildung 2.3: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 1 und der relativen Größenanpassung auf der händisch erstellen map (siehe Anhang).

Knoten 10 hat zum Beispiel ein Verhältnis von ca. 0.5 zu seinem Wert, während Knoten 6 ein Verhältnis von ca. 1.4 zu seinem Wert hat.

Je genauer die Größenanpassung der Knoten ist, desto geringer fällt natürlich dieser Effekt aus. Siehe im Vergleich dazu Abbildung ??, da die relative Größenanpassung deutlich genauer ist, ist der Effekt hier auch deutlich geringer. Knoten 10 hat Größe 40 und Knoten 6 hat Größe 41, was fast ähnlich Groß ist.

Persönlich würde ich sagen, dass dieser Effekt für das herunter skalieren der Knoten gut ist, da sonst eine grundlegende eigenschaft der Darstellung verletzt wird, aber für das hoch skalieren der Knoten könnte man auch sagen, dass es wichtiger ist die Fläche der Knoten proportional zu ihren Werten zu halten, als die *verschwendete* Fläche aufzufüllen. Diese jetzt hier sehr subjektive Einschätzung wird aber nochmal genauer beläuchtet in der evaluation und vergleich.

2.2.4 Mehrfache Berechnung

Die Grundlegende Idee dieser Erweiterung ist es, das Layout mehrfach zu berechnen und dabei die Fläche der Knoten immer weiter anzupassen. Dies wird so lange wiederholt, bis sich die Fläche der Knoten nicht mehr ändert oder eine maximale Anzahl an Iterationen erreicht ist.

2.2.5 Fazit

Immernoch straight forward, es gibt aber noch Probleme

Warum besonders den squarify algorithmus betrachtet und nicht zum beipsiel pivot oder circle?? -> weil diese Andere Ziele haben -> noch mehr begründen anhand der definierten Ziele

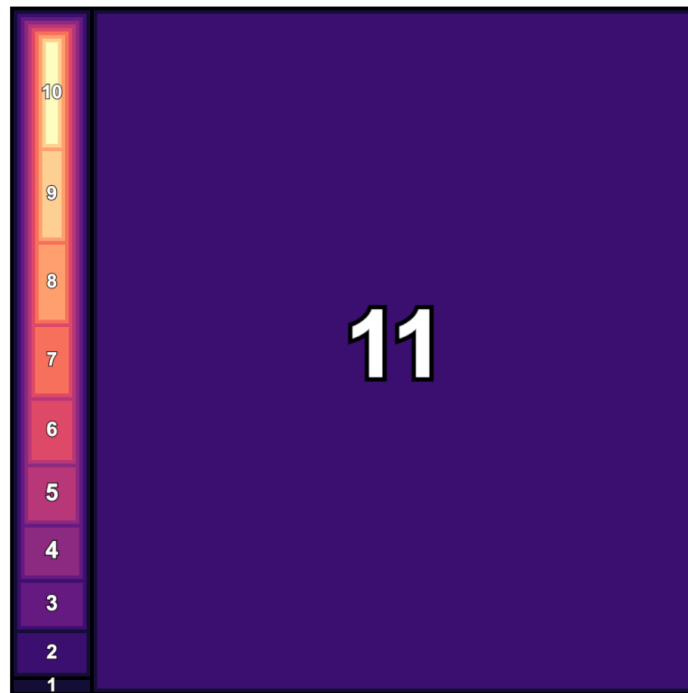


Abbildung 2.4: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 1 und der einfachen Größenanpassung auf der händisch erstellen map (siehe Anhang) und der Skalierung der Knoten.

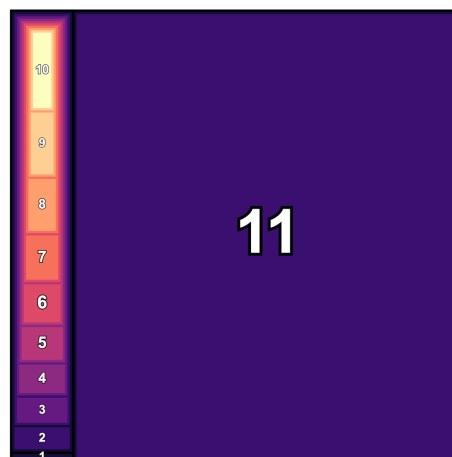


Abbildung 2.5: Treemap Layout generiert mit dem Squarify Algorithmus nach Abschnitt ?? mit einem Abstand von 1 und der relativen Größenanpassung auf der händisch erstellen map (siehe Anhang) und der Skalierung der Knoten.

SCHLUSS

3.1 EVALUATION

Evaluation anhand eines echten beispiels mit verschiedenen Layouts. In diesem Abschnitt werden 1. die verschiedenen Treemap-Algorithmen verglichen und anhand von konkreten Kriterien evaluiert, um den besten Algorithmus für die Visualisierung von Codequalitätsmetriken zu finden. 2. Werden alle vorgestellten Layouts anhand eines konkreten Beispiels aus der Praxis evaluiert. Dies wird gemacht, indem eine echte, bereits durchgeführte, Software-Analyse nochmal mit den verschiedenen Layouts visualisiert wird. Dann wird verschiedenen unerfahrenen Personen diese verschiedenen Visualisierungen gezeigt und gefragt, wie sie die Software-Qualität einschätzen würden. Wenn das möglichst nahe an die Experten Einschätzung kommt, so wie es in der Analyse durchgeführt wurde, dann ist das Layout gut geeignet.

3.2 KRITERIEN FÜR TREEMAP LAYOUTS

Um diese fünf Aspekte zu erreichen, definieren wir sechs Kriterien für das 2D layout, die diese Aspekte messbar machen. Diese Kriterien sollen helfen, die Qualität der Visualisierung zu bewerten und zu vergleichen. Die Kriterien sind:

- **Abstände:** Abstände zwischen den Knoten verbessern die Übersichtlichkeit und die visuelle Komplexität.
- **Platznutzung:** Es sollte so wenig Fläche wie möglich ohne Informationsgehalt bleiben. Als gegenbeispiel kann man die Order-Knoten sehen, wie sie in der CodeCity Arbeit beschrieben wurden, bei denen die Fläche der Knoten nicht proportional zur Anzahl der Zeilen im Code ist, wodurch die Fläche an sich keinen Informationsgehalt mehr hat und außerdem viel leere Fläche entsteht.
- **Knoten sichtbarkeit:** Es sollte keine Knoten geben, die aufgrund von Abständen oder anderen Gründen nicht sichtbar sind. Dieses Ziel spielt speziell auf das in abschnitt ?? beschriebene Problem ab.
- **Zeitaufwand:** Die Generierung des Layouts sollte in einem angemessenen Zeitrahmen erfolgen, um eine schnelle Visualisierung zu ermöglichen. Dies verbessert die Skalierbarkeit und generelle Nutzbarkeit der Visualisierung.
- **Seitenverhältnis:** Um die visuelle Komplexität zu reduzieren und die Verständlichkeit zu erhöhen, sollte das Seitenverhältnis der Knoten möglichst nahe bei 1:1 liegen. Dies verbessert die Lesbarkeit der Knoten und macht es einfacher, die Informationen zu erfassen.
- **Flächengröße:** Um die Korrelation mit dem Code zu gewährleisten, sollte die Fläche der Knoten proportional zum Metrikwert sein. (Hier ist noch unklar, ob das nur für Blätter gilt oder für alle Knoten)

- **Stabilität:** Die Knoten sollten bei Änderungen die Position und Größe beibehalten, um eine stabile Visualisierung zu gewährleisten.

In dieser Arbeit sollen space filling approaches analysiert werden und speziell darauf untersucht werden, wie sie sich eignen für die definierten Anforderungen. Wie gut wird was erfüllt? Wann sollte man was anwenden? Kann eine gute Kombination aus verschiedenen Ansätzen gefunden werden? Bisher wurden im Grundlagenteil vor allem die Splitting Algorithmen vorgestellt, aber es gibt natürlich auch andere Ansätze, die verfolgt werden können, um Treemap Layouts zu generieren. Zum Beispiel gibt es Bin Packing oder Optimierungs Algorithmen. In dieser Arbeit sollen auch diese Ansätze betrachtet werden, um zu sehen, ob sie für die Visualisierung von Codequalitätsmetriken in einer space filling layout approach geeignet sind.

“Especially for the node weights, we assess if the algorithms use the weight values to scale the sizes of leaf nodes. Typically, the size of parent nodes is adjusted to the spatial consumption of the child nodes.”[SLD20, S. 3] - Scheibel et al. sagen also, für sie ist es nicht so wichtig, dass die Größe der Elternknoten proportional zu den Metrikwerten der Knoten ist.

Dies soll getestet werden auf Basis von verschiedenen öffentlichen Repositories, die von kleinen bis großen Codebasen reichen. Als Metrik für die Fläche soll der Einfachheit halber die Anzahl der Zeilen verwendet werden.

3.3 FAZIT

In dieser arbeit konnte natürlich ein eine endliche Anzahl an Ansätzen und Ideen untersucht werden. In Zukunft ist bestimmt noch viel kreativität und Forschungspotential in diesem Bereich. Außerdem vielleicht sogar komplett neue Ansätze, die nicht auf der Stadt-Metapher basieren sondern vielleicht auf planeten oder anderen Metaphern, auch wenn zu vermuten ist, dass dabei die übersichlichekeit und einfachheit leiden könnte. Die in diser arbeit untersuchten Ansätze sind alle sehr einfach mit basic formen und daher übersichtlich.

Teil II

APPENDIX

anhang/artificial.cc.json

LITERATUR

- [Sof] [Online; accessed 17-June-2025]. 2019. URL: <https://www.study-smarter.de/studium/informatik-studium/softwareentwicklung/softwaremetriken/>.
- [Iso] 2024. URL: <https://www.iso25000.com/index.php/en/iso-25000-standards/iso-25010>.
- [Wor] [Online; accessed 17-June-2025]. 2025. URL: <https://www.gitcl ear.com/four/%5Fworst/%5Fsoftware/%5Fmetrics/%5Fagitating/%5Fdevelopers>.
- [D3t] [Online; accessed 19-May-2025]. 2025. URL: <https://d3js.org/d3-hierarchy/treemap>.
- [Atz+21] Daniel Atzberger, Tim Cech, Merlin Haye, Maximilian Söchting, Willy Scheibel, Daniel Limberger und Jürgen Döllner. “Software Forest: A Visualization of Semantic Similarities in Source Code using a Tree Metaphor”. In: Feb. 2021, S. 112–122. DOI: [10.5220/0010267601120122](https://doi.org/10.5220/0010267601120122).
- [BHVWoo] Mark Bruls, Kees Huizing und Jarke J Van Wijk. “Squarified treemaps”. In: *Data Visualization 2000: Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization in Amsterdam, The Netherlands*. Springer. 2000, S. 33–42.
- [EH98] Institute of Electrical und Electronics Engineers (Hrsg.) *IEEE Std 1061-1998: IEEE Standard for a Software Quality Metrics Methodology*. Kapitel 2. Definitions, S. 2. New York: IEEE, 1998. ISBN: 1-55937-529-9.
- [HF94] Tracy Hall und Norman Fenton. “Implementing software metrics—the critical success factors”. In: *Software Quality Journal* 3 (1994), S. 195–208.
- [JS91] Brian Johnson und Ben Shneiderman. *Tree-maps: A space filling approach to the visualization of hierarchical information structures*. Techn. Ber. UM Computer Science Department; CS-TR-2657, 1991.
- [KMoo] C. Knight und M. Munro. “Virtual but visible software”. In: *2000 IEEE Conference on Information Visualization. An International Conference on Computer Visualization and Graphics*. 2000, S. 198–205. DOI: [10.1109/IV.2000.859756](https://doi.org/10.1109/IV.2000.859756).
- [KHA10] Nicholas Kong, Jeffrey Heer und Maneesh Agrawala. “Perceptual Guidelines for Creating Rectangular Treemaps”. In: *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2010). URL: <http://vis.stanford.edu/papers/perception-treemaps>.

- [LFo8] Hao Lü und James Fogarty. “Cascaded treemaps: examining the visibility and stability of structure in treemaps”. In: *Proceedings of graphics interface 2008*. 2008, S. 259–266.
- [Lu+17] Liangfu Lu, Shiliang Fan, Maolin Huang, Weidong Huang und Ruolan Yang. “Golden Rectangle Treemap”. In: *Journal of Physics: Conference Series* 787.1 (2017), S. 012007. DOI: [10.1088/1742-6596/787/1/012007](https://doi.org/10.1088/1742-6596/787/1/012007). URL: <https://dx.doi.org/10.1088/1742-6596/787/1/012007>.
- [Lud] Jochen Ludewig. *Wie gut ist die Software?* URL: <https://elib.uni-stuttgart.de/server/api/core/bitstreams/d40bec3f-ba78-495f-9c44-4e54db9daa30/content>.
- [MFM03] Andrian Marcus, Louis Feng und Jonathan I. Maletic. “3D representations for software visualization”. In: *Proceedings of the 2003 ACM Symposium on Software Visualization*. SoftVis ’03. San Diego, California: Association for Computing Machinery, 2003, 27–ff. ISBN: 1581136420. DOI: [10.1145/774833.774837](https://doi.org/10.1145/774833.774837). URL: <https://doi.org/10.1145/774833.774837>.
- [Rei95] Steven P. Reiss. “An Engine for the 3D Visualization of Program Information”. In: *Journal of Visual Languages & Computing* 6.3 (1995), S. 299–323. ISSN: 1045-926X. DOI: <https://doi.org/10.1006/jvlc.1995.1017>. URL: <https://www.sciencedirect.com/science/article/pii/S1045926X85710178>.
- [SLD20] Willy Scheibel, Daniel Limberger und Jürgen Döllner. “Survey of treemap layout algorithms”. In: *Proceedings of the 13th international symposium on visual information communication and interaction*. 2020, S. 1–9.
- [Sch19] Gast Carina Schmitz. *Wenn Software auf Qualität (s-Metriken) trifft — Teil 2 - MEDtech Ingenieur GmbH*. [Online; accessed 17-June-2025]. 2019. URL: <https://medtech-ingenieur.de/wenn-s-software-auf-qualitaet-s-metriken-trifft-teil-2/>.
- [VWW99] J.J. Van Wijk und H. Van de Wetering. “Cushion treemaps: visualization of hierarchical information”. In: *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis’99)*. 1999, S. 73–78. DOI: [10.1109/INFVIS.1999.801860](https://doi.org/10.1109/INFVIS.1999.801860).
- [VK17] Jeffrey Voas und Rick Kuhn. “What Happened to Software Metrics?” In: *Computer* 50.5 (2017), 88–98. DOI: <https://doi.org/10.1109/mc.2017.144>. URL: <https://tsapps.nist.gov/publication/get%5Fpdf.cfm?pub%5Fid=922615>.
- [WLo7] Richard Wettel und Michele Lanza. “Visualizing Software Systems as Cities”. In: *2007 4th IEEE International Workshop on Visualizing Software for Understanding and Analysis*. 2007, S. 92–99. DOI: [10.1109/VISSOF.2007.4290706](https://doi.org/10.1109/VISSOF.2007.4290706).

- [WPo4] Autoren der Wikimedia-Projekte. *Beschreibung von Eigenschaften einer Software durch Zahlen*. [Online; accessed 17-June-2025]. 2004. URL: <https://de.wikipedia.org/wiki/Softwaremetrik>.
- [Wit18] Frank Witte. "Metriken für die Softwarequalität". In: *Metriken für das Testreporting: Analyse und Reporting für wirkungsvolles Testmanagement*. Wiesbaden: Springer Fachmedien Wiesbaden, 2018, S. 63–70. ISBN: 978-3-658-19845-9. DOI: [10.1007/978-3-658-19845-9_8](https://doi.org/10.1007/978-3-658-19845-9_8). URL: https://doi.org/10.1007/978-3-658-19845-9_%5F8.
- [YM98] P. Young und M. Munro. "Visualising software in virtual reality". In: *Proceedings. 6th International Workshop on Program Comprehension. IWPC'98 (Cat. No.98TB100242)*. 1998, S. 19–26. DOI: [10.1109/WPC.1998.693276](https://doi.org/10.1109/WPC.1998.693276).