Jens Gerken
Mat. 01/473804
Universität Konstanz

Bachelor-Arbeit

Evaluation eines Metadatenbrowsers

Liste vs. Leveltable

Jens Gerken

1. Gutachter: Prof. Dr. Harald Reiterer
2. Gutachter: Prof. Dr. Rainer Kuhlen

Universität Konstanz
FB Informatik und Informationswissenschaft

März 2004

Jens Gerken
Mat. 01/473804
Universität Konstanz

Kurzfassung (deutsch)

In den letzten Jahren konnte sich im Bereich der Mensch-Computer Interaktion ein Forschungszweig deutlich herauskristallisieren: Die Erforschung und Entwicklung von visuellen Suchsystemen oder allgemeiner, von Informationsvisualisierungswerkzeugen. Mittlerweile werden die ersten Produkte kommerziell erfolgreich umgesetzt [IBM04, hum04], was nicht zuletzt darauf schließen lässt, dass für die Zukunft in diesem Bereich weitere Entwicklungen zu erwarten sind. Allerdings macht sich oftmals das Fehlen von Evaluationen, welche diese neuen Formen der Informationsdarstellung mit den bisher verwendeten Techniken vergleichen, in Hinblick auf eine mögliche Kommerzialisierung negativ bemerkbar. Schließlich reicht es nicht, die Vorteile der visuellen Darstellungsmöglichkeiten zu proklamieren, vielmehr sollten sie sich mit Hilfe derartiger Evaluationen auch beweisen lassen.

Vor diesem Hintergrund ist diese Bachelor-Arbeit zu sehen, welche als zentralen Bestandteil die Präsentation der Evaluationsergebnisse eines Performance Tests im Rahmen des VisMeB Projekts – eines visuellen Metadaten Browsers – beinhaltet. Mit Hilfe dieses Tests sollte überprüft werden, ob die in VisMeB verwendete, tabellenbasierte Darstellung von Suchergebnissen dem Benutzer gegenüber einer herkömmlichen, listenbasierten Darstellung Vorteile hinsichtlich der Bearbeitungsgeschwindigkeit bietet.

Darüber hinaus werden sowohl die Methodik als auch die Ergebnisse weiterer Usability Tests näher beleuchtet, welche während des Projekts Anwendung fanden. Ein besonderes Augenmerk wurde dabei den Validitätsproblemen zu Teil. Als Ausgangspunkt und somit ebenfalls Bestandteil dieser Arbeit, diente hierzu eine umfangreiche Exkursion in die Methoden der Sozialwissenschaften, welche in vielen Bereichen als Referenz für Usability Test-Methoden betrachtet werden können.

Abstract (english)

During the last couple of years, one field of research within the Human-Computer Interaction was able to appear on a more recognizable level: the research and development of Visual Seeking Systems, or in a more general perspective Information Visualization Tools. Nowadays some of them even have been commercialized [IBM04, hum04], which leads to the conclusion, that there is still more to come. But despite those successful products, the development of such tools still lacks an essential part: Evaluations, which prove that those new forms of visualizing information are really superior to traditional approaches. This Bachelor thesis tries to close this gap by focusing on the presentation of an evaluation of VisMeB – a Visual Metadata Browser. The main aspect of this evaluation was to prove or disprove the advantage of a table-based visualization (VisMeB) of search results over the traditional list-based view, for example used by several popular search engines, such as Google.
In addition, there are also the methods and results of several other usability tests conducted during the development of VisMeB presented. Therefore a theoretical background is being build up, by taking a closer look at the methods used in social sciences, which should be seen as a reference system for usability methods.

Inhaltsverzeichnis

1 Einleitung

Getreu dem Sprichwort „ein Bild sagt mehr als tausend Worte“ hat sich das Erscheinungsbild heutiger Software und vor allen Dingen des Internets in den letzten Jahren entscheidend gewandelt. Dank moderner Breitbandverbindungen und deutlich gestiegener Rechenleistung können vermehrt grafische Hilfsmittel anstatt tristem Text verwendet werden – sei es rein zur optischen Aufwertung einer Webseite oder eines Programms oder aber, um komplexe Sachverhalte ansprechend darstellen und leichter verständlich vermitteln zu können. Interessant ist dabei der Umstand, dass sich, unbeeindruckt von dieser Entwicklung, sowohl professionelle Recherche-Systeme als auch Online Suchmaschinen größtenteils weiterhin mit der Ergebnisdarstellung in Text-lastigen Listen begnügen. Seit sich die Informationsvisualisierung jedoch als eigener Forschungszweig innerhalb der Mensch-Computer Interaktion herauskristallisieren konnte, werden zunehmend Visualisierungswerkzeuge entwickelt, welche diesen Umstand zu ändern versuchen. Auch wenn Forschungsprojekte in diesem Bereich weiterhin dominieren, so konnten in den letzten Jahren auch einige Produkte, beispielsweise der Attribute Explorer [IBM04] oder Infozoom [hum04], kommerzielle Erfolge feiern. Auf dem heiß umkämpften Markt der Online Suchsysteme, welcher momentan größtenteils von *Google* beherrscht wird, konnte sich, abgesehen von wenigen Ausnahmen mit bescheidenen Erfolgen [Kart04], allerdings noch kaum ein derartiges Produkt durchsetzen. Ein offensichtliches Manko, das vielen Visualisierungswerkzeugen nach wie vor anhaftet, ist oftmals das Fehlen von Evaluationen, welche die Leistungsfähigkeit des Produktes der von herkömmlichen Suchsystemen gegenüberstellen und somit diese auch objektiv sicherstellen. Insbesondere die bei Online Suchsystemen zumeist verwendete Listendarstellung, wurde nur in wenigen Fällen einem direkten Vergleich unterzogen (siehe Kapitel 6). Somit ist es auch wenig verwunderlich, dass sich ein Paradigmen Wechsel, weg von der listenbasierten Darstellung hin zu neuen Formen der Ergebnisvisualisierung, bislang nicht abzuzeichnen scheint.

Dennoch ist die Frage durchaus berechtigt, ob diese, seit vielen Jahren kaum veränderte, Präsentation der Suchergebnisse heutzutage, unter Berücksichtigung der Entwicklungen im Bereich der Informations-Visualisierungswerkzeuge, noch zeitgemäß ist und den Benutzer schnellstmöglich zum Ziel führt. Im Rahmen der Entwicklung von VisMeB wurde versucht, mittels eines umfangreichen Performance Tests, dieser Fragestellung nachzugehen. Die Ergebnisse sind Hauptbestandteil dieser Arbeit und es wird daher in Kapitel 7 näher darauf eingegangen.

VisMeB [VisM04] ist ein Forschungsprojekt der AG Mensch-Computer Interaktion des Fachbereichs Informatik und Informationswissenschaften unter der Leitung von Prof. Dr. Harald Reiterer an der Universität Konstanz. Das wesentliche Ziel des Projektes ist, mithilfe eines Visuellen Metadaten Browsers den Benutzer unabhängig von der Anwendungsdomäne bei der Suche und Extraktion von relevanten Daten aus einer großen Datenmenge effizient zu unterstützen [Koen03]. Als Grundkonzept dient dem System eine auf dem Supertable Konzept aufbauende, tabellenbasierte Darstellung, welche

dem Benutzer mehrere Detailstufen bietet – die so genannte *Leveltable*. Zusätzlich wurden zahlreiche Visualisierungstechniken eingebunden, wie etwa *2D & 3D-Scatterplot, Circle Segment View*, aber auch eine weitere Variante der *Leveltable*, die so genannte *Granularity-Table*. Eine genaue Vorstellung dieser Visualisierungstechniken findet sich im Anschluss an diese Einleitung in Kapitel 2.

Das Projekt VisMeB basiert auf zwei Vorgänger Projekten, welche die Entwicklung entscheidend prägten und ermöglichten. Zum einen das EU-Projekt INSYDER [INSY04], welches eine Anwendung im Umfeld der so genannten Business Intelligence Systems zur Verfügung stellt. Es soll Unternehmen bei der Suche und Analyse von Informationen aus dem World Wide Web durch geeignete Visualisierungen unterstützen [RMMH00].

Das Nachfolger Projekt INVISIP [INVI04], welches im Rahmen der Information Society Technologies Programme der EU gefördert wurde, soll Standortentscheidungen begleiten und nachfolgende Prozesse sowie beteiligte Parteien unterstützen. Zu diesem Zweck wurde in der AG Mensch-Computer Interaktion ein Metadaten Browser zur Suche auf Geo-Metadaten und zur Ergebnisvisualisierung entwickelt.

VisMeB basiert letztendlich auf diesem Metadaten Browser, wurde jedoch um weitere Visualisierungen erweitert und deutlich flexibler, was die Art der möglichen Anwendung betrifft, gestaltet. So können nicht nur Geo-Metadaten visualisiert werden, sondern auch beispielsweise Produktkataloge, Film- oder Bibliotheksdaten. Durch den generischen Aufbau ist eine Anbindung an zahlreiche unterschiedliche Datenbanken möglich.

Die Entwicklung von VisMeB wurde bereits von Beginn an von zahlreichen Evaluationen und Usability Tests begleitet, um die Konformität des Metadatenbrowsers mit den Erkenntnissen des Usability Engineering zu gewährleisten [Mann02, Jett03]. Diese Arbeit umfasst, neben dem bereits angesprochenen Performance Test, auch die Ergebnisse zweier weiterer Usability Tests. Diese verfolgten dabei das Ziel, Usability Schwächen aufzudecken und Redesign-Vorschläge zu liefern und somit die weitere Entwicklung nach Usability Gesichtspunkten aktiv sicherzustellen. Die Ergebnisse dieser beiden Tests finden sich in Kapitel 5 wieder.

Zur Sicherstellung der Validität der Usability Tests, wurden allerdings im Vorfeld der Tests zunächst theoretische Vorarbeiten geleistet, welche sich eingehend mit den Methoden der Sozialwissenschaften beschäftigten. Diese können in vielen Dingen als Referenz von Usability Test Methoden gesehen werden und liefern somit entscheidende Erkenntnisse bezüglich der Validität und auch des jeweiligen Testdesigns. Kapitel 4 beschäftigt sich eingehend mit diesem Thema.

Weiterhin wurden im Rahmen dieser Arbeit auch Evaluationen von weiteren Informations-visualisierungswerkzeugen in einer State of the Art Analyse untersucht. Ziel war es hier, nicht nur einen Überblick, über die Leistungsfähigkeit dieser Tools zu erhalten, sondern auch die Methodik der jeweiligen Evaluationen kritisch zu betrachten. Die Ergebnisse hierzu finden sich in Kapitel 6.

Jens Gerken
Mat. 01/473804
Universität Konstanz

Zu Beginn stehen zwei Kapitel, welche für das grundsätzliche Verständnis dieser Arbeit Voraussetzung sind. Zum einen werden in Kapitel 2 wie bereits erwähnt die Visualisierungen und Funktionsweisen von VisMeB genauer erläutert und zum anderen wird in Kapitel 3 eine kurze statistische Einführung angeboten. Diese beschränkt sich auf die, zum Verständnis der Auswertung des Performance Tests in Kapitel 7, notwendigen Informationen.

2 VisMeB – ein visueller Metadaten-Browser

Die grundlegenden Konzepte, welche zu der Entwicklung von VisMeB führten und den dort eingesetzten Visualisierungen zu Grunde liegen, wurden bereits in zahlreichen Veröffentlichungen eingehend betrachtet [KRML03], [GHRM02], [KMRE02], [RMMH00]. Dieses Kapitel soll somit keine erneute theoretische Abhandlung dieser Konzepte, sondern vielmehr eine anschauliche Einführung in VisMeB darstellen. Ziel ist es, mit Hinblick auf die Präsentation der Evaluationsergebnisse im weiteren Verlauf dieser Arbeit, ein Verständnis dieser zu ermöglichen.

Im Folgenden werden anhand von Screenshots von VisMeB einige ausgewählte Visualisierungen und deren Funktionsweise vorgestellt. Dabei diente eine, während der Tests verwendete, Version von VisMeB als Ausgangsbasis.

2.1 Leveltable

Die Leveltable ist die grundlegende Visualisierung von VisMeB. Vereinfacht gesagt können sämtliche weiteren Visualisierungen als Ergänzung zu ihr gesehen werden. Innerhalb der Leveltable werden die Ergebnisse einer Suchanfrage auf einer Datenbank in einer tabellarischen Form dargestellt.

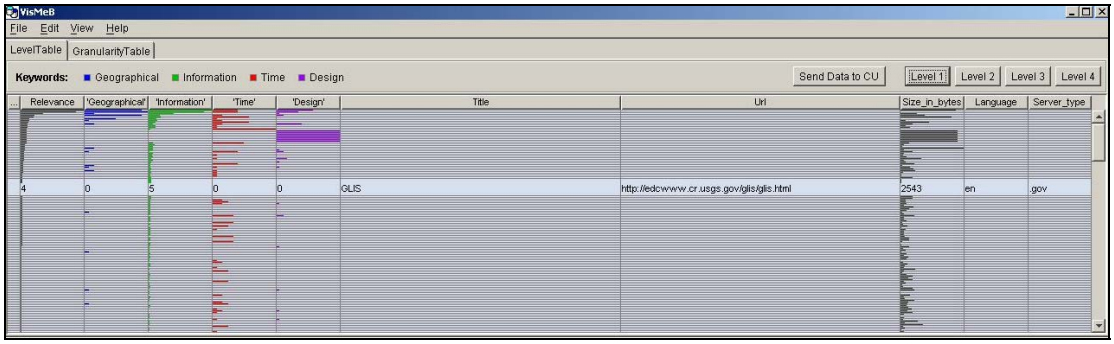


Abbildung 2.1: Leveltable Level 1

Abbildung 2.1 zeigt die Leveltable, wie sie dem Benutzer nach Starten der Suche erscheint. Wie der Name *visueller Metadaten Browser* von VisMeB schon andeutet, werden in der Tabelle nicht die Daten selbst – in diesem Beispiel wären das Webseiten – sondern vielmehr beschreibende Daten, so genannte Metadaten oder Meta-Attribute angezeigt. Dies können beispielsweise vom System errechnete Relevanzwerte sein oder weitere, für die Entscheidung wichtige Attribute wie Titel, Größe oder Sprache. Um dem Nutzer einen größtmöglichen Überblick zu ermöglichen, bietet die Leveltable in Level 1, welches in Abbildung 2.1 zu sehen ist, eine Supertable ähnliche Darstellung. Alle Treffer werden sehr eng aufeinander aufgelistet, so dass eigentlich keine Detailinformationen mehr sichtbar sind. Mit Hilfe von Bar-Charts können jedoch trotzdem die Treffer bezüglich mancher Metadaten, etwa

der Relevanz, mit einander verglichen werden. Mittels eines Fish-Eye ähnlichen Effekts, können die Detailinformationen zu einzelnen Treffern sichtbar gemacht werden. Hierzu muss die Maus nur über die entsprechende Zeile der Tabelle bewegt werden. Zusätzlich bietet die Leveltable noch drei weitere Levels an, welche dem Benutzer sukzessive mehr Informationen anbieten und ihn letztendlich zum eigentlichen Objekt der Suche, in diesem Fall eine Webseite, hinführen.

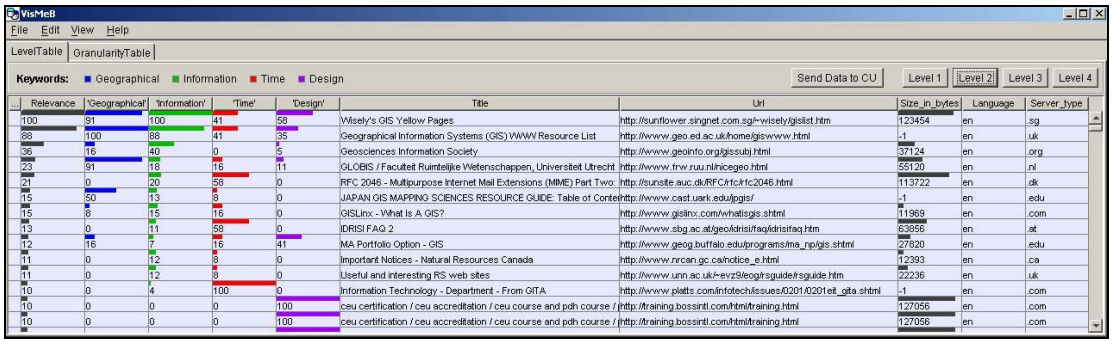


Abbildung 2.2: Leveltable Level 2

Sobald der Benutzer auf Level 4 gewechselt hat, blendet sich im unteren Bereich des Bildschirms die Browserview ein. In dieser wird nun das komplette HTML Dokument angezeigt (allerdings in dieser Version noch nicht korrekt formatiert). Darüber hinaus ist in der Tabelle selbst ein neues Visualisierungstool namens *Detailed Relevance Curve* integriert, mit dessen Hilfe der Text des Dokumentes effektiv und effizient nach relevanten Bereichen durchforstet werden kann.

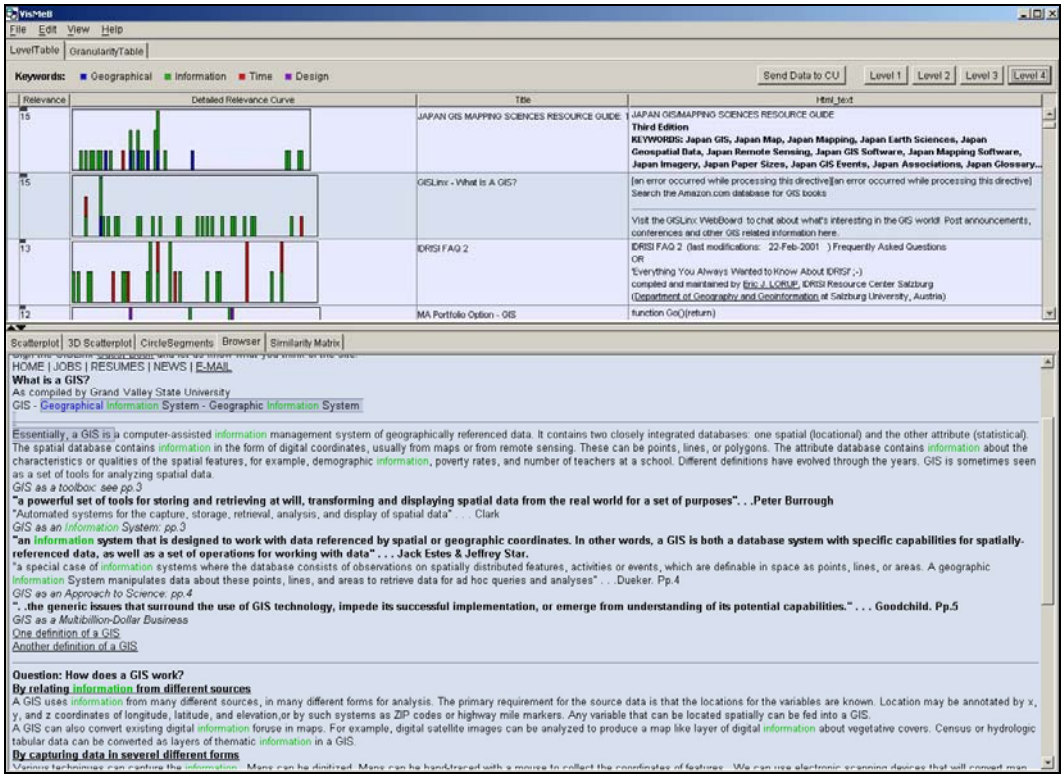


***Abbildung 2.3*:** Leveltable Level 4 & Browserview

Ein weiteres Feature der Tabellendarstellung ist die Möglichkeit, nach jeder Spalte sortieren zu können, um somit schnell zusammenhängende Daten erfassen zu können.

2.2 Granularity-Table

Die Granularity-Table ist eine Variante der Leveltable, mit dem Ziel, den Wechsel der Modalitäten – von der Präsentation der Ergebnisse in Tabellenform, hin zu dem letztendlich den Benutzer interessierenden Dokument – nochmals zu verringern. Um dies zu ermöglichen, bietet diese Form der Darstellung sechs verschiedene Detailstufen, wovon jede auch nur für einzelne Treffer geändert werden kann. Weiterhin wird das Dokument an sich nicht mehr in einer zusätzlichen Ansicht wie der Browserview dargestellt sondern ebenfalls in der Tabellendarstellung.

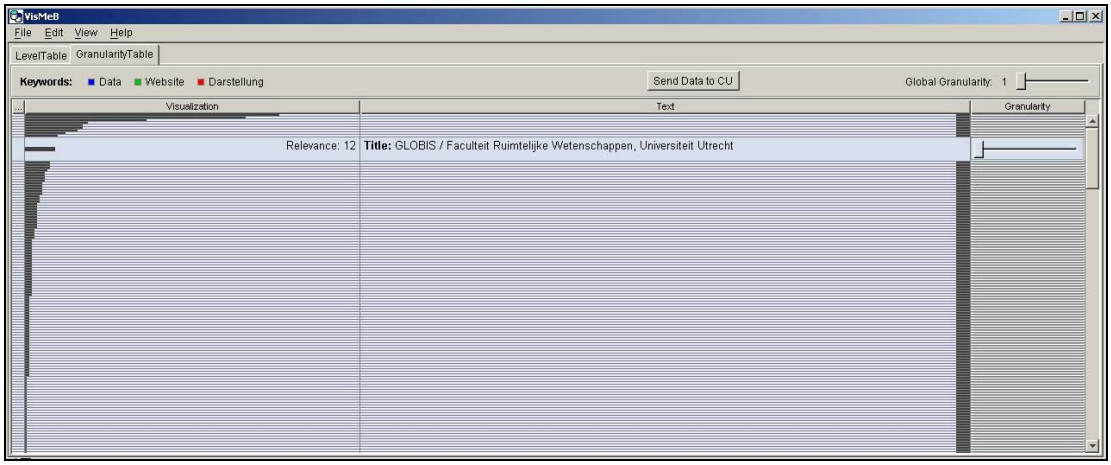


Abbildung 2.4: Granularity-Table Stufe 1

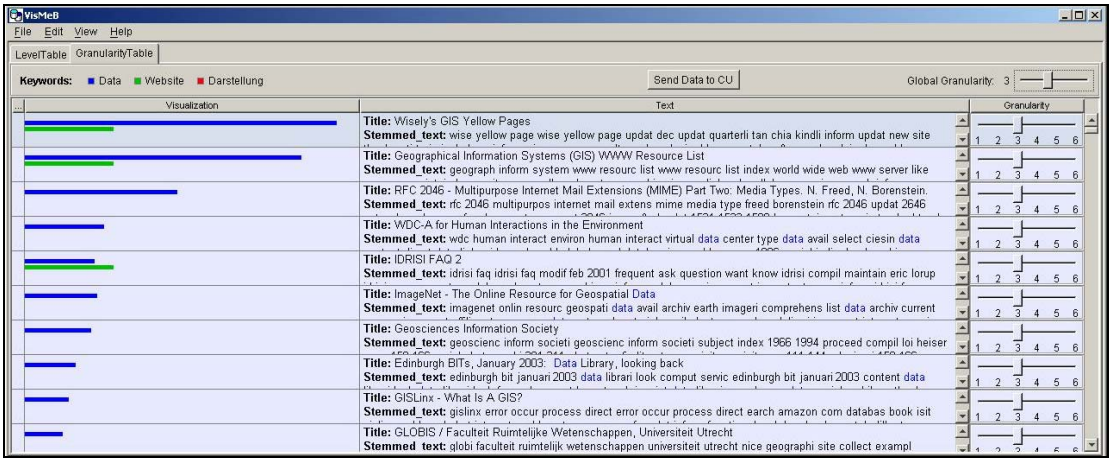


Abbildung 2.5: Granularity-Table Stufe 3

Die grundsätzliche Bedienung ist mit der Leveltable identisch.

2.3 2D-Scatterplot

Der 2D-Scatterplot wird in den Level 1-3 der Leveltable und innerhalb jeder Stufe der Granularity-Table standardmäßig als Visualisierung im unteren Bereich von VisMeB angezeigt. Nur in Level 4 der Leveltable wechselt hier die Ansicht automatisch zur Browserview. Der Scatterplot dient dazu, Zusammenhänge zwischen den Dokumenten zu erkennen, in dem sowohl auf der X-Achse als auch auf der Y-Achse ein Meta-Attribut abgetragen wird. Beispielsweise könnte somit überprüft werden, ob die als relevant eingestuften Dokumente durchweg auch die größten sind.

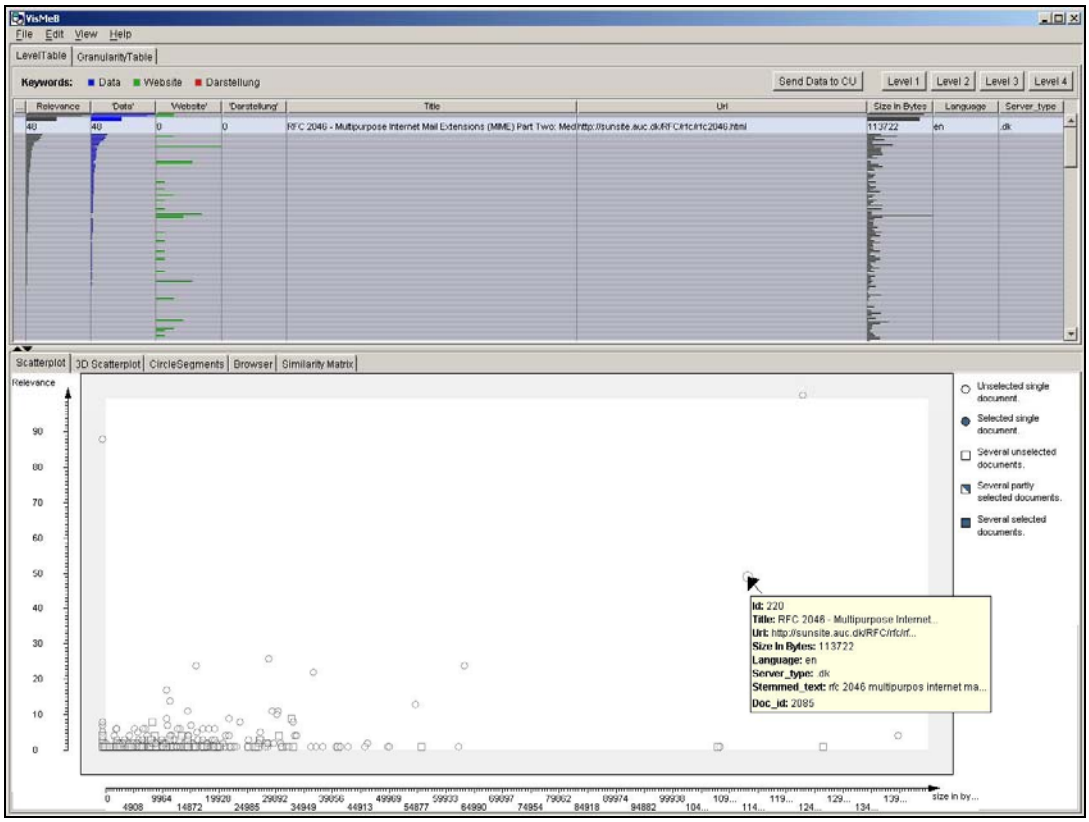


Abbildung 2.6: 2D-Scatterplot mit Leveltable Level 1

Der Scatterplot ist dabei direkt mit der Leveltable/Granularity-Table durch so genanntes *Brushing & Linking* verbunden. Wird also ein Objekt im Scatterplot mit der Maus berührt oder markiert, so wird dieses auch in der Tabellensicht in den Fokus gerückt beziehungsweise markiert.

2.4 Weitere Visualisierungen

Zum Zeitpunkt der Evaluationen waren bereits weitere Visualisierungen integriert, welche jedoch noch in der Entwicklung steckten und aus diesem Grund während der Usability Tests weitestgehend außen vor gelassen wurden und an dieser Stelle auch nur kurz erwähnt werden sollen. Dazu zählt der Circle

Segment View, welcher, basierend auf *Dynamic Query* und *Query Preview* Techniken, mittlerweile sowohl als integrierte Visualisierung in VisMeB zum Einsatz kommt, als auch als Alternative zu einer formularbasierten Suchanfrage.

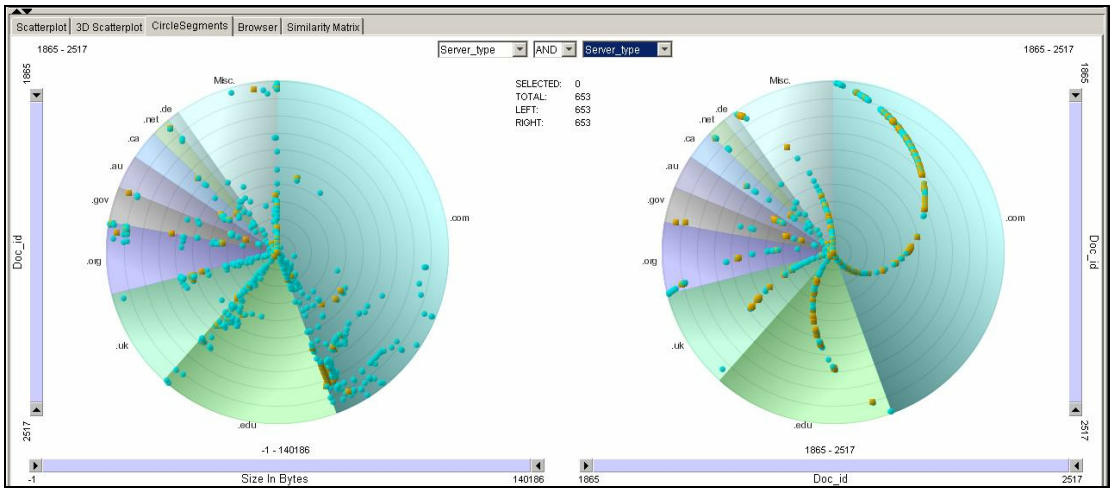


Abbildung 2.7: Circle Segment View

Weiterhin wurde auch ein 3D-Scatterplot integriert, welcher durch das Hinzufügen einer 3. Dimension dem Benutzer die Möglichkeit bietet, tiefere und komplexere Zusammenhänge zu erkennen.

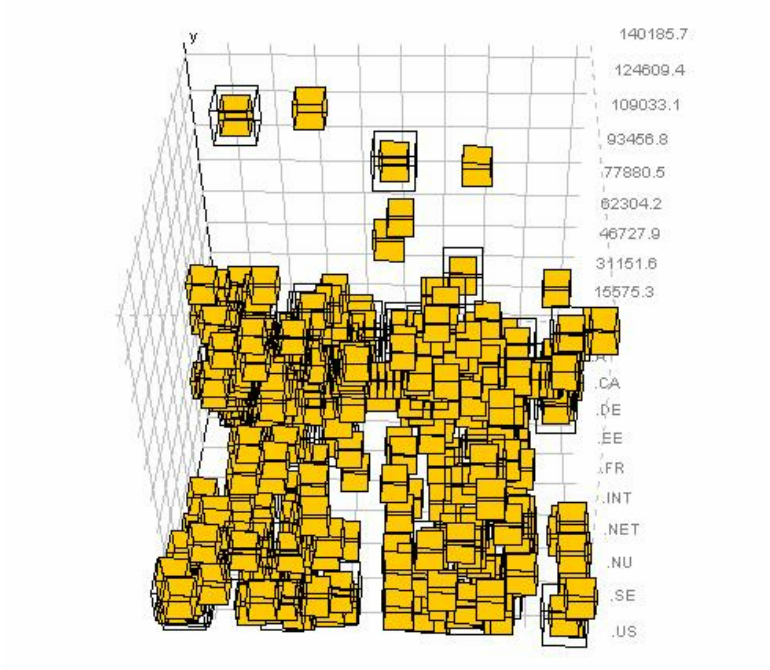


Abbildung 2.8: 3D-Scatterplot

Jens Gerken
Mat. 01/473804
Universität Konstanz

3 Statistische Einführung

Die Validität eines Usability Tests hängt zu einem Großteil, neben seiner korrekten Durchführung, von einer adäquaten Auswertung ab. Im Falle eines Performance Tests, zu welchem der in Kapitel 7 folgende Vergleich *Liste vs. Leveltable* zweifelsfrei zählt, ist eine statistische Auswertung unabdingbar.

Allerdings muss eine derartige Auswertung durchaus mit Bedacht angegangen werden – ein gedankenloses Verwenden scheinbar sinnvoller Formeln kann zwar, oberflächlich betrachtet, zu interessanten Ergebnissen führen, diese können unter Umständen jedoch von keinerlei Signifikanz sein. Aus diesem Grund ist es unablässig bereits vor dem eigentlichen Usability Test, also vor dem Experiment, genau zu bestimmen, mit Hilfe welcher Analysemethoden die Auswertung erfolgen soll und aus welchem Grund diese gewählt werden.

Grundsätzlich muss zunächst einmal bestimmt werden, was überhaupt das Ziel des Tests ist – dieses sollte bereits vor dem Entwurf der Fragen vorliegen, um zu gewährleisten, dass diese auch wirklich Ergebnisse hinsichtlich dieses Ziels liefern. Beim Vergleich zweier Systeme hinsichtlich der Performance könnte das Ziel beispielsweise sein herauszufinden, welches System schneller ist. Für die Auswertung muss dieses Ziel als so genannte Null-Hypothese formuliert werden. Diese wird durch den Test überprüft und kann gegebenenfalls verworfen werden. Beim oben angesprochenen Performancevergleich zweier Systeme könnte eine Null-Hypothese folgendermaßen formuliert werden: „Zwischen System 1 und System 2 besteht hinsichtlich der Performance kein Unterschied“. Die Null-Hypothese geht also grundsätzlich davon aus, dass zwischen abhängiger und unabhängiger Variable kein Effekt besteht. Eben diese Variablen müssen nun auch noch genau bestimmt und klassifiziert werden. Die unabhängige Variable ist diejenige, die einen Effekt verursacht und seitens des Versuchsleiters variiert werden kann. Im obigen Beispiel wäre dies das verwendete System 1 oder 2. Die abhängige Variable wird von der unabhängigen Variable beeinflusst – zumindest ist das die Vermutung, die dem Test vorausgeht und welche dieser letztendlich überprüfen soll. In diesem Fall wäre das die benötigte Zeit der Versuchspersonen. Es ist ungemein wichtig, dass zwischen abhängiger und unabhängiger Variable auch ein logischer Zusammenhang besteht, da ansonsten die Ergebnisse zwar auf dem Papier aussagekräftig erscheinen mögen, letztendlich aber keine Bedeutung haben.

Es ist auch durchaus möglich, sowohl mehrere, abhängige Variablen, als auch mehrere unabhängige Variablen zu bestimmen, wobei vor allem letzteres die Analyse sehr viel komplexer werden lässt und auch die Aussagekraft der Ergebnisse abschwächen kann.

Welche Analysemethode überhaupt einzusetzen ist, hängt letztlich von dem Skalenniveau der Variablen ab. Es werden hierbei vier verschiedene Skalenniveaus unterschieden. Das niedrigste Skalenniveau besitzen nominal skalierte Variablen. Dies bedeutet, dass eine Variable mehrere Ausprägungen hat, welche allerdings alle gleichwertig sind. Man sagt auch, dass Nominalskalen eine Klassifizierung

qualitativer Eigenschaftsausprägungen darstellen [BEPW00]. Beispiele für nominal skalierte Variablen sind Geschlecht (Ausprägung: männlich, weiblich) oder Religion (Ausprägung: katholisch, evangelisch, sonstiges). Wie bereits gesagt, kann zwischen den Ausprägungen zwar unterschieden, jedoch keine Rangfolge gebildet werden. Genau das ist bei ordinal skalierten Variablen möglich. Hier wird nicht nur zwischen verschiedenen Ausprägungen unterschieden, sondern es werden diese auch noch mit Hilfe ordinaler Zahlen in eine Rangfolge zueinander gesetzt. Es ist dabei allerdings zu beachten, dass keine Abstände zwischen den einzelnen Ausprägungen berücksichtigt werden. Beispielsweise könnte man eine persönliche Liste der zehn besten Filme aller Zeiten erstellen, wobei der beste Film den Rang 1 erhält und der „schlechteste“ den Rang 10. Um welchen Faktor der Film auf Rang 1 aber wirklich besser ist, als derjenige auf Rang 10, wird nicht festgelegt – es steht nur fest, dass er besser ist. Diese beiden Skalenniveaus können zu den nicht-metrischen Skalenniveaus zusammengefasst werden, da auf Variablen dieses Typs keine arithmetischen Rechenoperationen möglich sind.

Die nächst höhere Skalierung gehört bereits zu den metrischen Skalenniveaus – die Intervallskala. Hierbei ist jeder Skalenabschnitt gleich groß, weswegen auch die Abstände eine Rolle spielen. Beispielsweise ist die Celsius-Skala intervallskaliert. Der Unterschied zum letzten und höchsten Skalenniveau, der Verhältnis-Skala, besteht darin, dass kein natürlicher Nullpunkt existiert, welcher sich im Sinne von „nicht vorhanden“ für das entsprechende Merkmal interpretieren lässt. Somit können zwar Additionen und Subtraktionen auf intervallskalierte Daten angewandt werden, jedoch keine Multiplikation und Division. Existiert jedoch ein natürlicher Nullpunkt, etwa bei physikalischen Größen wie Länge, Gewicht oder Geschwindigkeit, aber auch bei Einkommen oder Zeitmessungen, ist die Variable verhältnisskaliert, was dazu führt, dass sämtliche arithmetischen Rechenoperationen möglich sind.

Um nun die zu verwendende Analysemethode zu bestimmen, müssen sowohl die abhängige, als auch die unabhängige Variable nach einem der Skalenniveaus klassifiziert werden. Anhand folgender Tabelle (*Abbildung 3.1)* lässt sich dann die passende Methode erkennen [BEPW00]:

		Unabhängige Variable	
		Metrisches Skalenniveau	Nominales Skalenniveau
Abhängige Variable	Metrisches Skalenniveau	Regressionsanalyse, Varianzanalyse	Varianzanalyse, Student T-Test
	Nominales Skalenniveau	Diskriminanzanalyse, logistische Regression	Kontigenzanalyse

Abbildung 3.1: Methodentabelle

Sowohl bei klassischen Experimenten in der Sozialwissenschaft, als auch bei Usability Tests, wird sich zumeist die *Varianzanalyse* bzw. der *Student T-Test* als geeignete Analysemethode herauskristallisieren. In obigem Beispiel hat die unabhängige Variable die Ausprägungen System 1 und System 2, es wird also zwischen beiden System unterschieden, jedoch ohne eine Wertung zu beinhalten. Die Variable ist

somit nominal skaliert. Die abhängige Variable ist die gemessene Zeit, welche verhältnisskaliert ist und damit metrisches Skalenniveau besitzt. Trotzdem sollte genau überprüft werden, welches Skalenniveau die Variablen besitzen, die bei dem Experiment untersucht werden, um wirklich die passende Analysemethode auszuwählen.

3.1 Varianzanalyse

Grundsätzlich dient die *Varianzanalyse* der Untersuchung eines möglichen Effekts, den eine unabhängige Variable auf eine abhängige Variable auswirken könnte. Im Weiteren soll folgendes Beispiel verwendet werden: Ein Kinoeigentümer möchte die optimale Form der Werbung (unabhängige Variable – nominal skaliert) feststellen, um seine Besucherzahlen (abhängige Variable – metrisch skaliert) zu maximieren. Zur Auswahl stehen zum einen Anzeigen in den lokalen Tageszeitungen und zum anderen Radiowerbung bei den Lokalsendern. Um beides zu vergleichen, schaltet er einen Monat lang nur die Werbung in den Tageszeitungen und im nächsten Monat nur im Radio. Dabei wertet er jeden Tag die Besucherzahlen aus. Es wird nun sowohl jeweils vom ersten und zweiten Monat, als auch von der Gesamtdauer die mittlere Besucheranzahl pro Tag ausgerechnet. Die *Varianzanalyse* untersucht nun, inwieweit die einzelnen Tageswerte der Besucherzahlen von diesen Mittelwerten abweichen. Dabei wird zwischen erklärter Abweichung und nicht erklärter, das heißt zufälliger Abweichung, unterschieden. Erstere meint dabei die Abweichung, welche durch die unterschiedliche Werbung verursacht wird und Zweitere die zufällige Abweichung durch andere Faktoren, die nicht bekannt sind, von denen jedoch angenommen wird, dass sie in beiden Monaten gleich auftreten. Vereinfacht gesagt bedeutet nun eine im Verhältnis zu der erklärten Abweichung, kleine zufällige Abweichung, dass die Unterschiede bei den Besucherzahlen zwischen den beiden untersuchten Monaten mit einer höheren Wahrscheinlichkeit auf die unterschiedliche Werbung zurückzuführen sind. Je größer dagegen diese zufällige Abweichung ist, desto unsicherer ist es, ob die Unterschiede wirklich mit der Werbemethode zusammenhängen [BEPW00].

Der *Student T-Test*, welcher im Bereich der Usability Tests ebenfalls oftmals Verwendung findet, führt in den meisten Fällen zu den gleichen Ergebnissen, da er einen Spezialfall der *Varianzanalyse* darstellt. Dieser erfordert zwingend, dass nur je eine abhängige und eine unabhängige Variable vorhanden sind und letztere auch nur genau zwei Ausprägungen besitzt. Sollte die unabhängige Variable mehr als zwei Ausprägungen besitzen – beispielsweise drei verschiedene Formen der Werbung – so können zwar prinzipiell mehrere *T-Tests* ebenfalls zu einem Ergebnis führen, dabei steigt jedoch die Fehlerwahrscheinlichkeit sehr stark an, weswegen grundsätzlich die *Varianzanalyse* zu empfehlen ist [Stock96].

3.2 Konfidenzintervall

Ziel der *Varianzanalyse* ist es herauszufinden, ob die vorher definierte Null-Hypothese verworfen werden kann oder beibehalten werden muss. Da ein Effekt allerdings nur mit einer gewissen Wahrscheinlichkeit festgestellt werden kann, kann auch die Null-Hypothese nur mit einer gewissen Wahrscheinlichkeit verworfen werden. Diese wird mit dem Konfidenzintervall festgelegt. Im Allgemeinen üblich sind Konfidenzintervalle von 90%, 95% oder 99%. Stellt man in obigem Beispiel mittels der *Varianzanalyse* fest, dass bei einem Konfidenzinzervall von 95% der Unterschied zwischen beiden Werbemethoden signifikant ist und entscheidet sich somit dafür, die Null-Hypothese zu verwerfen, so bleibt dennoch eine 5% Irrtumswahrscheinlichkeit – genannt Signifikanzniveau – dass dies ein Fehler war und die Unterschiede rein zufällig sind. Ein höheres Konfidenzintervall, beziehungsweise ein niedrigeres Signifikanzniveau drückt also ein aussagekräftigeres Ergebnis aus. Generell kann jeder noch so kleine Gruppenunterschied signifikant sein, wird nur das Signifikanzniveau hoch genug gewählt. Deswegen ist es unabdinglich, dass dieses grundsätzlich mit angegeben wird [FKPT99].

3.3 SPSS Ergebnis - Tabelle

SPSS ist ein weitläufig eingesetztes Statistik-Programm, welches die Analyse deutlich vereinfacht. Da dieses auch bei dem in dieser Arbeit vorgestellten Performance Test von VisMeB Verwendung fand, wurde zur besseren Illustration auch jeweils die Ergebnis-Tabelle der *Varianzanalyse* mit aufgeführt. Diese soll im Folgenden kurz erklärt werden, um deren Lesen und Verstehen zu ermöglichen *(Abbildung 3.2)*. Dabei wird nur auf die entscheidenden Merkmale näher eingegangen [Gar03]:

ONEWAY ANOVA

	Quadratsumme	df	Mittel der Quadrate	F	**Signifikanz**
Zwischen den Gruppen	31823,106	1	31823,106	2,869	**,101**
Innerhalb der Gruppen	310592,835	28	11092,601		
Gesamt	342415,941	29			

Abbildung 3.2: SPSS ANOVA Tabelle

Zwischen den Gruppen: Diese Zeile beinhaltet die Daten bezüglich der erklärten Abweichung.

Innerhalb der Gruppen: Diese Zeile beinhaltet die Daten bezüglich der zufälligen Abweichung.

Quadratsumme: Variation, hier wird die jeweilige Abweichung aufsummiert und quadriert angegeben.

df: degree of Freedom = Freiheitsgrade; Berechnung erfolgt für erklärte Abweichung „Anzahl der Gruppen minus eins“ und für die zufällige Abweichung „Anzahl der Gruppen multipliziert mit den Teilnehmern pro Gruppe minus eins“

Mittel der Quadrate: Varianz; Wird nach dem Schema „Quadratsumme geteilt durch Freiheitsgrade“ berechnet.

F: Ergebnis der F-Statistik; Berechnet wird dieser Wert „Mittel der Quadrate zwischen den Gruppen geteilt durch das Mittel der Quadrate innerhalb der Gruppen“. Mittels einer F-Tabelle kann dieser Wert interpretiert werden.

Signifikanz: Um dem Benutzer den Blick in die F-Tabelle zu ersparen, wird zusätzlich ein Signifikanzwert ausgegeben. Um zu überprüfen, ob der gemessene Gruppenunterschied signifikant ist, muss dieser Wert nun nur noch mit dem vorher festgelegten Signifikanzniveau verglichen werden. Beträgt dieses beispielsweise 5% (also ein Konfidenzintervall von 95%), so muss der Signifikanzwert kleiner als 0,05 sein, wenn der Gruppenunterschied signifikant sein soll. In *Abbildung 3.2* wäre somit der Unterschied nicht signifikant, da 0,101 größer als 0,05 ist.

3.4 Ausreißer und Extremwerte

Bei jedem Experiment kann es zu Ausreißern kommen, welche im schlimmsten Fall, das Ergebnis verzerren. In obigem Kinobeispiel könnte etwa eine *Herr der Ringe-Nacht* mit allen drei Teilen an einem Tag zu deutlich höheren Besucherzahlen führen. Dieser Anstieg wäre jedoch völlig unabhängig von der verwendeten Werbemethode und könnte somit die zufälligen Fehler erhöhen und eventuell einen Effekt verdecken. Bei solchen Extremfällen können Ausreißer relativ leicht „logisch“ erkannt werden. Oftmals ist diese Grenze allerdings nicht allein mit gesundem Menschenverstand zu erkennen, weswegen eine formale Definition notwendig ist. Um dies kurz zu erklären, müssen zunächst einige Begrifflichkeiten eingeführt werden [FKPT99]:

Median: Der *Median*, welcher auch *Q50* genannt wird, teilt alle Datenpunkte in genau zwei Teile. Somit befinden sich 50% der Datenpunkte oberhalb und 50% unterhalb des Medians. In obigem Beispiel werden hierfür zunächst alle Tage nach der Höhe der Besucherzahlen sortiert – allerdings für beide Monate separat. Anschließend wird bei genau der Hälfte der Tage in dieser Rangfolge der Median gesetzt. Somit waren die Besucherzahlen bei der einen Hälfte der Tage höher und bei der anderen Hälfte niedriger als bei diesem festgelegten Median. Der Median muss dabei nicht zwingend genau auf einem

Datenpunkt liegen, da dies bei gerader Fallzahl nicht möglich ist, sondern kann sich auch zwischen den beiden mittleren Punkten befinden und nimmt dann genau den Mittelwert dieser an.

Q25/Q75: Zusätzlich wird nun auch noch bei 25% der Tage und bei 75% der Tage (zu beachten ist jeweils, dass diese Einteilung ebenfalls in der, nach Besucherzahlen sortierten, Rangfolge geschieht) ein Punkt gesetzt, genannt *Q25* beziehungsweise *Q75*. Das *Q* steht hierbei für *Quartil*. Somit befinden sich oberhalb von *Q25* 75% der Tage und unterhalb von *Q25* die restlichen 25% – analog dazu bei *Q75*.

IQR: Damit wird der Bereich von *Q25* bis *Q75* bezeichnet, der so genannte „Interquartils Range“. In ihm befinden sich genau 50% der Datenpunkte, also die Hälfte der Tage. Die Länge des *IQR*, also wie viele Besucher er umfasst, wird somit mit *Q75-Q25* berechnet.

Boxplot: Um diese definierten Punkte visualisieren zu können, wird zumeist ein *Boxplot* verwendet *(Abbildung 3.3)*.

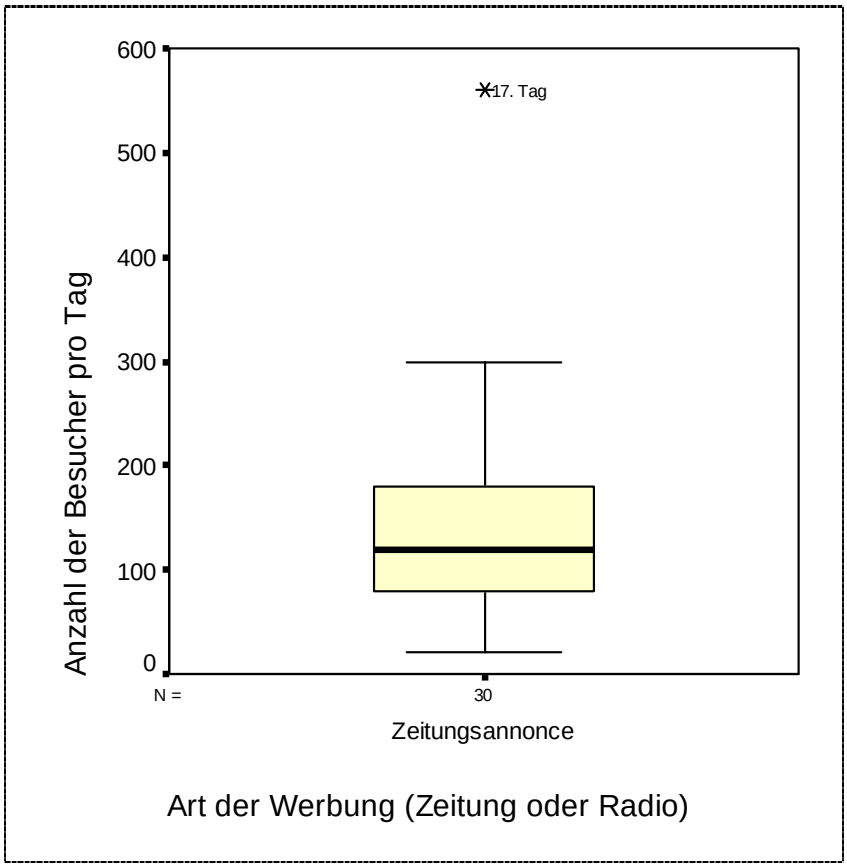


Abbildung 3.3: Boxplot

Der große gelbe Kasten, die Box des Boxplots, umfasst dabei den *IQR* und der waagrechte, etwas dickere, schwarze Balken, entspricht dem *Median*. An den Kasten schließen sowohl oben als auch unten die *Whisker* an. Sie haben eine maximale Länge von *1,5 x IQR*, enden aber in jedem Fall direkt mit einem Datenpunkt, was in diesem Fall einem beobachteten Tag entspricht. Der *Boxplot* zeigt auch sehr

schön die Verteilung innerhalb des *IQR* auf. In dieser Abbildung zeigt sich beispielsweise, dass die Besucherzahlen unterhalb des *Medians* deutlich näher beieinander liegen als die darüber.

Sobald sich ein Punkt außerhalb der *Whisker* des *Boxplots* befindet, also einen Wert größer als *Q75+1,5 x IQR* (Ausreißer nach oben) beziehungsweise *Q25 – 1,5 x IQR* (Ausreißer nach unten) annimmt, wird er als Ausreißer bezeichnet. Ein Extremwert unterscheidet sich davon, dass dieser *3 x IQR* außerhalb der Box liegt. In *Abbildung 3.3* trifft das auf den 17. Tag zu, an welchem beispielsweise die besagte *Herr der Ringe-Nacht* stattgefunden haben könnte.

Generell ist es möglich, auf diese Art erkannte Ausreißer bei der Auswertung außen vor zu lassen und somit das Verzerren des Ergebnisses durch diese zu verhindern. Allerdings wird durch diesen Eingriff natürlich auch wiederum das Ergebnis beeinflusst, weswegen eine solche Handlung nicht unüberlegt durchgeführt werden sollte. Beispielsweise sind nach dem Entfernen von Ausreißern die Gruppen eventuell nicht mehr gleich groß. Bei komplexen Designs, mit mehreren unabhängigen Variablen, führt das zu einer zumeist nur unbefriedigend lösbaren Problematik.

3.5 Between Subjects Design vs. Within Subjects Design

In obigem Beispiel wird die Versuchsgruppe, wobei diese bei einem derartigen Experiment nicht explizit eingeschränkt ist, zuerst der Werbemethode A (Anzeige in Zeitung) und anschließend der Werbemethode B (Radiospot) ausgesetzt. Dieses Testdesign wird *Within Subjects Design* genannt, da auf die Teilnehmer alle Ausprägungen der unabhängigen Variablen nacheinander einwirken. Im Gegensatz dazu wäre es auch möglich, die Teilnehmer vorher in zwei Gruppen aufzuteilen (in obigem Beispiel allerdings in der Praxis nur schwer umsetzbar) und beispielsweise der Gruppe 1 nur die Anzeige in der Tageszeitung zugänglich zu machen und der Gruppe 2 nur den Radiospot. Eine zeitliche Trennung der Werbemethoden wäre dann nicht mehr notwendig. Anschließend könnten die Besucherzahlen beider Gruppen verglichen werden. Dieses Testdesign wird *Between Subjects Design* genannt. Vorteil des letzteren ist, dass dadurch keine Lerneffekte möglich sind und auch Störeinflüsse eher für beide Gruppen vergleichbar auftreten. Beispielsweise würde die angesprochene *Herr der Ringe-Nacht* für beide Werbemethoden auftreten und somit das Ergebnis nicht verzerren. Allerdings ist hierfür eine höhere Anzahl an Teilnehmern notwendig, da durch die Aufteilung in zwei Gruppen nur die Hälfte der Datensätze für jede Werbemethode vorliegt. Wie das Kinobeispiel zeigt, ist eine derartige Aufteilung in der Praxis auch nicht immer möglich, bei Laborexperimenten ist es aber zumeist machbar. Das *Within Subjects Design* hingegen gibt sich mit deutlich weniger Teilnehmern zufrieden, da hier nicht durch die Gruppenaufteilung Datenwerte verloren gehen. Allerdings besteht die Gefahr von Lerneffekten und zeitlichen Störeinflüssen. Es gibt verschiedene Techniken diesen entgegenzuwirken, welche jedoch hier nicht weiter aufgeführt werden sollen. Generell ist es in der Literatur umstritten, welches der beiden Verfahren wirklich das Beste ist und hängt nicht unerheblich von dem jeweiligen Testsetting und auch der Vorliebe des Versuchsleiters ab [Gar03, Lane03].

Jens Gerken
Mat. 01/473804
Universität Konstanz

3.6 Likert-Skala

Eine *Likert-Skala* wird zumeist in Fragebögen verwendet. Auf ihr sollen die Teilnehmer ihre Zustimmung oder Ablehnung gegenüber einer Aussage oder Frage ankreuzen. Gebräuchlich sind hierbei 5-Punkt-Likert-Skalen und 7-Punkt-Likert-Skalen. Die fünf beziehungsweise sieben Punkte werden dabei waagrecht unter der Frage angeordnet und mit eins bis fünf (bzw. sieben) nummeriert. Eine eins bedeutet dabei im Allgemeinen, dass der Teilnehmer der Aussage überhaupt nicht zustimmt und eine fünf (sieben), dass er ihr hingegen vollkommen zustimmt. Mit dem Ankreuzen der goldenen Mitte (drei bzw. vier) drückt der Teilnehmer aus, dass er weder zustimmend noch ablehnend der Aussage gegenüber eingestellt ist. In *Abbildung 3.4* ist eine 7-Likert-Skala als Beispiel abgebildet.

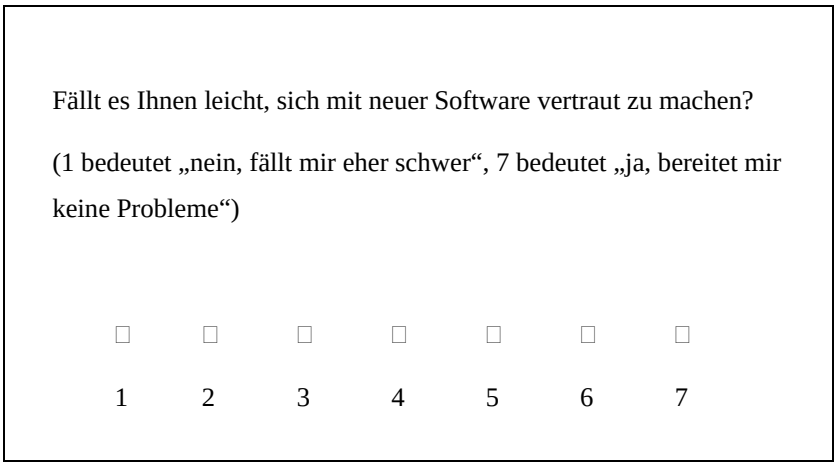


Abbildung 3.4: Likert-Skala

4 Validität von Usability Testmethoden

Ein großes Problem von Usability Tests ist die Frage, wie valide solche Tests überhaupt sind. Für das Management eines Unternehmens kostet ein Usability Test in erster Linie Geld, wobei oftmals der *Return of Income,* also ob dieses Geld sich wirklich rechnet, nicht klar ersichtlich ist. Insbesondere ist dieses fraglich, wenn nicht sichergestellt werden kann, dass ein Usability Test überhaupt brauchbare, auf die Realität übertragbare Ergebnisse liefert, also valide ist. Somit stellt sich die Frage, wie die Validität wirklich sichergestellt werden kann. Um hier mögliche Lösungsansätze zu erhalten, empfiehlt sich ein Blick über den Tellerrand des Usability Testing hinaus. Denn die Methoden, auf denen Usability Tests aufbauen, sind zum Großteil keine eigene, unabhängige Entwicklung, sondern basieren auf den langjährig erforschten Methoden der Sozialwissenschaften. In diesen ist das Problem der Validität auch nur allzu bekannt und wurde dementsprechend auch bereits umfangreich erforscht. Dabei lassen sich die Methoden grundsätzlich in zwei unterschiedliche Ansätze unterteilen: Die quantitativen Methoden und diesen voran das quantitative Experiment, sowie die qualitativen Methoden, welche in den vergangenen Jahrhunderten oftmals einen schweren Stand hatten, jedoch in den letzten Jahrzehnten wieder vermehrt zum Einsatz kommen [Klei94a]. Im Folgenden sollen zunächst die Methoden der Sozialwissenschaft umfangreich betrachtet werden, um anschließend Rückschlüsse auf die Usability Methoden, welche im Verlauf des Entwicklungsprozesses von VisMeB zum Einsatz kamen, möglich zu machen.

4.1 Sozialwissenschaftliche Methoden

4.1.1 Das Quantitative Experiment [Att95]

Das quantitative Experiment kann auf eine lange Tradition zurückblicken und gilt dementsprechend als weitestgehend erforscht. Grundlage des quantitativen Experiments ist die Hypothesenprüfung. Es wird also, ähnlich wie bei statistischen Tests, zunächst eine Hypothese aufgestellt, welche durch das Experiment überprüft werden sollen.

Ein quantitatives Experiment versucht im Idealfall einen kausalen Zusammenhang zwischen genau einer unabhängigen und einer abhängigen Variable zu zeigen. Dies wird erreicht, indem mittels der Veränderung der unabhängigen Variablen ein Effekt nachgewiesen wird – sich die abhängige Variable also allein durch die Variation der unabhängigen Variable ebenfalls verändert. Um diese Veränderung dann eindeutig der unabhängigen Variablen zuschreiben zu können, sollten beide Variablen vorher isoliert werden. Es müssen also jegliche Dritt- oder Störvariablen entweder komplett aus dem Versuchsaufbau ausgeschlossen werden oder, falls das nicht oder nur schwer möglich ist, in das Design integriert werden.

Ebenfalls muss bei der Erstellung der Hypothese darauf geachtet werden, dass die unabhängige Variable im Rahmen des Experiments überhaupt variiert werden kann.
Als letztes sei hier noch die Wiederholbarkeit genannt, also die Möglichkeit, das Experiment zu einem anderen Zeitpunkt erneut durchzuführen und dabei die gleichen Ergebnisse zu erhalten, welche bei einem quantitativen Experiment gefordert wird.

Es gibt verschiedene Formen des quantitativen Experiments, welche im Folgenden ausführlich beschrieben und erklärt werden. Genauere Betrachtung finden hierbei auch mögliche Validitätsprobleme.

4.1.1.1 Das Laborexperiment

Das meist verbreitete, quantitative Experiment ist wohl das Laborexperiment. Hierbei sind die Kontrolle von Drittvariablen und das Ziel der Wiederholbarkeit am einfachsten zu erreichen. Im Folgenden soll an einem konkreten Beispiel eine mögliche Vorgehensweise bei einem Laborexperiment geschildert werden.

Das Laborexperiment – Ablauf an einem praktischen Beispiel

Ein Forscher stellt die Vermutung auf, dass Menschen die rauchen, in Stresssituationen einen erhöhten Zigarettenkonsum aufweisen. Diese These formuliert der Forscher als so genannte Null-Hypothese:

In Stresssituationen ist bei Rauchern KEIN erhöhter Zigarettenkonsum festzustellen.

Um hier Missverständnisse zu vermeiden, müssen die verwendeten Begriffe eindeutig definiert und entweder als unabhängige oder abhängige Variable deklariert werden. In diesem Beispiel könnte man in Hinsicht auf das Testsetting (im Anschluss) die folgenden Definitionen wählen:

Stresssituation:	Eine Person, die unter zeitlichem Druck mathematische Aufgaben lösen muss.
Zigarettenkonsum:	Anzahl der gerauchten Zigaretten in einem festgelegten Zeitraum
Raucher:	Eine Person, die täglich (also regelmäßig) zwischen 15 und 20 Zigaretten raucht. Die genauen Zahlen sind hierbei nicht wichtig, solange sie vorher festgelegt werden.
Abhängige Variable:	Zigarettenkonsum

Unabhängige Variable: Stresssituation mit den Ausprägungen: *Stresssituation nicht vorhanden* und *Stresssituation vorhanden*

Zunächst werden die Teilnehmer, im Folgenden Versuchspersonen (kurz VP) genannt, in zwei Gruppen zufällig aufgeteilt, die Kontrollgruppe und die Versuchsgruppe. Erstere wird mit der Ausprägung der unabhängigen Variable *Stresssituation nicht vorhanden* getestet, wohingegen auf die Versuchsgruppe die unabhängige Variable mit der Ausprägung *Stresssituation vorhanden* während des Tests einwirken wird. Beide Gruppen werden für das Experiment in separate aber identische Räume geführt. Jede VP setzt sich an einen Platz, an welchem sich neben Schreibmaterial auch Getränke, Aschenbecher, Feuerzeuge und die bevorzugte Zigarettenmarke in ausreichender Menge befinden. Um eine Selbstbeeinflussung der VP zu verhindern, erfahren die Teilnehmer nicht, dass die Untersuchung des Zigarettenkonsums in Stresssituationen der eigentliche Grund für das Experiment ist.

Beide Gruppen erhalten nun einige mathematische Aufgaben, welche in beiden Gruppen identisch sind. Die Teilnehmer der Kontrollgruppe werden jedoch nicht unter Zeitdruck gesetzt, sondern dürfen in der vorgegebenen Gesamtzeit, so viele Aufgaben bearbeiten, wie sie mögen. Den Teilnehmern der Versuchsgruppe wird für jede Aufgabe ein striktes Zeitlimit gesetzt, um so eine künstliche Stresssituation herzustellen. Das Zeitlimit sollte möglichst so gewählt sein, dass es wirklich nur sehr schwer möglich ist, innerhalb diesem die Aufgabe zu lösen.

Nach Ablauf der Zeit werden nun in beiden Gruppen die gerauchten Zigaretten gezählt und anschließend verglichen. Mit Hilfe einer statistischen Analyse – in diesem Fall beispielsweise einer Varianzanalyse, wird nun überprüft, ob der gemessene Unterschied signifikant ist. Falls ja, kann die Null-Hypothese zu Gunsten einer der beiden Gruppen verworfen werden. Falls nein muss sie bis auf weiteres beibehalten werden.

Ein Laborexperiment zeichnet sich vor allem durch die Künstlichkeit der Versuchsanordnung aus. Durch diese künstliche Situation ist eine Kontrolle von Dritt- und Störvariablen optimal möglich. Dementsprechend dürfen sich die Kontroll- und Experimentalgruppe nur durch die Variation der unabhängigen Variable unterscheiden. Nur dann ist anschließend ein eindeutiger Kausalschluss möglich.

4.1.1.2 Weitere Varianten des quantitativen Experiments

I. Feldexperiment

Der Ort der Untersuchung wird hierbei aus dem Labor heraus, hin zu dem angestammten Platz des zu untersuchenden Gegenstandes, also in dessen natürlicher Umgebung verlegt. Ansonsten wird auch hier versucht in dieser natürlichen Umgebung eine unabhängige Variable zu verändern um einen Kausalschluss zu erreichen und eine Hypothese zu prüfen.

Aufgrund oftmals nur schwer zu kontrollierender äußerer Einflüsse, ist dies aber ungleich schwerer. Allerdings sind die Ergebnisse oftmals besser auf die Realität übertragbar.

II. Simultanexperiment

Es werden mehrere Gruppen gleichzeitig untersucht beziehungsweise beeinflusst.

III. Sukzessives Experiment

Es existiert keine Kontrollgruppe – lediglich eine Gruppe wird sowohl vor als auch nach der Veränderung der unabhängigen Variable untersucht. In manchen Fällen kann eine Kontrollgruppe überflüssig oder nicht zwingend notwendig sein. Dies wird auch als *Within Subjects Design* (siehe Kapitel 3.5) bezeichnet.

4.1.1.3 Validität quantitativer Experimente [Ross02]

Worin liegen nun die Schwierigkeiten bei quantitativen Experimenten, welche Probleme können auftreten und wie valide sind solche Tests?

4.1.1.3.1 Repräsentativität der Ergebnisse

Hier ist zunächst die Auswahl der Versuchspersonen von Bedeutung. Im optimalen Fall stellt die Auswahl einen verkleinerten Ausschnitt der jeweiligen Grundgesamtheit dar. Mit Grundgesamtheit ist im Allgemeinen zunächst die Menge aller Menschen gemeint. Da diese im Normalfall aber nur bedingt von Interesse ist, wird der Begriff Grundgesamtheit auch für Teilmengen verwendet, die für den Test relevant sind. In obigem Beispiel ist dies die Menge der Raucher, die durchschnittlich 15-20 Zigaretten am Tag rauchen. Die Auswahl an Versuchspersonen sollte somit, wenn möglich, ein repräsentativer Schnitt der 15-20 Zigaretten pro Tag rauchenden Bevölkerung sein. Die Definition der Grundgesamtheit entscheidet also, für welche Personengruppe das Ergebnis letztendlich signifikant ist. Oftmals ist es jedoch nicht möglich, allein durch ein Merkmal hier einschränkend vorzugehen. Selbst wenn die Auswahl noch lokal, auf zum Beispiel eine Stadt, begrenzt werden würden, wäre es kaum möglich eine Gruppe von Rauchern zu wählen, welche der gewählten Grundgesamtheit in allen Facetten gleicht. Da letztendlich auch nur Freiwillige teilnehmen können, wird die Auswahl weiter erschwert. In der Praxis führt dies dazu, dass oftmals Versuchspersonen dort gesucht werden, wo Gruppen von Menschen anzutreffen sind. Studenten, Soldaten und Arbeiter in großen Betrieben sind hierbei beliebte Pools. Die Verteilung innerhalb der Versuchspersonen entspricht dann jedoch nur noch bedingt der vorher definierten Grundgesamtheit.

Was heißt das nun für die Repräsentativität der Ergebnisse? Grundsätzlich sind ohne weitere Untersuchungen die Ergebnisse eines Experiments aus oben genannten Gründen meistens nicht repräsentativ für die entsprechende Grundgesamtheit, dürfen also nicht verallgemeinert werden. Da diese Situation natürlich das ganze Experiment ad absurdum führen würde, können, wie bereits

angedeutet, weitere Überlegungen und Untersuchungen angestellt werden, um eine Übertragbarkeit doch zu ermöglichen. Zunächst muss untersucht werden, inwieweit sich die Versuchspersonen von einer wirklich repräsentativen Auswahl unterscheiden. Daran anschließend, ob diese Unterschiede überhaupt Einfluss auf das Ergebnis haben. Hierbei kann unterschieden werden, ob der aufgetretene Effekt nur in seiner Höhe differieren könnte, oder ob er in der definierten Grundgesamtheit so eventuell gar nicht auftreten könnte.

In ersterem Fall können die Ergebnisse zwar auf die Grundgesamtheit ausgedehnt werden, jedoch nur als relative Aussage. In diesem Beispiel also: *Raucher die sich in Stresssituationen befinden, haben einen erhöhten Zigarettenkonsum.* Was *erhöht* bedeutet, kann und darf hierbei nicht absolut definiert werden! Für den Fall, dass der Verdacht besteht, dass der entdeckte Effekt so bei der Grundgesamtheit gar nicht auftreten könnte, besteht die einzige Möglichkeit, die Ergebnisse doch noch verwertbar zu machen darin, die Grundgesamtheit weiter einzugrenzen. Es müssen hierbei die Merkmale klassifiziert werden, die dazu führten, dass die Ergebnisse so in einer wirklich repräsentativen Auswahl nicht auftreten. Anschließend kann dadurch eine neue, vermutlich aber kleinere Grundgesamtheit definiert werden, auf welche die Ergebnisse übertragbar sind.

Letztendlich sind für die Übertragbarkeit aber noch weitere Faktoren ausschlaggebend, welche im Folgenden aufgeführt sind.

4.1.1.3.2 Validität des Experiments

Die Validität bezeichnet die Gültigkeit eines Experiments. Misst die Untersuchung/das Instrument wirklich das, was es messen soll? Sind die Ergebnisse verwendbar? Um hier eine Antwort zu erhalten, wird der Begriff zunächst weiter differenziert in interne Validität und externe Validität.

I. Interne Validität

Die interne Validität gibt Auskunft darüber, ob der beobachtete Effekt wirklich eindeutig der Veränderung der unabhängigen Variablen zugeschrieben werden kann. Sie ist also für die Kontrolle der internen Versuchssituation zuständig. Je künstlicher ein Versuchsaufbau ist (im Optimalfall Laborexperiment), desto einfacher lassen sich eventuelle Störeinflüsse kontrollieren und desto höher ist die interne Validität.

II. Externe Validität

Die externe Validität gibt Auskunft darüber, inwieweit sich der beobachtete Effekt auf die Grundgesamtheit generalisieren lässt. Eine hohe externe Validität stellt also sicher, dass die gemessenen Werte nicht nur innerhalb des Versuchsaufbaus zustande kommen, sondern auch in der realen Welt auftreten. Es geht hierbei vor allem darum festzustellen, inwieweit der Versuchsaufbau die Ergebnisse beeinflusst und somit eine direkte Übertragung auf die wirkliche Welt verhindert.

III. Interne Validität contra externe Validität

Externe und interne Validität sind eng miteinander verknüpft. Zwischen ihnen herrscht eine Art Antagonismus. Eine hohe interne Validität ist zwar wünschenswert, um einen eindeutigen Kausalschluss zu ermöglichen, jedoch wird dadurch die Situation auch unnatürlich, was dazu führt, dass die Ergebnisse nicht mehr ohne weiteres in die Realität außerhalb der Versuchsanordnung übertragen werden können. Kurz gesagt also: *Je höher die interne Validität desto niedriger die externe Validität und umgekehrt.*

IV. Stör- und Drittvariablen

Wie bereits mehrfach gesagt, können Störfaktoren das Ergebnis beeinflussen. Das Problem dabei ist, dass es quasi unmöglich ist, Störvariablen vollkommen aus dem Testsetting zu verbannen. Vielmehr liegt oftmals das Ziel darin, die Störvariablen konstant zu halten und dadurch in das Testsetting zu integrieren (passive Integration). Eine aktive Integration ist in manchen Fällen auch möglich und fördert zumeist die Natürlichkeit des Experiments, erhöht also die externe Validität. Allerdings wird es dadurch schwerer die Ergebnisse zu interpretieren, da oftmals kein eindeutiger Kausalschluss mehr zulässig ist. Durch standardisierte Testabläufe und kleinere Testgruppen können Störvariablen, beziehungsweise deren Einfluss, weiter reduziert werden.

Das Auftreten von Drittvariablen ist äußerst ungünstig. Der Fehler liegt hierbei meistens in der Aufstellung der Hypothese, wenn bei der Identifizierung der unabhängigen und abhängigen Variablen eine vermutete Kausalität mit einer Korrelation verwechselt wird.

Allerdings sind monokausale Zusammenhänge in der Sozialwissenschaft auch äußerst selten. Somit existiert fast immer zumindest eine weitere Drittvariable, die Einfluss auf die abhängige Zielvariable nimmt. Die Hypothese muss aus diesem Grund mit sehr viel Bedacht gewählt werden, um möglichst schon von vorneherein Drittvariablen, die nicht ausgeschlossen werden können, direkt zu integrieren. Ein gänzlicher Ausschluss dieser würde zwar die interne Validität erhöhen, eine Integration im Gegensatz dazu jedoch die externe. Es ist nicht immer leicht hier einen angemessenen Kompromiss zu finden!

V. Demand Characteristics und Forced Exposure

Demand characteristics bezeichnet eine Problematik, die die externe Validität beeinflussen kann. Es geht hierbei weniger um die Künstlichkeit, beispielsweise einer Laborsituation, sondern vielmehr darum, dass die Versuchspersonen im Allgemeinen wissen, dass sie an einem Experiment teilnehmen. Aus ethischen Gesichtspunkten sollten sie das auch, jedoch kann es ihr Verhalten beeinflussen. Oftmals verhalten sich Versuchspersonen in Testsituationen anders als in ihrem normalen Umfeld. Sie versuchen eventuell das „wahre“ Versuchsziel zu erkennen – selbst wenn dieses ihnen mitgeteilt wurde.

Ebenfalls wichtig für die externe Validität ist das so genannte F*orced Exposure.* Diese Problematik wird in der Literatur nur sehr selten behandelt, stellt dies doch das Grundprinzip des Experiments in Frage. Während eines Versuchs können die Versuchspersonen letztendlich nur den Anweisungen des Versuchsleiters Folge leisten. In obigem Beispiel müssen sie zum Beispiel mathematische Aufgaben lösen, auch wenn sie dazu vielleicht gar keine Lust haben und im Moment lieber etwas anderes tun würden. Der einzige Ausweg besteht im Abbruch des Experiments – allerdings scheuen sich viele, diesen Weg zu gehen, um zum Beispiel in den Augen der restlichen Teilnehmer aber auch in denen des Versuchsleiters nicht als Versager dazustehen. In gewissem Sinne kann zumindest bei einigen Versuchspersonen die Situation eintreten, dass sie sich zu dem Experiment gezwungen fühlen, obwohl sie sich freiwillig gemeldet haben. In obigem Beispiel tritt dieser Punkt nicht so stark auf, da eine realistische Stresssituation zumeist ebenfalls die Handlungsmöglichkeiten stark einschränkt. Nichts desto trotz ist es sehr schwierig, dieser Problematik gerecht zu werden, da zum Beispiel eine Lockerung des Versuchsablaufs unkontrollierbare Störvariablen mit ins Spiel bringen würde.

VI. Zufällige und Systematische Fehler

Zufällige und systematische Fehler sind eng mit dem Auftreten von Störvariablen verbunden. Die Gefahr besteht darin, dass eine Störvariable eventuell unentdeckt bleibt oder einfach nicht zu verhindern ist. Falls es sich um Einzelversuche handelt (die Versuchspersonen also Experimental und Kontrollgruppe zugeordnet werden, aber einzeln/isoliert das Experiment absolvieren), kommt es im Normalfall zu zufälligen Fehlern. Das heißt, dass zwar ein solcher Fehler das Ergebnis in seiner Genauigkeit beeinflusst, es aber nicht in eine bestimmte Richtung verändert. Je mehr zufällige Fehler auftreten, desto ungenauer werden die Ergebnisse. Eventuell kann dann sogar ein vorhandener Unterschied zwischen Kontroll- und Experimentalgruppe verdeckt werden.

Zu systematischen Fehlern kommt es, wenn die Versuchspersonen in Gruppen getestet werden. Tritt hier eine Störung auf, beeinflusst das unter Umständen sofort die gesamte Gruppe. Die Ergebnisse verändern sich dementsprechend nicht nur in der Genauigkeit, sondern werden verzerrt. Somit verringern systematische Fehler die interne Validität eines Experiments. Je kleiner die Gruppen sind, desto geringer wird die Auswirkung des Störfaktors auf das Ergebnis. Im besten Fall führt dies wiederum zu Einzelversuchen und der damit verbundenen Gefahr von zufälligen Fehlern. Allerdings sind Einzelversuche oftmals aus wirtschaftlichen Gründen nicht realisierbar, da sie zu viel Zeit in Anspruch nehmen.

VII. Lern- und Reifungseffekte

Selbst wenn die finanziellen Mittel und die Zeit für Einzeltests zur Verfügung stehen, können diese nicht vorbehaltlos eingesetzt werden. Da derartige Tests zwangsläufig über

einen längeren Zeitraum stattfinden müssen, sind sie anfällig für so genannte Lern- und Reifungseffekte. Diese können sowohl beim Versuchsleiter, als auch bei den Versuchspersonen auftreten. Der Versuchsleiter erwirbt über die Zeit mehr Routine und passt sein, zu Beginn vielleicht noch etwas unsicheres Auftreten mit der Zeit an. Versuchspersonen, die an einem späteren Test teilnehmen, fühlen sich deshalb vielleicht besser betreut oder einfach nicht so unwohl, wenn der Versuchsleiter ein sympathisches Auftreten besitzt. Lerneffekte bei Versuchspersonen können hingegen auftreten, wenn diese beispielsweise von Bekannten, welche bereits an dem Experiment teilgenommen haben, Details erfahren und somit schon eine Vorstellung haben, was auf sie zukommt. Dies kann dann beispielsweise die oben angesprochenen *Demand Characteristics* verstärken.

VIII. Versuchsleitereffekte

Wie eben bereits angemerkt kann der Versuchsleiter ebenfalls das Experiment beziehungsweise die Ergebnisse beeinflussen. Je nach Art des Experiments hat der Versuchsleiter die Aufgabe den Versuchspersonen den Test zu erklären, die Messungen durchzuführen, Störvariablen auszugrenzen, den Versuchspersonen die Aufgabe zu präsentieren und für eventuelle Hilfestellungen oder Rückfragen zur Verfügung zu stehen. Er ist somit das einzige Bindeglied zwischen Experiment und Versuchsperson. Dadurch besteht aber auch die Gefahr, dass der Versuchsleiter die Versuchspersonen in irgendeiner Hinsicht beeinflusst und somit das Ergebnis systematisch verzerrt. Versuchsleitereffekte lassen sich grundsätzlich in drei Kategorien unterteilen:

- Effekte die auf physischen oder sozialen Merkmalen des Versuchsleiters basieren. Hierzu können beispielsweise ein ausgeprägter Dialekt oder aber auch physische Attraktivität (oder auch das Gegenteil) zählen.
- Lern- und Reifungseffekte, wie bereits im vorigen Abschnitt beschrieben
- Effekte, die sich aus den persönlichen Erwartungen des Versuchsleiters an das Experiment ergeben. Eventuell vermutet oder erhofft der Versuchsleiter bereits ein bestimmtes Ergebnis der Messungen und versucht unbewusst, die Versuchspersonen dahingehend zu beeinflussen, dass dieses auch eintritt.

Die Lösung dieser Probleme ist nicht ganz einfach und letztendlich nicht ohne Kompromiss möglich. Eine Möglichkeit besteht darin, den Ablauf des Experiments weitestgehend zu standardisieren. Zum Beispiel sollten die Instruktionen für die Versuchspersonen schriftlich ausgehändigt werden, mit dem Ziel, dass der Kontakt zwischen Versuchsleiter und Versuchsperson auf ein Minimum reduziert wird. Hierdurch steigt jedoch wieder die Künstlichkeit der Situation und die Teilnehmer fühlen sich eventuell unwohl.

Lern- und Reifungseffekte können durch intensive Schulungen der Versuchsleiter minimiert werden. Dadurch gewinnen diese die nötige Erfahrung um von Beginn an, zum Beispiel auf Nachfragen der Teilnehmer, richtig reagieren zu können.

Um eine persönliche Bindung des Versuchsleiters mit dem Ziel des Experiments zu verhindern, sollte er im Idealfall dieses gar nicht vollständig kennen. Er sollte somit auch mit der Forschungsobjekt an sich nichts zu tun haben, sondern möglichst ein externer Experte sein.

Versuchsleitereffekte sind allgemein ein sehr großes Problem, welchem allerdings oftmals zu wenig Beachtung geschenkt wird.

IX. Ethische Bedenken

Ethische Bedenken bei quantitativen Experimenten sind keinesfalls von der Hand zu weisen. Es gibt einige unrühmliche Beispiele in der Geschichte, die deutlich machen, dass es oftmals ein Drahtseilakt sein kann, was moralisch vertretbar ist und was nicht (Beispiel *Stanford Prison Experiment*, verschiedene Gehorsamkeitsexperimente [Mil74]). Letztendlich muss man sich immer bewusst sein, dass Menschen keine Laborratten sind und dementsprechend auch nicht ihre persönliche Freiheit und Selbstbestimmung eingeschränkt werden darf. Fragwürdig werden Experimente beispielsweise immer dann, wenn die Versuchspersonen über den wahren Zweck der Untersuchung hinweggetäuscht werden. Hier sollte spätestens am Ende ein so genanntes D*ebriefing* erfolgen, das den Teilnehmern das wahre Ziel enthüllt.

Abschließend ist zu sagen, dass die Validitätsprobleme weitestgehend erforscht sind und auch Lösungsmöglichkeiten existieren. Jedoch ist jede Lösung wiederum ein Tradeoff – es gibt kein Geheimrezept für das perfekte Experiment, welches absolut repräsentative Ergebnisse liefert und nicht angezweifelt werden kann.

4.1.2 Qualitative Methoden

Im Gegensatz zu dem quantitativen Experiment, haben qualitative Methoden nicht das Ziel der Hypothesenprüfung. Vielmehr sollen Gegenstände oder Sachverhalte zunächst untersucht werden, um daraus dann eventuell Hypothesen generieren zu können.

4.1.2.1 Teilnehmende Beobachtung [AMR89], [May90]

Bei einer teilnehmenden Beobachtung befindet sich der Forscher nicht in einer rein beobachtenden Position, sondern nimmt selbst aktiv an der sozialen Situation teil, in die der zu untersuchende Gegenstand eingebettet ist. Es wird also eine persönliche Beziehung zu den Beobachteten aufgebaut – man nimmt an deren Leben teil, während man im Hintergrund Daten sammelt. Dadurch erhofft man

sich genauere Einblicke in die Struktur des Gegenstands zu erlangen, sozusagen die Innenperspektive erleben zu können. In manchen Fällen ist eine strukturelle Erschließung nur über diese Technik möglich.

Die Methodik der teilnehmenden Beobachtung ist nur teilweise standardisiert. Es sollte zwar ein Beobachtungsleitfaden erstellt werden und dieser sollte auch theoriegeleitet sein, der Beobachter/Forscher muss diesen Leitfaden aber weder immer parat haben, noch muss er ihn strikt befolgen. Vielmehr sollen mit Hilfe dieses Leitfadens die Erstellung einheitlicher Protokolle ermöglicht werden, um die Ergebnisse mehrerer beteiligter Forscher besser vergleichen zu können. Die Schwierigkeit einer teilnehmenden Beobachtung liegt darin, dass die Forscher einen Weg finden müssen, um sich in die soziale Umgebung integrieren zu können, ohne abgelehnt oder als Störung empfunden zu werden. Nach Mayring [May90] ist diese Methode auch sehr gut geeignet, um durch ihre explorative Form Hypothesen zu generieren.

4.1.2.2 Das Interview [Schäf95], [May90], [BD95]

Da die reine Beobachtung für sich genommen, sei sie auch teilnehmend, oft nicht ausreicht, wird zusätzlich auf eine weitere Form qualitativer Forschung zurückgegriffen – das Interview. Im Gegensatz zu Befragungen in der quantitativen Forschung, in der die Befragten oftmals nur die Auswahl aus mehreren vorgegebenen Antworten haben, sind qualitative Interviews deutlich offener und unstrukturierter gestaltet. Sie sollen dem Befragten die Möglichkeit geben, zu sagen, wie er persönlich über einem Sachverhalt denkt. Der Unterschied zeigt sich auch in der Art der Befragung – während quantitative Befragungen meist schriftlich anhand von Fragebögen ablaufen, vertraut die qualitative Forschung auf die verbale Form.

Mittlerweile gibt es eine Vielzahl von Interviewtechniken, wobei die meisten sich jedoch nur in dem Grad der Standardisierung unterscheiden. Zunächst also eine kurze Übersicht über die grundlegenden Vorgehensweisen und Techniken. Prinzipiell wird in der Vorbereitungsphase der interessierende Gegenstandsbereich unter theoretischen Gesichtspunkten aufgearbeitet. Anschließend werden die Fragen formuliert, je nach verwendeter Methode mehr oder weniger konkret. Diese werden in einem so genannten Interviewleitfaden zusammengestellt.
Die nun folgende Pilotphase, in der einige Testpersonen interviewt werden, soll neben der Sicherstellung der Qualität der Fragen vor allem dazu dienen, dem Interviewer die notwendige Erfahrung für die anschließenden „richtigen“ Befragungen zu geben. Für die Erstellung der Fragen und des Leitfadens, sollten einige grundsätzliche Dinge beachtet werden:

- Die Fragen sollten verständlich, möglichst einfach und nicht zu lang formuliert sein.
- Die Fragen dürfen die befragte Person nicht überfordern, dementsprechend muss darauf geachtet werden, dass kein Wissen implizit vorausgesetzt wird, welches der Befragte eventuell gar nicht aufweist.

- Die Fragen dürfen nicht suggestiv sein, also nicht von vorneherein eine bestimmte Antwort nahe legen.
- Die Eingangsfragen sollten möglichst leicht und unkontrovers zu beantworten sein, um dem Befragten den Einstieg zu erleichtern
- Fragen haben oftmals einen Ausstrahlungseffekt auf die nachfolgenden Themen, zum Teil können diese überflüssig werden oder müssen während des Interviews abgeändert werden – je nach Technik kann hier mehr oder weniger flexibel gehandelt werden.

Im Folgenden werden zwei Interviewtechniken, das strukturierte Interview und das problemzentrierte Interview, näher vorgestellt.

I. Das strukturierte Interview

Das strukturierte Interview, auch zum Teil als standardisiertes Interview bekannt, versucht eine möglichst hohe Standardisierung zu erreichen. Dementsprechend müssen nicht alle Fragen offen gehalten werden, sondern können durchaus auch geschlossen sein. Um das strukturierte Interview einsetzen zu können, müssen bereits Informationen über den interessierenden Gegenstandsbereich vorliegen. Durch seine Verwendung können oftmals Anhaltspunkte für das Vorhandensein bestimmter Variablen erlangt werden.

II. Das problemzentrierte Interview

Das problemzentrierte Interview stellt ein offenes, halbstrukturiertes Verfahren dar. Es soll einem offenen Gespräch möglichst nahe kommen, zugleich aber auf ein bestimmtes Problem (meist gesellschaftlich begründet) fokussiert/zentriert sein. Der Interviewer und der Befragte sollten möglichst eine offene, gleichberechtigte Beziehung aufbauen.

Es existieren drei unterschiedliche Fragetypen, die in einem problemzentrierten Interview verwendet werden:

- *Sondierungsfragen* dienen dazu, einen Einstieg in die Thematik zu geben. Sie sollten dem Interviewer Informationen darüber liefern, inwieweit den Befragten das Thema überhaupt interessiert und welche subjektive Bedeutung es für ihn hat.
- *Leitfadenfragen* dienen dazu, die wichtigsten Themenbereiche vorher bereits abzustecken. Sie sollten möglichst offen gestellt werden, um ein normales Gespräch zu ermöglichen. Im Optimalfall sollte der Befragte gar nicht merken, dass gerade eine vorher bereits festgelegte „Frage“ gestellt wurde.
- *Ad hoc Fragen* sind vorher nicht festgelegte Fragen oder Themen, die aber aufgrund der Gesprächsentwicklung von Bedeutung sein können. Hier muss der Interviewer spontan reagieren können, wenn er auf vorher zwar nicht bedachte, aber doch interessante Aspekte trifft.

4.1.2.3 Gruppendiskussionen – Focus-Groups [May90], [Schäf95]

Interviews bieten die Möglichkeit, sehr genau auf einzelne Personen eingehen zu können. Allerdings sind viele Meinungen und Einstellungen stark an soziale Zusammenhänge gebunden. Diese können am besten in einer Gruppenbefragung erforscht werden. Besonders geeignet sind sie bei der Untersuchung von Vorurteilen und Ideologien. Im Normalfall bringt eine direkte Frage nach, beispielsweise antisemitischen Vorurteilen kaum Ergebnisse, da die Befragten hier nicht offen antworten. Durch die Gruppendynamik können aber Diskussionen entstehen, in denen eventuell vorhandene Meinungen und Einstellungen zu diesem Thema viel offener zu Tage treten. Eine gut geführte Gruppendiskussion vermag die psychischen Sperren zu schwierigen Themen zu überwinden. Allerdings liegt hierin auch eine große Gefahr solcher Diskussionen. Die Teilnehmer werden eventuell zu Äußerungen gedrängt, die sie eigentlich gar nicht tätigen wollten.

Zu Beginn wird der Gruppe meistens ein so genannter Grundreiz, beispielsweise eine besonders kontroverse These präsentiert. Darauf aufbauend entwickelt sich dann die weitestgehend frei ablaufende Diskussion. In bestimmten Fällen kann es sinnvoll sein, am Ende eine so genannte Metadiskussion durchzuführen. Hier kann der Diskussionsleiter Fragen stellen, die darauf abzielen, ob die Teilnehmer ihre Ansichten auch wirklich frei äußern konnten und wie sie sich dabei gefühlt haben.

In Bezug auf die Planung einer solchen Gruppendiskussion gibt es einige Punkte, die es zu beachten gilt:

- *Diskussionsthema:* Hier kann unterschieden werden zwischen eng umschriebenen und wenig strukturierten Themen. Weiterhin ist von Belang, inwieweit die Teilnehmer persönlich von dem Thema betroffen sind, sprich inwieweit sie motiviert sind, darüber zu diskutieren.
- *Gruppengröße:* Optimal gilt eine Gruppengröße von 5-15 Personen. Je größer die Gruppe, desto weniger Sprechzeit hat ein Einzelner, jedoch umso mehr verschiedene Meinungen können auch existieren.
- *Zusammensetzung der Gruppe:* Um eine Diskussion zu ermöglichen, in der jeder Teilnehmer das Gefühl hat, den Diskussionspartnern ebenbürtig zu sein, sollten die Personen in Bezug auf so genannte soziodemographische Merkmale möglichst homogen ausgewählt werden. Zu den interessanten Aspekten gehören hier zum Beispiel Art der Ausbildung, Sachkompetenz in Bezug auf das Diskussionsthema, etc.
- *Bekanntheit der Mitglieder der Gruppe:* Hier kann unterschieden werden, ob die Diskussionsteilnehmer sich aus ihrem sozialen Umfeld bereits kennen, oder ob die Gruppe nur aufgrund der Untersuchung zusammengeführt wurde – eine so genannte ad-hoc Gruppe. Prinzipiell haben beide Vor- und Nachteile, je nach Themengebiet lässt sich hier aber keine generalisierende Empfehlung geben.
- *Meinungsverteilung:* Damit überhaupt eine Diskussion zustande kommt, sollten die Meinungen zu dem Themengebiet möglichst vielfältig sein

- *Schweiger:* Damit sind Teilnehmer gemeint, die sich gar nicht oder nur sehr selten zu Wort melden. Die Gründe können hier vielfältig sein, angefangen bei der Persönlichkeit des Teilnehmers, bis hin zu dem Problem, dass er mit dem Thema vielleicht nicht so viel anfangen kann. In kleineren Gruppen und in solchen, in denen sich die Teilnehmer kennen, ist die Anzahl der Schweiger im Allgemeinen kleiner.
- *Verhalten des Diskussionsleiters:* Der Diskussionsleiter hat mehrere Möglichkeiten die Diskussion zu führen. Er kann zum einen die formale Gesprächsleitung übernehmen, also beispielsweise die Steuerung, wann wer mit Reden an der Reihe ist, um hier ein Durcheinander bei hitzigen Diskussionen zu vermeiden. Er kann aber auch aktiver in die Diskussion eingreifen, sei es nur durch das Einbringen von themenrelevanten Stichworten, um bisher nicht angesprochene Themenbereiche zu erfassen oder aber durch die aktive Teilnahme an der Diskussion. Hierzu ist es ratsam nicht direkt mitzudiskutieren, jedoch von Zeit zu Zeit weitere Reizargumente einzubringen, um die Diskussion aktiv zu steuern.

Wenn möglich, sollte die Diskussion auf Video und/oder Tonband aufgezeichnet werden um die Auswertung zu erleichtern. Ebenfalls vorteilhaft kann das Einbringen eines „stillen Beobachters“ sein. Dieser beteiligt sich nicht an der Diskussion sondern achtet vielmehr auf die Gestik und Mimik der Teilnehmer, sowie sonstige Auffälligkeiten.

4.1.2.4 Validität qualitativer Methoden [Schäf95], [May90]

In Bezug auf die Gütekriterien herrscht bei der qualitativen Forschung Uneinigkeit darüber, ob diese in Relation zu denen der quantitativen Forschung gesehen werden müssen oder ob die qualitative Forschung nicht besser eigene definieren sollte.

I. Validität

Die Validität besitzt in der qualitativen Forschung einen besonders hohen Stellenwert, da sie selbst an sich den Anspruch stellt, besonders gegenstandsangemessen vorzugehen. Somit existieren auch hier einige Aspekte, die es zu beachten gilt. Beispielsweise ist es fraglich, inwieweit die Äußerungen in einem Interview wirklich authentisch sind, oder ob der Interviewer vielleicht unbewusst auf den Befragten einwirkt und so die Antworten verzerrt. Diese Verzerrungen lassen sich letztendlich sowohl an dem Interviewer, als auch an dem Befragten festmachen:

- *Verzerrung durch den Interviewer:* Probleme entstehen hier, wenn der Interviewer an entscheidenden Stellen nicht nachhakt, nicht auf den Befragten eingeht, es eventuell versäumt eine für den Befragten angenehme Atmosphäre zu schaffen. Verhindert werden können diese Fehler größtenteils durch intensive Schulung der Interviewer.

- *Verzerrung durch den Befragten:* Diese Art der Verzerrung kann entstehen, wenn der Befragte, aus unterschiedlichsten Gründen nicht bereit ist, wahrheitsgemäß zu antworten. Es ist nicht leicht eine solche Situation als Interviewer zu erkennen und angemessen darauf zu reagieren. Wie oben bereits erwähnt, können die Schaffung einer vertrauensvollen Atmosphäre oder auch die einer möglichst transparenten Untersuchungssituation – damit der Befragte sich wirklich sicher ist, dass er den wahren Grund der Befragung kennt – mögliche Gegenmaßnahmen sein.

II. Interne und Externe Validität

- *Interne Validität:* Als wichtigste Maßnahme zur Sicherung der internen Validität gilt in der qualitativen Forschung die Konsensbildung. Sobald mehrere Personen sich auf die Glaubwürdigkeit und den Bedeutungsgehalt der Ergebnisse einigen können, kann dies als Hinweis auf seine Validität aufgefasst werden. Die Konsensbildung kann hierbei nicht nur zwischen mehreren Forschern stattfinden sondern auch zwischen Forscher und erforschter Person (kommunikative Validierung) oder zwischen Forschern und so genannten Laien (argumentative Validierung).
- *Externe Validität:* Wie bei der quantitativen Methodik geht es bei der Frage der externen Validität darum, inwieweit die Ergebnisse sich auf die Wirklichkeit übertragen lassen, inwieweit sie verallgemeinerbar sind. Hier muss in der qualitativen Methodik unterschieden werden zwischen den hypothesengenerierenden Verfahren (z.B. qualitatives Experiment, [Klei86]) und den hypothesenprüfenden Verfahren (z.B. Grounded Theory, [GS84]). In ersterem Fall kann die externe Validität weitestgehend vernachlässigt werden, da es dem Verfahren nicht darauf ankommt, dass die Ergebnisse verallgemeinerbar sind. Es werden ja keine Hypothesen überprüft, vielmehr soll ein Gegenstand erst erkundet, beziehungsweise exploriert werden, um seine Struktur zu erfassen. Es erfolgt also keine Prüfung auf Verallgemeinerbarkeit – somit ist die externe Validität auch nicht von Bedeutung.

  Hypothesenprüfende Verfahren sind eher selten in der qualitativen Methodik. Als Beispiel könnte man hier die des *Theoretical Sampling* in der *Grounded Theory* nennen, auf welches jedoch an dieser Stelle nicht näher eingegangen werden soll [May90]. Letztendlich muss in diesem Fall eingestanden werden, dass die Ansprüche der quantitativen Forschung an die externe Validität von diesen Verfahren nur unzureichend erfüllt werden.

III. Ethik

Ähnlich wie bei der quantitativen Forschung existieren auch bei der qualitativen einige Bedenken bezüglich der ethischen Angemessenheit. Beispielsweise kommt es oft vor, dass bei einer teilnehmenden Beobachtung, die beobachteten Personen über die wahren Absichten der Forscher getäuscht werden. In Interviewsituationen, die nicht als solche zu erkennen

sind, gibt der Befragte eventuell sehr persönliche Dinge preis. Selbst wenn ein reguläres Interview stattfindet, kann beispielsweise ein missbilligender Blick des Interviewers zu einem massiven Vertrauensverlust bei der befragten Person führen. Oftmals kann das Eintreten von negativen Konsequenzen nicht von vorneherein ausgeschlossen werden.

IV. Eigene Gütekriterien der qualitativen Forschung

Auch wenn die Gütekriterien quantitativer Forschung weitestgehend übertragbar sind, gibt es Bewegungen innerhalb der qualitativen Forschung, die auf die Definition von eigenen Gütekriterien drängen. Dies hängt vor allen Dingen damit zusammen, dass beispielsweise das „schlechte“ Abschneiden qualitativer Verfahren bei Gütekriterien, wie Reliabilität oder externer Validität, diese in ein schlechtes Licht rückt und für Außenstehende schnell den Eindruck der nicht oder unzureichenden Wissenschaftlichkeit erweckt wird. Mayring [May90] schlägt zum Beispiel eine Unterteilung in sechs eigenständige Gütekriterien vor[1]. Diese lassen sich jedoch zum Großteil wieder den bereits bekannten Gütekriterien der quantitativen Forschung zuordnen, allerdings breiter und detaillierter aufgeschlüsselt.

4.1.3 Zusammenfassung: Methoden der Sozialwissenschaft

Abschließend bleibt zu sagen, dass sowohl qualitative als auch quantitative Methoden Vor- und Nachteile haben. Die qualitativen Methoden müssen in Bezug auf die Gütekriterien Schwächen einräumen – insbesondere eignen sie sich nur unzureichend für die Hypothesenprüfung. Das müssen sie allerdings auch nicht, denn hier liegt das Spezialgebiet der quantitativen Methoden. Die qualitativen Methoden eignen sich vielmehr dazu, überhaupt erst einmal Hypothesen zu entwickeln. Durch ihre größere Gegenstandsnähe erlauben sie zudem eine bessere Integration der gesamten Situation – es werden keine „Störfaktoren“ ausgeblendet, die vielleicht für die Struktur des Gegenstands wichtig sind.
Ethische Bedenken können sowohl bei qualitativer als auch bei quantitativer Forschung zu Recht geäußert werden und sollten bei der Planung auf jeden Fall eine Rolle spielen.
Es spricht letztendlich nichts gegen eine friedliche Koexistenz beider Ansätze, von der die gesamte Forschung profitieren würde. Bei der Auswahl von qualitativen und quantitativen Methoden sollte diese, wenn möglich, so gegenstandsbezogen wie nur möglich erfolgen. Auch die Entwicklung von Verfahren, die sowohl qualitative, als auch quantitative Aspekte beinhalten, ist nur positiv entgegenzusehen. Dadurch könnten eventuell die jeweiligen Schwächen kompensiert und die Stärken gebündelt werden.

[1] *Verfahrensdokumentation, Argumentative Interpretationsabsicherung, Regelgeleitetheit, Nähe zum Gegenstand, Kommunikative Validierung, Triangulation*

4.2 Usability Test-Methoden

Eine Verwandtschaft zwischen Usability Test Methoden und den Methoden der Sozialwissenschaften ist unübersehbar vorhanden. Beispielsweise finden sich in der Literatur ebenfalls qualitative Methoden, die schon allein aufgrund der, zumeist sogar identischen, Bezeichnung an Interviewtechniken, teilnehmende Beobachtungen – beispielsweise im Rahmen einer Kontextanalyse[2] – oder Focus-Groups erinnern. Und auch das quantitative Experiment der Sozialwissenschaften findet sich in Form von Performance Tests im Bereich des Usability Engineering wieder.

Somit stellt sich die Frage, ob die dortigen Erkenntnisse hinsichtlich der Validität problemlos auf das Usability Engineering übertragen werden können. Dies muss allerdings zunächst verneint werden. Der Grund hierfür liegt zum einen in der Tatsache, dass der Gegenstand der Untersuchung bei Usability Tests im Gegensatz zu den aufgezeigten Methoden der Sozialwissenschaften nicht der Mensch ist. Vielmehr interagiert dieser mit dem eigentlich zu untersuchenden Gegenstand, dem Software-Produkt. Zum anderen findet, im Gegensatz zu den sozialwissenschaftlichen Methoden, keine strikte Trennung zwischen qualitativen und quantitativen Methoden statt. Vielmehr gibt es einige Verfahren, welche sowohl quantitative also auch qualitative Daten liefern.

Nichts desto trotz können viele der Lehren, welche aus der Sozialwissenschaft bereits bekannt sind, in angepasster Form auf Usability Test Methoden übertragen werden.

Im Folgenden wird die Methodik der Usability Tests, welche im Entwicklungsprozess von VisMeB durchgeführt wurden und deren Ergebnisse Bestandteil dieser Arbeit sind (siehe Kapitel 5 & 7), ausführlich vorgestellt. Soweit möglich wird hierbei auch immer die Verwandtschaft zu den Methoden der Sozialwissenschaften beleuchtet, mit besonderem Augenmerk auf mögliche Validitätsprobleme.
Für einen umfassenderen Überblick über Usability Testmethoden und deren Verwandtschaft zu den Methoden der Sozialwissenschaften sei auf die Seminararbeit „Validität und Aussagekraft von Usability Test Methoden“ verwiesen [Ger03].

4.2.1 Focus-Groups/Gruppendiskussionen [Niel97], [McNam99]

Focus-Groups Tests können generell während des gesamten Entwicklungsprozesses Anwendung finden. Je nach Stand der aktuellen Entwicklung kann die Zielsetzung variieren. Zu Beginn können solche Tests etwa den Charakter eines Brainstormings haben, wohingegen zu späteren Zeitpunkten das Aufdecken von Usability Problemen vorrangig ist. Wie in den Sozialwissenschaften ist es ein rein qualitatives Verfahren, welches sich somit zumeist nicht dazu eignet, abschließend die Güte eines Produktes sicherzustellen und zu bewerten.

[2] *DATech Prüfverfahren zu DIN EN ISO 9241-10/-11, Anhang C*

In der Praxis stellt der Versuchsleiter zunächst eine Teilnehmergruppe zusammen. Dabei sollten wenn möglich Endnutzer und Usability Experten, aber auch ein oder zwei Programmierer vertreten sein. Die Gruppengröße sollte dabei 10 Teilnehmer nicht überschreiten, da ansonsten die Konsensbildung sehr schwer fallen kann. Da die Betrachtung des gesamten Software-Produkts zumeist deutlich zu umfangreich wäre, legt der Versuchsleiter im Voraus einen gewissen Fokus fest, in welchem sich später die Diskussion bewegen soll. Zusätzlich entwickelt er einige Leitfragen um sicherzustellen, dass kritische Stellen der Software auf jeden Fall Beachtung finden. Es ist darauf zu achten, dass diese Fragen nicht suggestiv sind und die Teilnehmer nicht zu Ja/Nein Antworten drängen (Beispiel: „Denken Sie, dass die gewählten Farben in diesem Menü nicht etwas zu grell sind?“ – vielmehr: „Was halten sie von der verwendeten Farbgebung in diesem Menü?“).

Während des Tests selbst, befinden sich die Teilnehmer und der Versuchsleiter gemeinsam in einem Raum. Wenn möglich, sollte die gesamte Diskussion zu Auswertungszwecken auf Video aufgezeichnet werden. Sollte das nicht möglich sein, ist eine weitere Person, welche die Diskussion möglichst ausführlich protokolliert, notwendig. Da in den meisten Fällen keiner oder nur wenige der Teilnehmer die Software bereits kennen, sollte zu Beginn eine Systemdemonstration der Diskussion vorausgehen. Weiterhin ist es essentiell wichtig, dass das System während der gesamten Diskussion zur Verfügung steht. Sollte dies nicht möglich sein, so müssen zumindest Screenshots der relevanten Bereiche vorhanden sein. Eine völlig freie Diskussion ohne jegliches Anschauungsmaterial wäre nur im Extremfall möglich und auch nur, wenn alle Teilnehmer mit dem Software-Produkt sehr vertraut sind. Da Endnutzer und Usability Experten vertreten sein sollten, ist dies zumeist aber nicht der Fall.

Die Diskussion sollte, wenn möglich, nicht länger als 90-120 Minuten dauern und zumindest eine Pause enthalten. Eine längere Dauer ist nur sinnvoll, falls gegen Ende der anvisierten Zeit klar ersichtlich ist, dass die Qualität der Diskussion noch nicht nachgelassen hat und alle Teilnehmer noch engagiert mitarbeiten.
Als Einstieg sollte der Versuchsleiter eine relativ freie Frage verwenden, zu der, im besten Fall, jeder der Teilnehmer etwas zu sagen hat. Zu Beginn tritt der Versuchsleiter als Moderator auf, der beispielsweise auch entscheiden kann, wer als nächstes etwas sagen darf. Dies ist vor allen Dingen bei größeren Gruppen sinnvoll. Anschließend sollte er aber weitestgehend in den Hintergrund treten. Seine Aufgabe besteht nun darin, eher zurückhaltend zu agieren und darauf zu achten, dass die Diskussion in geordneten Bahnen verläuft. Es kann auch durchaus vorkommen, dass die vorher festgelegten Leitfragen völlig ohne Zutun des Versuchsleiters zur Sprache kommen. Andererseits können auch vorher nicht bedachte Themen angesprochen werden. In diesem Fall liegt es in der Hand des Versuchsleiters zu entscheiden, ob er die Diskussion an dieser Stelle unterbricht und mit Hilfe einer Leitfrage wieder auf den eigentlichen Fokus lenkt, oder ob er sich durch dieses, bisher nicht bedachte Thema, eventuell weitere Erkenntnisse erhofft.

Ziel der Diskussion sollte stets sein, bezüglich eines Usability Problems einen Konsens zu erzielen. Falls beispielsweise zu der Menüfarbgebung jeder Teilnehmer eine andere Meinung besitzt und diese

während der Diskussion auch nicht ändert, hilft das später den Entwicklern nicht weiter. Darüber hinaus ist die Konsensbildung, wie auch bei den qualitativen Methoden der Sozialwissenschaften, Grundvoraussetzung für die Sicherstellung der Validität. Eine Gruppendiskussion bietet hierbei den Vorteil, diesen Konsens bereits während der Diskussion erzielen zu können und somit auch Endnutzer wirklich dabei mit einzubeziehen und zu berücksichtigen – sofern sie denn in der Gruppe vertreten sind.

Ebenfalls analog zu den Sozialwissenschaften ist das korrekte Verhalten des Versuchsleiters ungemein wichtig. Dieser sollte, wenn möglich, bereits Erfahrung aufweisen oder speziell geschult worden sein. Insbesondere mögliche Endnutzer lassen sich durch falsche Fragestellungen sehr leicht beeinflussen. Dies liegt zumeist daran, dass diese sich eventuell mit der Thematik nur bedingt auskennen und somit die Argumentation des Versuchsleiters schwer hinterfragen können. Weiterhin suchen Benutzer immer noch viel zu oft den Fehler bei sich selbst, anstatt die mangelhafte Software verantwortlich zu machen. In diesem Fall fällt dem Versuchsleiter auch durchaus noch eine psychologische Aufgabe zu – die Teilnehmer zu offener Kritik zu ermutigen und im Falle der Anwesenheit von Entwicklern einen hier möglichen Konflikt zu verhindern.

Abschließend ist zu sagen, dass der Einsatz von Focus-Groups während der Entwicklung nur zu begrüßen ist. Der Aufwand ist aufgrund der geringen notwendigen Teilnehmerzahl überschaubar und auch hinsichtlich Vorbereitungszeit und Dauer der Auswertung bietet diese Form des Usability Testing ein enorm gutes Kosten/Nutzen Verhältnis. Ein weiterer Vorteil liegt darin, dass durch die Flexibilität der Gruppendiskussion oftmals nicht nur Probleme angesprochen oder aufgedeckt werden. Vielmehr entwickeln sich im Anschluss daran zumeist konkrete Verbesserungsvorschläge, welche aufgrund des Diskussionscharakters oftmals weit über vage Ideen hinausgehen.

4.2.2 Heuristische Evaluation [Niel94]

Die heuristische Evaluation ist eine Entwicklung der Usability Forschung, welche nicht auf sozialwissenschaftlichen Methoden basiert. Auch die eindeutige Zuordnung zu einem Methodentyp ist nur schwer möglich, da sowohl qualitative als auch quantitative Daten erhoben werden können. Von Seiten der Usability Experten wird sie zumeist als Inspektionsmethode kategorisiert, welche etwas abseits der Tests mit Benutzern steht. Im Gegensatz zu diesen testen hierbei keine Endnutzer, sondern Experten das Produkt. Dabei handelt es sich zumeist um Usability Experten, jedoch sind auch Experten anderer Fachrichtungen möglich, beispielsweise aus dem Bereich der kognitiven Psychologie.

Ziel der heuristischen Evaluation ist das Aufdecken und Kategorisieren von Usability Problemen. Weiterhin sind zudem auch Re-Design Vorschläge möglich.

Ähnlich wie bei einer Gruppendiskussion führt der Versuchsleiter die Teilnehmer zunächst in die Software ein. Dabei können Kontextszenarios hilfreich sein, um einen umfassenden Überblick über die Funktionen zu ermöglichen. Anschließend untersuchen die Experten die Software einzeln und ohne

Kontakt zu einander oder dem Versuchsleiter. Sie orientieren sich dabei an einer vorher definierten Liste von Usability Kriterien – den so genannten Heuristiken. Jakob Nielsen, welcher die heuristische Evaluation geprägt hat, hat eine Liste der 10 wichtigsten Heuristiken erstellt, an welcher auch heute noch die meisten Heuristischen Evaluationen angelehnt sind [Niel94b]. Da diese 10 Heuristiken zumeist aber immer noch recht offen ausgelegt sind, kann es lohnenswert sein, auf konkretere Fragebögen zurückzugreifen, welche ebenfalls auf diesen Heuristiken aufbauen. Die XEROX Corporation hat beispielsweise einen derartigen Fragebogen[3] entwickelt. Die Experten kategorisieren jedes gefundene Usability Problem in *minor*, *major* und *catastrophe*. Dadurch kann bei der Auswertung auch gleich eine Prioritätsliste der Mängel oder Fehler erstellt werden. Weiterhin können sie auch direkt Verbesserungsvorschläge notieren.

Die Dauer für die eigentliche Inspektion sollte mit etwa zwei Stunden veranschlagt werden, eventuell muss dazu der Fokus auf bestimmte Bereiche der Software beschränkt werden. Anschließend ist es sinnvoll, mit allen Teilnehmern noch eine Nachbesprechung durchzuführen. Hier können besonders brisante Probleme direkt angesprochen werden, weswegen diese Sitzung zu Auswertungszwecken auf Video aufgenommen werden sollte.

Jakob Nielsen konnte in empirischen Untersuchungen feststellen, dass 3-5 Experten ausreichen, um etwa 75% der Usability Probleme aufzudecken und daher den optimalen Kosten/Nutzen-Faktor darstellen. Somit kann ein solcher Test relativ kostengünstig und zeitschonend durchgeführt werden, was ihn, ebenso wie die Gruppendiskussion, zu mehrfachem Einsatz innerhalb eines Entwicklungsprozesses prädestiniert.

Hinsichtlich der Validität ist zu beachten, dass der Fragebogen nicht von den Entwicklern selbst zusammengestellt werden sollte. Ansonsten besteht hier die Gefahr, dass bereits eine unterbewusste Vorauswahl getroffen wurde, welche Aspekte der Software durch den Fragebogen wirklich abgedeckt werden und daraus resultierend, welche Usability Probleme überhaupt entdeckt werden können. Standardisierte Fragebögen haben allerdings den Nachteil, oftmals zu abstrakt zu sein und gerade bei Teilnehmern, welche keine Usability Experten sind, für Verwirrung zu sorgen. Der Ablauf sollte zudem weitestgehend standardisiert sein, so dass alle Experten die gleichen Voraussetzungen haben.

Die Ergebnisse der heuristischen Evaluation sollten durchaus kritisch betrachtet werden. Aus Entwicklersicht ist es beispielsweise problematisch, dass die Methode einzig und allein darauf abzielt, Probleme und Fehler festzustellen, aber nicht auch die positiven Aspekte zu nennen, welche von „echten“ Nutzern eher geäußert werden. Weiterhin ist zu beachten, dass Experten keine Endnutzer sind. Die Praxis hat gezeigt, dass der unbedarfte Nutzer, welcher zumeist geringere Erfahrung mit Software hat, oftmals eine völlig andere Vorgehensweise offenbart und beispielsweise vorher als heikel angesehene Stellen problemlos umschifft, jedoch dafür an anderer Stelle hängen bleibt.

[3] *http://www.stcsig.org/usability/topics/articles/he-checklist.html, online am 15.03.2004*

Die heuristische Evaluation kann also Tests mit realen Benutzern keineswegs ersetzen, sondern sollte vielmehr im Vorfeld zu diesen angesetzt werden, um die größten Usability Schwächen zu identifizieren und beseitigen.

4.2.3 Performance Testing [DR99], [Usab03]

Bei den bisher beschriebenen Methoden konnten die Endnutzer entweder die Software nur sehen und darüber diskutieren (Focus-Groups) oder aber sie waren gar nicht in den Test involviert (Heuristische Evaluation). Wirkliche Schwachstellen einer Software können aber oft nur aufgespürt werden, wenn potentielle Endnutzer wirklich mit dem Produkt auch arbeiten. Hierzu dient der „klassische“ Usability Test. Bei diesem werden mindestens fünf Versuchspersonen ausgewählt, welche möglichst aus dem Bereich der Zielgruppe stammen sollten. In Einzeltests, welche zumeist in einem „Usability Labor“ stattfinden, bearbeitet nun jeder Teilnehmer vorgegebene Aufgaben an dem zu testenden System. Je nachdem zu welchem Zeitpunkt der Entwicklung ein solcher Test stattfindet, sind die Zielsetzung und damit auch der Ablauf unterschiedlich. Grundsätzlich sollte in jedem Fall zumindest ein lauffähiger Prototyp vorhanden sein, mit welchem die Versuchspersonen ohne Behinderung, beispielsweise durch ständige Abstürze oder das Fehlen elementarer Features, arbeiten können. Weiterhin sollte der Test möglichst standardisiert ablaufen, wobei der Grad der Standardisierung ebenfalls variieren kann. Findet der Test während der Entwicklung statt, also nicht abschließend oder sogar nach Fertigstellung des Produktes, so liegt das Hauptaugenmerk auf der Gewinnung qualitativer Daten – also das konkrete Aufdecken von Usability-Schwachstellen. Aus diesem Grund sind für solche Tests auch kleinere Versuchsgruppen von 5-10 Teilnehmern völlig ausreichend. Um eventuell auftretende Probleme und Schwierigkeiten bei der Benutzung des Produktes besser nachvollziehen zu können, werden die Benutzer gebeten, während dem Test ihre Gedanken laut zu artikulieren. Diese Technik wird *Thinking Aloud* genannt und ist ein vielfach angewandtes Verfahren des Usability Testings. Dient der Test einzig und allein dem Finden von Usability Schwächen, so ist zwar trotzdem ein standardisierter Testablauf wünschenswert, jedoch sind mögliche Stör- oder Drittvariablen nicht so kritisch zu betrachten, da während dem Test keine Hypothese geprüft wird. Oftmals werden aber zusätzlich auch quantitative Daten erhoben, zumeist in Form der zeitlichen Dauer, welche die Benutzer zur Lösung der Aufgaben benötigen. Dies ist beispielsweise der Fall, wenn verschiedene Produktversionen verglichen werden sollen oder aber überprüft werden soll, inwiefern die zu Beginn der Entwicklung definierten quantitativen Usability-Goals erreicht werden. In diesem Fall wird also ähnlich wie beim quantitativen Experiment in den Sozialwissenschaften eine Hypothese aufgestellt und mit Hilfe des Tests überprüft. Dementsprechend gewinnen auch mögliche Validitätsprobleme an Bedeutung.

Mit Hilfe von Fragebögen können bei jeder Form des „klassischen“ Usabilty-Tests zusätzliche Daten gewonnen werden, beispielsweise inwieweit den Benutzern die Farbgebung gefallen hat. Es können aber auch offene Kritikpunkte und Anregungen auf diese Weise erfasst werden. Sollen die Daten hierbei

quantitativ erhoben werden, so wird zumeist eine 5- oder 7-Likert-Skala verwendet (siehe Kapitel 3.6). Die qualitativen Daten werden in Freifeldern gewonnen.

Findet der Test allerdings am Ende oder im Anschluss an den Entwicklungsprozess statt, so sollte das Auffinden von Usability-Schwächen nicht mehr das primäre Ziel sein. Vielmehr geht es nun prinzipiell nur noch darum, quantitative, also vergleichbare Daten zu gewinnen. Aus diesem Grund wird diese Art von Test im Folgenden als *Performance Testing* bezeichnet. Im Vordergrund steht hier zum Beispiel der direkte Vergleich mit Konkurrenzprodukten. Aber auch das Überprüfen der quantitativen Usability-Goals kann ein wichtiges Anliegen sein, um im Notfall die Auslieferung noch stoppen und das Produkt nochmals überarbeiten zu können. In jedem Fall sollten die gewonnenen Daten für die Zielgruppe repräsentativ sein. Um dies zu erreichen, sollte zu allererst die Teilnehmerzahl deutlich erhöht werden, da bei lediglich fünf Teilnehmern keine sinnvolle statistische Auswertung möglich ist. Eben jene gewinnt nun deutlich an Bedeutung, da nun das Überprüfen einer Hypothese Ziel des Tests ist.
Weiterhin muss das Testsetting grundsätzlich genauer durchdacht werden. Beispielsweise muss zwischen *Within-Subjects Design* (in den Sozialwissenschaften auch sukzessives Experiment genannt) und *Between-subjects Design* gewählt werden. (Kapitel 3.5)

Hinsichtlich der Anwendung von *Thinking-Aloud* Techniken ist bei *Performance Testing* Vorsicht geboten. Zwar wird dies zum Teil auch bei diesem empfohlen, allerdings konnten, beispielsweise bei Vorab-Tests im Umfang des VisMeB Performance Evaluation (Kapitel 7), deutliche Leistungseinbrüche bei der Verwendung von *Thinking Aloud* festgestellt werden.
Die Hypothesenprüfung, der standardisierte Ablauf, die statistische Auswertung, all dies sind deutliche Hinweise auf die methodische Nähe zum quantitativen Experiment der Sozialwissenschaften. Inwieweit das auch auf die dort bekannten Validitätsprobleme zutrifft, soll im Folgenden betrachtet werden.

4.2.3.1 Validität und Aussagekraft – Performance Testing

Grundsätzlich können die meisten Erkenntnisse bezüglich der Validität des quantitativen Experiments in den Sozialwissenschaften (im Folgenden mit q.E.S. abgekürzt) auf das Performance Testing übertragen werden.

I. Repräsentativität der Ergebnisse

Die Auswahl der Versuchspersonen ist wie beim q.E.S auch hier entscheidend, falls die Ergebnisse auf eine Grundgesamtheit verallgemeinert werden sollen. Falls die Versuchspersonen beispielsweise alle gute bis sehr gute Computerkenntnisse hatten, die Software aber auch von Einsteigern benutzt werden soll, können hinsichtlich dieser Zielgruppe fehlerhafte Ergebnisse entstehen. Einsteiger haben oftmals gänzlich andere Probleme als erfahrene Computer-Benutzer. Marketingaussagen unterscheiden hier oftmals jedoch nicht, denn diese sollen so allgemein wie nur möglich formuliert sein („mit unserem Produkt kommen sie doppelt so schnell ans Ziel wie mit dem der Firma XY“). Weiterhin

werden fast immer absolute Werte verwendet – dass dies eigentlich fast nie möglich ist, ist ebenfalls von dem q.E.S. bekannt. Auch quantitative Usability-Goals werden oftmals absolut formuliert und verlangen somit auch von den Ergebnissen des Performance Testings absolute Werte. Es muss in jedem Fall klar darauf geachtet werden, dass dies für die definierte Grundgesamtheit, also die Zielgruppe wirklich zutrifft. Erleichternd kommt hier allerdings hinzu, dass nicht beliebig viele Merkmale der Teilnehmer das Ergebnis beeinflussen, sondern zumeist die Computererfahrung und die Sachkenntnis bezüglich des Anwendungsgebiets der Software entscheidend sind. Somit ist es deutlich einfacher, eine repräsentative Gruppe auszuwählen.

II. Interne Validität und externe Validität

Wie beim q.E.S. herrscht zwischen interner und externer Validität auch hier ein Antagonismus. Es muss je nach Software-Produkt sehr sorgfältig überlegt werden, welche der beiden wichtiger ist. Beispielsweise könnte eine hohe interne Validität bei einer Lotsensoftware, bei welcher die schnelle und problemlose Benutzung extrem wichtig ist, dazu führen, eben jene gewünschten Ergebnisse zu erhalten – allerdings ist es fraglich, ob diese Ergebnisse auch in der Realität erzielt werden können, in welcher sich der Lotse zumeist in einer Stresssituation befindet und sich nicht in dem Maße auf die Bedienung der Software konzentrieren kann, wie bei einem solchen Test. In diesem Fall wäre also eine möglichst realitätsnahe Situation besser geeignet, die dann aber gegebenenfalls schlechtere Zahlen liefern würde – was sich wiederum schlechter vermarkten lässt.

III. Stör- und Drittvariablen, zufällige Fehler

Stör- und Drittvariablen haben beim *Perfomance Testing* eine etwas andere Bedeutung. Da eigentlich grundsätzlich Einzeltests Anwendung finden, ist die Gefahr durch solche ungewollten Einflüsse das Ergebnis zu verzerren geringer einzuschätzen. Trotzdem sollten sie Beachtung finden, da die Anzahl an Versuchspersonen meistens nicht derart groß ist, dass hier viele Fehler abgefangen werden können. Während bei einem q.E.S. solche Störvariablen meistens komplett ausgeschlossen werden, ist es beim *Performance Testing* am wichtigsten, dass alle Teilnehmer die gleichen Bedingungen haben. Es kommt also darauf an, mit was die Ergebnisse verglichen werden sollen. Wenn in dem gleichen Testlauf auch das Konkurrenzprodukt getestet wird, sind vorhandene, aber konstante Störeinflüsse verkraftbar. Wenn allerdings die absolute Höhe der Ergebnisse wichtig ist, sollten solche unerwünschten Einflüsse naturgemäß ebenfalls weitestgehend unterbunden werden.

IV. Demand Characteristics und Forced Exposure

Demand Characteristics sind beim *Performance Testing* weniger wichtig, da den Versuchsteilnehmern das Versuchsziel vorher klar dargelegt wird und die Versuchsperson auch nicht Gegenstand des Tests ist. Trotzdem ist es eine unnatürliche Situation und die

Personen verhalten sich dementsprechend eventuell anders. Solange dadurch nicht ihre Leistungsfähigkeit beeinflusst wird, ist das jedoch vernachlässigbar.

Das Problem des F*orced Exposure* ist auch bei *Performance Testing* vorhanden, jedoch gilt auch hier, dass es nur kritisch wird, falls dadurch die Leistungsfähigkeit der Versuchsperson beeinträchtigt wird. Man sollte aus diesem Grund der Versuchsperson zu Beginn des Tests eindeutig erklären, dass sie jede Aufgabe und auch den ganzen Test jederzeit abbrechen kann. Zumindest das Abbrechen einzelner Aufgaben wird in der Praxis auch des Öfteren wahrgenommen.

V. Versuchsleitereffekte

Der Versuchsleiter hat eine ähnliche Rolle inne, wie bei einem q.E.S.. Allerdings gibt es einige Unterschiede, welche sich wie folgt äußern:

Da nicht das Verhalten der Versuchspersonen analysiert wird, sondern ihre Leistungsfähigkeit mit der Software, sind Effekte aufgrund physischer oder sozialer Merkmale eher unbedeutend. Wichtiger ist, dass der Versuchsleiter für eine angenehme Atmosphäre sorgt und der Versuchsperson glaubhaft vermitteln kann, dass nicht sie getestet wird, sondern die Software. Dies ist in der Praxis oftmals gar nicht einfach, da viele Teilnehmer sich trotzdem beobachtet und überwacht fühlen, wenn neben einer Videokamera noch ein Protokollant und der Versuchsleiter im Raum sitzen – eine Laborsituation, in welcher die Versuchsperson allein im Raum sitzt, und nur durch Kameras beobachtet wird, mag hier in gewissen Fällen helfen, kann aber auch die Problematik durch die erhöhte Künstlichkeit noch verstärken.

Lern- und Reifeeffekte sind ebenfalls nicht außer Acht zu lassen. Es hängt hier von dem Grad der Standardisierung ab, in wie weit sie von Bedeutung sein können. Da der Versuchsleiter aber unter anderem auch für Hilfen zuständig ist und im Allgemeinen allein entscheidet, wann eine solche zu geben ist, hat er außerordentlichen Einfluss auf die Ergebnisse. Eine Schulung ist hier in jedem Fall wünschenswert.

Wenn möglich sollten externe Versuchsleiter verwendet werden, die nicht aktiv an der Entwicklung des getesteten Produkts beteiligt sind. Da der Versuchsleiter, wie oben bereits erwähnt, unter anderem für die Hilfen verantwortlich ist, könnte er auch hier unterbewusst das Ergebnis verfälschen, wenn er persönlich an einem bestimmen Ergebnis interessiert ist. Die Entwickler selbst sollten prinzipiell weder als Versuchsleiter noch als Protokollant in Erscheinung treten. Es fällt ihnen erfahrungsgemäß sehr schwer, ruhig zu bleiben, wenn die Versuchsperson bei einfachen Aufgaben verzweifelt oder wenn ein Feature nicht gefunden oder missbilligt wird.

Jens Gerken
Mat. 01/473804
Universität Konstanz

4.2.4 Zusammenfassung: Validität von Usability Test-Methoden

Validitätsprobleme sind offenkundig auch bei Usability Test-Methoden vorhanden und sollten dementsprechend beachtet werden. Die Methoden der Sozialwissenschaft können hier wichtige Anhaltspunkte liefern, auf welche Aspekte besonders geachtet werden muss und wann Konflikte entstehen können. Denn trotz aller Bemühungen, einen validen Test durchzuführen, sollte man sich immer vor Augen halten, dass dies in jedem Fall nur mit Abstrichen möglich ist. Die Welt lässt sich nicht auf die Größe eines Labors reduzieren und somit kann es auch nicht möglich sein, Ergebnisse aus einem solchen Labor auf die Welt vollständig und fehlerlos zu übertragen.

Hinsichtlich der Evaluationen, die im Rahmen der Entwicklung von VisMeB durchgeführt wurden, konnten die Erkenntnisse bezüglich der Validität, wie in diesem Kapitel erklärt, aktiv mit eingebunden und berücksichtigt werden. Insgesamt wäre es wünschenswert, wenn diesem Thema auch in der Literatur in Zukunft mehr Beachtung geschenkt werden würde, um auf lange Sicht eine eigenständige Methodik zu entwickeln, welche über das Adaptieren der Methoden aus den Sozialwissenschaften hinausgehen könnte und sollte.

5 Evaluation des VisMeB Prototypen

Die Entwicklung des VisMeB Prototypen war durchweg geprägt von formativen Evaluationen. Hierzu gehörten bereits in frühem Stadium umfangreiche User-Tests, heuristische Evaluationen, Webumfragen und weitere methodische Eingriffe. Die während des INSYDER Projektes durchgeführten Usability-Tests sind in der Arbeit von Thomas Mann [Mann02] näher erläutert, wohingegen die Tests zu Beginn des INVISIP Projektes Teil der Bachelor Arbeit von Christian Jetter sind [Jett03]. Im Umfang dieser Projekt-Arbeit, welche die späte Entwicklung von VisMeB begleitete, wurden drei weitere Usability-Tests durchgeführt. Zum einen ein Focus Group Test, welcher mittels der qualitativen Methodik, Usability-Schwächen offenbaren und Redesign-Vorschläge ermöglichen sollte. Weiterhin eine heuristische Evaluation, welche ebenfalls Usability-Schwachstellen aufzeigen sollte und darüber hinaus den Stand der Lösung von zuvor identifizierten Usability-Problemen, welche bereits in einer im September 2002 durchgeführten heuristischen Evaluation [Jett03] entdeckt wurden, möglich machen sollte. Der abschließende Performance-Test hatte zum Ziel, den generellen Nutzen einer *Leveltable-Darstellung* im Vergleich zu einem listenbasierten System anhand eines quantitativen Tests zu überprüfen. Die Ergebnisse dieses Tests finden sich aufgrund des Umfanges separat in Kapitel 7.

5.1 Focus-Groups Test

Der Test fand im April 2003 statt. Verwendet wurde die zu diesem Zeitpunkt aktuellste Entwicklungsversion von VisMeB. Im Fokus stand dabei das Konzept der *Leveltable/Granularity-Table*, wobei sich die Diskussion aber auch auf andere Bereiche wie den *Circle Segment View* oder das *Assignment Tool* erstreckte. Im Folgenden sind die Testvoraussetzungen und Ergebnisse detailliert wiedergegeben.

5.1.1 Testsetting

An dem Test nahmen folgende fünf Personen teil:

Werner König
Christian Jetter
Inga Reeps
Philipp Liebrenz
Sebastian Rexhausen

Alle fünf Personen waren mit VisMeB vertraut und in unterschiedlichen Bereichen an dem Projekt beteiligt. Werner König und Philipp Liebrenz waren für die Programmierung des 3D Scatterplots zuständig, Christian Jetter und Inga Reeps im Bereich der Usability Evaluation und Sebastian

Rexhausen kümmerte sich um die Datenbankanbindung. Es wurde bewusst niemand ausgewählt, welcher aktiv an der Entwicklung der zu diskutierenden Themen beteiligt war/ist. Zum einen um zu verhindern, dass die Personen bereits eine zu stark vorgeprägte Meinung haben könnten und zum anderen um ein absolut ehrliches und offenes Gespräch zu ermöglichen, bei dem nicht aufgrund der Anwesenheit des Entwicklers keine allzu harsche Kritik geäußert wird.

Der Test fand in einem kleinen Konferenzraum statt, ausgestattet mit Beamer, Kamera und Notebook. Der VisMeB Prototyp wurde somit per Beamer an die Leinwand projiziert. Aufgrund von organisatorischen Problemen mit der Videokamera, verzögerte sich der Test um 40 Minuten in welchen die Teilnehmer warten mussten.

Die Diskussion startete mit einer kleinen Einführung in VisMeB. Anschließend begann der Test mit einer Einstiegsfrage, die bewusst relativ offen gehalten wurde, um schnell eine Diskussion herbeiführen zu können.

„Worin seht ihr die generellen Unterschiede zwischen Leveltable und Granularity-Table, wo liegen die jeweiligen Vorteile, wo die Nachteile – bietet diese Aufteilung einen echten Mehrwert“

Wie sich recht schnell zeigte, waren die Teilnehmer sehr eifrig bei der Sache, weswegen seitens des Versuchsleiters zunächst quasi gar nicht eingriffen wurde – die meisten Leitfragen ergaben sich bereits aus dem Gespräch oder bedurften nur noch eines kleinen Hinweises. Da ein Focus-Groups Test ein qualitatives Testverfahren ist, dessen Ergebnisse natürlich stark von der Einstellung der Teilnehmer abhängen, werden im Folgenden lediglich die Kritikpunkte und Ideen aufgezeigt und erklärt, welche sich während der Diskussion ergaben. Die hier angelegte Gliederung entspricht nur zum Teil den Fragen und dient mehr der Übersichtlichkeit.

5.1.2 Unterschiede zwischen Leveltable und Granularity-Table

Die *Leveltable* bietet nach einhelliger Meinung auf den ersten Blick deutlich mehr Informationen, die erste Stufe der *Granularity-Table* wirkt sehr nach Platzverschwendung, da nur drei Spalten angezeigt werden. Auch scheint die *Leveltable* übersichtlicher zu wirken und intuitiver in ihrer Funktion.

Weiterhin scheint die Farbgebung der *Leveltable* diese interessanter wirken zu lassen – zumindest auf den ersten Blick.

Die grundsätzlichen Unterschiede zwischen Level- und *Granularity-Table* – also der unterschiedliche Ansatz von Browsing-basiertem bzw. analytischem Suchen, wurde von den Teilnehmern nicht in der Art erkannt, dass hier wirklich eine klare Unterscheidung möglich wäre – vielmehr biete die *Granularity-Table* eben noch andere Visualisierungen, die in der *Leveltable* nicht vorhanden seien. Eine klare Definition inwieweit die unterschiedlichen Konzepte in der jetzigen Version noch vorhanden sind,

wäre wünschenswert – dann könnte diese auch in Usertests überprüft werden, was nach diesem Focus-Groups Test in jedem Fall sinnvoll erscheint.

5.1.3 Vor- und Nachteile der einzelnen Stufen der Leveltable

I. Level 1

Wie bei der *Granularity-Table* kam zunächst auch hier der Kritikpunkt, dass aufgrund der SuperTable Ansicht, der tatsächliche Informationsgehalt in Level1 zu geringe wäre. Dadurch wurde dann ein weiterer Punkt angesprochen: Ist Level 1 als Einstieg überhaupt geeignet? Es wurde hierbei die Frage aufgeworfen, ob eine Darstellung wie beispielsweise in Level 2 nicht erwartungskonformer wäre, da sie sich mehr an einer konventionellen Listendarstellung in Internetsuchmaschinen orientiere. Der Nutzen von Level 1, gerade bei großen Datenmengen um eine gewisse Übersicht zu erhalten, wurde allerdings nicht bezweifelt – inwieweit ein Einstieg mit Level 2 sinnvoller wäre, ist sicher gerade bei großen Datenmengen kontrovers zu sehen. Auch hier könnte mit Usertests diese These überprüft werden.

Ein Problem, das sämtliche Levels, sowohl in der *Leveltable* als auch *Granularity-Table*, betrifft: ein Indikator, nach welcher Spalte gerade sortiert ist, fehlt völlig.

II. Level 2

Grundsätzlich wurde an Level 2 wenig kritisiert. Einzig der etwas große Abstand zwischen den einzelnen Zeilen wirke hinsichtlich des Vergleichs der Relevanzbalken zwischen den Dokumenten etwas störend. Wie bereits erwähnt, wurde zum Teil die Meinung geäußert, dass Level 2 dem Benutzer einen besseren, da vertrauteren Einstieg bieten könnte. Es wurde der Vorschlag vorgebracht, dass je nach Anzahl der vorhandenen Dokumente (bei vielen Level 1, bei wenigen Level 2) das jeweilige Level standardmäßig zuerst angezeigt werden sollte – ob dies allerdings erwartungskonform ist und vom Nutzer verstanden wird, ist sicherlich auch eher zweifelhaft.

III. Level 3

Level 3 bietet als Neuerung die Relevance Curve, welche allerdings nicht interaktiv verwendbar ist. Diese wurde daraufhin auch von allen Teilnehmern vor allem hinsichtlich der erweiterten Version in Level 4 als relativ nutzlos empfunden. Hinzukommt, dass auch der Sinn einer derartigen Visualisierung erst durch die Version in Level 4 wirklich klar hervortritt. Diesbezüglich stellte sich nun die Frage, inwieweit Level 3 eine Daseinsberechtigung hat und ob Level 3 und 4 nicht zusammengelegt werden könnten. Allerdings wurde hinsichtlich dieser Idee angemerkt, dass in diesem Fall eventuell der Sprung von Level 2 zu dem neuen Level 3/4 zu groß sein könnte.

IV. Level 4

Level 4 wurde allgemein als sehr sinnvoll und gut zu benutzen empfunden. Allerdings war zunächst nicht klar, was die Länge der X-Achse der Detailed Relevance Curve für eine Bedeutung hat, da hier keine Beschriftung oder ähnliches vorhanden ist.

5.1.4 Assignment Tool

In Zusammenhang mit der Diskussion um die einzelnen Levels kam auch das Assignment Tool zur Sprache, da zum Teil die Belegung der einzelnen Spalten kritisiert wurde. Dieses dient dazu, die Zuordnung der Datenbankfelder in VisMeB zu manipulieren. Beispielsweise kann durch dieses Tool die Spaltenbelegung für jeden Level der Leveltable konfiguriert werden. Das Vorhandensein des Assignment Tools wurde zwar begrüßt, die Handhabung aber für den Normalbenutzer als zu kompliziert empfunden. Es wurde vorgeschlagen, ein abgespecktes Assignment Tool direkt in die Visualisierung zu integrieren, um beispielsweise wie bei Microsoft Excel oder auch Microsoft Outlook direkt die Belegungen der einzelnen Spalten verändern zu können. Hinsichtlich des Assignment Tools wurde vor allem das nicht Vorhandensein einer „what you see is what you get“-Manipulation kritisiert.

5.1.5 Vor- und Nachteile der einzelnen Stufen der Granularity-Table

I. Stufe 1

Wie einleitend bereits erwähnt, wirke nach Ansicht der Teilnehmer die erste Stufe der *Granularity-Table* etwas leer und informationsarm, da nur drei Spalten angezeigt werden und darüber hinaus durch den Lense Effekt auch in diesen, auf den ersten Blick keine wirklichen Informationen angezeigt würden. Andererseits wirke die Darstellung dadurch auch recht übersichtlich und auf das wesentliche – die Gesamtrelevanz und den Titel – beschränkt. Die Bezeichnung „Visualization“ der Spalte für die Gesamtrelevanz wurde jedoch als verwirrend empfunden.

II. Stufe 2

Die Teilnehmer kritisierten hierbei vor allem die nun veränderte Visualization Spalte, in welcher nun nicht mehr die Gesamtrelevanz zu sehen war, sondern die kumulierten Einzelrelevanzen. Dabei sei nicht klar ersichtlich, was die Länge des Balkens insgesamt nun aussage, da eine Sortierung nach Gesamtrelevanz in der ersten Stufe (welche beim Wechsel zur 2. übernommen wird) deutlich zeige, dass die Länge des kumulierten Balkens nicht die Gesamtrelevanz darstelle. Hinzukommt, dass eine derartige Balkendarstellung nur sehr wenig Informationsgehalt bietet, da die Balkenlängen kaum verglichen werden können.

III. Stufe 3/4

Die Teilnehmer bemängelten hier die zu geringen Unterschiede zwischen den beiden Stufen. Weiterhin sorgte wieder die Visualization Spalte für Verwirrung, die auf den ersten Blick die Einzelrelevanzen der Suchterme widerspiegeln sollte. Allerdings stimmten weder die Reihenfolge der Farben noch die Anzahl der Balken mit den Suchtermen überein.

Weiterhin wurde bei der Beschriftung der Balken vorgeschlagen, ein zusätzliches Prozentzeichen der Relevanzzahl anzuhängen, um dies dem Nutzer zu verdeutlichen.

IV. Stufe 5/6

Erneut entstand die Frage, inwieweit hier eine Aufteilung in 2 Stufen wirklich sinnvoll sei. Die Visualisierung in Stufe 5 wurde als weniger selbsterklärend als die Detailed Relevance Curve der *Leveltable* empfunden und sorgte für Diskussion hinsichtlich der Bedeutung einiger Details. Die Volltext bzw. Abstract Darstellung in der Tabelle selbst wurde als eher kritisch betrachtet, da das Scrollen nur sehr schwerfällig möglich war und auch allgemein ein Scrollen innerhalb einer Tabellenzelle als eher störend empfunden wurde.

5.1.6 Grundsätzliche Problematik Granularity-Table

Die Teilnehmer hatten zum Teil große Probleme den unterschiedlichen Ansatz zwischen *Granularity-Table* und Level Table zu verstehen und nachvollziehen zu können. Insbesondere die *Granularity-Table* wirkt nach Aussage der Teilnehmer, was die Zusammenstellung der einzelnen Stufen angeht, zum Teil sehr einfallslos und aufgesetzt. Das Konzept einer nahezu stufenlosen Veränderung des Informationsgehalts konnte nicht vermittelt werden. Das Konzept der *Granularity-Table* sollte somit hinsichtlich der momentanen Umsetzung überprüft und eventuell angepasst werden.

5.1.7 Circle Segment View

Keiner der Teilnehmer war in der Lage, die genaue Funktionsweise des CSV zu erklären. Er sei keineswegs selbsterklärend und auch nicht intuitiv zu bedienen. Die Farbverläufe innerhalb der Kreisabschnitte wirkten eher störend, da sie nicht radial verlaufen, sondern eher auf Eckpunkte ausgerichtet scheinen.

5.1.8 Ausblick Filterfunktionen

Das Fehlen von Filtermöglichkeiten wurde von allen Teilnehmern bemängelt und sollte dementsprechend möglichst bald umgesetzt werden

5.1.9 Zusammenfassung: Ergebnisse des Focus-Groups Test

Die Ergebnisse zeigen deutlich, dass das Konzept der Level/*Granularity-Table* zum damaligen Zeitpunkt noch nicht zu vollster Zufriedenheit umgesetzt war. Insbesondere die Aufteilung der einzelnen Level/Granularitätsstufen sorgte für Diskussion. Interessanterweise ergaben sich bei der Diskussion auch etliche Redesign Vorschläge. Dies ist wahrscheinlich auf den Umstand zurückzuführen, dass alle Beteiligten an der Entwicklung von VisMeB in gewisser Hinsicht beteiligt waren und somit eine persönliche Motivation zur Fehlerausbesserung empfanden. Nachteilig für die Aussagekraft des Tests ist allerdings, dass keine Endnutzer oder zumindest entwicklungsfremde Teilnehmer anwesend waren. Dadurch sind viele Vorschläge oder Kritikpunkte hinsichtlich der Aufteilung der Level/Granularitätsstufen nur mit Bedacht zu sehen und letztendlich nur durch einen weiteren Test mit realen Benutzern zu validieren. Nichts desto trotz konnten für die weitere Entwicklung viele Erkenntnisse gewonnen werden, welche diese positiv beeinflussten.

5.2 Heuristische Evaluation

Wie eingangs kurz angeführt, wurde im Rahmen des INVISIP/VisMeB Projektes bereits im Anfangsstadium eine heuristische Evaluation durchgeführt. In zwei weiteren Testsessions im April und Mai 2003 wurde nun im späten Entwicklungsstadium nochmals eine solche Evaluationsmethode angewandt. Insgesamt nahmen an den beiden Testsessions sechs Experten teil. Die erste Gruppe bestand dabei aus drei Mitarbeitern des Fachbereichs „Datenbanken und Visualisierungen“ der Universität Konstanz. Die Teilnehmer besaßen somit allesamt nachhaltige Kenntnisse im Bereich der visuellen Suchsysteme. In dem zweiten Testdurchgang im Mai 2003 wurden die gleichen Experten zu Rate gezogen, welche schon bei der heuristischen Evaluation, die im September 2002 von Christian Jetter durchgeführt wurde, teilgenommen hatten. Somit konnte bei diesen auf eine Einführung in das System verzichtet werden. Weiterhin ergab sich die Möglichkeit, die gewonnenen Daten mit der vorigen Evaluation zu vergleichen. Für den Test wurde die von XEROX[4] entwickelte Checkliste verwendet, welche auf den von Jakob Nielsen [Niel94b] identifizierten Heuristiken aufsetzt. Allerdings mussten einige Bereiche, wie zum Beispiel Fragen bezüglich der Hilfefunktionalität ausgeklammert werden, da sie zum Zeitpunkt des Tests nicht auf VisMeB anwendbar waren. Da dieser Fragebogen jedoch, trotz der Einschränkungen, bereits bei der ersten heuristischen Evaluation erfolgreich angewendet wurde, empfahl sich die erneute Verwendung.

5.2.1 Testdurchführung

Insbesondere bei dem ersten Test, bei welchem die Visualisierungsexperten teilnahmen, wurden diese zunächst umfassend eingeführt. Hierzu zählte eine umfangreiche Systempräsentation sowie eine

[4] *http://www.stcsig.org/usability/topics/articles/he-checklist.html, online am 15.03.2004*

ausführliche Erklärung des Tests und Ihrer Aufgabe. Im Anschluss an den Test wurde eine kurze Gruppendiskussion geführt, in welcher die Teilnehmer ihre Empfindungen und Entdeckungen äußern konnten.

Die 2. Gruppe bestand aus Usability Experten des Fachbereichs „Mensch Computer Interaktion“ der Universität Konstanz. Die Teilnehmer waren allesamt mit der Methodik der heuristischen Evaluation vertraut und kannten auch bereits den XEROX Fragebogen. Auf eine ausführliche Systempräsentation wurde dementsprechend verzichtet.

Jeder Teilnehmer war während der Testdauer von den anderen Teilnehmern und dem Testleiter isoliert in einem Raum untergebracht.

5.2.2 Testergebnisse

Zunächst ist zu sagen, dass insbesondere die Qualität der Ergebnisse der ersten Gruppe leider mangelhaft ist. Dies ist nicht direkt den Experten anzulasten, vielmehr der Methodik der heuristischen Evaluation und einer offensichtlich doch zu kurzen Einführung. Es scheint, dass die Methodik für *Nicht-Usability* Experten nicht auf Anhieb adaptierbar ist. Beispielsweise versäumten die Teilnehmer, die gefundenen Usabilty Probleme genauer zu kategorisieren und auch zu spezifizieren. Es wurde zwar im Vorfeld deutlich darum gebeten, allerdings war den Teilnehmern die Wichtigkeit dieser Kategorisierung wohl nicht bewusst, weswegen sie nur sporadisch vorgenommen wurde. Allerdings konnten trotzdem einige interessante Ergebnisse gewonnen werden.

Aufgrund der Erfahrung der Teilnehmer der 2. Testgruppe, traten diese Probleme verständlicherweise nicht auf. Die im Folgenden auszugsweise präsentierten Ergebnisse, stützen sich somit vornehmlich auf diese Experten (als Experte 4-6 bezeichnet). Die Auswahl der Kategorien wurde aufgrund der Anzahl der gefundenen Usability Probleme getroffen.

5.2.2.1 Visibility of System Status

	Experte 4	Experte 5	Experte 6
Minor	1	1	4
Major	2	1	3
Catastrophe	0	1	0

Abbildung 5.1: Einstufung der Usability Probleme – Visibility of System Status

Erstaunlich ist die Tatsache, dass die Experten zum Großteil völlig unterschiedliche Probleme identifizierten. Eine Übereinstimmung gab es lediglich bei der Beanstandung der Reaktionszeit des Systems, welche durchgehend als zu lang empfunden wurde. Hinsichtlich der Einstufung wurde zweimal die Kategorie *Major* gewählt und einmal *Catastrophe*. Die Visualisierungsexperten

beanstandeten in der abschließenden Diskussion ebenfalls die zu langsame Antwort- und Reaktionszeit des Systems.

Ein weiterer Kritikpunkt, welcher zum einen als *Major* eingestuft wurde und zum anderen auch zweifach genannt wurde, betraf das Verhalten des Systems beim Umschalten zwischen den Levels. In diesem Fall scrollte die Tabelle wieder automatisch ganz nach oben. Gerade bei großen Datenbeständen, bei welchen schon etwas Aufwand notwendig ist, um ein bestimmtes Dokument zunächst einmal zu finden, wurde dies als sehr störend empfunden. In diesem Fall musste bei einem Levelwechsel das Dokument oftmals erneut gesucht werden, da es sich nun nicht mehr im Fokus befand. (2 x Major)

Weiterhin wurde beanstandet, dass inaktive Menüpunkte, beispielsweise bei dem Kontextmenü des Scatterplots, nicht grau unterlegt wurden. (1 x Major)

In Hinblick auf den 3D Scatterplot wurde zum einen ebenfalls die sehr träge Handhabung und Reaktionszeit bemängelt und zum anderen die für die Legende verwendeten Icons kritisiert, da diese nicht konsistent zum restlichen System waren. (1 x Major, 1 x Minor)

Als letzter Kritikpunkt sei hier noch die mangelhafte Umsetzung der Sortierfunktion innerhalb der Spalten der *Leveltable* erwähnt. Es war zum einen nicht ersichtlich, nach welcher Spalte sortiert war und zum anderen konnte auch nicht die Sortierreihenfolge erkannt werden. (1 x Minor)

5.2.2.2 Consistency and Standards

	Experte 4	Experte 5	Experte 6
Minor	3	3	3
Major	3	5	2
Catastrophe	0	0	0

Abbildung 5.2: Einstufung der Usability Probleme – Consistency and Standards

Die hier aufgedeckten Schwachstellen stehen zum Großteil in Zusammenhang mit der Farbgebung. Zum einen wurde von allen drei Experten beanstandet, dass zur Aufmerksamkeitsgewinnung zu viele Farben verwendet wurden. Dies zeigte sich besonders deutlich in der *Browserview* – hier wurden alle Suchbegriffe in ihrer jeweiligen Farbe hervorgehoben, wodurch einige Texte nur noch ein einziges Farbenwirrwarr darstellten. (3 x Major)

Weiterhin wurde angemerkt, dass auch allgemein zu viele unterschiedliche Farben verwendet wurden, welche darüber hinaus noch zum Teil sehr nahe beieinander im Farbspektrum angesiedelt waren. (1 x Minor, 1 x Major)

Ebenfalls in Zusammenhang mit den Farben wurde das Fehlen einer Legende sowohl im Circle Segment View als auch im 3D Scatterplot kritisiert. (1 x Major)

Darüber hinaus wurde das Fehlen von Hotkeys kritisiert und das Nichtvorhandensein einer horizontalen Scrollmöglichkeit innerhalb der Level-/Granularity-Table. (Hotkeys: 1 x Minor, 1 x Major; Scrolling: 1 x Minor, 1 x Major)

Wie bereits zu erwarten, kritisierten die Visualisierungsexperten die Farbgebung ebenfalls deutlich. Die Kritikpunkte waren grundsätzlich dieselben, wobei deutlicher das Fehlen einer Legende bemängelt wurde.

5.2.3 Zusammenfassung: Ergebnisse der Heuristischen Evaluation

Insgesamt konnte vor allen Dingen im Vergleich zu den Ergebnissen der ersten heuristischen Evaluation [Jett03] ein deutlich positiveres Fazit gezogen werden. Viele der damals kritisierten Usability Mängel waren mittlerweile behoben. Der Test konnte aber auch deutlich zeigen, dass das Einbeziehen von externen Experten, welche keine Usability Fachkenntnis aufweisen, zwar zusätzliche, interessante Ergebnisse hervorbringen kann, ohne eingehende Schulung dieser allerdings eher kritisch zu betrachten ist. Ein Focus-Groups Test wäre in diesem Fall wohl die effektivere Testmethode gewesen.

6 Analyse der Methodik ausgewählter Evaluationen

Trotz all der technologischen Fortschritte in den letzten Jahren und der damit verbundenen, nun deutlich erhöhten grafischen Komplexität, welche mit einem modernen PC dargestellt werden kann, dominieren im Bereich der Recherche – vor allen Dingen bei Online Suchmaschinen und professionellen Recherchesystemen – weiterhin, zumeist rein textbasierte Systeme. Allerdings werden die Vertreter so genannter visueller Suchsysteme nicht müde, die Vorteile ihrer jeweiligen Visualisierungen anzupreisen und langsam aber sicher scheint sich dafür auch ein Markt zu etablieren – dies zeigt sich insbesondere bei Gebieten, die etwas abseits der reinen Suchfunktionalität liegen. Beispielsweise bieten Datamining Programme, also Software die Unternehmen bei der Entscheidungsfindung helfen soll, indem z.B. große Mengen Kundendaten auf Zusammenhänge analysiert werden, immer öfter auch verschiedene Visualisierungen an. Allgemein können solche Programme unter *Informationsvisualisierungs-Werktzeugen* zusammengefasst werden, welche im Folgenden kurz als *Visualisierung* bezeichnet werden sollen.
Grundsätzlich stellt sich die Frage, ob eine Visualisierung gegenüber einer beispielsweise textbasierten Darstellung überhaupt Vorteile bietet. Gerade im Bereich der Online Suchsysteme hat sicherlich nicht umsonst mit *Google* ein Unternehmen ein Quasi-Monopol inne, welches weiterhin auf eine rein textbasierte Darstellung setzt. Kommerzielle Produkte, welche mit Visualisierungen arbeiten, verzichten zudem oftmals darauf, Evaluationsstudien über die Leistungsfähigkeit ihrer Visualisierung im Vergleich zu herkömmlicher Software – sofern sie überhaupt durchgeführt wurden – zu veröffentlichen. Nicht selten wird von Seiten der Marketingabteilung die Visualisierung an sich schon als Feature und Vorteil gegenüber einem Konkurrenz- oder auch Vorgängerprodukt angepriesen.

Diese Analyse soll einerseits einen kurzen Einblick in die Ergebnisse verschiedener Evaluationen zu Visualisierungen ermöglichen und anderseits die Methodik dieser Evaluationen näher beleuchten. Da die Unterschiede zwischen verschiedenen Visualisierungen, sowohl was die Optik als auch den konkreten Einsatzzweck betrifft, oftmals sehr groß sind, sind generalisierende Aussagen bezüglich der Leistungsfähigkeit dieser nur sehr schwer möglich. Vor allen Dingen verhindern aber zum Teil völlig unterschiedliche Ansätze bei der Evaluierung der Systeme einen marktübergreifenden Vergleich. Aus diesem Grund wurde das Hauptaugenmerk auf die Methodik der Evaluationen gelegt.
Im Folgenden wird zunächst jeweils die Visualisierung kurz vorgestellt und im weiteren Verlauf das Testdesign, wie von den Autoren veröffentlicht, beschrieben. Daraufhin folgt ein Fazit dieses Testdesigns, welche als subjektive Bewertung des Autors dieser Arbeit zu betrachten ist. Gleiches gilt für die Auswertung der jeweiligen Evaluationsergebnisse. Auch hier stehen zuerst die Ergebnisse, wie von den Autoren veröffentlicht, und anschließend, in dem wieder als Fazit betitelten Abschnitt, die subjektive Bewertung mit Schwerpunkt auf die jeweilige statistische Auswertung. Im Fall von DEViD wurde nur ein allgemeines Fazit gezogen, welches sowohl auf Testdesign als auch Auswertung Bezug nimmt.

6.1 InfoZoom

InfoZoom ist ein kommerzielles Produkt, welches von der Firma *humanIT* [hum04] weiterentwickelt und vertrieben wird. Die Einsatzmöglichkeiten sind sehr vielfältig, angefangen bei Katalogsystemen bis hin zu der Möglichkeit, einen schnellen Überblick über alle Geschäftsfelder eines Unternehmens zu erlangen. Es wäre auch durchaus denkbar, es als Frontend eines Document Retrieval Systems einzusetzen, wobei hier, ähnlich wie bei VisMeB, die detaillierte Analyse von Metadaten interessant wäre. Im Internet stehen unter der Adresse des Herstellers[5] einige Demoversionen bereit, mit deren Hilfe man sich einen ersten Eindruck des Systems machen kann. Auf den ersten Blick wirkt Infozoom recht verwirrend. Es wird eine Art Tabellendarstellung verwendet, auf der auf Anhieb jedoch nicht sehr viel zu erkennen ist (*Abbildung 6.1*). Wie der Name bereits suggeriert, besteht allerdings die Möglichkeit in die Tabelle hineinzuzoomen und somit die benötigten Informationen aufzuschlüsseln (*Abbildung 6.2*). In den Zeilen werden hierbei die verschiedenen, vorhandenen Attribute des jeweiligen Themenkomplexes angezeigt. Beispielsweise bei dem auf der Seite angebotenen Uhren Katalog Management Beispiel[6], der jeweilige Hersteller, Preis oder Typ. Die Spalten sind für jede Zeile einzeln definiert und entsprechen den möglichen Ausprägungen der Attribute. In diesem Fall sind das zum Beispiel alle Herstellernamen, die möglichen Preise oder vorhandenen Uhrentypen wie Herren- oder Damenuhr. Um zunächst eine Übersicht zu ermöglichen, werden die Ausprägungen kategorisiert und erst je nach Zoom Stufe vollständig einzeln sichtbar. Um die Navigation zu erleichtern, existieren, wie von Internetbrowsern bekannt, *Vor-* und *Zurück-Buttons*. Hinsichtlich einer möglichen Umsetzung als Metadatenbrowser könnten in den Zeilen die verschiedenen Meta-Attribute aufgereiht werden und in den Spalten die jeweiligen Ausprägungen.

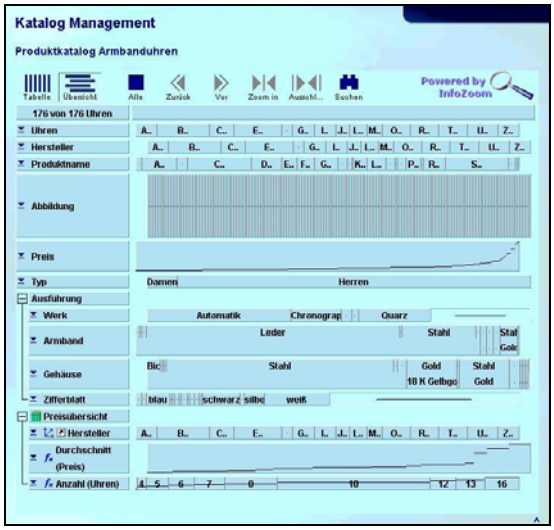


Abbildung 6.1: InfoZoom - Eingangsscreen

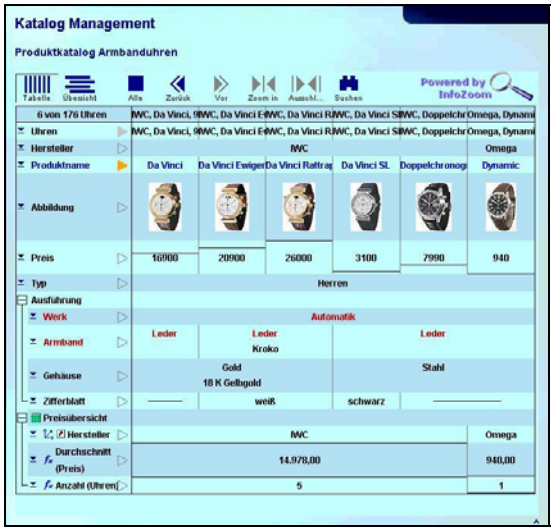


Abbildung 6.2: InfoZoom – Detailzoom

[5] *www.humanit.de*, online am 15.03.2004

[6] *http://www.humanit.de/de/produkte_loesungen/products/iz/beispiele/index.html*, online am 15.03.2004

Erfreulich ist die Tatsache, dass zu InfoZoom mehrere, zum Teil recht umfangreiche Evaluationen getätigt wurden und die Ergebnisse auch öffentlich zugänglich sind.

6.1.1 Callahan, Koenemann – InfoZoom vs. hierarchisches Katalogsystem [CK00]

Ewa Callahan von der Indiana University und Jürgen Koenemann von humanIT konzentrierten sich bei ihrer Evaluation aus dem Jahre 2000 auf den möglichen Einsatz von InfoZoom als „Online Produkt Katalog“, womit kurz gesagt das Katalogsystems eines Onlineshops gemeint ist. Herkömmliche Systeme, wie sie auch heute noch eingesetzt werden, basieren meistens auf Baumstrukturen und langen Listen der einzelnen Produkte, wobei ein Klick auf das jeweilige Produkt detaillierte Informationen ermöglicht. Weiterhin wird zumeist eine Suchfunktion angeboten, welche entweder Freitextsuche ermöglicht oder per Formular die Auswahl von Kriterien anbietet. Die Autoren haben im Vorfeld der Evaluation zunächst 16 dieser online verfügbaren Systeme untersucht und anschließend in Anlehnung an diese, eine für den Test letztendlich verwendete Version eines solchen hierarchischen Produktkatalogs entwickelt. Als Datenbasis für beide Systeme, also sowohl InfoZoom als auch den hierarchischen Produktkatalog, wurde eine Autodatenbank herangezogen, welche 1690 verschiedene Fahrzeuge mit jeweils über 20 verschiedenen Attributen beinhaltete.

I. Test Design

Die Auswahl der 26 Versuchspersonen erfolgte relativ zufällig mit Hilfe von Flyern. Grundvoraussetzung für die Teilnahme war eine gewisse Internet Erfahrung, welche sich allerdings auf die Aussagen der Teilnehmer stützte. Interessant für die späteren Ergebnisse ist sicherlich der Umstand, dass 16 der Teilnehmer von den Autoren als erfahrene Computerbenutzer eingestuft werden konnten, weitere sechs nutzten den Computer immer noch relativ häufig und wurden somit als durchschnittliche Benutzer eingestuft und lediglich vier wurden als unregelmäßige und somit unerfahrene Benutzer klassifiziert. Erfahrung mit Einkäufen im Internet hatten lediglich sieben der Teilnehmer und weitere sechs gaben an, per Internet sich schon einmal über Produkte informiert zu haben.

Für den Test wurden die Teilnehmer in zwei Gruppen aufgeteilt, wobei trotz zufälliger Zuteilung die eben angesprochenen Unterschiede in der Computernutzung in beiden Gruppen in etwa gleich verteilt auftraten. Gruppe 1 arbeitete während dem Test mit InfoZoom und Gruppe 2 mit dem hierarchischen Katalogsystem, demzufolge ein *Between Subjects Design* (siehe Kapitel 3.5). Die Tests fanden nicht in einem Usability Labor statt, sondern differierten in ihren Lokalitäten, abhängig von der jeweiligen Versuchsperson und deren Vorlieben – beispielsweise also an deren Arbeitsplatz oder zu Hause. Um den Nachteil einer doch recht ungewohnten Visualisierung im Falle von InfoZoom etwas auszugleichen, erhielten die Teilnehmer dieser Gruppe eine kurze Einführung in die Software. Es wurde

dabei jedoch ein anderer Datenbestand zur Demonstration verwendet und darüber hinaus darauf geachtet, dass dadurch keine direkten Lösungsvorschläge für die späteren Aufgaben vorgestellt wurden. Insgesamt wurden den Teilnehmern neun Aufgaben gestellt, welche das Finden von speziellen Produkten bei gegebenen Attributen, das Finden von Attributen zu gegebenen Produkten und das Vergleichen von Produkten bzw. von verschiedenen Attributen eines Produktes beinhalteten. In die Zusammenstellung und Auswahl der Aufgaben flossen dabei bereits durchgeführte Interviews mit potentiellen Benutzern und die Erfahrungen von den, im Vorfeld durchgeführten, Untersuchungen bekannter Online Katalogsysteme mit ein.

Während des Tests wurde nach jeder Aufgabe kurz unterbrochen und die Versuchspersonen mussten drei kurze Fragen hinsichtlich ihrer Zufriedenheit im Allgemeinen, ihres Eindrucks von der benötigten Zeit zur Lösung der Aufgabe und einer Einschätzung ihrer Effizienz, jeweils auf einer 7-Likert-Skala, beantworten.

II. Fazit Test Design

Mögliche Kritikpunkte sind die etwas zufällige Auswahl der Teilnehmer und die jeweils unterschiedlichen Umgebungsbedingungen, je nachdem wo der Test durchgeführt wurde. Auch wenn hier argumentiert werden kann, dass die Teilnehmer in der für sie gewohnten Umgebung im Allgemeinen leistungsfähiger und vielleicht weniger nervös sind, so kann auch genau dieser Umstand die Ergebnisse verfälschen, da manche Teilnehmer schnell dazu neigen, ein derartiges Experiment nicht mehr mit der notwendigen Ernsthaftigkeit zu betrachten und dadurch der Test zu informell gerät. Dazu könnten auch die Unterbrechungen nach jeder Aufgabe beitragen, wobei allerdings die hierbei gewonnenen zusätzlichen Daten nicht zu vernachlässigen sind und eine interessante Möglichkeit darstellen, die subjektive Einschätzung der Teilnehmer besser zu erfassen und in die Auswertung mit einfließen zu lassen.

III. Auswertung der Testergebnisse

Die gewonnenen Daten wurden mit Hilfe von T-Tests statistisch ausgewertet. Neben der jeweils benötigten Zeit wurde sowohl die Fehlerrate, als auch die zwischen den einzelnen Aufgaben, beziehungsweise am Ende mit Hilfe eines Post Test Fragebogen, subjektiv gewonnenen Daten derartig ausgewertet.

Bei der Auswertung konnte letztendlich festgestellt werden, dass die Benutzer von InfoZoom (Gruppe 1) zwar insgesamt gleich viele Aufgaben wie die Benutzer des hierarchischen Katalogsystems (Gruppe 2) falsch gelöst hatten, jedoch zumindest bei jeder zu einer Lösung kamen, wohingegen bei Gruppe 2 insgesamt neun Teilnehmer vorzeitig eine Aufgabe abbrechen mussten und somit weder eine richtige noch eine falsche Lösung angeben

konnten. Beide Gruppen hatten Probleme mit der fälschlichen Annahme, dass je nach Aufgabe die Daten schon richtig nach der gesuchten Variable sortiert seien. Weiterhin hatte die Gruppe 2 besonders mit Aufgaben zu kämpfen, bei denen Attribute sowohl mit einer unteren als auch einer oberen Schranke eingeschränkt werden mussten (z.B.: Preis von 5000 DM bis 10000 DM). Ebenfalls bereiteten Pulldown Menüs Probleme, bei welchen ein zusätzlicher Scrollbalken dazu diente, um alle Auswahlmöglichkeiten zu sehen. Die Teilnehmer der Gruppe 1 bevorzugten das Abzählen der einzelnen Produktbilder, anstelle der Produkt ID's, was insofern zu Problemen führte, da zum Teil das gleiche Bild für mehrere Produkte verwendet wurde.

Die Performance Auswertung zeigte recht deutlich, dass die InfoZoom Benutzer im Durchschnitt signifikant schneller (Konfidenzintervall 99%) waren, als die Benutzer des hierarchischen Katalogsystems. Eine Analyse der zwischen den Aufgaben gestellten Frage, inwieweit die Benutzer mit Ihrer Leistung zufrieden waren, ergab, dass diese mit der tatsächlichen benötigten Zeit korrelierte, sprich je länger ein Benutzer für eine Aufgabe benötigte, desto unzufriedener war er mit seiner Leistung – gleich ob die Aufgabe aufgrund des Umfanges grundsätzlich mehr Zeit zur Lösung benötigte. Dies trat sowohl bei der Visualisierung als auch bei dem hierarchischen Katalogsystem gleichermaßen auf.

IV. Fazit statistische Auswertung

Die statistische Auswertung wurde ebenfalls relativ umfangreich gestaltet, wobei die T-Test Auswertung der Likert Skalen nicht unproblematisch ist. Es ist durchaus umstritten, ob eine Likert Skala als metrisches Messniveau angesehen werden kann [BEPW00], da die Abstände zwischen den Auswahlmöglichkeiten von den Teilnehmern eventuell nicht als gleichwertig eingestuft werden. Die Praxis zeigt hier beispielsweise, dass Extremwerte oftmals gemieden werden. In diesem Fall wäre ein T-Test nicht möglich, da dieser zwingend metrisches Messniveau für die abhängige Variable voraussetzt.

Ansonsten ist die genaue Aufschlüsselung in abhängige und unabhängige Variable zu begrüßen. Eine Varianzanalyse wäre aufgrund ihrer größeren Flexibilität eventuell wünschenswerter gewesen, ist jedoch in diesem Fall, da jeweils nur eine abhängige und unabhängige Variable verwendet wurden, wobei letztere zudem auch nurzweiAusprägungen aufzuweisen hatte, nicht zwingend erforderlich.

Die Korrelation der tatsächlichen Zeit mit den persönlichen Empfindungen der Teilnehmer hätte durchaus größere Bedeutung zugemessen werden können: Daraus könnte sich beispielsweise die These aufstellen lassen, dass zumindest in diesem Fall die Visualisierung nicht oder nur begrenzt dazu beiträgt, den Nutzer auch bei langwierigen oder schwierigen Aufgaben besser zu motivieren und zufrieden zu stellen, was allerdings häufig auf Grund des proklamierten Joy of Use, den eine Visualisierung oftmals per se bieten soll, angenommen

wird. Eine gezielte Untersuchung dieses Umstandes wäre für die Zukunft in jedem Fall wünschenswert.

6.2 NIRVE

6.2.1 Sebrechts, Cugini – 2D vs. 3D vs. Text Retrieval Interface [SVMCL99]

Ziel dieser Evaluation aus dem Jahr 1999 war der Vergleich dreier verschiedener Visualisierungen. Zu diesem Zweck wurde das Visualisierungstool NIRVE untersucht, welches von dem National Institute of Standards & Technology[7] entwickelt wurde und mehrere verschiedene Darstellungsmöglichkeiten unterstützt, darunter eine dreidimensionale Sphärendarstellung, eine flache 2D Sicht und eine rein textuelle Darstellung. NIRVE visualisiert in der 3D Sphärenansicht eine Art Weltkugel, auf dessen Oberfläche einzelne Dokumenten *Cluster* angezeigt werden (*Abbildung 6.3*).

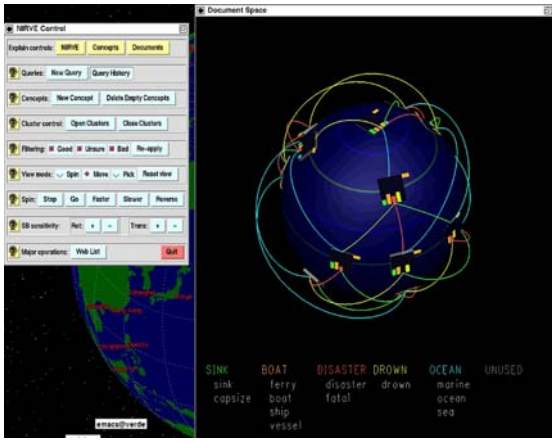


Abbildung 6.3: NIRVE 3D Sphere –Übersicht

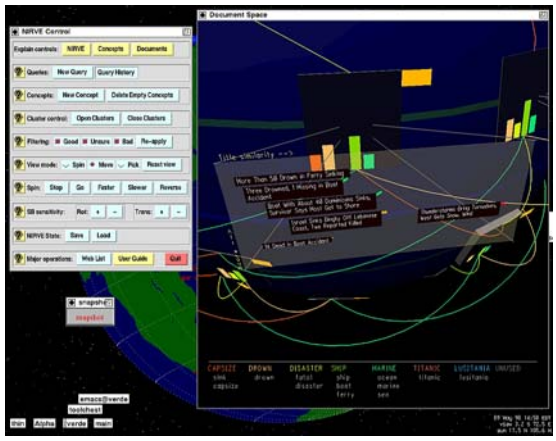


Abbildung 6.4: NIRVE 3D Sphere - Detailansicht

Am Nordpol befindet sich dabei der Dokumenten *Cluster*, welche zu allen eingegebenen *Concepts* relevant ist. Ein *Concept* ist dabei ein Zusammenschluss mehrerer Suchbegriffe, wobei jedes *Concept* eine bestimmte Farbe zugewiesen bekommt. Je weiter südlich man sich bewegt, desto weniger *Concepts* werden in den *Clustern* mit einbezogen, wobei somit letztendlich alle möglichen Kombinationen abgedeckt werden. Weiterhin sind die einzelnen *Cluster* noch mit Linien verbunden, welche jeweils die Farbe des *Concepts* annehmen, welches bei einem der beiden *Cluster* nicht vorhanden ist und diese somit unterscheidet. Der Benutzer hat die Möglichkeit, per Klick auf ein *Cluster*, sich dieses genauer anzusehen. Dazu wird der Inhalt des *Clusters* in einem zweidimensionalen Rechteck, welches im

[7] *NIST, http://www.nist.gov/, online am 15.03.2004*

Vordergrund erscheint, angezeigt (*Abbildung 6.4*). Die Dokumente sind dort nach Relevanz und Ähnlichkeit zueinander angeordnet. Der Inhalt eines einzelnen Dokuments wird in einem externen Webbrowser dargestellt. Zusätzlich zu dieser 3D Ansicht beinhaltet das Programm noch ein *Control Panel*, welches beispielsweise das Starten einer neuen Suche oder das Zusammenstellen eines neuen *Concepts* ermöglicht. Dieses *Control Panel* ist zudem auch in der 2D- und der textuellen Darstellung enthalten. In ersterer (*Abbildung 6.5*) wurde die dreidimensionale Sphäre einfach platt gedrückt und bietet somit alle *Cluster* auf einen Blick. Das generelle Prinzip wurde allerdings beibehalten. Die textuelle Darstellung (*Abbildung 6.6*) hingegen stellt sämtliche Ergebnisse als Liste dar. Dabei wurde das Prinzip der *Cluster* und *Concepts* jedoch ebenfalls beibehalten, was dazu führt, dass das *Cluster* mit allen möglichen *Concepts* zu oberst in der Liste zu finden ist. Die einzelnen Dokumente werden einfach innerhalb der *Cluster* aufgelistet. Zusammenhänge zwischen den Dokumenten oder den *Clustern* werden in dieser Ansicht nicht dargestellt.

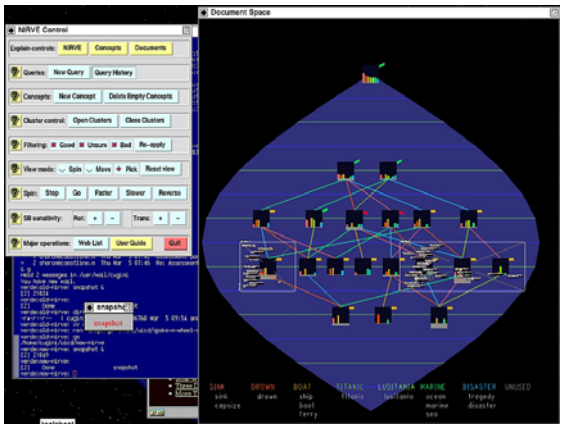


Abbildung 6.5: NIRVE 2D - Darstellung

Abbildung 6.6: NIRVE Text - Darstellung

I. Testdesign

Insgesamt nahmen 15 Versuchspersonen an dem Test teil. Neun davon wurden von den Autoren als Anfänger hinsichtlich ihrer Computer Erfahrung eingestuft und weitere sechs als Experten, welche bereits Erfahrungen sowohl mit grafischen Benutzeroberflächen als auch mit Retrieval Systemen aufweisen konnten. Die Teilnehmer wurden quasi-zufällig auf drei Versuchsgruppen verteilt, so dass in jeder davon drei Anfänger und zwei Experten vorhanden waren. Jede Gruppe arbeitete nur mit einer der Visualisierungen, also ebenfalls ein *Between Subjects Design*. Etwas ungewöhnlich ist die verwendete Hardware. Es kamen hier zwei verschiedene Systeme zum Einsatz. Die als Einsteiger eingestuften Versuchspersonen arbeiteten an einer Silicon Graphics Indy Workstation mit 100 MHz ohne hardwarebeschleunigte OpenGL Unterstützung. Den Experten hingegen wurde eine schnellere Silicon Graphics Onyx Workstation zur Verfügung gestellt, welche eine hardwarebeschleunigte OpenGL Unterstützung bietet und somit eine höhere Rechengeschwindigkeit bei Grafikanwendungen aufweist. Der Grund für diese Aufteilung

wird von den Autoren nicht genannt. Die Hauptunterschiede zwischen beiden Systemen zeigten sich naturgemäß bei der Verwendung der 3D Oberfläche.

Der Test selbst wurde sehr umfangreich gestaltet. Jede Versuchsperson musste zu insgesamt sechs Themengebieten jeweils 16 Aufgaben bewältigen. Um hier in einem vernünftigen Zeitrahmen zu bleiben, wurden an drei aufeinander folgenden Tagen jeweils zwei Themengebiete abgedeckt, wobei für ein Gebiet in etwa 45 - 75 Minuten benötigt wurden. Die Testaufgaben umfassten beispielsweise das Suchen eines bestimmten Dokumentes, von welchem der Titel oder das Thema bekannt war, weiterhin auch das Suchen nach ähnlichen Dokumenten oder nach Clustern, welche bestimmte Themengebiete behandelten. Der Versuchsleiter erklärte jedem Teilnehmer zunächst die grundlegenden Funktionen anhand von ausgewählten Beispielen. Darüber hinaus wurde die Bearbeitung des ersten Themengebiets als Trainings Session genutzt, in welcher die Teilnehmer Fragen stellen konnten und der Versuchsleiter konkrete Hilfestellung leistete. Der Versuchsleiter gab auch bei jedem Test die festgelegten Suchanfragen ein, auf welche die Teilnehmer keinen Einfluss hatten.

Während dem Test wurden die Aktivitäten der Versuchspersonen auf Video und Audio aufgezeichnet. Zudem wurden die Teilnehmer gebeten ihre Gedanken fortlaufend laut zu äußern (*Thinking Aloud*). Für jede Aufgabe wurde ein Zeitfenster von drei Minuten zur Verfügung gestellt. Falls die Aufgabe bis dahin nicht gelöst werden konnte, wurde seitens des Versuchsleiters ein Tipp gegeben und eine zusätzliche Minute Zeit. Zwischen jeder Aufgabe wurde den Teilnehmern die Möglichkeit zu einer fünfminütigen Pause gewährt.

II. Fazit Test Design

Positiv hinsichtlich des Test Designs ist sicherlich der große Umfang von sechs umfangreichen Testdurchläufen mit verschiedenen Themenkomplexen zu bewerten. Ebenfalls die dadurch mögliche Trainingssession ist methodisch sehr zu begrüßen. Allerdings offenbaren sich auch einige Schwächen. An erster Stelle sei hier die geringe Anzahl an Teilnehmern genannt. Insbesondere die Tatsache, dass aufgrund des *Between Subjects Designs* drei getrennte Versuchsgruppen gebildet wurden, lassen 15 Teilnehmer und damit fünf pro Gruppe für einen Performance Test doch sehr wenig erscheinen. Weiterhin ist die Aufteilung der Teilnehmer auf die beiden zur Verfügung stehenden Rechner strittig zu sehen – hier hätte eher dafür gesorgt werden müssen, dass die drei Gruppen gleichmäßig auf beide Systeme verteilt werden oder nur die Gruppe, welche die textbasierte Darstellung testete, mit dem schwächeren System arbeitet. Das explizite Auffordern zum *Thinking Aloud* ist ebenfalls nicht ganz unproblematisch bei einem Experiment, welches hauptsächlich die benötigte Zeit in den Vordergrund stellt. Tests im Umfang der Evaluation von VisMeB, welche in Kapitel 7 ausführlich vorgestellt wird, zeigten recht deutlich, dass manche

Benutzer so viel zu erzählen haben, dass ihre Leistung dadurch deutlich geschmälert wird. Das vorgegebene Drei-Minuten Zeitlimit soll nach Aussage der Autoren dazu beitragen, dass das Hauptaugenmerk wirklich auf der Effizienz der Lösung der Aufgaben liegt. Dies steht zum einen im Gegensatz zu dem vorher geforderten *Thinking Aloud*, und zum anderen könnte dadurch durchaus ein gewisser Stressfaktor für die Teilnehmer entstehen, welcher eigentlich in jedem Fall zu vermeiden ist, da dieser eine zusätzliche Störvariable darstellt, welche die Ergebnisse verzerren kann.

III. Auswertung der Testergebnisse

Aufgrund einiger Probleme bei der Bearbeitung von drei der jeweils 16 Testaufgaben, wurden diese bei der Auswertung nicht berücksichtigt.

Die statistische Auswertung der quantitativen Daten erfolgte durch eine Varianzanalyse. Als abhängige Variable diente die benötigte Zeit, wobei jedoch verschiedene Sichtweisen angeführt wurden. Zunächst wurde die Gesamtperformance der drei Visualisierungen verglichen, wobei sich herausstellte, dass die textbasierte Darstellung signifikant schneller war als die 2D beziehungsweise die 3D Darstellung (Konfidenzintervall 95%). Die Autoren begründen diese Tatsache mit der möglichen Vertrautheit der Benutzer mit einer derartigen Darstellung. Dass diese These durchaus zutreffen könnte, zeigt die genauere Betrachtung hinsichtlich der Lerneffekte. Da jeder Benutzer insgesamt sechs recht umfangreiche Testsessions absolvierte, können hier durchaus Rückschlüsse gezogen werden. So zeigt sich, dass die 3D Visualisierung im Laufe des Experiments den größten Performance Zuwachs verzeichnen konnte. Die 2D Visualisierung konnte sich ebenfalls verbessern und verdrängte bei den letzten beiden Sessions die textbasierte Lösung, welche sogar leichte Einbußen zu verzeichnen hatte, von dem ersten Platz. Die Autoren gaben allerdings nicht an, ob dieser Vorsprung signifikant ist. Ein weiteres Indiz für die Gewöhnungsbedürftigkeit des 3D Interfaces sehen die Autoren darin, dass nach der ersten Session, welche wie bereits erwähnt als Testsession fungierte, bei dieser Visualisierung die benötigte Zeit zunächst anstieg, wohingegen die Benutzer des 2D Interface und der textbasierten Lösung sofort deutliche Verbesserungen hinsichtlich der Performance zeigten.

Weiterhin wurde untersucht, ob innerhalb der beiden Gruppen der Experten und Anfänger Unterschiede zwischen 2D und 3D Visualisierung zu erkennen waren. Bei ersteren konnte ein solcher nicht festgestellt werden, wohingegen die Anfänger bei der Benutzung der 3D Visualisierung zwar langsamer, aber nicht signifikant langsamer waren. Einen möglichen Grund für diesen Unterschied sehen die Autoren in der Verwendung der langsameren Rechner bei der Gruppe der Anfänger, worunter hauptsächlich die Performance der 3D Visualisierung zu leiden hatte.

Als letzten Punkt der quantitativen Analyse untersuchten die Autoren die unterschiedlichen Aufgabentypen. Aus den insgesamt verbliebenen 13 Aufgaben wurden acht verschiedene Kategorien gebildet und die jeweiligen Zeiten zwischen den Visualisierungen wieder per Varianzanalyse untersucht. Es stellte sich dabei heraus, dass einige Aufgabentypen klar eine bestimmte Visualisierung favorisierten und die dort gemessenen Unterschiede zum Teil auch signifikant waren.

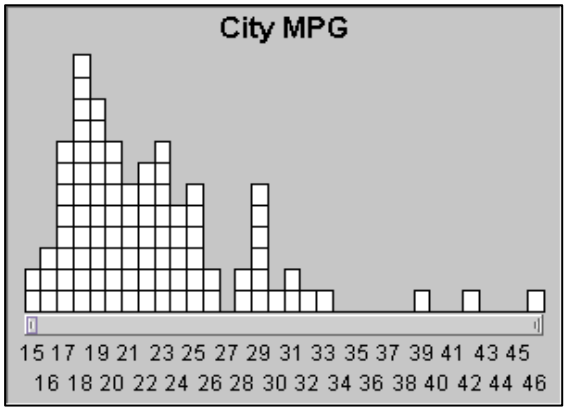
Die qualitative Auswertung konzentrierte sich auf spezielle Eigenschaften der Visualisierungen, leider ohne generalisierende Aussagen bezüglich der Stärken und Schwächen sowie der Akzeptanz dieser drei verschiedenen Visualisierungen zu tätigen. Die Autoren ziehen allerdings auch ein ähnliches Fazit - die Evaluation orientiere sich sehr an den speziellen Eigenschaften der Visualisierungen und für generalisierende Aussagen sei eine Studie, die sich mehr auf die unterschiedlichen Prinzipien solcher Visualisierungen konzentriert, von Nöten. Allerdings kommen sie auch zu dem Schluss, dass die Vor- und Nachteile jeweils sehr stark von den Aufgabentypen, den Benutzern und auch von der verwendeten Hardware abhängen, dass also zumindest in dieser Studie keine der Visualisierungen sich wirklich herauskristallisieren konnte. Die Vorteile der textbasierten Lösung könnten beispielsweise bei größeren Dokumentenmengen (im Test waren diese auf 100 beschränkt) deutlich geringer ausfallen, da sich hier ein einfaches Durchscrollen als ineffizient erweisen könnte.

IV. Fazit Auswertung

Ein sehr positiver Aspekt bei der statistischen Auswertung ist die recht umfangreiche Betrachtung der Performance Ergebnisse, welche weit über eine Analyse der Gesamtdauer hinausgeht. Insbesondere wurde das umfangreiche Testdesign mit mehreren Testsessions optimal ausgenutzt um auch Lerneffekte genauer zu betrachten. Da auch zusätzlich die Daten hinsichtlich der beiden verwendeten Rechnerkonfigurationen untersucht wurden, konnten die Autoren selbst den Schluss ziehen, dass hier das schwächere System die Daten bei der 3D Visualisierung verfälscht haben könnte. Die Aufspaltung in verschiedene Aufgabentypen ermöglichte den Autoren ebenfalls die interessante Schlussfolgerung, dass jede Visualisierung Vor- und Nachteile bietet, was bei einer reinen Betrachtung der Gesamtzeiten, nicht möglich gewesen wäre. Die verwendete Varianzanalyse zeigt statistisches Verständnis, da aufgrund der Tatsache, dass die unabhängige Variable (Typ der Visualisierung) drei Ausprägungen aufwies, die Kombination mehrerer T-Tests zwar ebenfalls als Analyse möglich gewesen wäre, als Folge daraus aber, wie in der statistischen Einführung kurz angesprochen (Kapitel 3.1), die Fehlerwahrscheinlichkeit sehr stark angestiegen wäre. Inwieweit die Post-Test Fragebögen überhaupt ausgewertet wurden, geht aus dem Evaluationsreport leider nicht hervor.

6.3 Attribute Explorer

Der Attribute Explorer wurde von Robert Spence [STWB94] entwickelt und kann wohl als, zumindest auf den ersten Blick, recht ungewöhnliche Visualisierung betrachtet werden. Sie ermöglicht die Suche und Auswahl aufgrund der Einschränkung gewisser Attribute. Beispielsweise könnten beim Kauf eines Autos die Attribute Hubraum, Motorleistung und Verbrauch eine kaufentscheidende Rolle spielen. Ein formularbasierter Ansatz würde dem Benutzer ermöglichen, per Minimum- und Maximum-Felder hier Filter zu setzen. Allerdings hat dies oftmals den Nachteil, dass der Benutzer nicht sofort sieht, wie sich das Ändern, beispielsweise des Maximalpreises, auf die Treffermenge auswirkt. Der Attribute Explorer versucht dieser Problematik gerecht zu werden. Für jedes Attribut wird ein Histogramm dargestellt, auf welchem auf der X-Achse die verschiedenen Ausprägungen des jeweiligen Attributes und auf der Y-Achse die einzelnen Objekte, also beispielsweise die Autos, abgebildet sind.
In *Abbildung 6.7* wird auf der X-Achse das Attribut *Treibstoffverbrauch im Stadtverkehr* (Meilen pro Gallone) angegeben.



Mittels Schieberegler kann der Benutzer auf der X-Achse direkt manipulierend eingreifen und beispielsweise den maximalen Verbrauch einschränken. Die Objekte, welche nun außerhalb dieser Beschränkung liegen, werden farblich markiert. Die eigentliche Idee besteht nun darin, dass mehrere dieser Histogramme, intern miteinander verbunden sind.

Abbildung 6.7: Attribute Explorer Einzel-Histogramm

Wird nun also ein Attribut eingeschränkt, werden in allen Histogrammen die Objekte farblich markiert, welche diese eine Eigenschaft nicht erfüllen (*Abbildung 6.8*). Durch die Einschränkung mehrerer Attribute kann somit die in Frage kommende Ergebnismenge schnell und präzise eingeschränkt werden (*Abbildung 6.9*). Vor allen Dingen kann problemlos untersucht werden, inwieweit sich kleine Änderungen auf die Anzahl der in Frage kommenden Objekte auswirkt. Das bei Formularanfragen, bei welchen der Nutzer zunächst viele Attribute einschränken kann, häufig auftretende Problem der 0 Treffer Ergebnisse kann somit weitestgehend vermieden werden. Weiterhin besteht die Möglichkeit, dass Objekte, welche beispielsweise nur eines der geforderten Attribute nicht erfüllen, farblich noch einmal abgegrenzt dargestellt werden. Dies ist natürlich auf quasi beliebig viele Zwischenstufen (allerdings kleiner der Attributanzahl) erweiterbar.

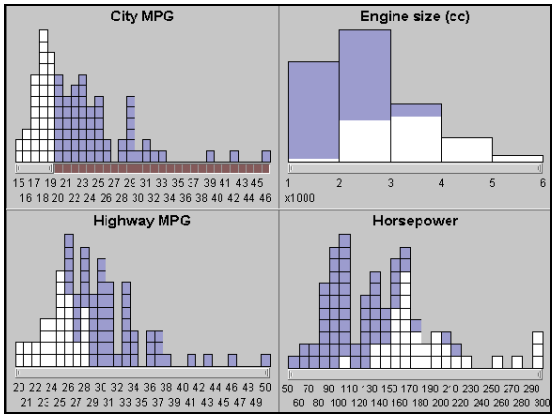


Abbildung 6.8: ein Attribut eingeschränkt (City MPG)

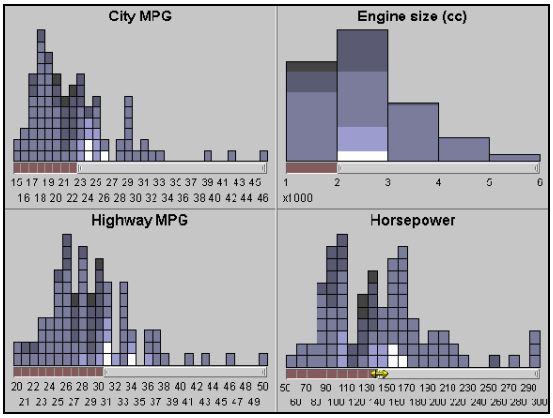


Abbildung 6.9: mehrere Attribute eingeschränkt

6.3.1 English, Garret & Pearson - Travellite Evaluation [EGP01]

Travellite ist eine webbasierte Anwendung zur Erstellung individueller Reiseführer. Es wurde im Rahmen einer Masterarbeit an der Berkeley University entwickelt. Der Benutzer hat hierbei die Möglichkeit aus einer Vielzahl von Reisemerkmalen, angefangen bei dem Ziel der Reise, über das Hotel bis zu weiteren Details wie Mietwagenservice, die für ihn relevanten auszuwählen und einzugrenzen, um am Ende einen persönlichen Reiseführer zu erhalten, welcher beispielsweise auf einen Palm herunter geladen werden kann. Dabei wurde prinzipiell ein formularbasierter Ansatz verwendet. Jedoch zogen die Autoren auch eine Visualisierung in Form des Attribute Explorers in Betracht, welcher mittlerweile von IBM weiterentwickelt und vertrieben wird. Um den Nutzen eines möglichen Einsatzes zu überprüfen, wurde im Jahr 2001 ein Performance Test durchgeführt, bei welchem sich der Attribute Explorer zum einen gegen ein formularbasiertes Interface und zum anderen gegen ein so genanntes *dynamic query interface* behaupten musste. Dieses zeigt dem Nutzer, im Gegensatz zu Ersterem, die Auswirkungen von Änderungen an der Attributauswahl hinsichtlich der Gesamttreffer, unmittelbar auf derselben Seite an.

I. Testdesign

Insgesamt nahmen an dem Experiment 13 Versuchspersonen teil, wobei lediglich 12 davon später in die Auswertung mit einflossen. Diese setzten sich jeweils zur Hälfte aus Männern und Frauen zusammen und wurden aus Studenten und Absolventen der UC Berkeley ausgewählt. Hinsichtlich des Umgangs mit Visualisierungen gaben dabei drei Teilnehmer an, nicht zu wissen was eine Visualisierungs-Software sei, sechs konnten bereits Erfahrungen aufweisen und zwei konnten sich zumindest etwas darunter vorstellen.

Für den Test wurde eine Restaurant Datenbank verwendet, in welcher die Versuchspersonen mit Hilfe des jeweiligen Interfaces das passende Restaurant anhand vorgegebener Kriterien finden sollten. Insgesamt musste dabei jeder Teilnehmer neun Aufgaben bewältigen, drei pro Interface, was auch bedeutet, dass jeder Teilnehmer mit allen Darstellungen arbeiten musste

(*Within Subject Design*). Dabei wurde bei jeder Darstellung jeweils die Komplexität der Aufgaben schrittweise erhöht – zunächst drei Attribute, dann sechs und zuletzt neun Attribute. Die Autoren erstellten für jeden Komplexitätslevel drei sehr ähnliche, aber in Hinblick auf das jeweilige Ergebnis unterschiedliche, Aufgaben. Diese wurden dann randomisiert und den Versuchspersonen zugeteilt, so dass letztendlich jeder Teilnehmer alle neun Aufgaben bewältigen musste, durch das Variieren der Reihenfolge eventuelle Lerneffekte aber minimiert werden sollten. Als letztes mussten die Teilnehmer noch einen vierten Aufgabentyp bearbeiten, welcher nicht direkt die einzuschränkenden Attribute vorgab sondern freies Arbeiten mit der jeweiligen Darstellung ermöglichte.

Zusätzlich wurden die Versuchspersonen in zwei Gruppen aufgeteilt, wobei eine Gruppe die Möglichkeit hatte, sich längere Zeit mit dem Attribute Explorer vertraut zu machen, bevor der Test begann. Hierdurch sollte überprüft werden, ob eine längere Eingewöhnungszeit einen positiven Effekt auf die Ergebnisse hat. Zwischen den einzelnen Testaufgaben wurden die Teilnehmer jeweils kurz befragt, inwieweit sie Vertrauen in ihre Lösung der Aufgabe hinsichtlich der Korrektheit hätten. Mit Hilfe eines Post Test Fragebogen wurden weitere subjektive Ergebnisse gewonnen.

II. Fazit Testdesign

Der größte Schwachpunkt des Test Designs ist sicherlich die mangelhaften Ausführungen in dem veröffentlichten Evaluationsreport. Es geht beispielsweise nur sehr bedingt daraus hervor, wie die Aufgaben wirklich ausgesehen haben und wie die genaue Zuordnung geschah. Der genaue Testablauf wurde ebenso wenig festgehalten und auch der Testaufbau ist nicht näher erklärt. Inwieweit die Randomisierung der Aufgaben die auftretenden Lerneffekte eines W*ithin Subject Designs* ausgleichen konnten, ist auch nur zu erahnen, da aus dem Text nur sehr unklar hervorgeht, wie diese Randomisierung genau vonstatten ging. Weiterhin erscheinen drei Aufgaben pro Interface doch recht wenig, um hier aussagekräftige Ergebnisse erzielen zu können. Durch die zusätzliche Aufteilung der Teilnehmer in zwei Gruppen (mit Training und ohne Training) wurde von vorneherein das Ergebnis wiederum abgeschwächt, da sich dadurch letztendlich nur die Daten von jeweils sechs Teilnehmern des Attribute Explorers mit den jeweils 12 Datensätzen der beiden formularbasierten Darstellungen vergleichen ließen. Darüber hinaus kann eine kurze Schnupperphase, bei welcher der Teilnehmer die Länge auch noch selbst bestimmen konnte und bei welcher keinerlei Einführung eines Versuchsleiters geschah kaum als Trainingsphase betrachtet werden. Hier wäre eine Beschränkung auf eine der beiden Gruppen sinnvoller gewesen.

III. Auswertung der Ergebnisse

Die Autoren geben nicht an, mit Hilfe welcher statistischer Analysemethode die drei Darstellungen hinsichtlich der benötigten Zeit zur Lösung der Aufgaben verglichen wurden.

Allerdings wurde herausgefunden, dass die benötigte Zeit bei der Arbeit mit dem Attribute Explorer doppelt so lang war, wie bei den beiden formularbasierten Darstellungen und dieser Unterschied auch signifikant sei. Weiterhin schien dieser Unterschied auch noch größer zu werden, je höher die Aufgabenkomplexität war. Zwischen den beiden Gruppen konnten hinsichtlich des Trainings keine signifikanten Unterschiede festgestellt werden – der Attribute Explorer scheint also nur bedingt durch längere Einarbeitungszeit zu profitieren. Je länger die Teilnehmer aber mit dem Attribute Explorer arbeiteten, desto mehr Vertrauten hatten sie in ihre eigenen Ergebnisse. Allerdings ist auch dieses Vertrauen im Vergleich zu den beiden formularbasierten Interfaces geringer, wobei die Autoren nicht angeben, mit Hilfe welcher Skala dieses subjektive Empfinden gemessen wurde.

Zusätzlich zu diesen Performance Messungen wurde auch noch die *Precision* (Anzahl relevanter Treffer/Anzahl aller Treffer) und der *Recall* (Anzahl relevanter Treffer/all in der Datenbank vorhandenen relevanten Treffer) gemessen, wobei die Autoren hier entdeckten, dass beide Werte bei Verwendung des Attribute Explorers 13% niedriger ausfielen. Zu berücksichtigen ist hier allerdings, dass diese Unterschiede nicht signifikant waren. Darüber hinaus bemerken die Autoren, dass aufgrund der Aufgabenstellungen *Recall* und *Precision* eventuell nicht so aussagekräftig seien, weswegen noch die absolute Fehleranzahl verglichen wurde. Bei Verwendung des Attribute Explorers traten insgesamt acht Fälle auf, bei denen der jeweilige Teilnehmer die Aufgabe nicht korrekt lösen konnte – bei den beiden formularbasierten Lösungen nur jeweils 3.

Die qualitative Auswertung des Testablaufs und der Post-Test Fragebögen ergab, dass die meisten Versuchspersonen von dem Attribute Explorer zunächst sehr verwirrt waren und mit einigen Eigenschaften wie Farbgebung und Bedienung nur bedingt zu recht kamen. Letztendlich schnitt sowohl hier, als auch bei den Performance Messungen das *dynamic query interface* am Besten ab. Dementsprechend kommen die Autoren zu dem Schluss, dass die Integration des Attribute Explorers nur als zusätzliche Visualisierung Sinn machen könnte und ansonsten das *dynamic query interface* zu verwenden sei.

IV. Fazit der statistischen Auswertung

Wie bereits zuvor, liegt der größte Kritikpunkt in der veröffentlichten Studie. Es geht nicht einmal daraus hervor, mit Hilfe welcher statistischer Analysemethode die Systeme verglichen wurden. Weiterhin werden weder genaue Zeiten noch elementare Dinge wie das zu Grunde liegende Signifikanzniveau oder Standardabweichungen genannt. Die Autoren geben zwar Veränderungen der Varianz an, jedoch werden auch hier keinerlei Zahlen oder wenigstens Dimensionen genannt. Die Betrachtung von *Precision* und *Recall* macht in diesem Zusammenhang ebenfalls keinen Sinn, wie die Autoren letztendlich auch selbst zugestehen, was durch den Hinweis, dass eine reine Fehleranalyse wohl aussagekräftiger sei,

deutlich wird. Nichtsdestotrotz wird im Fazit ein weiteres Mal auf die Unterschiede bei diesen beiden Messgrößen hingewiesen. Als kurzer Hinweis soll hier nur angemerkt sein, dass *Precision* und *Recall* als Messgrößen bei der Bewertung von Retrieval Systemen verwendet werden. Da die Aufgaben jedoch sehr restriktiv mit genauen Attributsvorgaben gestellt wurden, ist das hier mitnichten der Fall. Vielmehr sind niedrige *Precision* oder *Recall* Werte hier nur auf Fehler der Benutzer zurückzuführen.

Letztendlich bleibt festzuhalten, dass die Ergebnisse dieser Evaluation nur sehr kritisch betrachtet werden können, da die Methodik, soweit sie beschrieben ist, deutliche Schwächen offenbart. Und auch der gedankenlose Umgang mit Messgrößen wie *Precision* und *Recall* wertet die Ergebnisse letztendlich ab, vor allen Dingen da die statistische Analysemethode für die Auswertung der Zeiten nicht einmal angegeben wurde.

6.4 DEViD [Eibl00]

DEViD wurde in den Jahren 1996-1999 von Maximilian Eibl entwickelt und im Rahmen seiner Dissertation mit dem Titel *Visualisierung im Document Retrieval* präsentiert. Im Gegensatz zu vielen anderen Visualisierungen hat sich M. Eibl auch intensiv mit dem Thema Retrieval Algorithmen beschäftigt und seine Visualisierung nicht nur als Frontend für bestehende Retrieval Systeme entwickelt, sondern als eigenständiges, professionelles Document Retrieval System.

DEViD ermöglicht es dem Benutzer, mehrere Suchterme einzugeben und jegliche mögliche boolesche Kombinationen dieser zu erkennen. Wie in *Abbildung 6.10* zu sehen, ist die Visualisierung horizontal aufgebaut. Jeder Suchterm erhält dabei ein winkelförmiges Symbol in einer entsprechenden Farbe. Die Eingabe der Suchterme erfolgt am linken Bildschirmrand. Sobald diese getätigt ist, breitet sich die Visualisierung nach rechts aus, wobei die verschiedenen Winkel, je nach Kombination der Suchterme, untereinander gereiht werden, bis alle möglichen Kombinationen erreicht sind – also auf Stufe 2 beispielsweise jeweils zwei Suchterme und auf der nächsten Stufe jeweils drei (*Abbildung 6.11*). Ganz rechts findet sich die Kombination aller Suchterme wieder. Die kleine Zahl gibt dabei jeweils an, wie viele Dokumente zu dieser Kombination gefunden wurden.

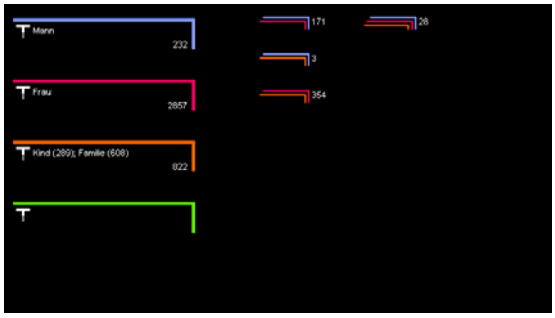


Abbildung 6.10: DEViD mit 3 Suchtermen

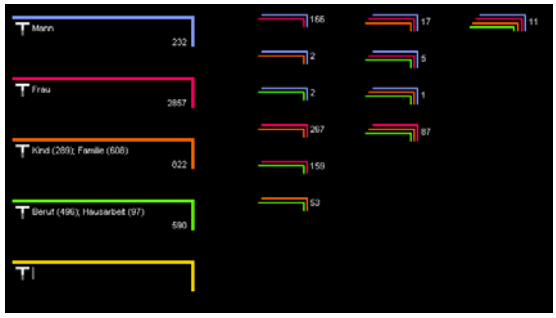


Abbildung 6.11: DEViD mit 4 Suchtermen

Weiterhin wird auch noch probabilistisches Retrieval, also quasi ein Ranking System, und die Erweiterung der Treffermenge mit Hilfe von vagem Retrieval unterstützt.

I. Testdesign

Das Testdesign der DEViD Evaluation ist wohl in dem Bereich der „Informationsvisualisierungs-Werktzeuge“ bislang einzigartig. Um dem eigenen Anspruch gerecht zu werden, nicht nur eine schöne Visualisierung, sondern auch ein schlagkräftiges Document Retrieval System zu sein, wurde auf das am Informations-Zentrum (IZ) Sozialwissenschaften in Bonn entwickelte GIRT (German Indexing and Retrieval Test Database) zurückgegriffen, welches seinerseits quasi das deutsche Gegenstück zur TREC Initiative (Test Retrieval Conference) des NIST (National Institute for Standards and Technology) darstellt. Hierbei werden standardisierte Retrieval Tests zur Verfügung gestellt, um die Vergleichbarkeit verschiedener Systeme zu ermöglichen, ohne bei jedem Test sämtliche Systeme neu testen zu müssen. GIRT bietet also beispielsweise eine standardisierte Datenbank, von externen Experten überprüfte und standardisierte Testaufgaben und einen standardisierten Versuchsablauf. Die oftmals existierende Gefahr, einen Test zu entwickeln, welcher die Fähigkeiten der eigenen Visualisierung besonders hervorhebt, konnte dadurch nahezu ausgeschlossen werden. Zum Zeitpunkt der Evaluation wurden im Rahmen des GIRT Projekts bereits zwei professionelle textbasierte Retrieval Systeme getestet. Zum einen das probabilistische Retrieval System freeWAIS und zum anderen das Freitextrecherchesystem Messenger. Als statistische Messgrößen dienten die Kenngrößen Precision und Recall, welche zur Bewertung von professionellen Information Retrieval Systemen standardgemäß Verwendung finden. Die benötigte Zeit für die einzelnen Aufgaben wurde zwar gemessen, da dies jedoch nicht Bestand der GIRT Spezifikationen ist und somit in dieser Hinsicht keine Daten von den beiden, zum Vergleich herangezogenen, Systemen vorlagen, konnte hier keine weitere statistische Analyse durchgeführt werden. An dem Test nahmen acht Versuchspersonen teil, ausnahmslos Sozialwissenschaftler am IZ Sozialwissenschaften in Bonn. Diese konnten als Retrievalexperten eingestuft werden und entsprachen somit der Testgruppe der früheren GIRT Tests.

Insgesamt wurden von allen acht Teilnehmern neun Testaufgaben bearbeitet, wobei davon die erste Aufgabenstellung innerhalb der Einführung als Probelauf diente, um den Teilnehmern die Möglichkeit zu geben, zunächst die Visualisierung kennen zu lernen. Die verbliebenen acht Aufgabenstellungen wurden dann ohne Eingriff des Testleiters bearbeitet, wobei die Reihenfolge variiert wurde, um einen Einfluss der Aufgabenreihenfolge auszuschließen. Jeder Testdurchlauf wurde sowohl handschriftlich protokolliert, als auch auf Video aufgezeichnet. Die Teilnehmer wurden dazu aufgefordert, ihre Gedanken während des Tests laut zu äußern. Nach Beendigung der Aufgaben wurde von jedem Teilnehmer noch ein Post-Test Fragebogen ausgefüllt, wobei hier sehr stark auf Anonymität Wert gelegt wurdet,

da die Teilnehmer, ebenso wie der Autor am IZ tätig waren und somit eventuelle „persönlichkeitsbezogene Verschiebungen im Namen der Kollegialität“ bei den Antworten vermieden werden sollten. Letztendlich dauerte ein Testdurchlauf für jeden Teilnehmer im Durchschnitt 3 1/4 Stunden.

II. Testergebnisse

Die Auswertung der Daten wurde sehr umfangreich gestaltet. Eine genauere Betrachtung würde zunächst eine umfangreiche Einführung in die Kenngrößen Recall und Precision erfordern, was an dieser Stelle zu weit führen würde. Zusammengefasst konnte sich DEViD allerdings erstaunlich gut gegen die klassischen und von den Teilnehmern täglich Benutzten Konkurrenten behaupten. Insgesamt kommt Maximilian Eibl sogar zu dem Schluss, dass die Teilnehmer mit Hilfe der Visualisierung qualitativ bessere Ergebnisse als mit freeWAIS und Messenger erzielten. Durch den zusätzlichen Einsatz des probabilistischen Retrievals (Ranking) konnte dies noch weiter gesteigert werden. Die Teilnehmer waren von der Darstellung des Systems ebenfalls sehr angetan und kamen größtenteils, trotz der nur kurzen Eingewöhnungszeit, gut damit zu recht. Rein subjektiv hatten zwei der Teilnehmer das Gefühl, schneller als mit den textbasierten Systemen arbeiten zu können, wobei zwei weitere hier eher eine gegensätzliche Meinung vertraten. Allerdings gaben alle Teilnehmer an, dass mit weiterer Übung die Performanz sicherlich deutlich verbessert werden könnte.

III. Fazit DEViD Testdesign und Testauswertung

Der hier getätigte Aufwand genügt sicherlich den höchsten Ansprüchen an die Evaluation eines professionellen *Document Retrieval Systems*. Die geringe Anzahl an Versuchspersonen wird letztendlich durch die Beschränkung auf Experten aufgewogen, da zwischen diesen die Unterschiede in der Leistung zumeist nicht so gravierend sind, wie bei einem breiter gefächerten Benutzerkreis. Weiterhin ist ein Test in diesem Umfang auch nur schwer mit mehr Teilnehmern möglich. Hinsichtlich der Leistungsfähigkeit der eigentlichen Visualisierung muss jedoch gesagt werden, dass das Fehlen von reinen Performancevergleichen doch zu bemängeln ist. Da im Rahmen der Evaluation allerdings die Konkurrenzsysteme nicht getestet wurden, sondern auf die Daten bereits getätigter Evaluationen im Rahmen des GIRT Projekts zurückgegriffen wurde, musste dieser Kompromiss seitens des Autors wohl eingegangen werden. Die subjektiven Empfindungen hinsichtlich der Leistung lassen aber zumindest darauf schließen, dass die Visualisierung nicht gravierend langsamer oder schneller zum Ergebnis führt. Das Herausragende dieser Evaluation ist aber sicherlich, dass durch die Berechnung von Precision und Recall wirklich die Güte der Ergebnisse gemessen wurde und somit mit den Testaufgaben ein Realeinsatz simuliert wurde. Dies ist bei den überwiegend unabhängig von der Datenbasis oder dem Retrievalsystem entwickelten Visualisierungen oft nicht möglich, da die Datenbasis und das Retrievalsystem qualitativ sehr hochwertig sein müssen, um hier sinnvolle Ergebnisse zu

erhalten. Durch die Kooperation mit dem GIRT Projekt war das in diesem Fall gegeben. Während also bei anderen Evaluationen oftmals noch Zweifel über die Leistungsfähigkeit der Visualisierung im realen Einsatz zurückbleiben, kann DEViD wohl gesichert als konkurrenzfähiges Produkt gelten.

Aufgrund des Umfanges der Evaluation konnte hier nur ein kleiner Einblick ermöglicht werden, für Interessierte lohnt sich in jedem Fall der Blick in die eingangs erwähnte Arbeit von Maximilian Eibl, in welcher sämtliche Details hierzu zu finden sind.

6.5 Zusammenfassung: Analyse der Methodik ausgewählter Evaluationen

Die Betrachtung der vier Evaluationen hat gezeigt, dass doch deutliche Unterschiede bezüglich der Methodik und auch des generellen Ansatzes einer derartigen Evaluation existieren. Zu begrüßen ist allerdings, dass zumindest drei der Evaluierungen statistisch valide Ergebnisse hervorgebracht haben. Kleinere Ungereimtheiten, wie die Auswertung von Likert-Skalen mit Hilfe von T-Tests schleichen sich aber auch hier ein. Weiterhin ist das insgesamt eher positive Abschneiden der Visualisierungen im Allgemeinen zu begrüßen.

Ein Aspekt welcher auch bei der Zusammenstellung dieser Evaluationen festgestellt wurde, ist die geringe Verfügbarkeit von öffentlich zugänglichen Evaluationen. Dies wurde auch fast durchgängig von den jeweiligen Autoren bemängelt. Somit ist es auch wenig verwunderlich, dass die Testdesigns sich zum Teil doch deutlich unterscheiden, da sich letztendlich noch kein „Industriestandard“ gebildet hat. Den Weg, den Maximilian Eibl mit Hilfe des GIRT Projekts hier eingeschlagen hat, wäre beispielsweise ein, zumindest für *Document Retrieval Systeme*, möglicher Lösungsansatz dieser Problematik. Für die Zukunft wäre es in jedem Fall wünschenswert, wenn hier einheitliche Konzepte entwickelt werden würden.

Jens Gerken
Mat. 01/473804
Universität Konstanz

7 VisMeB Performance Test: Liste vs. Leveltable

Obwohl die Entwicklung von VisMeB seit dem INSYDER Projekt ständig von Usability Tests begleitet wurde [Mann02, Jett03], so konnte doch eine Frage bislang nicht beantwortet werden: Bietet das Grundkonzept des Systems, also die Darstellung der Treffermenge in einer Art Supertable, welche mehrere Detailstufen bietet, gegenüber einer herkömmlichen, listenbasierten Darstellung Vorteile hinsichtlich der Performance? Sprich, kommt ein Nutzer durch die Arbeit mit der *Leveltable* schneller ans Ziel oder nicht. Um diese grundlegende Frage zu beantworten, wurde ein Performance Test konzipiert. Die in VisMeB integrierten Visualisierungen wie *2D- und 3D-Scatterplot oder Circle Segment View* sollten hierbei bewusst außen vorgelassen werden. Sie können den Benutzer zwar bei der Suche und Auswahl unterstützen, allerdings ist dazu einige Einarbeitungszeit notwendig. Ansonsten besteht vielmehr die Gefahr, die Testteilnehmer mit zu viel Funktionalität zu erschlagen und somit keine aussagekräftigen Ergebnisse zu erhalten.

Aufgrund der im Vorfeld bezüglich der Methoden der Sozialwissenschaften getätigten Recherchen, welche in Kapitel 4 wiedergegeben sind, wurde sowohl das Testdesign als auch die statistische Auswertung möglichst nahe an ein quantitatives Experiment angelehnt. Dementsprechend wurde zunächst die Null-Hypothese aufgestellt, welche durch den Test überprüft werden sollte:

„Hinsichtlich der Arbeitsgeschwindigkeit bietet die Visualisierung keine Vorteile gegenüber einer listenbasierten Darstellung."

Daraus ergaben sich die folgenden Variablen für den Test:

Unabhängige Variable: verwendete Visualisierung (Ausprägung *Leveltable* oder Liste)

Abhängige Variable: zur Bearbeitung vorgegebener Aufgaben benötigte Zeit in Sekunden

Zusätzlich wurden weitere Analysen bezüglich Fehlerhäufigkeit und der Art der aufgetretenen Benutzerfehler getätigt, welche jedoch lediglich als Ergänzung zu sehen sind. Das Hauptaugenmerk lag klar auf der Überprüfung der oben genannten Null-Hypothese.

Sowohl das Testdesign als auch die Ergebnisse, dieses im Sommer 2003 durchgeführten Performance Tests, werden im Folgenden ausführlich behandelt.

7.1 Testaufbau

Der Test fand in einem fünfwöchigen Zeitraum – beginnend Mitte Juni 2003 und endend Mitte Juli 2003 – statt. Als Usability Labor fungierte ein dafür extra eingerichtetes, ehemaliges Büro des Fachbereichs Informatik und Informationswissenschaften an der Universität Konstanz. In diesem befanden sich während der Tests der Testrechner, eine Videokamera, der Versuchsleiter sowie ein Protokollant. Für den Testrechner kam ein Intel Pentium 4 PC von DELL mit 1.6 GHz Prozessor, sowie einem 19“ TFT Bildschirm zum Einsatz. Die Auflösung des TFT betrug 1280x1024 Bildpunkte. Die Videokamera wurde neben dem Bildschirm auf einem Stativ aufgestellt, um die Mimik der Versuchsperson von vorne zu filmen. Auf eine zusätzliche Screencam musste aufgrund zu geringer Rechenleistung verzichtet werden. Da dies die quantitativen Daten für den Performance Test nicht beeinflusste, war dieser Umstand jedoch zu verschmerzen. Der Versuchsleiter saß auf einem Stuhl rechts neben der Versuchsperson. Schräg dahinter, mit freiem Blick auf den Bildschirm, befand sich der Protokollant. Aufgrund der Jahreszeit kam es zum Teil zu recht hohen Temperaturen in dem Raum, weswegen den Versuchspersonen stets etwas zu trinken zur Verfügung stand. Abgesehen von der Temperatur, welche während des Testzeitraums recht deutlich schwankte, waren die Bedingungen für alle Teilnehmer nahezu identisch. Es traten keine weiteren Einflüsse von außerhalb auf, welche die Ergebnisse hätten beeinflussen können.

Es wurde die zu Testbeginn aktuell lauffähige Version von VisMeB für den gesamten Test verwendet. Änderungen an der Software während dieses Zeitraums spielten somit ebenfalls keine Rolle. Als Vergleichssystem sollte, wie aus der Null-Hypothese ersichtlich, ein listenbasiertes System dienen. Da VisMeB standardmäßig keine listenbasierte Ansicht der Treffermenge unterstützt, wurde diese extra für diesen Test entwickelt. Sie wurde dabei nicht direkt in VisMeB integriert, vielmehr konnte bei Eingabe der Suchterme entschieden werden, ob VisMeB oder die listenbasierte Darstellung geladen werden sollte. Die Darstellung orientierte sich an bekannten Internetsuchmaschinen wie etwa *Google*. Wie in Abbildung 7.1 zu sehen, wurden pro Seite jeweils 10 Treffer angezeigt, geordnet nach der Relevanz. Zusätzlich wurden zu jedem Treffer weitere Meta-Attribute wie Dateigröße, Sprache und Server-Typ angezeigt. Die Auswahl orientierte sich hierbei an den in der *Leveltable* sichtbaren Meta-Attributen. Jedes Dokument konnte per Mausklick auf den sichtbaren Link in einem neuen Fenster angezeigt werden. Einschränkend ist hier zu bemerken, dass keine Suchfunktion innerhalb eines Dokumentes implementiert wurde. Da die getestete Version von VisMeB ebenfalls noch keine Stichwortsuche innerhalb eines Dokumentes ermöglichte, schien diese Entscheidung nahe zu liegen, um eine mögliche Verzerrung der Ergebnisse zu verhindern.

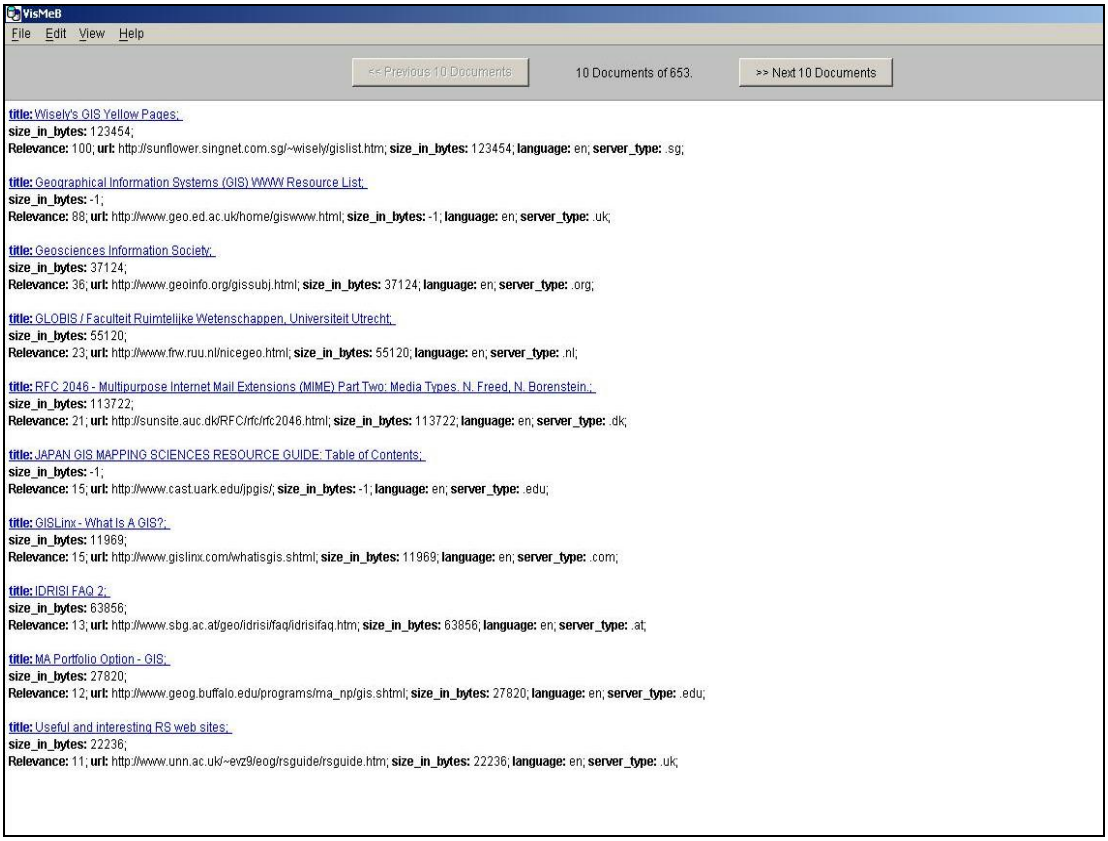


Abbildung 7.1: VisMeB Listendarstellung

7.2 Versuchspersonen

An dem Test nahmen insgesamt 32 Personen teil, von welchen letztendlich 30 in die Auswertung mit einbezogen wurden. Bei einer Versuchsperson war die Videoaufzeichnung unbrauchbar, weswegen nur die gemessenen Zeiten durch die Protokollanten zur Verfügung standen. Allerdings hatte die Erfahrung bereits gezeigt, dass diese für sich genommen zu ungenau waren. Eine weitere Person, welche bereits bei einigen Usability Tests teilgenommen hatte, verfälschte das Ergebnis durch sehr ausgeprägtes Thinking Aloud, welches sich zum Teil in längeren Erklärungspausen zeigte. Da dies während dem Test nicht bereits unterbunden wurde, konnten die hier gewonnenen Daten ebenfalls nicht in die Auswertung mit einbezogen werden.

Bei den Teilnehmern handelte es sich größtenteils um Studenten und Mitarbeiter der Universität Konstanz, beziehungsweise der Fachhochschule Konstanz. Es wurden aufgrund der Zielgruppe von VisMeB nur Personen ausgewählt, welche bereits eine gewisse Grund - Erfahrung im Umgang mit PCs und insbesondere Suchmaschinen besaßen. Jeder Teilnehmer musste zu Beginn einen Pre-Test Fragebogen ausfüllen, mit welchem zum einen die demographischen Daten erfasst werden sollte und zum anderen eine gewisse Einstufung der Teilnehmer hinsichtlich Erfahrung und persönlicher Affinität im Umgang mit Computern ermöglicht werden sollte.

Die ausführlichen Ergebnisse dieser Pre-Test Fragebögen sind in Anhang A zu finden. Im Folgenden werden die wichtigsten Punkte zusammenfassend aufgeführt.

Das Durchschnittsalter der Teilnehmer betrug 26 Jahre, wobei der jüngste Teilnehmer 21 Jahre und der älteste 41 Jahre war. Acht der 32 Personen waren weiblich, die restlichen 24 männlich. Nahezu alle konnten ein Abitur als höchsten schulischen Abschluss vorweisen (28). Weiterhin stammten 15 Personen aus dem Umfeld des Fachbereichs Informatik und Informationswissenschaften.

Alle Teilnehmer gaben an, bereits eine Internetsuchmaschine benutzt zu haben. Aufgrund der momentanen marktbeherrschenden Position wenig überraschend, ist hier die Angabe der am häufigsten genutzten Suchmaschine: *Google*. Auf einer 7-Likert-Skala kreuzten die Teilnehmer, auf die Frage, ob sie gerne mit Computern arbeiten würden (wobei eine 1 bedeutete: *arbeite sehr ungern mit Computern* und eine 7: *arbeite sehr gerne mit Computern*) im Schnitt eine 5,56 an, was doch eine recht positive Neigung gegenüber Computern ausdrückt. Die allgemeine Computererfahrung wurde durchschnittlich mit 5,09 angegeben und die Erfahrung mit Internetsuchmaschinen etwas niedriger mit 4,81 (in beiden Fällen: *1 = keinerlei Erfahrung* und *7 = sehr viel Erfahrung*). Viele der Versuchspersonen gaben an, jeden Tag sehr lange mit dem PC zu arbeiten, der Durchschnitt lag hier bei etwa fünf Stunden.

In Hinblick auf die angegebene momentane Beschäftigung können die Teilnehmer grob in zwei Gruppen unterteilt werden. Zum einen die Gruppe der Informatikstudenten und derjenigen, die in diesem Bereich arbeiten (im Folgenden „Informatiker“ genannt) und zum anderen die Gruppe der Nicht - Informatikstudenten und derjenigen, die auch nicht beruflich direkt damit zu tun haben (im Folgenden „Nicht-Informatiker“ genannt). Betrachtet man nun die Ergebnisse für beide Gruppen getrennt, ergibt sich das erwartete Bild: Die Gruppe der „Informatiker“ arbeitet deutlich länger mit dem PC (6,67 Stunden zu 2,95), ist der Arbeit mit diesem positiver zugeneigt (6,06 zu 5,00) und schätzt ihre persönliche Erfahrung auch höher ein (allgemeine Computererfahrung: 5,94 zu 4,13; Erfahrung mit Suchmaschinen: 5,53 zu 4,00). Zu beachten ist hierbei allerdings, dass Nicht-Informatiker eher dazu neigen, ihre persönlichen Kenntnisse in Hinblick auf den Umgang mit PCs zu unterschätzen. Bei der Auswertung der Testergebnisse wurde nicht weiter auf diese mögliche Unterscheidung in Informatiker/Nicht Informatiker eingegangen, da dies nicht Teil der Untersuchung war.

Zusammenfassend kann gesagt werden, dass die Teilnehmer der eingangs erwähnten Voraussetzung, eine gewisse Grund -Erfahrung aufzuweisen, alle entsprachen.

7.3 Testdesign

Für den Test wurde ein *Between Subjects Design* gewählt (siehe Kapitel 3.5). Die 32 Teilnehmer wurden dabei zufällig in zwei Gruppen aufgeteilt. Zum einen die Versuchsgruppe, welche die Aufgaben mit der *Leveltable* bearbeitete und zum anderen die Kontrollgruppe, welche den Test mit der Listendarstellung durchführte. Aufgrund kleinerer organisatorischer Probleme wurden die Teilnehmer

nicht völlig gleichmäßig verteilt. Die Versuchsgruppe beinhaltete 18 Teilnehmer, wohingegen die Kontrollgruppe nur aus 14 Teilnehmern bestand[8]. Um sicherzustellen, dass beide Gruppen wirklich vergleichbar sind und keine signifikanten Unterschiede aufweisen, wurde zunächst ein so genannter Baseline-Test durchgeführt. In diesem arbeiteten beide Gruppen mit der listenbasierten Darstellung, um eventuelle Unterschiede aufzuzeigen. In dem daran anschließenden Haupt-Test arbeitete die Kontrollgruppe erneut mit der listenbasierten Darstellung. Die Versuchsgruppe arbeitete hingegen mit der *Leveltable*. Aufgrund der, im Gegensatz zu der allen Teilnehmern bekannten Listendarstellung, nun völlig neuen Visualisierung, wurde jedem Teilnehmer dieser Gruppe zunächst ein Einführungsvideo gezeigt. Dieses erklärte die grundsätzlichen Funktionen der *Leveltable*, allerdings ohne konkrete, aufgabenbezogene Hinweise zu geben. Der Protokollant hatte während des Tests die Aufgabe, sowohl die Zeiten zur Bearbeitung der Aufgaben zu messen, als auch die Vorgehensweise zumindest in groben Zügen festzuhalten. Da es nicht möglich war, sämtliche Tests von ein und derselben Person protokollieren zu lassen, gab es hier einige Unterschiede in der Notation der Zeiten und Handlungen. Aus diesem Grund wurden die Zeiten zusätzlich mit Hilfe der Videoaufzeichnung überprüft und abgeglichen.

Nach Absolvierung aller Testaufgaben durften die Teilnehmer noch einen Post-Test Fragebogen ausfüllen. Allerdings waren davon nur die Teilnehmer der Versuchsgruppe betroffen, da der Fragebogen ihre subjektiven Empfindungen bezüglich dem Arbeiten mit der *Leveltable* abfragte. Da diese Teilnehmer ebenfalls mit der Listendarstellung im Verlauf des Baseline-Tests gearbeitet hatten, war es ihnen möglich hier auch Vergleiche zu ziehen und konkret Vor- und Nachteile zu benennen.

Die Auswertung der quantitativen Daten erfolgte mit Hilfe einer einfaktoriellen Varianzanalyse – eine so genannte OneWay ANOVA (siehe Kapitel 3.1). Die genauen Ergebnisse sind in Kapitel 7.5 ausführlich aufgeführt.

7.4 Testaufgaben

Für den Test wurde die GISWeb Datenbank verwendet, welche etwa 2000 gespeicherte Webseiten zum Thema GIS – Geographic Information System – enthält. Aus Performance und Relevanzgründen, wurden daraus etwa 300 Dokumente ausgewählt. Die Teilnehmer hatten keine Kenntnisse bezüglich dieser Thematik, weswegen die Aufgaben größtenteils keine inhaltliche Analyse der Dokumente

[8] *Dies hat keine Auswirkungen auf die Ergebnisse, da bei einer einfaktoriellen Varianzanalyse in solch einem Fall mit dem harmonischen Mittel aus beiden Gruppen gerechnet werden kann. Bei mehrfaktoriellen Varianzanalysen ist dies problematischer und erfordert deutlich komplexere Analysen. Vergleiche auch:*
http://www.uccs.edu/~lbecker/oneway.htm,
http://www.uvm.edu/~dhowell/gradstat/psych340/Lectures/Anova/anova2.html,
http://www.uwsp.edu/education/wkirby/stat/Lesson11Anova2.htm , alle online am 15.03.2004

erforderten. Das Verwenden einer anderen Datenbasis, welche die Teilnehmer eventuell inhaltlich mehr angesprochen hätte, war aus organisatorischen Gründen leider nicht möglich.
Beide Gruppen mussten jeweils 12 Aufgaben bearbeiten, wobei die Versuchsgruppe zusätzlich zu Beginn die Möglichkeit hatte, sich zunächst für kurze Zeit mit der neuen Darstellung vertraut zu machen. Die ersten 11 Aufgaben konnten innerhalb einer Suche gelöst werden. Der Suchterm wurde hierzu vorgegeben. Die letzte Aufgabe ermöglichte die Eingabe eines eigenen Suchterms.

Die Aufgaben waren prinzipiell für beide Gruppen identisch, jedoch mussten zum Teil einige kleinere Anpassungen vorgenommen werden[9]. Letztendlich lassen sich die Aufgaben in drei verschiedene Typen unterscheiden, welche im Folgenden näher erklärt werden. Die Aufteilung orientierte sich dabei an der von [Shnei98] vorgeschlagenen Typologie in *specific fact finding* und *extended fact finding*, welche sich auch bereits bei einer früheren INSYDER Evaluation bewährt hatte [Mann02]. Die ersten beiden Aufgabentypen können hierbei dem *specific fact finding* zugeordnet werden und der letzte Aufgabentyp dem *extended fact finding.* Der genaue Wortlaut aller Aufgaben ist in Anhang B einsehbar.

Aufgabentyp 1 – Dokumente suchen:

Dieser Typ umfasste Aufgaben, in welchen zum einen konkret nach einem Dokument mit einem bestimmten Titel gesucht werden sollte. Zum anderen wurden bestimmte Attribute, wie etwa Relevanz oder Sprache, einschränkend vorgegeben, anhand derer das entsprechende Dokument gefunden werden sollten. Zur Lösung der Aufgaben durften keine neuen Suchanfragen gestellt, sondern vielmehr musste die Ausgangssuche verwendet werden. Insgesamt fielen fünf Aufgaben in diese Kategorie.

Aufgabentyp 2 – Dokumente vergleichen:

Hierbei mussten mehrere Dokumente bezüglich vorgegebener Meta-Attribute wie Sprache oder Dateigröße miteinander verglichen werden. Diese Aufgaben wurden ebenfalls innerhalb der Ausgangssuche bearbeitet. Insgesamt zählten hierzu vier Aufgaben.

Aufgabentyp 3 – Dokumente inhaltlich untersuchen:

Dieser Aufgabentyp erforderte von den Teilnehmern den meisten Aufwand. Zum einen wurde hier nach bestimmten Meta-Attributen bezüglich eines vorgegebenen Dokumentes gefragt und zum anderen umfasste dieser Aufgabentyp auch die letzte Aufgabe, in welcher die Teilnehmer die Dokumente auf eine geforderte Information hin überprüfen mussten und hierfür auch eine neue Suchanfrage stellen konnten. Insgesamt zählten zu diesem Typ drei Aufgaben.

[9] *Diese beschränkten sich auf notwendige Umformulierungen. Beispielsweise musste bei der Leveltable zum Teil ergänzt werden, dass die Ergebnisse zunächst nach Relevanz sortiert werden müssen, da dies bei der Liste standardgemäß der Fall war.*

Jens Gerken
Mat. 01/473804
Universität Konstanz

7.5 Ergebnisse Liste vs. Leveltable

Für die Auswertung der gemessenen Zeiten zur Lösung der Aufgaben, wurde eine einfaktorielle Varianzanalyse verwendet, welche in Kapitel 3.1 näher erklärt wird. Aufgrund der Null-Hypothese stand dieser Teil der Auswertung auch klar im Fokus.

Im Folgenden werden zunächst die Ergebnisse des Baseline-Tests kurz vorgestellt, welcher dazu dienen sollte, sicherzustellen, dass keine signifikanten Unterschiede zwischen beiden Gruppen bestanden. Anschließend werden die Performance Ergebnisse des Haupt-Tests präsentiert, wobei hier unterschieden wird zwischen Gesamtzeit und benötigter Zeit pro Aufgabentyp. Weiterhin werden für jeden Aufgabentyp die größten Schwierigkeiten der Versuchspersonen herausgestellt und näher beleuchtet. Abschließend erfolgt eine Betrachtung der Post-Test Fragebogen-Ergebnisse, wobei hier unterschieden wird, zwischen quantitativen Ergebnissen und qualitativen Kommentaren der Teilnehmer.

Baseline-Test Ergebnisse: Liste vs. Liste

Wie bereits erwähnt, arbeiteten bei diesem Test alle Teilnehmer mit der listenbasierten Darstellung. Hierzu mussten 11 Aufgaben bearbeitet werden, welche sich von der Art an den Aufgaben des Haupt-Tests orientierten, allerdings natürlich keine direkten Hinweise zur Lösung dieser gaben. Für die Auswertung wurden schließlich 10 dieser Aufgaben verwendet, da es bei einer Aufgabe zu massiven Problemen aufgrund der Fragestellung kam und somit letztendlich zu wenige Daten für diese vorlagen.

Wie bereits erwähnt, wurden nur 30 Datensätze der 32 vorhandenen für die Auswertung verwendet – davon 17 Teilnehmer der Versuchsgruppe und 13 der Kontrollgruppe. Ein Vergleich der Gesamtzeiten ergibt folgendes Bild (*Abbildung 7.2*):

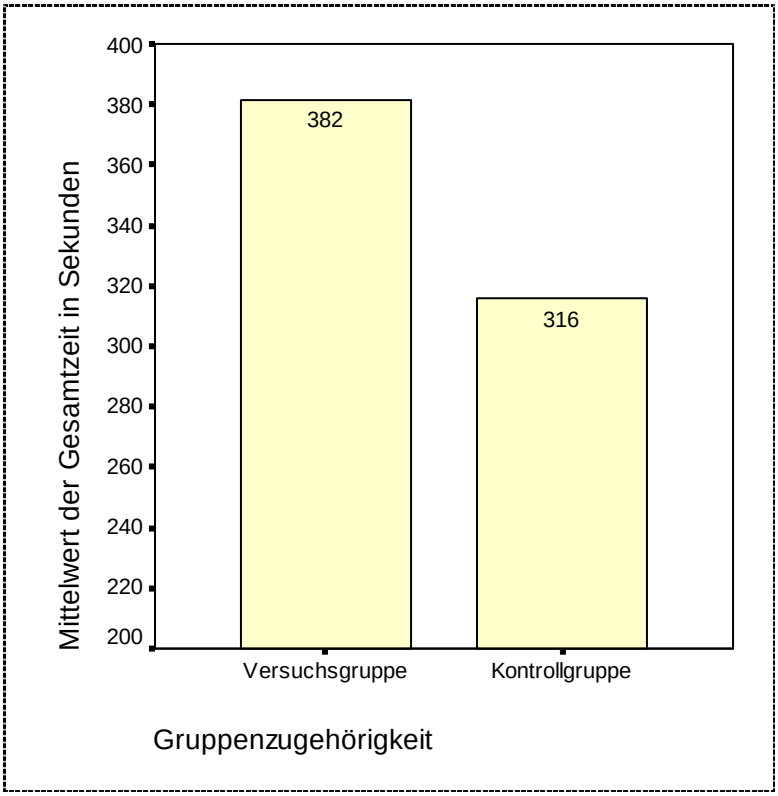


Abbildung 7.2:Vergleich der Gesamtzeiten für den Baseline-Test

Die Teilnehmer der Kontrollgruppe benötigten im Durchschnitt also 316 Sekunden zur Lösung der 10 Aufgaben, wohingegen die Teilnehmer der Versuchsgruppe etwas langsamer waren und 382 Sekunden benötigten. Die Varianzanalyse ergab jedoch, dass dies bei einem Konfidenzintervall von 95% keinen signifikanten Unterschied zwischen beiden Gruppen bedeutete (*Abbildung 7.3*) – sichtbar an dem Signifikanzwert von 0.101 > 0.05.

Gesamtzeit Baseline-Test

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	31823,106	1	31823,106	2,869	,101
Innerhalb der Gruppen	310592,835	28	11092,601		
Gesamt	342415,941	29			

Abbildung 7.3: Varianzanalyse zu Baseline-Test für die Gesamtzeiten

Eine Betrachtung der einzelnen Datensätze ergab bereits einen Verdacht auf mögliche Ausreißer, weswegen diese These mit Hilfe eines Boxplots überprüft werden sollte (*Abbildung 7.4*) Ein Datensatz gilt als Ausreißer, sobald er sich außerhalb von 1,5 x IQR befindet (siehe Kapitel 3.4).

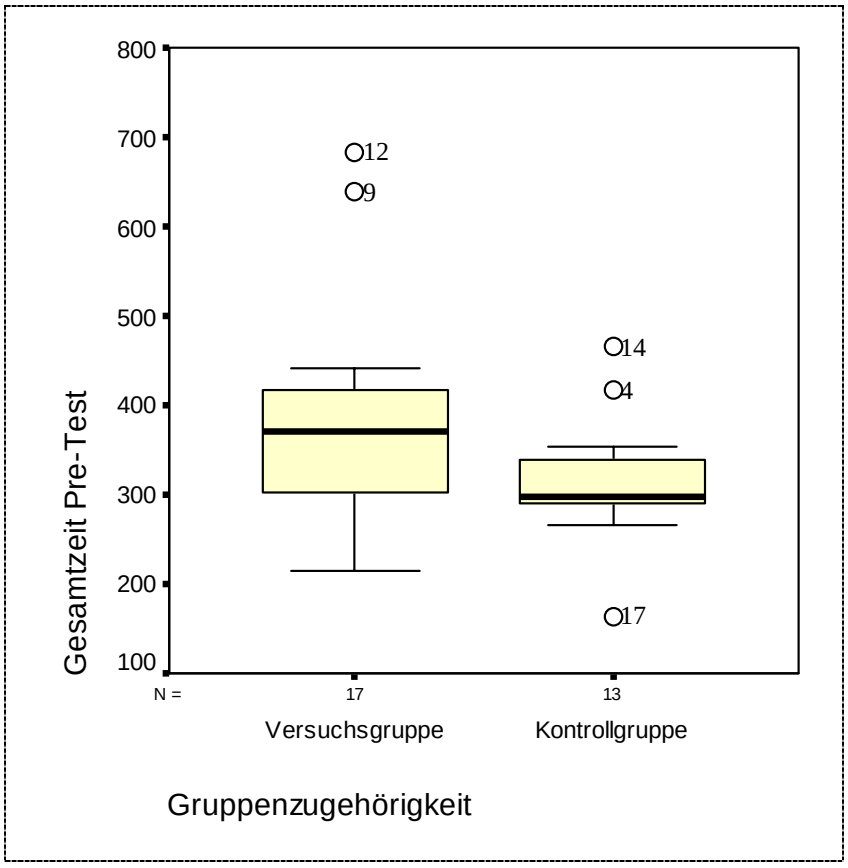


Abbildung 7.4: Boxplot zu Baseline-Test für die Gesamtzeiten

Es ist hier sehr schnell ersichtlich, dass insbesondere die Versuchsgruppe unter zwei deutlichen negativen Ausreißern zu leiden hat. Aus diesem Grund wurde eine weitere Varianzanalyse unter Ausschluss dieser Datensätze durchgeführt. Konkret wurden die Ergebnisse von Versuchsperson 9 und

12 bei der Versuchsgruppe und die von Versuchsperson 4, 14 und 17 der Kontrollgruppe im Folgenden nicht berücksichtigt (*Abbildung 7.5*).

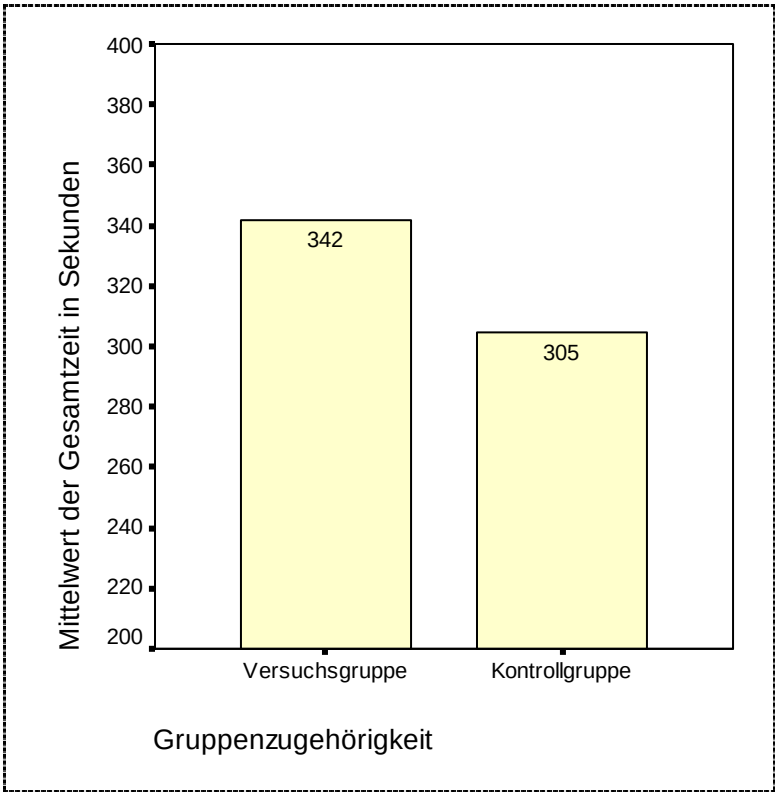


Abbildung 7.5: Vergleich der Gesamtzeiten für den Baseline-Test nach Entfernen der Ausreißer

Wie bereits erwartet sinkt die durchschnittliche Zeit der Versuchsgruppe deutlich ab von 382 auf 342 Sekunden, wohingegen die Zeit der Kontrollgruppe relativ konstant bleibt und nur geringfügig von 316 auf 305 Sekunden absinkt. Die Varianzanalyse bestätigt dieses Bild, da der Signifikanzwert mit 0,13 nun noch deutlicher über dem Grenzwert von 0,05 bei einem Konfidenzintervall von 95% liegt (*Abbildung 7.6*).

Gesamtzeit Baseline-Test ohne Ausreißer

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	8138,359	1	8138,359	2,468	,130
Innerhalb der Gruppen	75844,755	23	3297,598		
Gesamt	83983,114	24			

Abbildung 7.6: Varianzanalyse zu Baseline-Test nach Entfernen der Ausreißer

Der Baseline-Test konnte also deutlich zeigen, dass, trotz zufälliger Verteilung, zwischen beiden Gruppen kein signifikanter Unterschied bestand und diese somit in dieser Konstellation für den Haupt-Test geeignet waren.

7.5.1 Haupt-Testergebnisse: Liste vs. Leveltable

Für diesen Test arbeitete die Versuchsgruppe mit der *Leveltable*, wohingegen die Kontrollgruppe ein zweites Mal mit der Listendarstellung die Aufgaben bearbeitete. Zunächst sei gesagt, dass in die Wertung nur vollständig korrekt gelöste Aufgaben mit einflossen. Konnte ein Teilnehmer eine Aufgabe nicht, nur fehlerhaft oder nur durch Hilfestellung lösen, so wurde diese als *fehlender Wert* behandelt. Durch das Fehlen einer Aufgabe würde sich allerdings die Gesamtzeit verringern, was wiederum das Ergebnis verfälschen würde. Aus diesem Grund wurden diese *fehlenden Werte* mit Hilfe einer Mittelwertbestimmung aus den Datensätzen der anderen Teilnehmer zu der jeweiligen Aufgabe ergänzt. Konnte also beispielsweise Versuchsperson 9 die Aufgabe 5 nicht lösen, so wurde ihr dort die Durchschnittszeit der restlichen Teilnehmer ihrer Gruppe für diese Aufgabe zugeordnet. Dies kann die Ergebnisse natürlich leicht verzerren, ist aber für den Vergleich der Gesamtzeit unausweichlich notwendig. Wie bereits erwähnt mussten insgesamt 12 Aufgaben bearbeitet werden, welche auch alle in die Auswertung mit einbezogen wurden. Wie bei dem Baseline-Test flossen allerdings nur die Ergebnisse von 30 der 32 Teilnehmer in die Auswertung mit ein.
Im Folgenden sollen zunächst die Gesamtzeiten betrachtet und diese anschließend nach Aufgabentypen aufgeschlüsselt werden. Die Ergebnisse der einzelnen Aufgaben (hier wurden keine fehlenden Werte ersetzt), befinden sich in Anhang B.

7.5.1.1 Auswertung der Gesamtzeiten

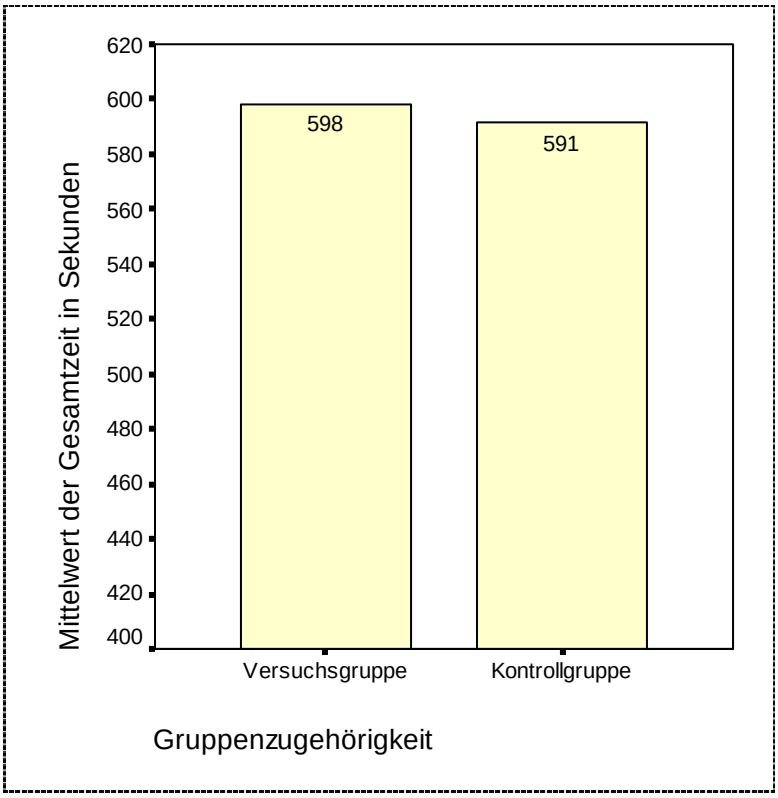


Abbildung 7.7: Vergleich der Gesamtzeiten für den Haupt-Test

Wie in *Abbildung 7.7* ersichtlich, unterscheiden sich die beiden Gruppen nur um lediglich sieben Sekunden und damit minimal. Die Varianzanalyse ergibt somit erwartungsgemäß keinen signifikanten Unterschied (*Abbildung 7.8*), erkennbar an dem Signifikanzwert von 0,931 > 0,05 bei einem Konfidenzintervall von 95%. Die Null-Hypothese kann für diesen, allgemeinsten Fall somit nicht widerlegt werden, sondern gilt vorläufig als bestätigt.

Gesamtzeit Haupt-Test

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	368,757	1	368,757	,008	,931
Innerhalb der Gruppen	1344691,152	28	48024,684		
Gesamt	1345059,909	29			

Abbildung 7.8: Varianzanalyse zu Haupt-Test für die Gesamtzeit

Allerdings zeigt ein Blick auf die Standardabweichung, dass hier doch deutliche Unterschiede zwischen beiden Gruppen bestehen. Sie beträgt innerhalb der Kontrollgruppe 115 Sekunden, wohingegen innerhalb der Versuchsgruppe eine Standardabweichung von 272 Sekunden existiert. Dies ist wiederum ein deutliches Indiz für mögliche Ausreißer. Ein Blick auf den Boxplot ergibt folgendes Bild (*Abbildung 7.9*):

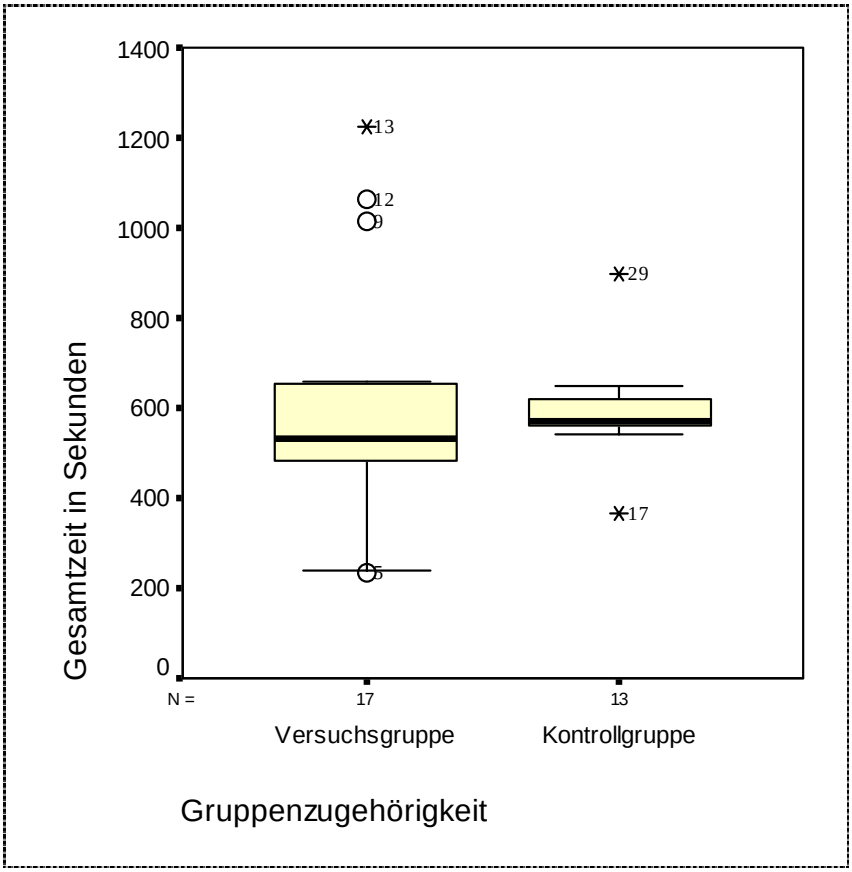


Abbildung 7.9: Boxplot zu Haupt-Test für die Gesamtzeit

Es zeigen sich bei beiden Gruppen wiederum Ausreißer und auch Extremwerte, jedoch scheinen sich diese bei der Kontrollgruppe gegenseitig auszugleichen, wohingegen die Versuchsgruppe eher darunter zu leiden scheint. Interessant ist die Tatsache, dass zwei der Ausreißer, Versuchsperson 9 und 12, bereits im Baseline-Test als Ausreißer identifiziert werden konnten. Allerdings ist das gehäufte Auftreten von Ausreißern innerhalb der Versuchsgruppe nicht weiter verwunderlich. Die Teilnehmer mussten mit einer völlig unbekannten Visualisierung arbeiten, was einige sicherlich in ihrer Leistungsfähigkeit einschränkte. Interessant ist an dieser Stelle die Frage aus dem Pre-Test Fragebogen: „Fällt es Ihnen leicht, sich mit neuer Software vertraut zu machen“, welche die Versuchsperson 13 mit einer 4 auf einer 7-Likert-Skala angab (1= nein fällt mir sehr schwer, 7 = ja, habe ich überhaupt keine Probleme damit) und damit innerhalb ihrer Gruppe der Informatiker deutlich unter dem Durchschnitt von 5,41 lag.

Ein Entfernen der Ausreißer, sowohl der positiven, als auch der negativen, ergibt das folgende, sicherlich etwas unerwartete Bild (*Abbildung 7.10*):

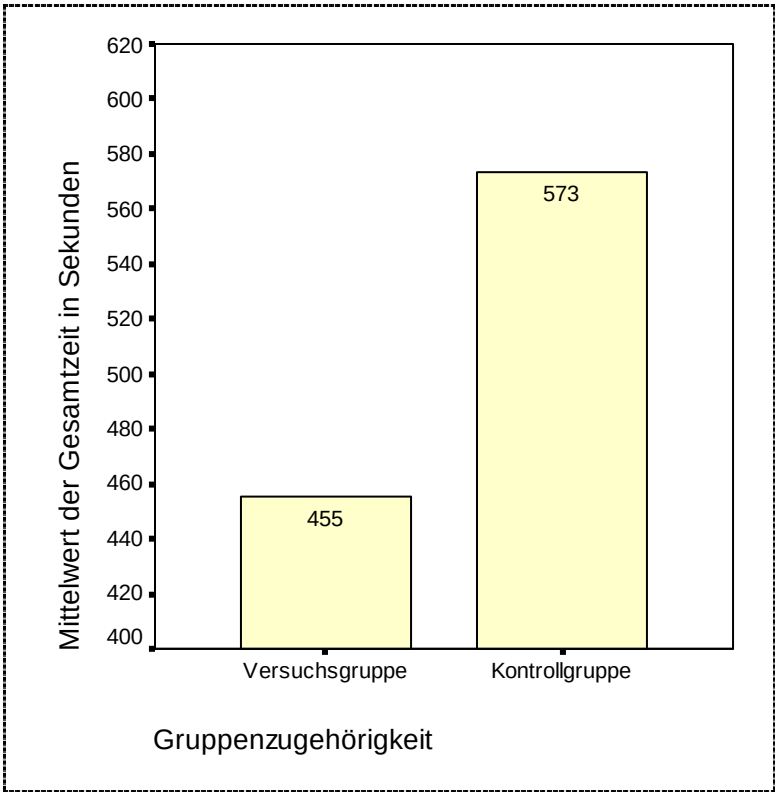


Abbildung 7.10: Vergleich der Gesamtzeiten für den Haupt-Test nach Entfernen der Ausreißer

Die Versuchsgruppe scheint nun mit lediglich 455 Sekunden statt 598 Sekunden deutlich schneller zu sein als die Kontrollgruppe, welche durch das Entfernen der Ausreißer die Durchschnittszeit nur von 591 auf 573 Sekunden senken konnte. Die Varianzanalyse bestätigt diesen deutlichen Unterschied ebenfalls (*Abbildung 7.11*).

Gesamtzeit ohne Ausreißer

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	83025,830	1	83025,830	12,998	**,002**
Innerhalb der Gruppen	140530,064	22	6387,730		
Gesamt	223555,894	23			

Abbildung 7.11: Varianzanalyse zu Haupt-Test nach Entfernen der Ausreißer

Der Signifikanzwert von 0,02 liegt deutlich unter den geforderten 0,05 bei einem Konfidenzintervall von 95%. Somit ist die Versuchsgruppe signifikant schneller als die Kontrollgruppe – die Null-

Hypothese kann verworfen werden. Zu beachten ist hier aber sicherlich, dass das Entfernen von Ausreißern nie unkritisch ist. Beispielsweise könnte das Entfernen nur eines Ausreißers schon dazu führen, dass sich der gesamte Boxplot verschiebt und ehemalige Ausreißer „verschluckt“. Dies ist insbesondere bei recht wenigen Datensätzen der Fall, was hier bei 13 beziehungsweise 17 Fällen auch zutrifft. Somit ist dieses Ergebnis sicherlich mit einem gewissen Vorbehalt zu betrachten, zeigt aber trotzdem auf, dass die Versuchsgruppe zumindest tendenziell schneller mit der *Leveltable* arbeitete als die Kontrollgruppe mit der listenbasierten Darstellung.

7.5.1.2 Auswertung nach Aufgabentypen

Wie bereits erwähnt, lassen sich die 12 Aufgaben in drei Typen unterteilen:

- Dokumente suchen
- Dokumente vergleichen
- Dokumente inhaltlich untersuchen

Neben der Betrachtung der Gesamtzeit, ist es somit sicherlich interessant, inwieweit zwischen beiden Gruppen Unterschiede hinsichtlich dieser Aufgabentypen bestehen um eventuelle Stärken und Schwächen der Systeme aufzudecken.

I. Dokumente suchen

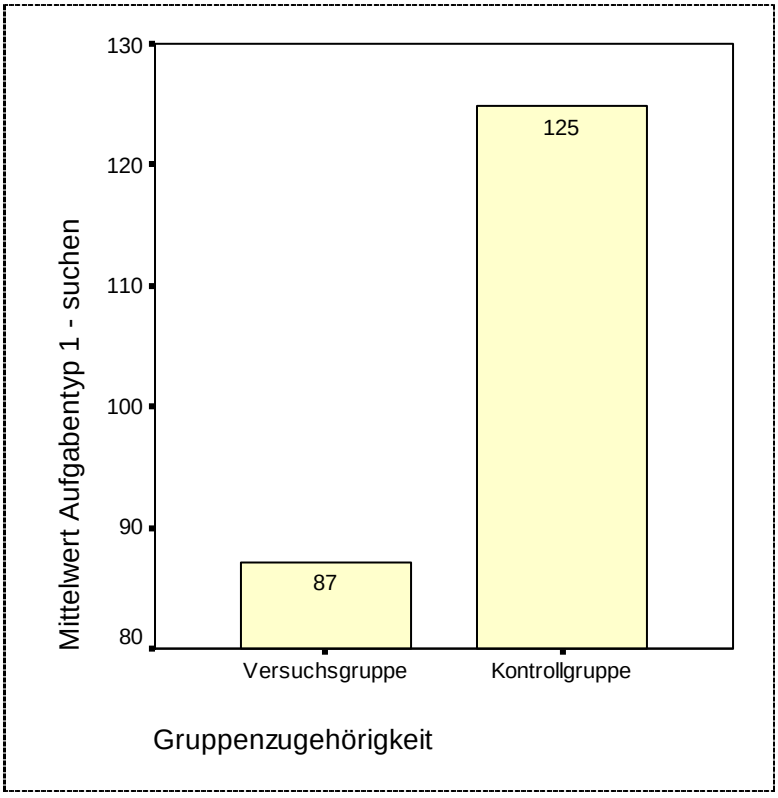


Abbildung 7.12: Vergleich der Gesamtzeiten für Aufgabentyp 1

Die Versuchsgruppe scheint hier (*Abbildung 7.12*) deutlich schneller die Aufgaben lösen zu können, was die Varianzanalyse auch bestätigen kann. Die Versuchsgruppe ist signifikant schneller als die Kontrollgruppe (Signifikanz Wert 0,008 < 0,05 bei Konfidenzintervall von 95%) (*Abbildung 7.13*).

Aufgabentyp 1 - suchen

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	10482,824	1	10482,824	8,262	**,008**
Innerhalb der Gruppen	35525,262	28	1268,759		
Gesamt	46008,086	29			

Abbildung 7.13: Varianzanalyse für Aufgabentyp 1

Der Grund für diesen deutlichen Vorsprung ist in der Tabellenfunktionalität der *Leveltable* und im speziellen den Sortierfunktionen zu suchen, welche sehr schnell das Eingrenzen nach bestimmten Meta-Attributen ermöglichen.

Mögliche Ausreißer konnten bei diesem Aufgabentyp bei keiner der beiden Gruppen entdeckt werden (*Abbildung 7.14*).

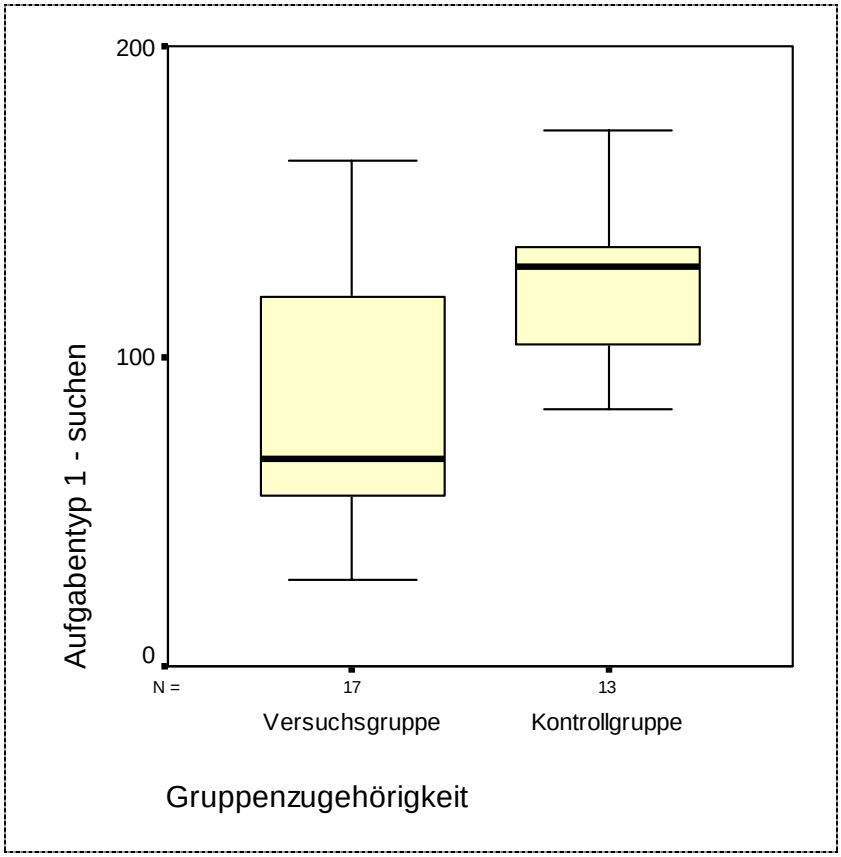


Abbildung 7.14: Boxplot für Aufgabentyp 1

II. Dokumente vergleichen

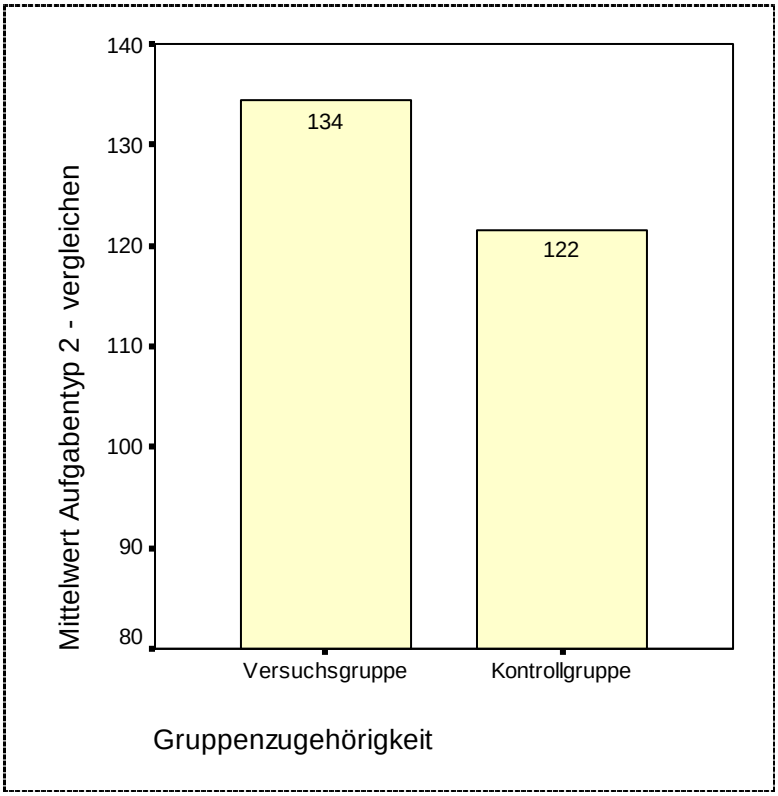


Abbildung 7.15: Vergleich der Gesamtzeiten für Aufgabentyp 2

Ein etwas anderes Bild zeigt sich bei dem 2. Aufgabentyp, bei welchem Dokumente miteinander verglichen werden sollten. Die Versuchsgruppe ist hier etwas langsamer als die Kontrollgruppe (*Abbildung 7.15*). Die Varianzanalyse ergibt allerdings, dass dieser Unterschied nicht signifikant ist → 0,437 > 0,05 bei einem Konfidenzintervall von 95% (*Abbildung 7.16*).

Aufgabentyp 2 - vergleichen

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	1214,333	1	1214,333	,621	**,437**
Innerhalb der Gruppen	54769,391	28	1956,050		
Gesamt	55983,724	29			

Abbildung 7.16: Varianzanalyse für Aufgabentyp 2

Ein möglicher Grund für diesen Gleichstand trotz der Sortierfunktionen der *Leveltable* könnte das Fehlen einer Filterfunktion zum Testzeitpunkt sein. Für das Vergleichen einer vorgegebenen Dokumentenmenge, musste diese zumeist erst mehrfach abgezählt werden.

Auch hier ergab die Untersuchung des Boxplots keine Hinweise auf Ausreißer (*Abbildung 7.17*).

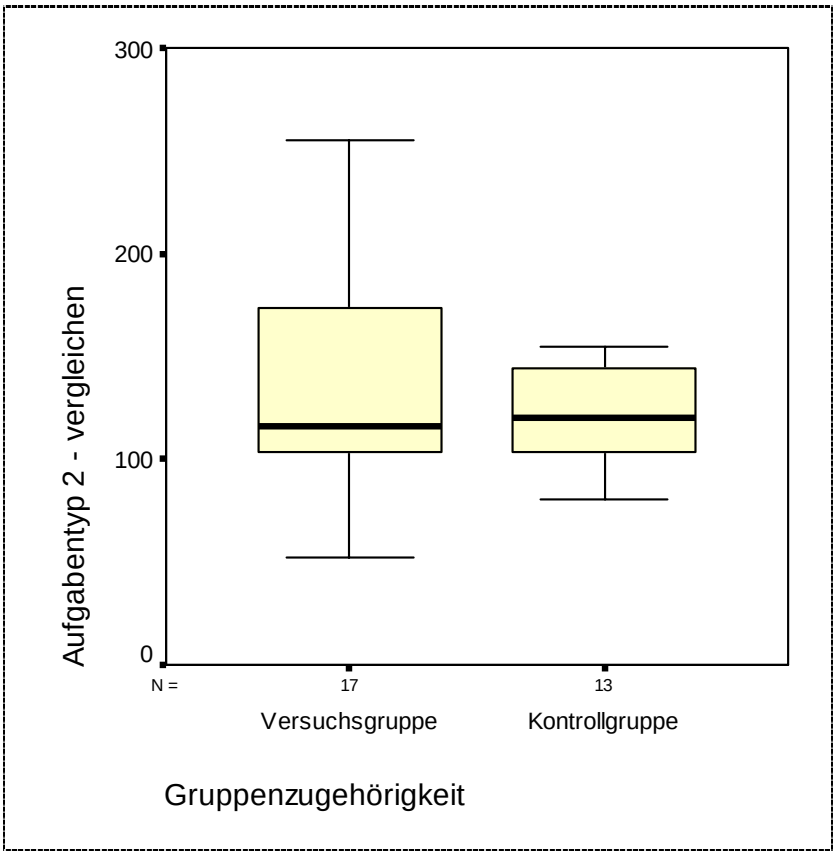


Abbildung 7.17: Boxplot für Aufgabentyp 2

Dieser Umstand ist in sofern interessant, da dies erwarten lässt, dass bei Aufgabentyp 3 die Streuung deutlich größer ist, da sich ansonsten die Ausreißer bei der Betrachtung der Gesamtzeit über alle Aufgaben nicht erklären lassen würde.

III. Dokumente inhaltlich untersuchen

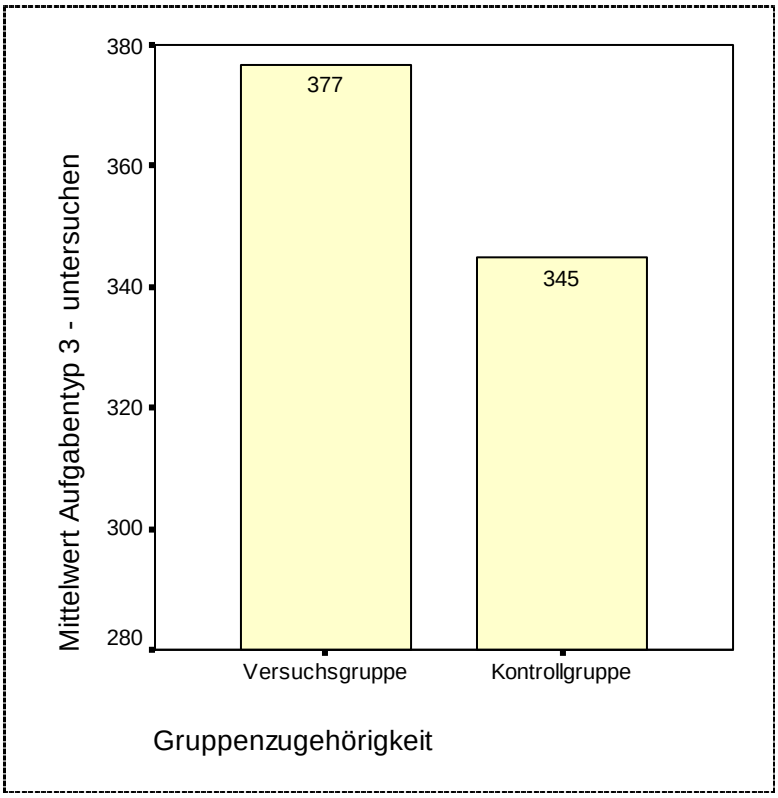


Abbildung 7.18: Vergleich der Gesamtzeiten für Aufgabentyp 3

Hier ergibt sich wiederum ein ähnliches Bild wie bei Aufgabentyp 2. Die Kontrollgruppe scheint zumindest leicht schneller zu sein (*Abbildung 7.18*). Die Varianzanalyse ergibt allerdings abermals keinen signifikanten Unterschied → 0,620 > 0,05 bei einem Konfidenzintervall von 95% (*Abbildung 7.19*).

Aufgabentyp 3 - untersuchen

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	7524,077	1	7524,077	,251	**,620**
Innerhalb der Gruppen	838064,773	28	29930,885		
Gesamt	845588,850	29			

Abbildung 7.19: Varianzanalyse für Aufgabentyp 3

Ein Blick auf die Standardabweichungen untermauert bereits den Verdacht, dass einige Ausreißer vorhanden sein könnten. Die Versuchsgruppe schwankt hier um 212 Sekunden, wohingegen die Standardabweichung innerhalb der Kontrollgruppe lediglich 98 Sekunden beträgt.

Der Boxplot (*Abbildung 7.20*) zeigt nun auch deutliche Ausreißer. Interessant ist hierbei, dass es sich genau um jene Ausreißer handelt, welche auch bei Betrachtung der Gesamtzeit entdeckt wurden. Da bei den anderen beiden Aufgabentypen keine Ausreißer zu finden waren, scheint allein der Aufgabentyp 3 für diese Abweichungen verantwortlich zu sein.

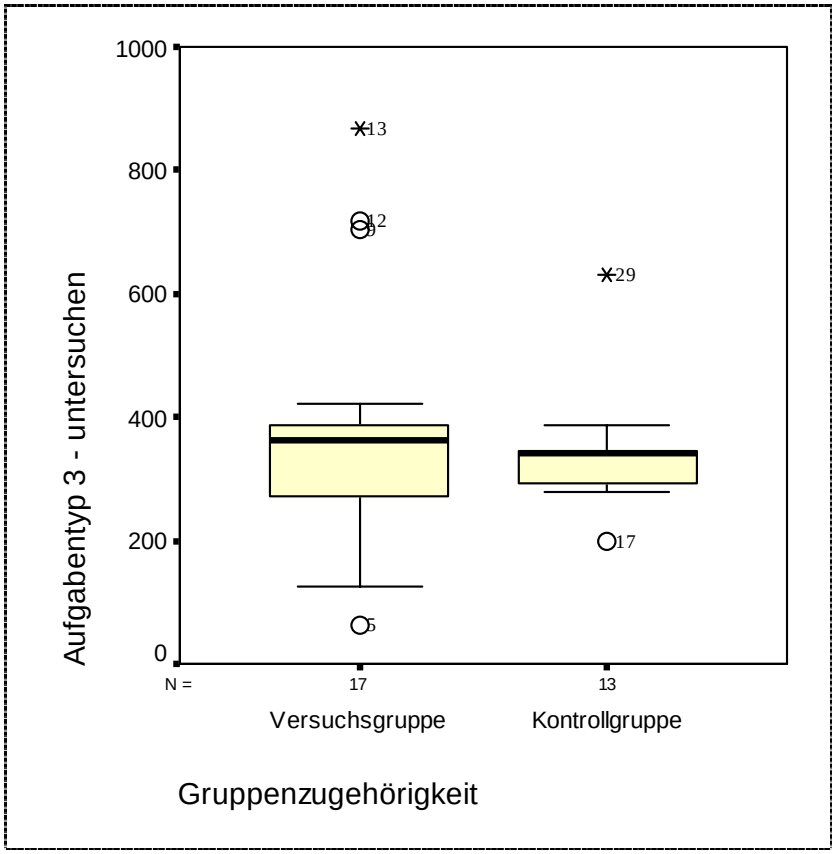


Abbildung 7.20: Boxplot für Aufgabentyp 3

Bei Entfernen der Ausreißer ergibt sich erwartungsgemäß ein umgekehrtes Bild. Die Versuchsgruppe ist nun schneller als die Kontrollgruppe und die Varianzanalyse ergibt, dass dieser Unterschied auch signifikant ist (*Abbildung 7.21 & 7.22*).

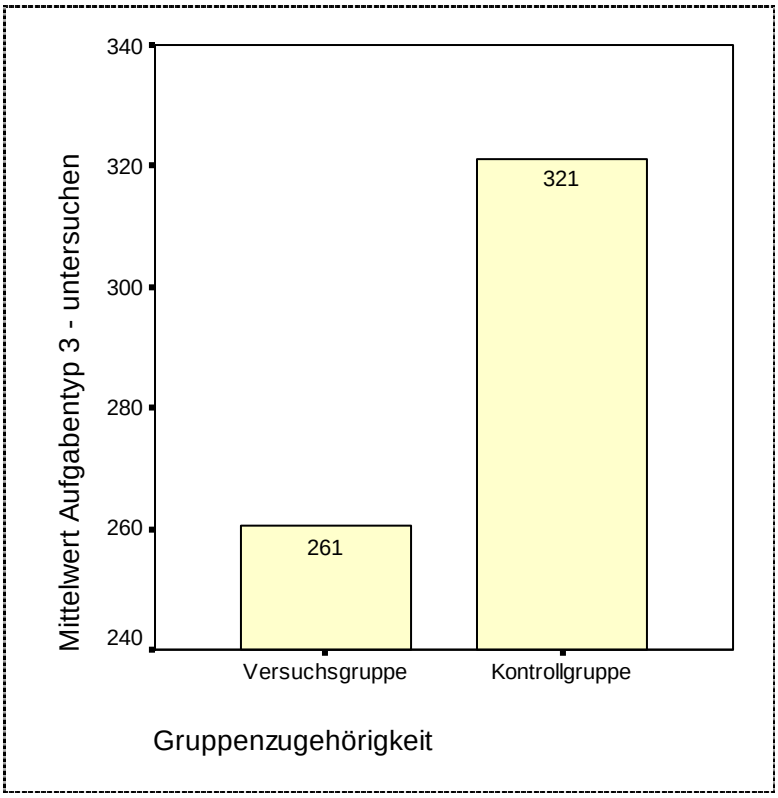


Abbildung 7.21: Vergleich der Gesamtzeiten für Aufgabentyp 3 nach Entfernen der Ausreißer

Aufgabentyp 3 - untersuchen

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	21918,844	1	21918,844	7,476	**,012**
Innerhalb der Gruppen	64503,239	22	2931,965		
Gesamt	86422,083	23			

Abbildung 7.22: Varianzanalyse für Aufgabentyp 3 nach Entfernen der Ausreißer

Die Auswertung nach Aufgabentypen konnte also die Stärken und Schwächen des Systems sehr deutlich herausstellen. So scheinen die Stärken der Leveltable recht deutlich bei Aufgabentyp 1, dem Suchen von Dokumenten innerhalb einer Ergebnismenge zu liegen und mit Vorbehalt auch bei Aufgabentyp 3, wenn die Ausreißer entfernt werden. Interessant ist insbesondere die Erkenntnis, dass dieser Aufgabentyp ganz allein für die entdeckten Ausreißer verantwortlich zu sein scheint. Um besser erkennen zu können, aus welchen Gründen, die Unterschiede innerhalb der einzelnen Aufgabentypen aufgetreten sind, sollen im Folgenden die aufgetretenen Fehler und Probleme der Teilnehmer genauer betrachtet werden.

7.5.2 Qualitative Fehleranalyse

Eine konkrete Fehleranalyse zu jeder Aufgabe findet sich in Anhang B, in welchem auch die einzelnen Aufgaben wortgetreu zu finden sind.

I. Aufgabentyp 1 – Dokumente suchen

Insgesamt konnten bei diesem Aufgabentyp die wenigsten Fehler und Probleme festgestellt werden. Wie bereits angesprochen können die Vorteile auf Seiten der *Leveltable* weitestgehend mit den Sortierfunktionen begründet werden. Allerdings führten diese bei einer der Aufgaben auch zu Problemen. Die Sortierung nach dem Titel des Dokuments beförderte Dokumente, welche mit einem Leerzeichen begannen (wovon leider einige in der Datenbank enthalten waren), ganz nach oben. Dadurch waren einige Teilnehmer verwirrt, da sie das Leerzeichen nicht entdeckten und sich wunderten, wieso das gesuchte Dokument bei dem entsprechenden Buchstaben nicht aufzufinden war. Weiterhin machte sich das Fehlen einer *History - Funktion* negativ bemerkbar – die Teilnehmer der Kontrollgruppe konnten signifikant schneller auf ein bereits gesuchtes Dokument erneut zugreifen.

II. Aufgabentyp 2 – Dokumente vergleichen

Insgesamt gesehen traten hier bereits deutlich mehr Fehler auf. Zwar hatten die Teilnehmer beider Gruppen quasi keine Probleme, die Aufgaben zu lösen, jedoch gab es einige falsche Antworten. Bei zwei der Aufgaben musste eine festgelegte Anzahl an Dokumenten hinsichtlich eines Meta-Attributes wie Größe oder Server-Typ miteinander verglichen werden. Für die Teilnehmer der Versuchsgruppe war es dazu notwendig, die Dokumente von Hand abzuzählen, da die *Leveltable* keine Nummerierung bot. Weiterhin war keine Doppelsortierung möglich, was einige der Teilnehmer aber scheinbar voraussetzten und somit das falsche Dokument auswählten. Diese beiden Problempunkte können wohl als Hauptkriterium dafür gesehen werden, dass die Kontrollgruppe zumindest bei einer dieser Aufgaben signifikant schneller war. Die beiden anderen Aufgaben erforderten kein Abzählen von einzelnen Dokumenten, weswegen hier wiederum die Vorteile auf Seiten der Versuchsgruppe liegen und diese bei einer der Aufgaben auch signifikant sind. Ebenfalls traten bei diesen auch deutlich weniger Fehler auf.

III. Aufgabentyp 3 – Dokumente inhaltlich versuchen

Wie bereits die deutlichen Ausreißer bei diesem Aufgabentyp vermuten lassen, kam es hier zu den meisten Fehlern. Problematisch zeigte sich hier vor allen Dingen das Themengebiet der Dokumente – Geographic Information Systems – mit welchem die Teilnehmer weitestgehend nicht vertraut waren. Dies führte in beiden Gruppen dazu, dass die Teilnehmer oftmals nur ungern ein Dokument durchlesen wollten und sich auch nur sehr schwer zu einer Antwort durchringen konnten. Sie waren sich teilweise einfach nicht sicher, ob sie die richtige Lösung gefunden hatten. Hierbei fällt auf, dass insbesondere die Teilnehmer der Kontrollgruppe die Aufgabe häufiger vorzeitig abbrachen, ohne das Dokument überhaupt genauer durchgelesen zu haben. Die Teilnehmer der Versuchsgruppe hingegen versuchten oftmals sehr intensiv, trotz der recht unübersichtlichen Dokumentendarstellung in der *Browserview*,

noch zu einer Lösung zu kommen. Dies könnte ein Hinweis darauf sein, dass die Visualisierung die Benutzer in ihrer Arbeit mehr motiviert, als eine herkömmliche listenbasierte Darstellung. Allerdings war dies nicht Teil dieser Untersuchung und ist somit nur als Hypothese zu betrachten, welche in einem separaten Test überprüft werden müsste.
Die Probleme mit der Thematik zeigten sich am deutlichsten bei der letzten Aufgabe. Bei dieser sollten die Teilnehmer eine Fragestellung beantworten, wofür sie selbständig mit Hilfe einer neuen Suchanfrage ein bestimmtes Dokument finden mussten, welches ihnen die benötigte Information zur Beantwortung der Frage lieferte. Innerhalb der Versuchsgruppe zeigten sich auch hier sehr deutlich die Mängel der Darstellung in der *Browserview* – zwar fanden fast alle Teilnehmer das richtige Dokument, jedoch brachen trotzdem sechs der 17 Personen die Aufgabe ab, ohne eine Lösung angeben zu können.

IV. Weitere Probleme bei der Benutzung der Leveltable

Neben den bereits angesprochenen Problemen konnten auch noch einige weitere identifiziert werden, welche nicht direkt die Leistung der Teilnehmer beeinflussten, jedoch trotzdem bezüglich der weiteren Entwicklung erwähnenswert sind.

Viele der Versuchspersonen hatten große Probleme bei der Bedienung der *Browserview* in Kombination mit der *Leveltable*. Die meisten Teilnehmer antizipierten, dass ein markiertes Dokument auch fest in der *Browserview* verankert sei. Dies muss jedoch explizit vom Benutzer per Kontextmenü veranlasst werden. Ansonsten führt das „überfahren“ eines anderen Dokumentes mit der Maus innerhalb der *Leveltable* unweigerlich dazu, dass nun dieses anstelle des Gewünschten in der *Browserview* angezeigt wird. Einige Teilnehmer bemerkten das Umspringen innerhalb der *Browserview* überhaupt nicht und untersuchten irrtümlich das falsche Dokument. Insbesondere unter Beachtung der Tatsache, dass alle Teilnehmer gute bis sehr gute Computerkenntnisse vorweisen konnten, ist hier in jedem Fall ein Umdenken dieser Funktionalität erforderlich.

Ebenfalls als problematisch erwiesen sich die fehlenden Filterfunktionen, die von einigen Teilnehmern explizit gewünscht wurden.

Hinsichtlich der Levelaufteilung in vier Stufen ergab der Test, dass insbesondere Level 3 von keinem Teilnehmer genutzt wurde. Mit großer Wahrscheinlichkeit ist das auf die Tatsache zurückzuführen, dass hier dem Benutzer kein wirklicher Mehrwert zu Level 2 geboten wurde und letztendlich Level 4 geeigneter war, um beispielsweise mit der Detailed Relevance Curve zu arbeiten. Aus diesem Grund wurde dieser Level im Anschluss an diesen Test neu designed und mit mehr Funktionalität versehen.

7.5.3 Post-Test Fragebogen Ergebnisse

Da mithilfe des Post-Test Fragebogen die Akzeptanz der *Leveltable* überprüft werden sollte, mussten lediglich die Teilnehmer der Versuchsgruppe einen solchen ausfüllen. Die Fragen können hierbei unterteilt werden in 7-Likert Skalen Fragen, welche sich zumindest bedingt auch quantitativ auswerten lassen und offenen Fragen, bei welchen die Teilnehmer Kommentare oder Anmerkungen frei eintragen konnten. Der genaue Wortlaut der Fragen befindet sich in Anhang C.

I. Quantitative Ergebnisse

Zunächst ist zu sagen, dass alle Teilnehmer mit der Sprache und Abgrenzung der Aufgaben gut zu recht kamen (Mittelwert bei 5,97 beziehungsweise 5,94).

Bezüglich der Visualisierung ergaben sich die folgenden Ergebnisse:

- Das Oberflächenlayout wurde als klar verständlich empfunden (Mittelwert von 5,61 auf 7-Likert-Skala, wobei *1 = Nein, das Layout war sehr unübersichtlich* und *7 = Ja, das Layout war sehr klar und übersichtlich*).
- Ebenso wurden die verwendeten Farben positiv bewertet – dies steht im Gegensatz zu den Ergebnissen der heuristischen Evaluation, in welcher die Farbgebung explizit kritisiert wurde (Mittelwert von 6,11 auf 7-Likert-Skala, wobei *1 = Nein, der Gebrauch der Farben war sehr unangenehm* und *7 = Ja, der Gebrauch der Farben war überaus angenehm*).
- Die Navigation wurde ebenfalls weitestgehend als intuitiv bewertet, allerdings nicht so deutlich wie die beiden oberen Punkte (Mittelwert 5,11 auf 7-Likert-Skala, wobei *1 = Nein, die Navigation war absolut nicht intuitiv* und *7 = Ja, die Navigation war überaus intuitiv*).
- Auf die Frage, ob sich die Teilnehmer manchmal verloren gefühlt hätten, waren die Antworten sehr unterschiedlich, was sich auch durch eine hohe Standardabweichung von 1,9 Punkten zeigt. Insgesamt ergab sich ein Mittelwert von 4,00 (7-Likert-Skala, wobei *1 = Nein, habe mich nie verloren gefühlt* und *7 = Ja, ich habe mich ständig verloren gefühlt*).

Weiterhin wurde gefragt, ob die Teilnehmer einerseits sich vorstellen könnten, mit der *Leveltable* täglich zu arbeiten und andererseits, ob sie sich Situationen vorstellen könnten, bei welchen eine solche Visualisierung gegenüber einer herkömmlichen Suchmaschine wie *Google* Vorteile bietet. Hier ergab sich ein klares positives Echo bezüglich der *Leveltable*. 14 der 18 Teilnehmer gaben an, sich vorstellen zu können mit einer solchen Visualisierung täglich zu arbeiten und alle Versuchspersonen waren der Meinung, dass es zumindest Situationen geben könnte, in welchem eine solche Visualisierung Vorteile biete. Dieses überaus positive Feedback ist doch recht erstaunlich für die recht kurze Testdauer, in welcher die Teilnehmer mit dem System arbeiten konnten.

II. Qualitatives Feedback

Zu den beiden letztgenannten Fragen konnten die Teilnehmer zusätzlich noch konkret beschreiben, in welchen Situationen die Visualisierung Vorteile bieten könnte und wieso sie das System in ihrer täglichen Arbeit verwenden würden. Weiterhin endete der Post-Test Fragebogen mit einer Frage nach allgemeiner Kritik am System. Die wichtigsten Anmerkungen der Teilnehmer werden im Folgenden im originalen Wortlaut wiedergegeben.

Anmerkungen zu der Frage, ob die Teilnehmer sich vorstellen könnten, mit dieser Visualisierung täglich zu arbeiten:

- Levelfunktionalität sehr gut (Antwort war JA)
- Höchstens wenn man anders (z.B. mit *Google*) nicht mehr weiterkommt. Man braucht mehr Klicks für die Suche (Antwort war NEIN)
- Nach Einarbeitungszeit ja, die Idee mit den Textpassagen gefällt mir (Antwort war JA)
- Eventuell nach mehr Einarbeitungszeit (Antwort war NEIN)

Anmerkungen zu der Frage, ob sich die Teilnehmer Situationen vorstellen könnten, in welchen die Visualisierung Vorteile gegenüber einer herkömmlichen Darstellung wie Google *bietet (sämtliche Antworten Ja):*

- leichteres Überblicken bei größerer Treffermenge
- Gute Auflistung der Relevanz der einzelnen Suchterme → erspart Durchklicken
- Passagensuche, Suchergebnisse vergleichen
- Sortierungsmöglichkeiten sehr gut
- Level 4 veranschaulicht, welche Worte zusammen auftauchen, macht google nicht – GUT

Generelle Kritik am System und weitere Anmerkungen:

- keine textbasierte FIND Option
- keine Nummerierung
- HTML Code in *Browserview* wirkt manchmal ein wenig kryptisch
- Keine Progress Bar bei Start der Suche, allgemein Feedback an den User zu gering
- Popup Hilfetexte z.B. zu den Levels, Kurzinfos wären gut
- Level 1 sehr unübersichtlich (Roll Over Effekt) – vielleicht unnötig?
- *BrowserView* zu unübersichtlich
- Etwas zu langsam
- Keine Doppelsortierung

Insgesamt ist auch hier das positive Feedback erstaunlich. Weiterhin konnten die bereits während dem Test beobachteten Probleme mit Hilfe dieser Fragen weitestgehend validiert werden.

Jens Gerken
Mat. 01/473804
Universität Konstanz

7.6 Kritik am Testdesign und Testablauf

Abseits der Kritik der Teilnehmer am System gibt es auch bezüglich des allgemeinen Testdesigns und des Testablaufs einige kritische Punkte zu bemerken. Zunächst sei hier noch einmal auf die Problematik der GISWeb Datenbank verwiesen, welche thematisch den Teilnehmern völlig unbekannt war. Um die Motivation der Teilnehmer auch bei komplexeren Aufgaben zu erhöhen, sollte hier für weitere Tests in jedem Fall eine Datenbasis verwendet werden, zu welcher die Teilnehmer auch einen persönlichen Bezug herstellen können, beispielsweise eine Filmdatenbank. Weiterhin war zu dem Testzeitpunkt noch ein recht einfacherer Retrievalalgorithmus implementiert, welcher nur bei bestimmten Suchanfragen – insbesondere bei der GISWeb Datenbank – wirklich sinnvolle Ergebnisse lieferte. Aus diesem Grund musste auf offene Retrievalfragen, bei welchen die Teilnehmer eigenständige Suchanfragen hätten stellen können, weitestgehend verzichtet werden.

In Hinblick auf das Testdesign ist der Baseline-Test im Nachhinein eher kritisch zu betrachten. Die Betrachtung der Ergebnisse konnte einen möglichen Lerneffekt innerhalb der Kontrollgruppe nicht vollends entkräften. Sollte dieser jedoch vorhanden gewesen sein, verstärkt das die Ergebnisse letztendlich nur noch weiter in Richtung Leveltable.

Bezüglich des Testablaufs muss kritisiert werden, dass durch wechselnde Protokollanten die Protokolle sich sehr stark unterschieden haben, weswegen auf eine nähere Effektivitätsanalyse verzichtet werden muss. Hierbei ist auch anzumerken, dass eine bereits im Vorfeld zu tätigende Standardisierung hinsichtlich dieser Fragestellung versäumt wurde.

Bezüglich der Hilfestellung scheint nachträglich ebenfalls eine zu geringe Standardisierung vorhanden gewesen zu sein. Eine feste Grenzzeit, bis zu welcher keine Hilfestellung gegeben wird, wäre hier vorteilhaft gewesen, da mit Hilfe beantwortete Aufgaben nicht in der Auswertung berücksichtigt wurden.

Jens Gerken
Mat. 01/473804
Universität Konstanz

7.7 Zusammenfassung: VisMeB Performance Test

Der Usability Test konnte eindrucksvoll zeigen, dass die Visualisierung in Form der *Leveltable* durchaus mit einer klassischen, listenbasierten Darstellung konkurrieren kann. Insbesondere unter Berücksichtigung dieser, für die meisten doch recht ungewohnten, Darstellung und der kurzen Einarbeitungszeit, gewinnt dieses Ergebnis an Aussagekraft. Die aufgestellte Null-Hypothese kann sowohl für einzelne Aufgaben als auch für ganze Aufgabentypen zu Gunsten der *Leveltable* verworfen werden. Unter Ausschluss der Ausreißer kann dieser Vorteil auf nahezu alle Aufgaben ausgeweitet werden. Da diese Ausreißer nur bei einem der Aufgabentypen auftreten, ist hier allerdings Vorsicht in Hinblick auf ein Entfernen dieser aus dem gesamten Test angebracht. Somit kann das Ergebnis betreffend der *Leveltable*, auch aufgrund des Teilnehmer-Feedbacks, als überaus positiv bewertet werden. Der Ruf nach einem Paradigmen Wechsel innerhalb der Suchsysteme, weg von der listenbasierten Darstellung, hin zu einem tabellarischen System scheint hinsichtlich dieser Ergebnisse durchaus berechtigt. Bezüglich weiterer Tests, wäre sicherlich die Auswirkung einer längeren und intensiveren Trainingsphase interessant. In diesem Fall könnten auch weitere Visualisierungen von VisMeB in den Test mit einbezogen werden, insbesondere *Circle Segment View* und *2D-Scatterplot*.

8 Ausblick

Die Ergebnisse der Usability Tests, welche im Rahmen der Entwicklung von VisMeB durchgeführt wurden und deren Ergebnisse Teil dieser Arbeit sind, haben deutlich gezeigt, dass zum einen durchaus eine Akzeptanz für tabellenbasierte Visualisierungen vorhanden ist und zum anderen, dass diese auch leistungsfähig genug sind, um die bisherigen listenbasierten Lösungen, beispielsweise im Bereich der Online Suchsysteme, in Frage zu stellen. Ähnliche Konzepte, welche ebenfalls eine tabellenbasierte Darstellung anbieten, wie etwa InfoZoom, zeigen zudem, dass die Ergebnisse dieser Evaluation kein Einzelfall sind. In Zeiten, da auch im Internet mit Hilfe von Java Applets und vergleichbaren Techniken umfangreiche Applikationen möglich sind, scheint somit ein Paradigmenwechsel, weg von der listenbasierten Darstellung traditioneller Suchsysteme, hin zu tabellenbasierten Darstellungen, nicht nur möglich, sondern auch notwendig, um der wachsenden Informationsflut gerecht zu werden. Interessant ist dabei, dass ein derartiger Wechsel nicht sprunghaft sondern durchaus schrittweise erfolgen könnte. In Zeiten, da es für Betreiber von Online Suchmaschinen immer schwieriger wird, sich durch bessere Ergebnisse von den Konkurrenten zu differenzieren, könnte eine tabellenartige Darstellung als Zusatzfunktion durchaus einen Markt beeinflussenden Vorteil bieten. Zudem offenbart ein Blick in die Suchfunktionalität des Windows Explorers, dass dieser bei der Präsentation der Ergebnisse ebenfalls bereits ein tabellenähnliches Format, mit umfangreichen Sortiermöglichkeiten nach verschiedenen Meta-Attributen, verwendet. Eine derartige Darstellung sollte somit auch dem unerfahrenen Benutzer nicht völlig unbekannt sein.

Mit Hilfe weiterer, öffentlich zugänglicher Evaluationen zu dieser Thematik, und natürlich auch der Entwicklung entsprechender Visualisierungswerkzeuge, könnten diese Thesen in Zukunft untermauert werden, so dass letztendlich nicht nur in der Forschung, sondern auch in der Praxis ein Paradigmenwechsel erfolgen kann.

In direktem Zusammenhang damit sollte die Entwicklung eines möglichst einheitlichen Testdesigns stehen. Dies könnte beispielsweise bei einer einheitlichen Definition von Aufgabentypen, welche allgemein als relevant für die Einstufung der Leistung eines visuellen Suchsystems anerkannt werden, beginnen. Weiterhin sollte der Umgang mit Faktoren, welche auf die Performance direkten Einfluss haben, wie etwa *Thinking Aloud* Techniken oder auch der Grad der Standardisierung des Tests allgemein festgelegt werden. Auf lange Sicht könnte somit ein Verfahren für Usability Tests visueller Suchsysteme, ähnlich dem GIRT, entwickelt werden, welches die Aussagekraft einer einzelnen Evaluation deutlich erhöhen würde. Angesichts der jetzigen Situation, in der jede Evaluation für sich allein betrachtet werden muss und somit oftmals nur innerhalb des speziellen Anwendungskontextes Gültigkeit besitzt, wäre dies ein enormer Fortschritt für die Bedeutung von Usability Tests auch im Allgemeinen als eigenständiges wissenschaftliches Forschungsgebiet.

9 Referenzen

9.1 Quellenverzeichnis

Statistik

[BEPW00] K. Backhaus, B. Erichson, W. Plinke, R. Weiber: „Multivariate Analysemethoden“, 9. Auflage, 2000

[FKPT99] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, Universität München: „Statistik – der Weg zur Datenanalyse“, 1999

[Gar03] G. David Garson, North Carolina State University: “Multivariate Analysis for Applied Social Science”, Work in progress, current status: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm, 2003, online am 15.03.2004

[Lane03] David M. Lane, Rice University: „HyperStat Online Textbook“,http://davidmlane.com/hyperstat/index.html, 2003, online am 15.03.2004

[Stock96] David W. Stockburger, Southwest Missouri State University: “Introductory Statistics: Concepts, Models and Applications” WWW Version 1.00: http://www.psychstat.smsu.edu/introbook/sbk00.htm, 1996, online am 15.03.2004

Sozialwissenschaften

[AMR89] R. Aster, H. Merkens, M. Repp: „Teilnehmende Beobachtung: Werkstattberichte und methodologische Reflexionen“, 1989

[Att95] Peter Atteslander: „Methoden der empirischen Sozialforschung“, 1995

[BD95] Jürgen Bortz, Nicola Döring: Forschungmethoden und Evaluation für Sozialwissenschaftler, 1995

[GS84] B. G. Glaser, A. L. Strauss „Die Entdeckung gegenstandsbezogener Theorie: Eine Grundstrategie qualitativer Sozialforschung“, 1984

[Klei86] Gerhard Kleining: „Das qualitative Experiment“, 1986

[Klei94a] Gerhard Kleining: „Qualitativ-heuristische Sozialforschung: Schriften zur Theorie und Praxis“, 1994a

[May00] Philipp Mayring: „Qualitative Inhaltsanalyse“, 2000

[May90] Philipp Mayring: „Einführung in die qualitative Sozialforschung“, 1990

[Mil74] S. Milgram: „Das Milgram-Experiment - Zur Gehorsamkeitsbereitschaft gegenüber Autorität“, 1974

[Ross02] Raphael Rossmann: Das Sozialwissenschaftliche Experiment, http://www.raphael-rossmann.de/exp.htm, 2002, online am 15.03.2004

[Schäf95] Jutta Schäfer: „Glossar qualitativer Verfahren“, Berliner Public Health Zentrum, 1995

Usability

[DR99] Joseph S. Dumas, Janice A. Redish: “A Practical Guide to Usability Testing”, 1999

[Ger03] Jens Gerken: „Validität und Aussagekraft von Usability Test Methoden“, Seminararbeit, Universität Konstanz, 2003

[KR97] Laurie Kantner, Stephanie Rosenbaum: „Heuristic Evaluation vs. Laboratory Testing“, 1997

[McNam99] Carter McNamara: “Conducting Focus Groups”, http://www.mapnp.org/library/grp_skll/focusgrp/focusgrp.htm, 1999, online am 15.03.2004

[Niel94] Jakob Nielsen: „How to conduct a Heuristic Evaluation”, http://www.useit.com/papers/heuristic/heuristic_evaluation.html, 1994, online am 15.03.2004

[Niel94b] Jakob Nielsen: “Ten Usability Heuristics”, http://www.useit.com/papers/heuristic/heuristic_list.html, 1994b, online am 15.03.2004

[Niel97] Jakob Nielsen: „The Use and Misuse of Focus Groups“, http://www.useit.com/papers/focusgroups.html, 1997, online am 15.03.2004

[Shnei98] Ben Shneiderman: „Designing the User Interface“, 1998

[Usab03] UsabilityNET EU Project: „Usability Resources for Practitioners and Managers”, http://www.usabilitynet.org, 2003, online am 15.03.2004

Evaluationspaper visueller Suchsysteme

[CK00] Ewa Callahan, Jürgen Koenemann: „A Comparative Usability Evaluation of User Interfaces for Online Product Catalogs“. Proceedings of the 2nd ACM conference on Electronic commerce, 2000.

[Eibl00] Maximilian Eibl: „Visualisierung im Document Retrieval“, 2000

[EGP01] Jennifer English, Kim Garrett & Sacha Pearson: „Visual Design Experiment“, 2001

[STWB94] Robert Spence, Lisa Tweedie, David Williams, Ravinder Bhoga: „The Attribute Explorer“, 1994

[SVMCL99] Marc M. Sebrechts, Joanna Vasilakis, Michael S. Miller, John V. Cugini, Sharon J. Laskowski: „Visualization of Search Results: A Comparative Evaluation of Text, 2D,

and 3D Interfaces". In Proceedings of 22nd ACM SIGIR conference on Research and development in information retrieval, 1999.

Visuelle Suchsysteme

[hum04] humanIT InfoZoom, Sehen – Wissen – Entscheiden, http://www.humanit.de/de/produkte_loesungen/products/iz/index.html, online am 15.03.2004

[IBM04] IBM Visual Attribute Explorer, A dynamic query mechanism, http://www.alphaworks.ibm.com/tech/visualexplorer, online am 15.03.2004

[Kart04] Kartoo – Meta-Suchmaschine, http://www.kartoo.com, online am 15.03.2004

Veröffentlichungen zu INSYDER/INVISIP/VisMeB

[GHRM02] S. Göbel, J. Haist, H. Reiterer, F. Müller: „INVISIP: Metadata-based Information Visualization Techniques to Access Geodata Archives and to Support the Site Plannung Process", 3. CO-DATA Euro-American Workshop, 2002, Juli 10-11, Paris, France

[INSY04] INSYDER, Internet Système de Recherche, Europäische Kommission, ESPRIT, Projekt Nr. 29232, http://www.insyder.com, online am 15.03.2004

[INVI04] INVISIP, Information Visualization for Site Planning, Europäische Kommission, Projekt Nr. IST-2000-29640, http://www.invisip.de, online am 15.03.2004

[Jett03] Christian Jetter: „Usability Evaluation im Rahmen von INVISIP", Bachelor-Arbeit, Universität Konstanz, 2003

[KMRE02] P. Klein, F. Müller, H. Reiterer, M. Eibl, „Visual Information Retrieval with the SuperTable + Scatterplot", Proceedings of the 6th International Conference on Information Visualisation (IV 02), IEEE Computer Society, 2002.

[Koen03] Werner Koenig: „Konzeption und Implementation eines 3D-Scatterplots zur Visualisierung von Metadaten", Bachelor-Arbeit, Universität Konstanz, 2003

[KRML03] P. Klein, H. Reiterer, F. Müller, T. Limbach, „Metadata Visualization with VisMeB", IV03, 7th International Conference on Information Visualization, London, 2003.

[Lieb03] Philipp Liebrenz: „Visualisierung von Multi-Data-Points in einem 3D-Scatterplot", Bachelor-Arbeit, Universität Konstanz, 2003

[Mann02] Thomas Mann: „Visualization of Search Results from the WWW", Dissertation, Universität Konstanz, 2002

[RMMH00] H. Reiterer, G. Mußler, T. Mann, S. Handschuh, „Insyder - An Information Assistant for Business Intelligence", Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, Athens, 2000.

[VisM04] VisMeB, A Visual Metadata Browser for Visual Data and Text Mining, AG Mensch-Computer Interaktion, Universität Konstanz, http://www.inf.uni-konstanz.de/iw_is/Forschung/vismeb/, online am 14.03.2004.

9.2 Abbildungsverzeichnis

10 Anhang

10.1 Anhang A: Pre-Test Fragebogen

Frage 1:

Wie viele Stunden am Tag benutzen Sie einen Computer?

Frage 2:

Fällt es Ihnen leicht sich mit neuer Software vertraut zu machen?

(1 bedeutet „nein, fällt mir eher schwer“, 7 bedeutet „ja, bereitet mir keine Probleme“)

Frage 3:

Wie schätzen Sie Ihre Erfahrung mit Computern allgemein auf einer Skale von 1-7 ein? (1 bedeutet „keine Erfahrung“, 7 bedeutet „sehr viel Erfahrung“)

Frage 4:

Wie würden Sie ihre Erfahrung mit Internet Suchmaschinen auf einer Skala von 1-7 einschätzen?

(1 bedeutet „keine Erfahrung“ und 7 bedeutet „sehr viel Erfahrung“)

Frage 5:

Wie würden Sie Ihre generelle Einstellung zu Computern auf einer Skala von 1-7 einschätzen?

(1 bedeutet „arbeite sehr ungern mit Computern“ und 7 bedeutet „arbeite sehr gerne mit Computern“)

Mittelwerte

GRUPPE	Alter	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5
Versuchsgruppe	25,56	5,06	5,33	5,50	5,33	5,83
Kontrollgruppe	26,29	4,77	4,21	4,57	4,14	5,21
Insgesamt	25,88	4,93	4,84	5,09	4,81	5,56

Weitere Merkmale der Versuchspersonen

Geschlecht

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	männlich	24	75,0	75,0	75,0
	weiblich	8	25,0	25,0	100,0
	Gesamt	32	100,0	100,0	

Schulbildung

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Gymnasium	28	87,5	87,5	87,5
	Realschule	1	3,1	3,1	90,6
	Sonstiges	3	9,4	9,4	100,0
	Gesamt	32	100,0	100,0	

Studium/Ausbildung/Beruf

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	BSc IE	7	21,9	21,9	21,9
	Dipl. Infwissenschaften	1	3,1	3,1	25,0
	Informatik	1	3,1	3,1	28,1
	Informatikkaufmann	1	3,1	3,1	31,3
	Jura	3	9,4	9,4	40,6
	Literaturwissenschaften/ Inf. Wiss.	1	3,1	3,1	43,8
	Maschinenbau	4	12,5	12,5	56,3
	MSc IE	5	15,6	15,6	71,9
	Pharmazie	1	3,1	3,1	75,0
	Physik/Englisch	1	3,1	3,1	78,1
	selbständig	2	6,3	6,3	84,4
	Sport/Mathe/Informatik	1	3,1	3,1	87,5
	Verwaltungswissenschaften	3	9,4	9,4	96,9
	wissenschaftlicher Mitarbeiter	1	3,1	3,1	100,0
	Gesamt	32	100,0	100,0	

Wie lang arbeiten Sie mit PCs

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Länger als ein Jahr	32	100,0	100,0	100,0

Welches Betriebssystem verwenden Sie?

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Win9x	6	18,8	18,8	18,8
	Win2k/XP	24	75,0	75,0	93,8
	Linux	2	6,3	6,3	100,0
	Gesamt	32	100,0	100,0	

Haben Sie Freunde mit besonderer Computererfahrung?

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Ja	31	96,9	96,9	96,9
	Nein	1	3,1	3,1	100,0
	Gesamt	32	100,0	100,0	

Benutzen Sie Internet Suchmaschinen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Ja	32	100,0	100,0	100,0
	Nein	0	0	100,0	100,0
	Gesamt	32	100,0	100,0	

Name der meistbenutzten Suchmaschine (Freitext)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Google	32	100,0	100,0	100,0

10.2 Anhang B: Performance Test Ergebnisse

10.2.1 Original Wortlaut der Testaufgaben

Aufgabe 1:

Da diese Aufgabe lediglich das Starten der Suche beinhaltete, wurde sie für die Auswertung nicht berücksichtigt.

Aufgabe 2:

„Suchen Sie das Dokument mit der höchsten Relevanz und nennen sie den Titel“

Fehleranalyse:
Kontrollgruppe: 1 Fehler
Versuchsgruppe: 0 Fehler

Kontrollgruppe:
Die Aufgabe wurde bis auf eine Ausnahme von allen Teilnehmern korrekt gelöst. Eine VP der Kontrollgruppe entschied sich für das 2. Dokument in der Liste, da ihrer Meinung nach der Titel mehr der Suchanfrage entsprach.

Aufgabe 3:
„Vergleichen Sie die Dateigröße der ersten 10 Dokumente (geordnet nach Relevanz – dieser Zusatz existierte nur bei Verwendung der Visualisierung) – welches ist das größte?“

Fehleranalyse:
Kontrollgruppe: 0 Fehler
Versuchsgruppe: 4 Fehler

Versuchsgruppe:
Zwei VP wählten einfach das falsche Dokument aus, zwei weitere sortierten die Tabelle nach Größe und wählten das insgesamt größte Dokument aus.

Aufgabe 4:
„In welcher Sprache ist dieses Dokument vorhanden und wie lautet der server_type?“

Fehleranalyse:
Keine Fehler

Aufgabe 5:
„Sie suchen ein deutschsprachiges Dokument. Überprüfen Sie ob in der Ergebnismenge eines vorhanden ist“ (bei der Listendarstellung wurde der Ergebnisraum auf 100 Dokumente eingeschränkt, da hier ansonsten ein zu großer Nachteil entstanden wäre)

Fehleranalyse:
Keine Fehler

Aufgabe 6:
„Vergleichen Sie den server_type der ersten 20 (nach Relevanz sortiert – Hinweis bei Versuchsgruppe) Dokumente. Welcher scheint hier vornehmlich vorhanden zu sein?“

Fehleranalyse:

Kontrollgruppe: 0 Fehler
Versuchsgruppe: 3 Fehler

Versuchsgruppe:
Wie bei der ähnlichen Aufgabe 3 hatte die Kontrollgruppe keinerlei Probleme. Bei der Versuchsgruppe hatten drei VP Probleme die Aufgabe richtig zu beantworten. Dabei ist auch eine VP, die bereits Aufgabe 3 nicht korrekt beantworten konnte. Die Versuchspersonen entschieden sich alle drei für den falschen server_type, obwohl .com relativ offensichtlich ist.

Aufgabe 7:
„Suchen Sie das Dokument – GISLinx – What is a GIS (Zusatz Kontrollgruppe: Es ist innerhalb der ersten 100 Dokumente enthalten)“

Fehleranalyse:
Kontrollgruppe: 1 Fehler
Versuchsgruppe: 2 Fehler

Kontrollgruppe:
Eine VP fand das Dokument nicht, erst nach Tipp auf der ersten Seite noch mal nachzusehen.

Versuchsgruppe:
2 VP fanden das Dokument nicht, da bei Sortierung nach Titel zu oberst die Dokumente stehen, welche mit einem Leerzeichen beginnen. Erst nach mehrmaligen Hinweisen auf diese Tatsache fanden beide VP das Dokument.

Aufgabe 8:
„Bietet das Dokument GISLinx – What is a GIS Informationen bezüglich des Suchterms „Design“ (ist das Wort darin enthalten)?“

Fehleranalyse:
Kontrollgruppe: 5 Fehler
Versuchsgruppe 2 Fehler

Kontrollgruppe:
Die Kontrollgruppe hatte deutliche Probleme die Aufgabe zu lösen. Die Dokumentenansicht der Listendarstellung bot keine Suchfunktion, weswegen das Dokument durchgelesen werden musste. Einige Teilnehmer hatten dazu keine Lust und brachen einfach sofort ab (VP6,VP14,VP18,VP29). Eine Versuchsperson (VP27) durchsuchte das Dokument zwar sehr lange (über zweieinhalb Minuten), konnte sich am Ende aber nicht entscheiden.

Versuchsgruppe:

Bei der Versuchsgruppe ergibt sich ein anderes Bild. Eine Versuchsperson (VP15) brach die Aufgabe nach 2 Minuten 40 Sekunden ab, ohne die Lösung zu nennen. Eine weitere (VP24) fand sich in der Browserview nicht zurecht - nach Tipp vom VL doch mal die Farbkodierung zu beachten, konnte sie die Aufgabe schließlich aber doch korrekt lösen.

Aufgabe 9:

„Vergleichen Sie alle Dokumente – welches ist das kleinste Dokument (ausgenommen Dokumente mit einem Wert von „–1“)?“ Anmerkung: Die Kontrollgruppe durfte sich auf die ersten 20 Dokumente beschränken, da der Aufwand hier ansonsten unverhältnismäßig hoch gewesen wäre

Fehleranalyse:
Kontrollgruppe: 2 Fehler
Versuchsgruppe: 2 Fehler

Kontrollgruppe:
2 VP (VP31 & VP6) entschieden sich für das falsche Dokument.

Versuchsgruppe:
Bei einer VP (VP1) war diese Aufgabe noch nicht in dieser Form im Test enthalten, die andere VP (VP9) suchte anstelle des kleinstem Dokument das größte und fand hier auch das richtige, insofern auch nur ein Missverständnis der Frage.

Aufgabe 10:

„Suchen Sie erneut das Dokument – GISLinx – What is a GIS“

Fehleranalyse:
Kontrollgruppe: 1 Fehler
Versuchsgruppe: 0 Fehler

Kontrollgruppe:
VP6 überspringt Aufgabe aus Versehen

Aufgabe 11:

„Finden Sie mindestens 5 Dokumente mit dem Server_type .edu“

Fehleranalyse:
Keine Fehler

Aufgabe 12:

„Versuchen Sie aus diesen das größte Dokument auszuwählen“

Fehleranalyse:

Kontrollgruppe: 3 Fehler

Versuchsgruppe: 1 Fehler

Kontrollgruppe:

3 Versuchspersonen (VP6, VP8 & VP20) entschieden sich für das falsche Dokument.

Versuchsgruppe:

1 Versuchsperson verstand die Aufgabe nicht und brach sie deswegen ab.

Aufgabe 13:

„Versuchen Sie heraus zu finden, welche Online Mapping Systeme von Geoscience Australia angeboten werden. Nennen Sie mindestens 3. Um die Aufgabe zu lösen können Sie eine neue Suchanfrage stellen.“

Fehleranalyse:

Kontrollgruppe: 5 Fehler

Versuchsgruppe: 6 Fehler

Kontrollgruppe:

Zwei Versuchspersonen brechen die Aufgabe nach einiger Zeit ab ohne eine Lösung zu nennen. Zwei weitere sind zunächst der Meinung, mehrere Dokumente finden zu müssen – nach Hilfe finden sie das richtige Dokument, können aber Aufgrund der Thematik nur raten. Wie die 5. VP nennen sie letztendlich zwei richtige Lösungen, brechen die Aufgabe aber dennoch ab, da sie nicht sicher sind.

Versuchsgruppe:

Zwei Versuchspersonen haben bereits Probleme das richtige Dokument zu finden und erhalten hierbei kleine Hilfestellungen (Dokument suchen, welches vom Titel bereits ähnlich der Suchanfrage ist). Eine der beiden VP kann daraufhin auch die Aufgabe noch korrekt lösen. Eine weitere VP findet das richtige Dokument, ist sich dessen aber nicht bewusst und weiß nicht richtig, was sie machen soll – kleiner tipp dass VP sich im richtigen Dokument befindet und nur die Systeme nennen soll hilft ihr soweit, dass sie die Aufgabe noch korrekt lösen kann. Drei weitere VP finden zwar das richtige Dokument, brechen dann aber ab ohne Lösung zu nennen.

10.2.2 Statistische Auswertung

Aufgabentypen:

1. Dokumente suchen (Aufgaben: 2, 5, 7, 10, 11)
2. Dokumente vergleichen (Aufgaben: 3, 6, 9, 12)
3. Dokumente inhaltlich untersuchen (Aufgaben: 4,8,13)

Zusammenfassung für Einzelaufgaben

	Kontrollgruppe (KG)	Versuchsgruppe (VG)	Unterschied signifikant?	Typ	n (KG/VG)
Aufgabe 2	14,1s	**8,2s**	ja	1	10 / 13
Aufgabe 3	**17,8s**	25,5s	ja	2	9 / 11
Aufgabe 4	4s	**2,4s**	ja	3	13 / 11
Aufgabe 5	73,8s	**17,3s**	ja	1	13 / 11
Aufgabe 6	28,4s	23,5	nein	2	11 / 11
Aufgabe 7	15,8s	33,2s	nein	1	13 / 10
Aufgabe 8	128s	**22,3s**	ja	3	11 / 7
Aufgabe 9	37,9s	**13,4s**	ja	2	12 / 9
Aufgabe 10	**2,1s**	15,7s	ja	1	13 / 10
Aufgabe 11	22,4s	**7,9s**	ja	1	13 / 11
Aufgabe 12	39,6s	49,6s	nein	2	12 / 8
Aufgabe 13	189s	234s	nein	3	7 / 6

Deskriptive Statistiken im Detail für Einzelaufgaben

	N	Minimum	Maximum	Mittelwert	Standardabweichung
Teil 2 - Aufgabe 2	29	1	30	11,03	7,953
Teil 2 - Aufgabe 3	26	4	85	23,81	15,425
Teil 2 - Aufgabe 4	30	1	8	3,60	1,793
Teil 2 - Aufgabe 5	30	2	101	42,07	30,101
Teil 2 - Aufgabe 6	27	11	135	34,00	27,862
Teil 2 - Aufgabe 7	27	2	116	26,22	25,328
Teil 2 - Aufgabe 8	23	1	226	61,78	66,511
Teil 2 - Aufgabe 9	26	5	65	25,35	15,484
Teil 2 - Aufgabe 10	29	1	60	9,93	12,180
Teil 2 - Aufgabe 11	30	2	45	14,20	10,733
Teil 2 - Aufgabe 12	26	5	120	44,35	27,058
Teil 2 - Aufgabe 13	19	51	705	290,95	207,930
Aufgabentyp 1 - suchen	30	28,00	173,00	103,4956	39,83071
Aufgabentyp 2 - vergleichen	30	52,00	255,00	128,8083	43,93715
Aufgabentyp 3 - untersuchen	30	63,00	866,00	362,9658	170,75783
Gesamtzeit in Sekunden	30	233,00	1224,27	595,2697	215,36336
Gültige Werte (Listenweise)	8				

Varianzanalyse im Detail für Einzelaufgaben

		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Aufgabe 2	Zwischen den Gruppen	190,28	1	190,28	3,25	,083
	Innerhalb der Gruppen	1580,68	27	58,54		
	Gesamt	1770,97	28			
Aufgabe 3	Zwischen den Gruppen	1204,96	1	1204,96	6,10	,021
	Innerhalb der Gruppen	4743,08	24	197,63		
	Gesamt	5948,04	25			
Aufgabe 4	Zwischen den Gruppen	9,13	1	9,13	3,04	,092
	Innerhalb der Gruppen	84,07	28	3,00		
	Gesamt	93,20	29			
Aufgabe 5	Zwischen den Gruppen	19926,40	1	19926,40	87,87	,000
	Innerhalb der Gruppen	6349,47	28	226,77		
	Gesamt	26275,87	29			
Aufgabe 6	Zwischen den Gruppen	973,34	1	973,34	1,27	,271
	Innerhalb der Gruppen	19210,66	25	768,43		
	Gesamt	20184,00	26			
Aufgabe 7	Zwischen den Gruppen	2184,07	1	2184,07	3,77	,064
	Innerhalb der Gruppen	14494,60	25	579,78		
	Gesamt	16678,67	26			
Aufgabe 8	Zwischen den Gruppen	48250,58	1	48250,58	20,65	,000
	Innerhalb der Gruppen	49071,33	21	2336,73		
	Gesamt	97321,91	22			
Aufgabe 9	Zwischen den Gruppen	2753,61	1	2753,61	20,40	,000
	Innerhalb der Gruppen	3240,28	24	135,01		
	Gesamt	5993,88	25			
Aufgabe 10	Zwischen den Gruppen	1260,71	1	1260,71	11,77	,002
	Innerhalb der Gruppen	2893,15	27	107,15		
	Gesamt	4153,86	28			
Aufgabe 11	Zwischen den Gruppen	1060,80	1	1060,80	13,03	,001
	Innerhalb der Gruppen	2280,00	28	81,43		
	Gesamt	3340,80	29			
Aufgabe 12	Zwischen den Gruppen	397,55	1	397,55	,53	,472
	Innerhalb der Gruppen	17906,34	24	746,10		
	Gesamt	18303,88	25			
Aufgabe 13	Zwischen den Gruppen	77359,53	1	77359,53	1,88	,189
	Innerhalb der Gruppen	700865,42	17	41227,38		
	Gesamt	778224,95	18			
Aufgabentyp 1 - suchen	Zwischen den Gruppen	10482,82	1	10482,82	8,26	,008
	Innerhalb der Gruppen	35525,26	28	1268,76		
	Gesamt	46008,09	29			
Aufgabentyp 2 - vergleichen	Zwischen den Gruppen	1214,33	1	1214,33	,62	,437
	Innerhalb der Gruppen	54769,39	28	1956,05		
	Gesamt	55983,72	29			

Aufgabentyp 3 - untersuchen	Zwischen den Gruppen	7524,08	1	7524,08	,25	,620
	Innerhalb der Gruppen	838064,77	28	29930,88		
	Gesamt	845588,85	29			
Gesamtzeit in Sekunden	Zwischen den Gruppen	368,76	1	368,76	,01	,931
	Innerhalb der Gruppen	1344691,15	28	48024,68		
	Gesamt	1345059,91	29			

10.3 Anhang C: Post-Test Fragebogen

Lediglich die Teilnehmer der Versuchsgruppe (18) mussten den Post-Test Fragebogen ausfüllen.

Für alle Fragen galt die Vorgabe:

1 bedeutet „nicht mit einverstanden“ und 7 bedeutet „Ja, das stimmt!“

Frage 1: War das generelle Layout der Oberfläche klar?

Frage 2: War der Gebrauch von Farben angenehm?

Frage 3: War die Navigation intuitiv?

Frage 4: Haben Sie sich manchmal „verloren“ gefühlt?

Frage 5: War auf einzelnen Seiten die Informationsflut zu hoch?

Frage 6: Boten einige Seiten zu wenig Informationen?

Frage 7: War die verwendete Terminologie verständlich?

Mittelwerte der Likert-Skalen

GRUPPE	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6	Frage 7
Versuchsgruppe	5,61	6,11	5,11	4,00	4,17	3,00	5,53

Frage 8: Könnten Sie sich vorstellen, mit dieser Visualisierung (Leveltable) täglich zu arbeiten?
Ergebnis: 14 x Ja, 4 x Nein
Frage 9: Könnten Sie sich Situationen vorstellen, in denen eine derartige Visualisierung (Leveltable) einer herkömmlichen, listenbasierten Darstellung (wie Google) überlegen ist?
Ergebnis: 18 x Ja