



Fachbereich 4: Informatik

Summative Evaluation zur Semantischen Annotation von Bildern

Masterarbeit
zur Erlangung des Grades
MASTER OF SCIENCE

vorgelegt von

Lydia Rebecca Weiland

Betreuer: Jun.-Prof. Dr. Ansgar Scherp, Web Science and Technologies,
Fachbereich Informatik, Universität Koblenz-Landau

Erstgutachter: Jun.-Prof Dr. Ansgar Scherp, Web Science and
Technologies, Fachbereich Informatik, Universität Koblenz-Landau

Zweitgutachterin: Dipl.-Inf. Tina Walber

Koblenz, im November 2012

Kurzfassung

Der Bereich der Annotation von digitalen Medien ist umfassend erforscht. Es gibt viele verschiedene innovative Ideen Annotationsprozesse zu gestalten. Den wahrscheinlich größten Anteil an verwandten Arbeiten hat die semiautomatische Annotation. Eine Eigenschaft aber ist allen Arbeiten gemein: Keine dieser Arbeiten untersucht, wie man ein Interface für einen Annotationsprozess optimieren kann, um den Nutzer zu unterstützen und folglich zufriedener zu machen. Die hier vorliegende Arbeit versucht im Rahmen einer Benutzerstudie die Usability Eigenschaften von verschiedenen Interface Varianten zu vergleichen, um anschließend Aussagen darüber machen zu können, welche Funktionen ein Annotationsinterface haben sollte und darf, um eben genau diese betreffenden Anforderungen zu erfüllen. Hierzu wurden verschiedene Ideen für Interfaces gesammelt, im Rahmen einer Diskussionsrunde mit Experten erörtert und letztendlich basierend auf den verbliebenen Ideen Variablen festgelegt, mit deren Kombination untereinander verschiedene Interfaces entstanden sind. Im Rahmen dieser Nutzerstudie konnte die Tendenz festgestellt werden, dass eine Autocompletion Funktion und Tag-Vorschläge während der Annotation hilfreich für den Nutzer sind. Außerdem konnte festgestellt werden, dass eine farbige Kennzeichnungen von Tag-Typen nicht störend auf den Nutzer wirken, sondern sich Ansätze in Richtung hilfreich erwiesen.

Ähnliche Tendenzen wurden für ein Layout mit nur 2 User Interface Elementen aus den Daten abgelesen. Des weiteren konnte mit Hilfe der Evaluation ein weiteres Beispiel dafür gefunden werden, dass es zum Teil erhebliche Unterschiede zwischen den Nutzern gibt, was intuitiv ist und was nicht. Für die Gestaltung von zukünftigen Annotationsinterfaces bedeuten die Erkenntnisse, dass weitere Nutzerstudien im Bereich der Annotationsinterfaces durchgeführt und bei der Gestaltung der Interfaces von klassischen Merkmalen, im Sinne der Nutzerzufriedenheit, abgewichen werden darf.

Abstract

The annotation of digital media is no new area of research, instead it is widely investigated. There are many innovative ideas for creating the process of annotation. The most extensive segment of related work is about semi automatic annotation. One characteristic is common in the related work: None of them put the user in focus. If you want to build an interface, which is supporting and satisfying the user, you will have to do a user evaluation first. Within this thesis we want to analyze, which features an interface should or should not have to meet these requirements of support, user satisfaction and being intuitive.

After collecting many ideas and arguing with a team of experts, we determined only a few of them. Different combination of these determined variables form the interfaces, we have to investigate in our usability study. The results of the usability leads to the assumption, that autocompletion and suggestion features supports the user. Furthermore coloring tags for grouping them into categories is not disturbing to the user, but has a tendency of being supportive. Same tendencies emerge for an interface consisting of two user interface elements. There is also an example given for the definition differences of being intuitive.

This thesis leads to the conclusion that for reasons of user satisfaction and support it is allowed to differ from classical annotation interface features and to implement further usability studies in the section of annotation interfaces.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Vereinbarung der Arbeitsgruppe für Studien- und Abschlussarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einver- ja nein standen.

Der Veröffentlichung dieser Arbeit im Internet stimme ich zu. ja nein

Koblenz, den 2. November 2012

Inhaltsverzeichnis

1 Einleitung	13
1.1 Motivation	13
1.2 Zielsetzung	14
1.3 Aufbau der Arbeit	15
2 Hintergrund	17
2.1 Manuelle Annotationsverfahren	19
2.2 Automatische Annotation	21
2.3 Kombination von automatischer und manueller Annotation	22
2.4 Benutzerstudien im Bereich der Annotationsinterfaces	29
3 Evaluationsdesign	33
3.1 Bestimmen der Hypothesen	33
3.1.1 Nullhypotesen	33
3.1.2 Unabhängige Variablen	33
3.1.3 Abhängige Variablen	33
3.1.4 Forschungshypothesen	34
3.2 Ablauf und Testbedingungen	34
3.3 Aufgabenliste	35
3.4 Datensätze	36
3.5 Datenerfassung	38
3.6 Fragebögen	39
3.7 Testumgebung	42
3.8 Implementation	52
4 Ergebnisse	55
4.1 Teilnehmer	57
4.2 Effektivität	57
4.3 Effizienz	69
4.4 Benutzerzufriedenheit	72
4.5 Diskussion	83

4.6	Schlussfolgerung	84
5	Fazit	87
5.1	Zusammenfassendes Fazit	87
5.2	Ausblick	88
6	Anhang A	91
6.1	Implementation	91
6.1.1	Basissystem Kreuzverweis	91
6.1.2	Lokaler Webserver	92
6.1.3	Kreuzverweis Server	92
7	Anhang B	95
7.1	Quellennachweise zu den Bildsets	95

Tabellenverzeichnis

3.1	Die vier verschiedenen Interfaces mit zwei unterschiedlichen Bildsets. Latin Square aus Bildsets, Typinformation und Layout	35
3.2	Die zehn Fragen der interfacebezogenen Fragebögen	41
3.3	Diskutierte und zum Teil begründet verworfene Ideen	45
4.1	Übersicht über die SPSS Variablen und ihre Bedeutungen	56

Abbildungsverzeichnis

2.1	Die 4 Voxalead Kategorien	19
2.2	Die 10 MPEG-7 Bestandteile	20
2.3	Das Caliph Interface	21
2.4	Autocompletion Gruppierungsarten	30
3.1	Bilderauswahl aus Bildset B1, Quellen siehe Anhang B	37
3.2	Bilderauswahl aus Bildset B2, Quellen siehe Anhang B	37
3.3	Frage 1 des Fragebogens mit Antwortskala	39
3.4	Frage 1 des Fragebogens nach Beantwortung mit Schieberegler	39
3.5	Das Kreuzverweis Webinterface	42
3.6	Darstellungsformen von Tag Clouds [LZT09]	43
3.7	Interface mit 3 User Interface Elementen und Vorschlägen	48
3.8	Interface mit 3 User Interface Elementen, Beschriftung auf Deutsch	49
3.9	Interface mit 2 User Interface Elementen, Beschriftung auf Deutsch	50
3.10	Informationsmeldung zu den Interfaces mit Typinformation	54
4.1	Mittelwert der gesetzten Tags eines Interfaces ohne Farbe	58
4.2	Mittelwert der gesetzten Tags eines Interfaces mit Farbe	58
4.3	Mittelwert der gesetzten Tags eines Interfaces in Bezug auf den Merger	59
4.4	Mittelwert der gesetzten Tags eines Interfaces ohne Farbe in Bezug zum Merger und den Bildsets	60
4.5	Mittelwert der gesetzten Tags eines Interfaces mit Farbe in Bezug zum Merger und den Bildsets	60
4.6	Test auf Normalverteilung mit Hilfe von Kolmogorov-Smirnov und Shapiro-Wilk	61
4.7	Signifikanzanalyse nach [ZVS07] für die Tag- und Kreuzverweis-Service-Variablen	63
4.8	Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei D_Sugg und Farbe_D_Sugg	64
4.9	Mittelwerte von D_Sugg und Farbe_D_Sugg in Bezug auf den Merger	64

4.10	Mittelwerte von D_Sugg in Bezug auf die Bildsets	65
4.11	Mittelwerte Farbe_D_Sugg in Bezug auf die Bildsets	65
4.12	Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei D_AutoComp und Farbe_D_AutoComp	66
4.13	Mittelwerte von D_AutoComp und Farbe_D_AutoComp in Bezug auf den Merger	67
4.14	Mittelwerte von D_AutoComp in Bezug auf die Bildsets	67
4.15	Mittelwerte von Farbe_D_AutoComp in Bezug auf die Bildsets	68
4.16	Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei D_Zeit und Farbe_D_Zeit	69
4.17	Mittelwerte von D_Zeit und Farbe_D_Zeit in Bezug auf den Merger	70
4.18	Mittelwerte von D_Zeit in Bezug auf die Bildsets	70
4.19	Mittelwerte von Farbe_D_Zeit in Bezug auf die Bildsets	71
4.20	Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei den Ergebnissen der Fragebögen	74
4.21	Signifikanzanalyse der Ergebnisse der Fragebögen	75
4.22	Mittelwerte von Reaktionszeit und Farbe_Reaktionszeit in Bezug auf den Merger	76
4.23	Mittelwerte von Bedienbarkeit und Farbe_Bedienbarkeit in Bezug auf den Merger	76
4.24	Mittelwerte von Funktionalität und Farbe_Funktionalität in Bezug auf den Merger	77
4.25	Mittelwerte von Hilfe und Farbe_Hilfe in Bezug auf den Merger	78
4.26	Mittelwerte von Überblick und Farbe_Überblick in Bezug auf den Merger	78
4.27	Mittelwerte von Eingabe und Farbe_Eingabe in Bezug auf den Merger	79
4.28	Mittelwerte von Erfolgreich und Farbe_Erfolgreich in Bezug auf den Merger	80
4.29	Mittelwerte von Bedienung und Farbe_Bedienung in Bezug auf den Merger	81
4.30	Mittelwerte von Lernen und Farbe_Lernen in Bezug auf den Merger	81
4.31	Mittelwerte von Fehlerkorrektur und Farbe_Fehlerkorrektur in Bezug auf den Merger	82

Kapitel 1

Einleitung

1.1 Motivation

Nicht nur in Smartphones findet man heute eingebaute Camcorder oder zum Teil hochauflösende Kameras, selbst in Spielzeugen für Kinder werden mittlerweile Geräte zum Aufzeichnen digitaler Inhalte eingebaut. Was nicht nur den Zugang zu, sondern auch die Erstellung von, digitalen Medien erleichtert hat. Ob Nachrichtensender, Amateurfotograf oder Blogger, jeder setzt in irgendeiner Form auf die digitalen Daten. Die inzwischen günstigen Preise für Speicher und dessen anscheinend undendliche Verfügbarkeit haben zur Folge, dass ein manuelles Aussortieren und zuordnen zu wiederauffindbaren Kategorien, wie man es zum Beispiel aus früheren Zeitungsarchiven kennt, von den wenigsten Nutzern, in einem schnelllebigen Jahrhundert, indem Zeit Geld bedeutet, praktiziert wird.

Als Beispiel für die Unmengen an Daten sei Googles Website Datenbank genannt, die 2005 mehr als 8 Milliarden Einträge fasste¹. Heute im Jahre 2012 gibt es keine Angaben mehr darüber welches Datenvolumen allein Google beherbergt. Diese Daten möchten aber nicht nur abgelegt, sondern auch - gegebenenfalls von anderen Nutzern - gefunden werden. Handelt es sich hierbei um Textdokumente ist eine schlagwortbasierte Suche, mit Hilfe von beispielsweise der Indizierung, kein Problem. Handelt es sich bei dem zu suchenden digitalen Medium allerdings um ein Bild oder ein Video, kann dieses nur wieder aufgefunden werden, wenn vorher zu jeder Datei Metainformationen, wie Tags oder RDF Schematas, definiert wurden. Versucht man abseits von einfachen Begriffen, wie Haus, Baum und Auto, Bilder zu suchen, erhält man selten zufriedenstellende Ergebnisse. Teilweise sind schlechte oder schlichtweg falsche Suchergebnisse auf fehlerhafte oder mangelnde durchsuchbaren Informationen zurück zu führen. Es gibt verschiedene Ansätze, digitale Medien mit Annotationen zu versehen, um ein qualitativ hochwertigeres

¹<http://www.neogrid.de/was-ist/Google>

Suchergebnis zu erzielen.

Die manuelle Annotation wurde von dem Umgang mit analogen Medien inspiriert. In der einfachsten Form der manuellen Annotation vergibt der Nutzer geeignete Dateinamen. In der komplexeren Form werden dem Bild Tags zugeordnet. Dieses Verfahren ist aber nicht nur sehr zeitintensiv, es birgt auch die Risiken der Rechtschreibfehler, dialektspezifischen Bezeichnungen eines Objektes und der zu geringen oder zu großen Anzahl an Schlagworten. Darüber hinaus ist eine Manipulation, wie sie oft mit Hilfe von Addwords vorgenommen wird nicht auszuschließen. Als Gegensatz zu den manuellen Ansätzen, gibt es automatische Verfahren. Bei automatischen Lösungen, werden unter anderem Bildverarbeitungs- und Pattern Recognition Algorithmen oder die Daten der GPS Informationen des Aufnahmeortes, der Kamera oder nutzerspezifische, wie Kalendereinträge, verwendet. Die semantische Lücke ist nur eines der möglichen Problemfelder.

Als Kompromiss zu manueller und automatischer Annotation hat man die semi-automatischen Annotationsverfahren konzipiert. Da Semi-automatische Ansätze das Optimum zwischen Genauigkeit und Effizienz darstellen, sind hier die meisten Arbeiten und Forschungsansätze zu finden. Auf Grund dieser Fülle, werden hier keine Beispiele genannt, sondern nur eine Gemeinsamkeit, die alle Arbeiten inne haben:

Obwohl der Nutzer deutlich in den Annotationsprozess eingebunden wird, gibt es keine Usability Studien in Bezug auf die Gestaltung des Interfaces.

1.2 Zielsetzung

Ziel dieser Masterarbeit ist, zu untersuchen in wie weit man einen Nutzer mit einer geeigneten Gestaltung eines Interfaces während des Annotationsprozesses unterstützen kann.

Es soll nicht nur die Frage der Effizienz, wie schnell man insgesamt mit einem bestimmten Interface annotieren kann, ob ein Interface den manuellen Anteil der Annotation beschleunigt, beantwortet werden, sondern auch die Frage, ob der Nutzer durch ein gewissen Layout oder durch Zusatzfunktionen motiviert ist, mehr qualitativ hochwertige Schlagworte zu setzen und auszuwählen, als es zum Beispiel mit einem anderen Layout der Fall ist.

Dafür wird eine summative Evaluation verschiedener Interfaces durchgeführt. Jeder Proband erhält jeweils nacheinander zwei Interfaces mit denen er mehrere Bilder annotieren muss. Nach jedem Interface wird ein Fragebogen und am Ende noch ein abschließender Fragebogen beantwortet werden.

Die Auswertung der Fragebögen und die vom Nutzer eingegebenen und über den Annotationsprozess gespeicherten Daten werden zur Analyse herangezogen, um sich entweder in ihrer Argumentation zu stützen oder aber kontroverse Ergebnisse

von subjektiver Einschätzung der Nutzer und objektiver Daten herauszuarbeiten. Während der Evaluation muss der Nutzer 20 Bilder annotieren, da im Anwendungsfall oft hunderte multimediale Inhalte annotiert werden und es Interface Eigenschaften gibt, die einem erst nach vermehrter Nutzung, als negativ auffallen. Dadurch, dass der Nutzer nach seiner Meinung befragt wird, kann ausgeschlossen werden, dass ein Interface durch bloße Maße zur Effizienz und Effektivität als hilfreich eingestuft wird, ohne es für den Nutzer letztendlich auch zu sein. Denn nur ein Nutzer, der zufrieden ist und sich mit der Handhabung eines Interfaces wohl fühlt, nutzt dieses gerne und viel.

1.3 Aufbau der Arbeit

Die Arbeit ist in fünf Teile untergliedert. Das erste Kapitel dient dem Einstieg in das Themengebiet der Annotation. Es wird ein Überblick darüber geboten, warum Annotation von digitalen Medien wichtig ist und worin die Motivation und das Alleinstellungsmerkmal für diese Arbeit liegt. Darüberhinaus dient das Unterkapitel Zielsetzung als Orientierung, in welche Richtung sich diese Arbeit entwickelt und damit auch als roter Faden und Begründung für ein bestimmtes Unterkapitel.

Da es sehr viele verschiedene Themengebiete im Bereich der Annotation und des Semantic Web gibt, musste mit einer Recherche über verwandte Arbeiten sichergestellt werden, dass es eine solche Arbeit, wie die hier vorliegende noch nicht gibt und, dass Ansätze und Ideen, die innerhalb dieser Arbeit verfolgt werden sollen, nicht schon durch andere dahingehend geprüft wurden, dass man sie schon im Voraus wieder ausschließen und verwerfen kann. Außerdem werden in diesem Kapitel einige Begrifflichkeiten und Namen erläutert, die in den nachfolgenden Kapiteln der Arbeit Verwendung finden. Das dritte Kapitel befasst sich mit der Evaluation. Angefangen bei der theoretischen Konzeptionierung, über die technischen Aspekte, bis hin zum Ablauf, sollen hier alle Themen, die mit der Evaluation in Zusammenhang stehen erläutert werden.

Im vorletzten Kapitel, werden alle Daten die innerhalb der Evaluation gewonnen werden konnten, untersucht und analysiert. Das fünfte und letzte Kapitel fasst die aus den vorherigen Kapiteln gewonnenen Erkenntnisse zusammen und gibt einen Ausblick auf weitere Aspekte, die basierend auf dieser Arbeit, untersucht werden könnten.

Kapitel 2

Hintergrund

Im folgenden Kapitel werden Namen und Begrifflichkeiten, die in der Arbeit Verwendung finden, erklärt. Neben Erklärungen zu den verschiedenen Annotationsarten, wird ein kurzer Überblick über verwandte Arbeiten gegeben.

Schlagworte

Schlagworte, Keywords oder auch Tags sind die Grundlage der Annotation. Es sind meist einzelne Wörter, die derart prägnant sind, dass beispielsweise Bilder mit einer gewissen Anzahl von Tags beschrieben werden können und sich damit, von anderen Bildern und ihren Schlagworten unterscheiden lassen¹.

Thesaurus

Ein Thesaurus ist vergleichbar mit einem Lexikon, nur dass es darüber hinaus auch noch Beziehungen einzelner Begriffe untereinander darstellt [Wer78]. Zu diesen Arten von Beziehungen, die ein Thesauri abbilden kann, gehören die Ähnlichkeits- und Synonymbeziehungen.

Thesauri bilden neben Taxonomien, Ontologien und Topic Maps, die Grundlage für das semantische Web und sind damit auch für Arbeiten im Bereich der semantischen Annotation interessant. Beispiele für Thesauri sind: Das Union List of Artist Names (ULAN)², welches neben den Namen von Künstlern auch diverse Informationen über diese beinhaltet. Das Art and Architecture Thesaurus (AAT)³, welches im Grunde eine Erweiterung des ULAN darstellt, da hier Kunstrichtungen und Informationen über diese in Beziehung zueinander gesetzt werden, welche letztendlich auch wieder zu den Künstlern führen.

¹http://lotse.uni-muenster.de/theologie/literatur_recherchieren_und_beschaffen/strategien_zur_literatursuche/exkurs_stichwort_schlagwort-de.php

²<http://www.getty.edu/research/tools/vocabularies/index.html>

³<http://www.getty.edu/research/tools/vocabularies/index.html>

Ontologie

Ontologien beschreiben komplexere Zusammenhänge zwischen Objekten, sodass letztendlich ein Netz aus Objekten und deren Relationen und Verbindungen zu anderen Objekten entstehen. Mit Hilfe vom Ressource Description Framework (RDF) oder der Web Ontology Language (OWL) ist es möglich, diese Zusammenhänge zu beschreiben. Als Beispiel sei hier WordNet⁴ genannt, welches in einem RDF und OWL Schema abrufbar ist.

Kreuzverweis

Kreuzverweis⁵ ist ein an der Universität Koblenz gegründetes Unternehmen, dass mit Smart Keywording eine Annotations-Erweiterung von Software bietet. Smart Keywording bietet neben Autocompletion Vorschlägen, auch Schlagwort-Vorschläge an, die basierend auf einem Relevance Feedback Algorithmus zu den bisher vom Nutzer eingegebenen Schlagworten generiert werden. Kreuzverweis bietet einen webbasierten Client, dessen Code auch unter GitHub zur Verfügung steht, mit dem man diese Funktionalität testen kann. Da es sich bei Smart Keywording um einen Service handelt, der eine gesicherte Verbindung zum Kreuzverweis Webserver Server voraussetzt, ist bei der Nutzung der Smart Keywording Erweiterung in eine eigene Software oder Applikation eine bestehende Verbindung zum Internet nötig.

Voxalead

Voxalead ist ursprünglich ein System, das Nachrichtenvideos mit Annotationen versieht, indem es sich der Speech-to-Text Funktionalität bedient [LTG11]. Um diese Annotationen noch übersichtlicher und für den Journalismus aufbereitet darzustellen, hat man die Tags, basierend auf den journalistischen W-Fragen, in vier Kategorien eingeteilt. Diese Kategorisierung wird visuell durch verschiedene Farben für jede einzelne Kategorie hervorgehoben. Die erste Kategorie ist die Organisation und wird rot eingefärbt. Die zweite Kategorie, die Personen, ist in Blau eingefärbt, die dritte, die Events, wird in Grün dargestellt und die vierte, die Orte, bekommen einen orangebraunen Hintergrund (Abbildung 2.1).

⁴<http://wordnet.princeton.edu/>

⁵<http://kreuzverweis.com/de/>

⁶Screenshot: <http://voxaleadnews.labs.exalead.com/>, Stand Oktober 2012

⁷<http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm>



Abbildung 2.1: Die 4 Voxalead Kategorien⁶

Verwandte Arbeiten

Das Forschungsfeld um Semantic Web ist sehr groß, daher gibt es nicht wenige Arbeiten im Bereich der Annotation oder annotationsnahen Themen, aber Studien im Bereich der Annotation gibt es eher wenige. Möchte man eine Arbeit im Bereich der Annotation von digitalen Medien verfassen, ist eine intensive Recherche über bereits vorhandene Arbeiten unerlässlich. Einige dieser Arbeiten, die zum Teil Inspiration für eine Idee waren und zum Teil Hinweis darauf, etwas anders zu gestalten, werden im Folgenden vorgestellt.

Annotationsverfahren

Es gibt verschiedene Verfahren Multimediadaten zu annotieren. Man unterteilt diese Verfahren generell in drei verschiedene Kategorien: Manuelle-, Semiautomatische- und Automatische-Verfahren.

2.1 Manuelle Annotationsverfahren

Die einfachsten Formen der manuellen Annotation bestehen in der Wahl eines geeigneten Namens für die Multimediadatei oder die Zuordnung zu einem Oberbegriff beziehungsweise Einsortierung in einen Ordner mit entsprechendem Namen. Da für die Benennung von Dateien aber nur eine endliche Anzahl von Zeichen

1.	MPEG-7 Systems
2.	MPEG7 Description Definition Language
3.	MPEG-7 Visual
4.	MPEG-7 Audio
5.	MPEG-7 Multimedia Description Schemes
6.	MPEG-7 Reference Software
7.	MPEG-7 Conformance Testing
8.	MPEG-7 Extraction and use of descriptions
9.	MPEG-7 Profiles and levels
10.	MPEG-7 Schema Definition

Abbildung 2.2: Die 10 MPEG-7 Bestandteile⁷

zugelassen sind und die Zuordnung zu Oberbegriffen oder Einsortierung in Ordner ebenfalls nur eine gewisse Anzahl von Schlagworten abdecken kann, schränken diese Verfahren zu sehr ein und werden komplexeren Daten wie beispielsweise Bildern, auf denen nicht nur ein Objekt zu sehen ist, nicht gerecht. Ein Einsatz im professionellen Bereich, bei dem mehrere tausend Bilder annotiert und in anderen Arbeitsschritten wieder aufgefunden werden müssten, ist daher nicht üblich.

Caliph und Emir

Caliph und Emir ist eine Abkürzung für Common And Light-weight Photo annotation und Experimental Metadata based Image Retrieval [Lux09] sind zwei Prototypen, die genau auf diesem Standart dem MPEG-7 aufbauen. Entwickelt wurden sie in Java und dienen hauptsächlich der Bildannotation. Außerdem dient das Tool Emir dem graphähnlichen Retrieval von semantischen Annotationen⁸. Diese beiden Prototypen unterstützen nicht nur die manuelle Annotation, sondern auch die automatische. Für eine bessere Übersicht, werden in diesem Abschnitt nur die manuellen Annotationsmöglichkeiten erläutert und im Abschnitt 2.2 die Automatischen. Mit Caliph ist es möglich, die W's der strukturellen Text Deskriptoren zu beantworten und ein Freitextfeld auszufüllen, bei dem man nicht nur einzelne Subjektive als Schlagworte setzen, sondern natürlichsprachliche Sätze benutzen kann (Abbildung 2.3). Daraüber hinaus bietet sich die Möglichkeit die Bildqualität anhand eines Quality Ranking mit einer Skala von Unsatisfactory über Fair bis Ex-

⁷<http://www.semanticmetadata.net/features/>

cellent zu bewerten. Diese Option dient der Möglichkeit nicht nur über inhaltliche Aspekte zu suchen, sondern, wie Metadaten auch, ebenfalls über Bildeigenschaften.

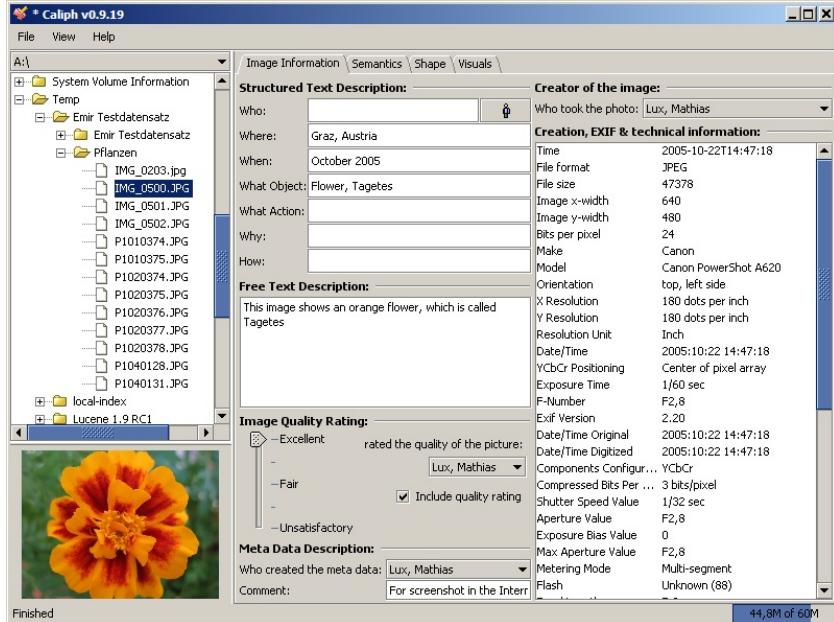


Abbildung 2.3: Das Caliph Interface⁹

2.2 Automatische Annotation

Auch im Bereich der automatischen Annotation gibt es sehr unterschiedliche Ansätze, die von kontextbasierten Daten über low-level Feature Extraktion bis hin zu Bildsemantiken, also den high-level Merkmalen reichen.

Bei den kontextbasierten Daten werden Metainformationen ausgelesen. Diese können zum Beispiel im Exchangable Image File Format (EXIF) vorliegen. Der EXIF Standard¹⁰ wurde in seiner zweiten Version im April 2002 von Japan Electronics and Information Technology Industries Association (JEITA¹¹) herausgegeben. Dieser besagt, dass Daten wie Brennweite, Belichtungszeit, Ort (sofern zum Zeitpunkt der Aufnahme ein GPS Signal vorlag), Datum und Uhrzeit (vollständige Liste als Bild/Tabelle) im Header eines JPEG oder TIFF komprimierten Bildes gespeichert werden. Durch den IPTC-NAA-Standard sind die fototechnischen Metadaten noch

⁹ Bildquelle: <http://www.semanticmetadata.net/images/screenshots/caliph-01.png>, Stand Juli 2012

¹⁰ <http://www.exif.org/specifications.html>

¹¹ http://www.jeita.or.jp/cgi-bin/standard_e/list.cgi?cateid=1&subcateid=4

um einige Informationen mehr, wie Urheber und Kommentare, erweiterbar. Adobe hat mit der Extensible Metadata Plattform (XMP) eine Möglichkeit geschaffen, EXIF und IPTC um weitere Metadaten, beispielsweise Änderungen am Foto, mit Hilfe des Resource Description Framework (RDF) zu ergänzen. Der Prototyp Caliph [Lux09] ist fähig die EXIF, IPTC und XMP Metadaten aus einem Foto auszulesen und dem Nutzer im Graphical User Interface zu präsentieren. Ein weiteres Schema zur Beschreibung von Metadaten stellen die fünfzehn Eigenschaften des Dublin Core Metadaten Element Set oder auch Dublin Core¹² dar, welche von der Dublin Core Metadata Initiative (DCMI) initiiert wurden.

Darüberhinaus zählen zu den Metainformationen auch die Inhalte der zugehörigen MPEG-7 Dateien. Der MPEG-7 Standart, auch bekannt unter dem Namen Multimedia Content Description Interface, wurde von der Moving Pictures Expert Group (MPEG) entwickelt und 2002 als ISO Standart festgelegt. Der ISO Standart MPEG-7¹³ ermöglicht es eine XML-Beschreibung einer Multimediadatei hinzuzufügen und bietet damit durch die höhere Komplexität, als die einfachen manuellen Annotationsverfahren, eine detailliertere Verschlagwortung. Ursprünglich wurde dieser Standart zur Archivierung von Videodateien ausgearbeitet. Nach und nach wurde den Entwicklern aber klar, dass die 10 Teile (Abbildung 2.2) des MPEG-7 sich aber auch auf andere Multimediainhalte anwenden lassen, da der Aufbau von MPEG-7 einer Toolbox gleicht und damit sehr flexibel auf verschiedene Anwendungsfälle reagieren kann.

Die Extraktion von Low-Level Merkmalen bedient sich Algorithmen der Bildverarbeitung. Mit ihnen ist es möglich zum Beispiel Farben zu erkennen oder Linien und Kanten zu detektieren und in Deskriptoren oder Merkmalsvektoren zu speichern. Auch hier bietet Caliph eine Möglichkeit zur Extraktion von Farbe und Kanten. Die Zusatzfunktion der Google Bildersuche zur Suche von ähnlichen Bildern arbeitet unter anderem mit solchen Low-Level Features, um auf diese Art ein künstliches Sehen erzeugen zu können und damit zu ermöglichen, zu einem hochgeladenen Bild weitere ähnliche Bilder zu finden¹⁴.

2.3 Kombination von automatischer und manueller Annotation

Die meisten Arbeiten sind im Bereich der semiautomatischen Annotationsverfahren zu finden. Automatische Verfahren verfolgen zwar oft innovative Ideen, allerdings haben sie eine geringere Qualität, als Verfahren, die auf der Kombination

¹²<http://dublincore.org>

¹³<http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm>

¹⁴<http://support.google.com/images/bin/answer.py?hl=de&answer=1325808>

von automatischer und manueller Annotation beruhen.

Im Idealfall sind bei semiautomatischen Verfahren gerade so viele Nutzereingaben nötig, dass Ergebnisse einer beispielsweise automatischen Merkmalsextraktion korrigiert werden können, ohne dass diese für den Nutzer zu zeitaufwändig oder lästig werden. Der automatische Teil wird also durch einen manuellen Anteil korrigiert oder erweitert. Baut der automatische Teil auf dem manuellen auf, so gibt der Nutzer beispielsweise initial einige Schlagworte ein und ein Algorithmus basierend auf Ontologien und semantischen Datenbanken, schlägt weitere passende Tags vor. Im Folgenden werden Beispiele für semiautomatische Annotationswerkzeuge vorgestellt.

Photocopain

Photocopain [THB⁺06] ist einer der Vertreter der semiautomatischen Annotationstools. Es hat die Fähigkeit Informationen über den Kontext in dem ein Foto aufgenommen wurde mit weiteren frei verfügbaren Quellen zu kombinieren, um so ein Grundvokabular von Tagvorschlägen zu generieren. Die Funktionen die Photocopain dem Nutzer zur Annotation, Informationsextraktion und als User Interface zur Verfügung stellt, wurden vom Advanced Knowledge Technologies Projekt¹⁵ (AKT) innerhalb einer sechs Jahre andauernden Forschungszeit entwickelt.

Der automatische Annotationsprozess greift ebenfalls auf die bereits im Abschnitt 2.2 beschriebenen Metainformationen, wie EXIF, IPTC und GPS, zu. Darüber hinaus können Kalenderdaten ausgelesen werden, um somit Fotos in einen bestimmten erzählenden Kontext zu setzen und darauf aufbauend weitere passende Annotationen vorschlagen zu können. Außerdem findet auch hier eine Merkmalsextraktion statt und mit Hilfe von Community Tagging auf Flickr.com wurden geeignete Datensätze als Trainingssätze für die Algorithmen zur Bildanalyse gefunden. Implementiert wurden bereits bekannte Algorithmen, wie beispielsweise die CIELab colour-map und die Hue, Intensity, Texture Map (HIT) [MC01], die der Gesichtserkennung dient, die aber um die bereits extrahierten Informationen angereichert wurden, um zu sehen, wie performant sich diese mit Hilfe der Zusatzinformationen verhalten. Um aber nicht eine ähnliche unsortierte Informationsflut, wie man sie oft im Internet vorfinden kann, an Schlagwörtern zu generieren, werden einige Daten benutzt, um die Tagauswahl auf ein bestimmtes Themengebiet einzuschränken. Beschäftigt man sich mit dem Hintergrundwissen in der Fotografie, weiß man, dass bestimmte Brennweiten typisch für Portraitaufnahmen oder Nahaufnahmen von Objekten sind. Auch wenn es sich bei einem Blatt um die gleiche Farbe, wie bei einem Baum handelt, lässt die Brennweite Rückschlüsse darüber zu, ob auf dem Foto wahrscheinlich eher der Baum oder nur das Blatt zu sehen ist [SJ08].

¹⁵<http://www.aktors.org/>

Im User Interface, welches auf dem AKTive Media System beruht, hat der Nutzer die Möglichkeit die Tagvorschläge zu bestätigen, zu löschen oder zu korrigieren oder Bildbereiche zu markieren, um diese mit Schlagworten zu versehen. Evaluiert wurde die Performance der Algorithmen.

Semantic Annotation of Image Collections

In [HSWW03] werden Kunstbilder mit einem semantischen Vorgehen annotiert. Hierzu stehen die Inhalte mehrerer Thesauri zur Verfügung, um die semantische Annotation zu unterstützen. Eines dieser Thesauri deckt ausschließlich den Bereich Kunst und Architektur (Art and Architecture Thesaurus, AAT) und ein anderes, The Union list of Artist names (ULAN), nur Informationen über ca 220.000 Künstler, ab. Das Iconclass Thesauri hilft mit seinen geordneten Konzepten visuelle Kontexte zu beschreiben. WordNet dient als Quelle für generelle Anfragen, da es in seiner lexikalischen Struktur nicht nur Nomen und Verben eingeordnet hat, sondern auch Synonyme und Adjektive. Alle Thesauri wurden mit ihren Relationen und Eigenschaften in das Resource Description Framework Schema (RDFS), eine Ontologiebeschreibungssprache, übersetzt.

Da es verschiedene Praktiken gibt, um Bilder zu beschreiben, werden in dem hier vorgestellten Werkzeug, die Visual Resource Association (VRA) Core Categories als Vorlage vorgegeben. Auf diese Weise existiert ein einheitlicher Standart zwischen dem annotieren von Bildern und der Art nach Bildern zu suchen. Die VRA ist eine Organisation, die sich unter anderem mit der Vereinheitlichung von Bildbeschreibungen beschäftigt hat, in deren Folge sie die Core Categories als Standart implementiert und etabliert haben, nicht zuletzt, weil die Core Categories auf dem Dublin Core Model aufbauen.

Im Interface des Tools wurden die 17 verschiedenen VRA Daten Elemente in drei Kategorien gruppiert und jeweils auf Registerkarten dargestellt. Die Datenelemente des VRA Annotations Templates wurden, dort wo passende Konzepte in den Thesauri existieren, mit ihnen verbunden, sodass ein Nutzer im Interface durch die Eltern- und Kindbäume eines Elementes browsen kann. Die Ontologien selber wurden ebenfalls untereinander verlinkt. Hierbei existieren verschiedene Arten von Verbindungen. Die Äquivalenz-Links arbeiten mit Synonymen und Singular/Plural Darstellungsunterschieden in den Ontologien, die Subklassen-Links unterstützen dort, wo es erhebliche Unterschiede in den Ontologiestrukturen gibt und, wo Äquivalenz-Links zu detailliert definieren. Als letzte Verbindung sind die domain-spezifischen Links zu benennen, die Relationen zwischen Konzepten in zum Teil unterschiedlichen Ontologien, als Domain-Range-Beziehungen, setzen. Darunter zählt zum Beispiel, dass eine Kunstrichtung von bestimmten Künstlern geprägt wurde.

Die Bilder, welche in der Evaluation des Tools benutzt wurden, hatten bereits eine

kurze, wenn auch nur halbstrukturierte, textuelle Annotation. Ein Parser machte es möglich diese Kurzannotationen in eine VRA adäquate Form zu bringen, um die Annotationen für den weiteren Verlauf verwenden zu können. Der Nutzer kann im Annotationsinterface bereits vorhandene VRA Elemente ändern oder dort welche eintragen, wo noch keine Einträge vorhanden sind. Hierzu gibt er in das Textfeld einen passenden Freitext ein. Es steht eine Autocompletion-Funktion zur Verfügung, sodass auch auf Vorschläge aus dieser zurückgegriffen werden kann. Zwei unterschiedliche Szenarien sind nach der Eingabe eines Tags möglich. Entweder wird direkt ein passendes Konzept in einer der Ontologien gefunden oder es wird eine Ergebnisliste mit den einzelnen Elternknoten der verschiedenen Thesauri präsentiert, aus denen der Nutzer dann das am besten passende Konzept auswählt. Neben der Annotation ist es umgekehrt auch möglich basierend auf den gleichen Darstellungsstrukturen zu suchen. Entweder gibt der Nutzer einen Begriff ein, woraufhin in allen VRA Elementen danach oder nach Synonymen, also nach Äquivalenz-Links gesucht wird, oder der Nutzer sucht sich zunächst aus der Annotationsvorlage ein VRA Element heraus und sucht nur in diesem Bereich nach einem Konzept.

Diese Arbeit vereinigt mehrere Wissensquellen, sodass ein Benutzer an zentraler Stelle mit semantischer Suche Ergebnisse erzielen kann. Sie verbindet außerdem den Annotationsprozess mit der Suche, was die Motivation für eine solchen Arbeit möglichen Probanden besser vor Augen führt. Allerdings wird auf zweckmäßige Interfaces zurückgegriffen und es liegen keinerlei Nutzerstudien über die Benutzung der Interfaces vor. Am interessantesten wäre hierbei die Frage, inwieweit ein ontologie-unerfahrener Nutzer mit der baumähnlichen Darstellung von Suchergebnissen zurecht kommen würde.

Interfaces for Networked Media Exploration and Collaborative Annotation

[ASSB05] ist erstellt worden mit dem Ziel den Austausch und die Suche von digitalen Medien innerhalb eines sozialen Netzwerkes zu vereinfachen. Daher verfolgt die Arbeit einen anderen Ansatz als bisherige Arbeiten, nämlich den des kollaborativen Annotierens.

Ein Nutzer kann Bilder über ein Webinterface annotieren. Er wird unterstützt durch low-level Merkmale und ein Toolkit. Es gibt verschiedene Felder, die ein Nutzer auszufüllen hat. Diese Felder, basieren auf den W-Fragen des Journalismus: Wer, wann, wo und was? Wenn ein Nutzer seine Annotationen gesetzt hat, bildet das System Beispiel Schlagwortmengen für jedes ausgefüllte Feld. Diese Beispilmengen dienen dann im zukünftigen Annotationsprozess weiteren Nutzern als mögliche Vorschläge. Dariüber hinaus werden demografische Daten aus dem Nutzerprofil des Probanden ausgelesen und als Annotationsinformation bereitgestellt.

Hat ein Nutzer bereits mehrere Bilder annotiert, berechnet ein Algorithmus im Hintergrund, wie oft welches Schlagwort benutzt wurde. Je häufiger ein Schlagwort benutzt wird, desto weiter vorne erscheint es in der Vorschlagsliste. ConceptNet [LS04] soll dem Nutzer weitere ähnliche Konzepte zur Auswahl präsentieren. Hierzu werden mit Hilfe von Distanzmaßen weitere Konzepte zu den bereits vorhandenen herausgesucht, beziehungsweise weitere Unter- und Überkategorien als Vorschläge angeboten. Durch die semantischen Beziehungen in ConceptNet ist es möglich, die Tag Vorschläge, welche von den Daten der Gruppe, also anderen Nutzern, mit denen man im sozialen Netzwerk verbunden ist, gewonnen werden, zu filtern. Die Vorschläge der Freunde können gewissen Konzepten zugeordnet werden und ebenso die selbst eingegebenen Tags. Findet sich eine Übereinstimmung der Konzepte, erscheinen die Vorschläge. Findet sich keine direkte Übereinstimmung, wird die semantische Distanz berechnet. Ist diese niedrig, werden die Schlagworte in die Vorschlagsliste mit aufgenommen, ansonsten werden die Vorschläge nur aus den bereits genannten demografischen Daten, eigenen häufig eingegeben Tags oder der low-level Merkmalen generiert. Auch diese Arbeit ist eng mit dem Zweck der Annotation verknüpft: Ein zweites Interface bietet die Möglichkeit nach allen Fotos zu suchen, auf der eine bestimmte Person abgebildet ist oder man lässt sich grafisch die Zusammenhänge zwischen Personen, Orten und Events darstellen. Es liegen keine Nutzerstudien zur Annotation mit diesem Interface vor. Allerdings kann festgehalten werden, dass auch hier wieder das eher zweckmäßige Interface und die andere Möglichkeit, Vorschläge zu gewinnen die Motivation für diese Arbeit darstellt.

Semi-Automatic Image Annotation in MiAlbum

Viele Arbeiten im Bereich der semiautomatischen Annotationen sind über Interfaces, bei denen der Nutzer ein beliebiges Bild, meist nach randomisierter Reihenfolge, mit der Aufgabe geeignete Schlagworte hinzuzufügen, präsentiert bekommt. Hierbei wird der Nutzer durch gewisse Features, Algorithmen oder Innovationen unterstützt, sodass er nicht den gänzlichen Annotationsprozess manuell durchlaufen muss. Sobald alle Bilder annotiert sind, bietet sich die Möglichkeit in einem weiteren Interface nach bestimmten Bildern zu suchen. Die Idee von MiAlbum [WDS⁺01] ist eine andere.

Zwar werden auch hier über automatische Verfahren initiale Annotationsvorschläge unterbreitet, allerdings werden die Bilder über eine Suchanfrage und das dann durch den Nutzer eingebene Relevance Feedback annotiert. Das Interface besteht aus drei verschiedenen Teilen. Der erste Teil ist ein Suchfeld, in das ein oder mehrere Schlagworte eingegeben werden können. Der zweite Teil ist ähnlich einer Bildergallerie aufgebaut, da hier die Ergebnisse zu der gestellten Suchanfrage präsentiert werden. Der dritte Teil besteht aus facebookähnlichen Like- beziehungsweise

Dislike-Buttons, die bei jedem Bild, welches in der Ergebnisgallerie angezeigt wird, unten rechts erscheinen. Über diese Buttons kann der Nutzer die Bilder, welche er als Ergebnisse auf eine Suchanfrage gezeigt bekommt, bewerten und definiert damit, ohne dass er etwas von dem Berechnungsprozess mitbekommt, das Relevance Feedback dieses bewerteten Bildes im Verhältnis zur Suchanfrage. Hierzu muss er den Like-Button anklicken, wenn er das Bild als geeignetes Ergebnis für seine Suchanfrage einschätzt, den Dislike-Button, wenn er das Bild als nicht passend einstuft und keinen der beiden Buttons, wenn der Nutzer sich nicht sicher ist oder das Bild, zwar nicht ideal, aber zumindest zur Hälfte zur Suchanfrage passt. Anfangs sind keine Annotationen vorhanden, daher werden nach einer Suchanfrage eine zufällige Auswahl an Bildern gezeigt. Zumeist gibt es mehr als Tausend Bilder in einer Bilddatenbank, die annotiert und durchsucht werden sollen, sodass die Anfangsergebnisliste reduziert ist auf eine überschaubare Menge, von beispielsweise 50 Bildern. Im Hintergrund berechnet ein Algorithmus, zu jedem Bild in der Bilddatenbank, das Ähnlichkeitsmaß. Dieses Ähnlichkeitsmaß basiert auf der Auswertung von low-level Merkmalen. Ein solches Vorgehen wird unter anderem auch bei der inhaltsbasierte Bildersuche (Content Based Image Retrieval, CBIR) benutzt. Sobald ein Nutzer die ersten Relevance Feedback Einschätzungen getroffen hat, wird für die Bilder eine allgemeine Ähnlichkeit berechnet. Diese erhält man durch die gewichtete Summe der Relevanzeinschätzungen der Schlagworte und der low-level Merkmale. Je häufiger ein Nutzer nun die Bilder positiv, neutral oder negativ zu einem Schlagwort respektive einer Suche einstuft, desto besser werden die darauffolgenden Ergebnisse. Mit Hilfe einer Ground Truth konnte die Performance und Genauigkeit des Algorithmus und der Bildersuche mit statistischen Daten belegt werden. Die Nutzerstudie zum Interface und zum Annotieren mit dem vorgestellten Tool, wird nur wage erwähnt und stützt sich auf die von den Probanden ausgefüllten Fragebögen. Es geht aus den Veröffentlichungen zum Beispiel nicht hervor, wieviele Probanden teilgenommen haben. Außerdem gibt es einige Interfaceeigenschaften, wie zum Beispiel eine sich ändernde Buttonbeschriftung und die Platzierung des Suchfeldes an der linken unteren Ecke, die gegen die allgemeinen Regeln der Nutzerfreundlichkeit sprechen, die durch eine Nutzerstudie gestützt oder aber durchaus hätten widerlegt werden können. Trotz interfacespezifischem Fragebogen, aus dem hervorgeht, dass die Bewertungsbuttons gut angenommen werden und, dass die Reihenfolge in der Bilder angezeigt werden, also das Relevance Feedback und seine Darstellung, für manche Nutzer unverständlich ist, lässt sich auch hier eine Konzentration auf die Untersuchung des Algorithmus feststellen.

Fazit Als Fazit kann festgehalten werden, dass die Like und Dislike-Buttons für Menschen, die soziale Netzwerke aktiv nutzen, intuitiv sein können, dass aber wei-

terhin an der Darstellung des Relevance Feedbacks gearbeitet werden muss, um auch dieses intuitiv verständlich zu machen.

Efficient Semantically-Aware Annotation of Images - Ratsnake

Neben der Möglichkeit Bilder als Ganzes zu annotieren, können durch die grafische Annotation auch Bildbereiche mit Schlagworten versehen werden. Die bekanntesten Tools in diesem Bereich sind beispielsweise das webbasierte LabelMe Tool [RTMF08] und die interaktiven Spiele ESP [AD04] und Peekaboom [ALB06]. Diese Arbeiten machen es möglich, Klassen von Objekten und Objekte in verschiedenen Szenerien und Positionen zu erkennen. Rapid Image Annotation with Snakes (Ratsnake) [IS11] ist eine Erweiterung dieser Ansätze für die Anwendung bei Bildserien, wie sie häufig im medizinischen Bereich vorzufinden sind. Der Nutzer muss auf diese Weise nur im ersten Bild den zu annotierenden Bildbereich mit einer Snake umschließen. Diesen Bereich kann der Nutzer festlegen, indem er mit der Mouse in das Bild zeichnet. Linien, die mit einer Mouse eingezeichnet werden, sind oft nicht glatt. Algorithmen der Bildverarbeitung, wie zum Beispiel ein Schwellwertverfahren, Kantendetektion und Kantenglättung, sorgen dafür, dass aus der grob eingezeichneten Snake eine glatte und an den ausgewählten Bildbereich angepasste Kontur wird. Startet der Nutzer die Annotation der Bildserie, vergleichen diese Algorithmen, den im Bild vorher markierter Bildbereich und die Position des Bereichs auf dem neuen Bild. Können auch in dem neuen Bild ähnliche Kanten gefunden werden, wird der markierte Bereich entweder übernommen oder auf die neuen Kanten angepasst. Zwar ist die Benutzung von Bildbereichen zur Annotation für die hier vorliegende Arbeit zu detailliert, aber für weiterführende Arbeiten, sollte die Unterstützung zum einen durch Bildverarbeitungsalgorithmen und zum anderen durch Annotationen von Bildbereichen betrachtet werden. Für die Interfacegestaltung lässt sich außerdem festhalten, dass das Einzeichnen eine andere Möglichkeit bietet den Nutzer mit einem Bild interagieren zu lassen, als herkömmliche Interfaces.

LabelMe

In [RTMF08] wird ebenfalls eine Methode zur Annotation einzelner Objekte in einem Bild beschrieben. Das Verfahren basiert aber auf einer grafischen Verschlagwortung. Es handelt sich bei LabelMe¹⁶ um ein webbasiertes System, da es zum einen Probleme, die auf den Computern der Nutzer durch verschiedenen Betriebs- und Softwaresystemen auftreten können, vermeiden möchte und zum anderen auf dem Prinzip der Kollaboration aufbaut. Ein Nutzer kann entweder eigene Bilder hochladen und annotieren oder mithelfen, bereits vorhandene Bilder und deren

¹⁶<http://labelme.csail.mit.edu/Release3.0/>

Annotationen zu erweitern oder zu korrigieren. Möchte der Nutzer ein Objekt in einem Bild annotieren, muss er das Objekt mit einem Polygon versehen. Dieses Polygon der Objekt-Kontur kann ein Nutzer erzeugen, indem er in das Bild klickt, um einen Startpunkt zu definieren und dann mit weiteren Punkten auf der Kontur des Objektes den Polygonzug schließt. Sobald der Polygonzug geschlossen ist, erscheint ein Textfeld, indem der Nutzer ein geeignetes Schlagwort eingeben kann. Da viele verschiedene Nutzer auch immer viele verschiedene Informationen bedeuten, gleicht LabelMe die Annotationen mit WordNet ab, um so Synonyme erkennen und, um Sub- oder Superklassen hinzufügen zu können. Auf diese Weise kann man dem Nutzer alle Freiheiten in Bezug auf seine Annotation lassen, ohne ihm vorschreiben zu müssen, wie allgemein oder speziell er annotieren darf. Neben den Schlagworten, können Nutzer auch die Kategorien, die WordNet Schlagworten zuordnet, verbessern, da es Schlagworte mit mehreren Bedeutungen in verschiedenen Kategorien (Beispiel: Bank, Möbelstück oder Gebäude) gibt. Es sind bereits mehrere tausend Bilder annotiert, wovon laut eigenen Angaben 2690 der Bilder 90% ihrer Pixel einer Annotation zuordnen können [RTMF08]. Durch diese große Datenmenge ist es möglich, nicht nur Bilder manuell zu annotieren, sondern auch basierend auf den bereits vorhandenen Annotationen und zugehörigen Polygonen, weitere Bilder annotieren zu lassen. Der Nutzer kann bei diesen Bildern ebenfalls die Polygone und Annotationen korrigieren oder die vorgeschlagene Annotation bestätigen.

2.4 Benutzerstudien im Bereich der Annotationsinterfaces

Label Positionierung in Formularen

Eine Untersuchung von UXMatters, einer Gemeinschaft, die sich mit Themen aus dem Bereich des User Experience auseinandersetzen, zeigte mit Hilfe eines Eye-Trackers, dass die Positionierung von Labels, also den Titeln von Textfeldern, Einfluss auf die Verständlichkeit und Effizienz, also der Bearbeitungszeit, eines Formulars hat.¹⁷ Hierfür wurden mit Probanden vier verschiedene Anordnungen und Designs untersucht:

1. Linksbündige Label links neben den Textfeldern (zweispalten Layout)
2. Rechtsbündige Label links neben den Textfeldern (zweispalten Layout)
3. Linksbündige Label oberhalb der Textfelder

¹⁷<http://www.uxmatters.com/mt/archives/2006/07/label-placement-in-forms.php>

4. Linksbündige fettgedruckte Label oberhalb der Textfelder

Den meisten Nutzen erhält man, wenn die Textfelder einen angemessenen Abstand zueinander haben und die zugehörigen Label sich direkt oberhalb und linksbündig davon befinden, aber ohne fettgedruckt zu sein. Ist ein Label linksbündig und direkt oberhalb des Textfeldes angeordnet, so bedarf es beim Nutzer weniger Sprünge des Blickes zwischen Textfeld und Label und zeitlich kürzere Sakkaden, die Zuordnungsmöglichkeit von Label und richtigem Textfeld werden erleichtert und die Bearbeitungszeit für ein ganzes Formular sinkt. Diese Erkenntnis floss sowohl bei der Planung der Annotationsinterfaces, als auch der Fragebögen mit ein, sodass sich die Beschriftungen der Textfelder und -boxen direkt links oberhalb befinden.

Autocompletion

Die Autocompletion besteht nicht nur aus einzelnen Vorschlägen, wie sie oft in Textprogrammen und Mobilen Geräten dargestellt werden, vielmehr ist sie eine Liste von möglichen passenden Optionen. Es existieren verschiedene Möglichkeiten, diese Liste von Vorschlägen der Autocompletion darzustellen. Eine besteht in der genannten Extraktion und Darstellung nur eines Vorschlags, eine andere ist die Darstellung mehrerer Vorschläge innerhalb einer Dropdown-Liste unterhalb eines Textfeldes. Die Arbeit „Organizing Suggestions in Autocompletion Interfaces“ [AHO⁺09] beschäftigt sich genau mit einer solchen Darstellung als Dropdown-Liste. Auch innerhalb einer Dropdown-Liste gibt es wiederum verschiedene Ansätze, die Vorschläge zu sortieren oder darzustellen. In der Arbeit wurden die alphabetische, die gruppierte und die zusammengesetzte Organisationsweise von Autocompletion Vorschlägen untersucht.



Abbildung 2.4: Autocompletion Gruppierungsarten [AHO⁺09]

Abbildung 2.4 zeigt diese verschiedenen Arten Autocompletion Vorschläge zu gruppieren anhand des Beispiels Kingston. Probanden hatten die Aufgabe mit

2.4. BENUTZERSTUDIEN IM BEREICH DER ANNOTATIONSINTERFACES31

diesen Autocompletion Interfaces Suchaufgaben zu absolvieren und die richtige Antwort aus der Autocompletion auszuwählen. Die verschiedenen Interfaces wurden als Within-Subject geprüft, die Inhalte der Autocompletion, also ob diese auf dem Thesaurus für geografische Namen (TGN) oder dem lexikalischen Thesaurus WordNet basiert, und die dazu passenden Suchaufgaben, wurden als Between-Group geprüft. Als Ergebnis dieser Studie lässt sich festhalten, dass eine Gruppierung innerhalb einer Autocompletion Dropdown-Liste nur dann vom Nutzer als hilfreich angesehen wird, wenn die Gruppierung für den Nutzer auf Anhieb verständlich ist. Dies ist zum Beispiel bei dem TGN der Fall. Bei geografischen Fakten sind Gruppierungen nicht zuletzt durch die geografische Vorerfahrung und Schulbildung intuitiv erfassbar. Die zusammengesetzte Gruppierung erwies sich durch das mehrstufige Menü als nicht sinnvoll, da hier sehr oft die falschen Antworten von Probanden ausgewählt wurden und die Zeit die es benötigt, um alle möglichen Antworten durchzugehen, zu lange ist. Für lexikalische Wörterbücher, die viele verschiedene Bereiche abdecken, ist die alphabetische Anordnung am hilfreichsten für den Nutzer. Da die Autocompletion Vorschläge für die in dieser Arbeit präsentierten Annotationsinterfaces, auf der DBpedia¹⁸, einer Abbildung der Infoboxen auf Linked Data, basieren, werden die Vorschläge innerhalb der Dropdown-Liste alphabetisch sortiert.

¹⁸<http://de.dbpedia.org/>

Kapitel 3

Evaluationsdesign

3.1 Bestimmen der Hypothesen

3.1.1 Nullhypotesen

Gerade durch die Analyse der verwandten Arbeiten, wurden einige Ideen für ein Annotationsinterface zusammengetragen. Da nicht alle diese Ideen sinnvoll erschienen, hat man mit Hilfe der Befragung von vier Experten einige zu untersuchende Gebiete von vornherein ausschließen können. Abschnitt 3.7 geht näher auf diese Ideen und Begründungen für ein Verwerfen ein.

Um die Evaluation zielgerichtet gestalten zu können, hat man sich auf folgende Nullhypotesen festgelegt:

3.1.2 Unabhängige Variablen

Es wurden folglich als unabhängige Variablen definiert:

1. Die 2 versus 3 User Interface Elemente
2. Die Farbdarstellung basierend auf Voxlead

3.1.3 Abhängige Variablen

Die abhängigen und damit mit Hilfe der Evaluation zu stützenden Variablen bilden sich aus:

1. Effizienz: Schnelligkeit eines Nutzers bei jedem Interface
2. Effektivität: Anzahl der gesetzten Tags pro Bild

3. Nutzung der angebotenen Services, also der Autocompletion- und Suggestion-Vorschläge
4. Benutzerzufriedenheit

3.1.4 Forschungshypothesen

Unter Zuhilfenahme der unabhängigen und abhängigen Variablen wurden die Forschungshypothesen definiert. Die Interfaces werden im Hinblick auf diese evaluiert und dienen der Widerlegung der Nullhypothesen.

1. Der Einsatz von nur 2 User Interface Elementen erhöht die Effizienz
2. Ein Interface Design mit 2 User Interface Elementen erhöht die Benutzerzufriedenheit
3. Das Verwenden von Farbe als Hinweis zur Typinformation beeinflusst die Effizienz positiv
4. Das Verwenden von Farbe als Hinweis zur Typinformation beeinflusst die Effektivität positiv
5. Das Verwenden von Farbe als Hinweis zur Typinformation erhöht die Benutzerzufriedenheit

Die Auswertung der Fragebögen, die im großen Maße die subjektive Empfindung eines Probanden widerspiegeln und damit bei der Bewertung der Nutzerzufriedenheit helfen, unterstützt ebenfalls bei der Beurteilung, ob die anderen, die Effizienz und Effektivität betreffenden, oben genannten Hypothesen bestätigt werden oder nicht. Des Weiteren werden zur Beurteilung der Erfüllung der Hypothesen, die mitgeloggten und aufgenommenen Daten ausgewertet. Zu diesen Daten gehören unter anderem die Zeit, die Häufigkeit der in Anspruch genommenen Services und die Quantität der gesetzten Schlagworte.

3.2 Ablauf und Testbedingungen

Es handelt sich bei der vorliegenden Arbeit nicht um einen Prototypen, der stufenweise verbessert werden soll, sondern um einen Prototypen, der den Vergleich und die Auswirkungen zwischen verschiedenen Interface Designs herausarbeiten soll. Daher erfolgt die Testdurchführung als summative Evaluation. Voraussetzung hierfür ist, dass alle Interfaces in ihrer Endfassung vorliegen. In dieser Evaluation werden zunächst die Layouts mit zwei und drei User Interface Elementen (2er und

3er) und die aktive bzw. inaktive Typinformation miteinander verglichen. Diese unabhängigen Variablen werden mit Hilfe eines Split-Plot Designs in einen Within- und einen Between-group Anteil aufgeteilt, damit ein möglichst breites Spektrum der Hypothesen mit einer realistischen Anzahl an Probanden abgedeckt werden kann. Der Within-group Anteil wird hierbei aus der unabhängigen Variable, ob die Typinformation aktiv oder inaktiv ist, gebildet. Der Between-group Anteil besteht aus der Anzahl der User Interface Elemente. Diese Zuteilung zu den Gruppen wurde gewählt, da bisher noch keine Arbeiten über den Grad der Beeinflussung, den die farbige Kennzeichnung von Typinformationen haben kann, vorliegen. Zusammen mit den zwei verschiedenen Bildsets (B1 und B2), die als zu testende Daten vorliegen, und die Permutation des Within-group Anteils, entstehen acht Kombinationsmöglichkeiten (Tabelle 3.1).

#	Bedingung 1	Bedingung 2
1.	2er ohne Typ mit B1	2er mit Typ mit B2
2.	2er mit Typ mit B1	2er ohne Typ mit B2
3.	3er ohne Typ mit B1	3er mit Typ mit B2
4.	3er mit Typ mit B1	3er ohne Typ mit B2
5.	2er ohne Typ mit B2	2er mit Typ mit B1
6.	2er mit Typ mit B2	2er ohne Typ mit B1
7.	3er ohne Typ mit B2	3er mit Typ mit B1
8.	3er mit Typ mit B2	3er ohne Typ mit B1

Tabelle 3.1: Die vier verschiedenen Interfaces mit zwei unterschiedlichen Bildsets. Latin Square aus Bildsets, Typinformation und Layout

3.3 Aufgabenliste

Dem Probanden werden nacheinander zwei Interfaces abwechselnd mit bzw. ohne Farbe präsentiert. Mit jedem Interface muss er jeweils zehn Bilder (d.h. B1 beziehungsweise B2) annotieren. Welche Interfaces und welche Bildsets in welcher Reihenfolge abgefragt werden, ist zufällig und wird für den Nutzer nicht sichtbar über den online Evaluations-Wizard gesteuert. Die Aufgabe ist, das Bild so gut wie möglich zu annotieren. Hierbei gibt es weder eine zeitliche Begrenzung, noch eine Mindestanzahl an zu setzenden Tags. Als Hintergrundszenario dient hierbei die Vorstellung über eine eigene umfangreiche Bilddatenbank, aus der man, nachdem man für jedes Bild die Verschlagwortung vorgenommen hat, die bestmöglichen Ergebnisse bei einer Suchanfrage erzielen möchte. Nach der Annotation von zehn Bildern, erscheint jeweils ein Fragebogen. Für jeden Proband also zwei Fragebögen,

da jeder Proband zwei Interfaces mit jeweils zehn Bildern nacheinander zu bearbeiten hat. Im dritten und letzten Fragebogen werden die demografischen Daten des Nutzers und seine Erfahrungen mit Annotationsinterfaces abgefragt.

3.4 Datensätze

Als Bilddatensätze während der Evaluation dienen jeweils zwei Sets mit je zehn Bildern. Das eine Bildset beinhaltet professionelle Fotos aus der Domäne Nachrichten. Diese Nachrichtenbilder sind aus September 2012 (Abbildung 3.1). Das andere Bildset besteht aus Semi-professionellen Fotos bis Amateuraufnahmen aus Flickr im Bereich Travel. Es beinhaltet zumeist berühmte Sehenswürdigkeiten (Abbildung 3.2). Die Nachrichtenbilder wurden zufällig von bekannten Nachrichtenportalen ausgewählt. Einige der Sehenswürdigkeiten des zweiten Datensatzes sind in weiter Entfernung oder aus der Luft aufgenommen worden, was den Vorteil birgt, dass selbst wenn ein Proband nicht die betreffende Sehenswürdigkeit kennt, er zumindest allgemeine Schlagworte, wie Fluss, Stadtzentrum oder Bahnhof setzen kann. Die Auswahl ist rein zufällig getroffen worden und erfolgte auf dem Portal Flickr. Jeder Proband erhält die gleichen zwei Bildersets, allerdings werden die Reihenfolgen je Bilderset randomisiert, sodass nie vorherzusehen ist, welches Bild als nächstes zu annotieren ist. Wie Tabelle 3.1 zu entnehmen ist, werden die Bildsets permutiert, sodass bestimmte Evaluationsergebnisse nicht auf die Kombination aus gewissem Bildset und Interface, sondern nur auf die Gestaltung des Interfaces zurück zu führen ist (Tabelle 3.1).

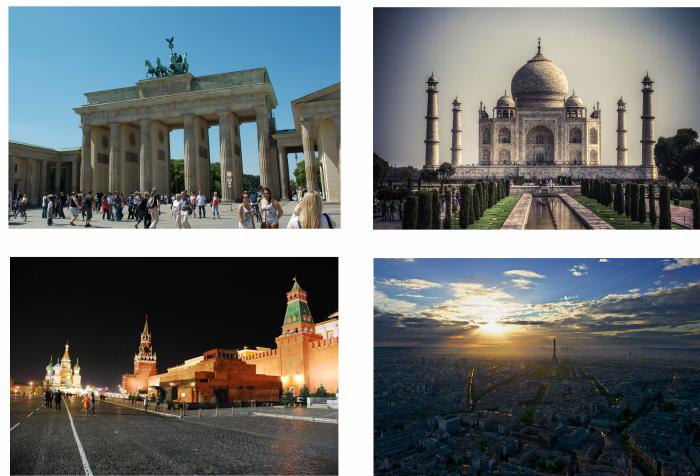


Abbildung 3.1: Bilderauswahl aus Bildset B1, Quellen siehe Anhang B



Abbildung 3.2: Bilderauswahl aus Bildset B2, Quellen siehe Anhang B

3.5 Datenerfassung

Vom Evaluations-Wizard bekommt jeder Nutzer eine ID zugewiesen, sodass auch bei Zugriff mehrerer Probanden zur gleichen Zeit die Messwerte eindeutig einem Nutzer zugeordnet werden können. Es wird pro Nutzer die Nummer der Gruppe, zu der ein Proband zugeteilt wurde (siehe Tabelle 3.1), ebenso gespeichert, wie der Startzeitpunkt an dem ein Nutzer die Evaluation beginnt. Darüberhinaus zu jedem Bild, der Bildname, die gesetzten Schlagworte, welche Schlagworte aus der Autocompletion und welche aus der Suggestion-Box ausgewählt wurden. Ein Zeitstempel macht es möglich bei der Auswertung Aussagen zur Effizienz machen zu können. Die Positionen der Schieberegler der Fragebögen werden als Werte zwischen 0 bis 100 gespeichert. Die zusätzlichen Antworten, die ein Nutzer bei jedem Fragebogen in das Textfeld eingeben kann, werden ebenfalls gesichert und mit in die Auswertung, gerade wenn es um die Frage der Benutzerzufriedenheit geht, einbezogen.

Quantitative Daten:

1. Zeit pro Interface
2. Anzahl der Tags
3. Anzahl der Tags, die aus der Autocompletion ausgewählt werden
4. Anzahl der Tags, die aus der Suggestion-Box ausgewählt werden

Werkzeuge:

1. Zeitstempel zu Beginn der Evaluation, je Bild und Fragebogen Format: Datum, Uhrzeit in Stunde, Minute und Sekunde
2. Mitloggen der Eingaben
3. Fragebögen: Interfacebetreffend und demografische Daten

Qualitative Daten:

1. Textfelder der Fragebögen
2. Persönliche Rückmeldung

3.6 Fragebögen

Oft werden Fragebögen mit einer 5-Punkte Likert Skala als mögliche Antwortoption erstellt. Hierbei muss sich der Nutzer auf eine von 5 möglichen Antwortoptionen festlegen. Diese Eigenschaft hat den Nachteil, dass sich kleine Unterschiede, wie der Benutzer sie empfindet, nicht darstellen lässt. Ein Proband, der alle Fragen mit einem mittleren Wert im Fragebogen ankreuzt, würde im persönlichen Gespräch dann vielleicht eröffnen, dass die Antwort auf die eine Frage eher zum positiven Bereich tendiert und eine andere eher zum negativen. Diese Nuancen gingen durch eine Einschränkung auf nur 5 Antwortoptionen verloren. Basierend auf den Studien von Funke [Fun10] sollte man gerade dann, wenn diese Messfehler vermieden werden sollen eine kontinuierliche Skala nutzen. Darüberhinaus fühlt sich der Nutzer damit weniger in seinen Antworten beeinflusst, da er feiner unterscheiden kann. Als Anhaltspunkt wurde neben den Beschriftungen am linken und rechten Rand der Skala, welche die positiven und negativen Antworten widerspiegeln, eine mittlere Beschriftung eingefügt (Abbildung 3.3). Abbildung 3.4 zeigt den positionierten Schieberegler, der erst dann sichtbar wird, wenn der Nutzer auf die Skala klickt, um die Frage mit Hilfe der Position des Schiebereglers zu beantworten.

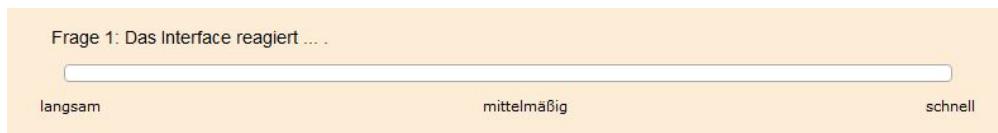


Abbildung 3.3: Frage 1 des Fragebogens mit Antwortskala



Abbildung 3.4: Frage 1 des Fragebogens nach Beantwortung mit Schieberegler

Dies soll deutlich machen, dass es nicht nur eine Entscheidung zwischen zwei Extremen, sondern auch einen neutralen, mittleren Bereich gibt. Der Fragenkatalog der Fragebögen baut auf den IsoMetrics nach Gediga und Hamborg [GH99] auf. Bei den IsoMetrics sind die Fragen in verschiedene Kategorien aufgeteilt, um alle Bereiche, die zur Untersuchung von graphischen Oberflächen interessant sind, abzudecken. Es werden folgende Kategorien unterschieden:

Aufgabenangemessenheit: Mit Fragen in dieser Kategorie möchte man Aufschluss darüber gewinnen, ob die Software den Nutzer unterstützt die gestellte Aufgabe zu lösen.

Selbstbeschreibungsfähigkeit: Ist die Software intuitiv verständlich und wenn nicht, gibt sie genügend weitere Informationen zur Erklärung ihrer Funktionen?

Steuerbarkeit: Diese Fragen klären, ob der Nutzer Kontrolle über die nächsten zu tätigen Schritte hat oder ob die Software den Nutzer in seinen Eingaben und Verhalten steuert.

Erwartungskonformität: Entspricht die Software den Erwartungen und Gebräuchen eines Nutzers

Fehlertoleranz: Wie zuverlässig arbeitet die Software auch wenn Fehler im Arbeitsablauf oder in den Eingaben vorliegen?

Individualisierbarkeit: Inwieweit lässt die Software Anpassungen durch den Nutzer zu, sodass der Nutzer diese auf seine Bedürfnisse, beispielsweise eine andere Schriftgröße anpassen kann?

Lernförderlichkeit: Kann ein Nutzer die Basisfunktionen schnell und eigenverantwortlich erlernen und ist es möglich tiefgehendere Funktionen mit Unterstützung durch die Software zu erlernen?

Es wurden auf Fragen aus dem Bereich der Individualisierbarkeit und Steuerbarkeit verzichtet, da diese Funktionen nicht in dem Evaluations-Wizard vorhanden sind. Waren Funktionen wie Individualisierbarkeit vorhanden, wäre es nicht mehr möglich die Ergebnisse der Probanden miteinander zu vergleichen. Aus den anderen Bereichen wurden insgesamt zehn Fragen ausgewählt. Diese Fragen wurden unter Berücksichtigung der Forschungshypothesen ausgewählt. Mit den Antworten der Probanden auf die Fragebögen soll untersucht werden ob H0 hinsichtlich der Benutzerzufriedenheit verworfen werden kann. Darüberhinaus stellen die Fragebögen einen elementaren Teil an Informationen bereit, der als Vergleichsgrundlage zwischen den einzelnen Interfaces genutzt werden kann. Die zehn Fragen und ihre Antwortskalen sind Tabelle 3.2 zu entnehmen.

#	Fragen und Antwortskalen (von - über - bis)
1.	Das Interface reagiert ... langsam - mittelmäßig - schnell
2.	Das Interface ist einfach zu bedienen. stimmt nicht - stimmt mittelmäßig - stimmt
3.	Das Interface bietet alle Funktionen, um die gestellten Aufgaben zu lösen. stimmt nicht - stimmt mittelmäßig - stimmt
4.	Das Interface bietet Hilfe oder Erklärungen auf Verlangen des Nutzers. stimmt nicht - stimmt mittelmäßig - stimmt
5.	Das Interface bietet einen guten Überblick über das Funktionsangebot stimmt nicht - stimmt mittelmäßig - stimmt
6.	Das Interface liefert genügend Informationen darüber, welche Eingaben nötig sind stimmt nicht - stimmt mittelmäßig - stimmt
7.	Das Interface lässt den Nutzer wissen, ob eine Eingabe erfolgreich war. stimmt nicht - stimmt mittelmäßig - stimmt
8.	Das Interface lässt sich durchgehend nach einem Prinzip bedienen stimmt nicht - stimmt mittelmäßig - stimmt
9.	Das Interface erfordert wenig Zeit zum Lernen. stimmt nicht - stimmt mittelmäßig - stimmt
10.	Korrekturen von Fehler sind ... nicht möglich - teilweise möglich - gut möglich

Tabelle 3.2: Die zehn Fragen der interfacebezogenen Fragebögen

3.7 Testumgebung

Als Testumgebung steht ein selbst implementierter webbasierter Evaluations-Wizard bereit. Dieser Wizard verwaltet die im Abschnitt 3.2 vorgestellten Kombinationen aus Interfaces und Bildsets. Die Rückmeldungen der Probanden zu den Interfaces erhält man zum einen über die ausgefüllten Fragebögen, mit der Option die Fragebögen durch ein Freitextfeld zu erweitern und zum anderen, indem die unter Abschnitt sec:Datenerfassung beschriebenen Daten jedes einzelnen Probanden erfasst und ausgewertet werden. Die Interfaces, welche die im Kapitel Hintergrund vorgestellten Kreuzverweis Services abrufen, sollen im Folgenden in ihrem Aufbau und ihrer Implementation, ebenso wie der Evaluations-Wizard und der technische Hintergrund von Kreuzverweis beschrieben werden

Analyse der Ideen und des Kreuzverweis Interfaces

Begonnen wurde mit der Analyse zur Anordnung und Nutzbarkeit des Kreuzverweis Webinterfaces. Die User Interface Elemente sind beim Kreuzverweis Webinterface untereinander angeordnet (Abbildung 3.5). Daher nahm man als Basis an, dass das zu annotierende Bild ebenfalls in diese vertikale Anordnung zu erst eingesortiert würde.

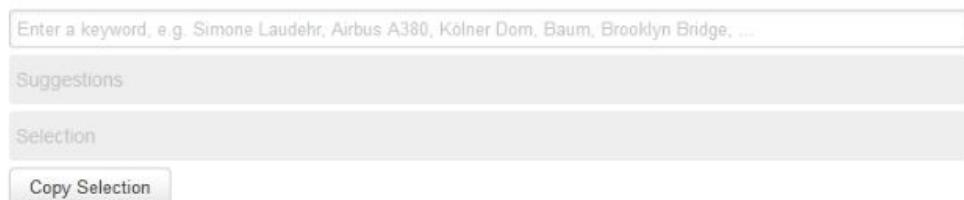


Abbildung 3.5: Das Kreuzverweis Webinterface¹

Daraus bildet sich als Reihenfolge in absteigender Anordnung: Das Bild, das Textfeld zum Eingeben der Annotations Tags, darunter die Vorschläge, im Folgenden Suggestions-Box genannt und zu Unterst die ausgewählten Tags, die sich in der Selections-Box befinden. Aus dieser Anordnung ergeben sich einige Probleme. Wenn das Bild nicht im Querformat vorliegt, dann muss je nach Auflösung gescrollt werden. Zu Scrollen stellt an sich kein Gegenargument dar. Viele Studien² belegen, dass das Scrollen entgegen vieler Vermutungen zum Beispiel auf Websites

¹Screenshot: <http://kreuzverweis.herokuapp.com/>, Stand Juli 2012

²<http://www.useit.com/alertbox/scrolling-attention.html>

intensiv genutzt wird. Das Gegenargument ist vielmehr die Voraussetzung scrollen zu müssen, um das Interface benutzen zu können. Folglich ein ständiger Wechsel zwischen Benutzung des Interfaces und damit des Annotationsprozesses und Bewältigung der äußereren Gegebenheiten des Interfaces. Darüber hinaus besteht die Gefahr bei schneller Benutzung die Suggestions und Selections-Felder miteinander zu vertauschen und somit viele Annotationstags nicht zu setzen, da man davon ausgeht, diese bereits ausgewählt zu haben.

Eine weitere Idee bestand in der Variation der Anordnung. Gespräche mit Experten und den Gründern von Kreuzverweis führten zu dem Ergebnis, dass die Anordnung der einzelnen UI Elemente innerhalb des zweispaltigen Layouts weniger interessant ist und die Resultate zwischen einem Layout mit links oder rechts angeordnetem Bild nur minimal variieren würden. Betrachtet man verwandte Arbeiten, sieht man die unterschiedlichsten Anordnungen und Handhabungen mit Annotationstags und die Darstellung von Relevance Scores. Eine Möglichkeit ist die Tag Cloud. Innerhalb derer werden nochmals verschiedene Darstellungsformen, seien es (Schrift-)Farbe der Tags und mit oder ohne button-ähnliche Umrandungen, und Anordnungsformen, ob rein alphabethisch, sequentiell, geclustert oder zirkulär, unterschieden. Hierbei kann die Größe der Schrift, des Buttons oder die Helligkeit der Farbe die Relevanz wiederspiegeln (Abbildung 3.6) [RGMM07], [LTG11].

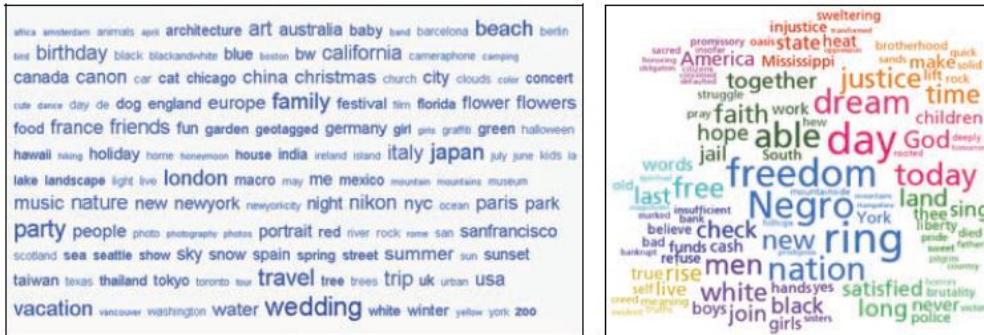


Abbildung 3.6: Darstellungsformen von Tag Clouds [LZT09]

[LZT09] hat mit Hilfe von Eyetracking Studien herausfinden können, dass die Aufmerksamkeit des Nutzers auf die großen Tags gelenkt werden und die kleinen zumeist gar nicht wahrgenommen und in den Auswertungen der anschließenden Fragebögen, um die subjektive Meinung der Nutzer zu erfahren, als spontan uninteressant eingestuft werden. Darüberhinaus konnte man den Eyetracking Ergebnissen entnehmen, welcher der Probanden es seitens seiner Muttersprache gewohnt ist von links nach rechts zu lesen: Diese haben neben der Bildschirmmitte, was im Besonderen bei der zirkulären Anordnung zutrifft, den linken oberen Bereich besonders oft und intensiv betrachtet.

Innerhalb der Expertenrunde haben wir auch die Kombinations- und Gestaltungsmöglichkeiten der einzelnen Userinterface-Elemente diskutiert. Gegen eine Anordnung und Handhabung der Tags als Tag Cloud sprachen nicht nur die bereits genannten Erkenntnisse, dass man durch die Anordnung und die Gestaltung der Tags zu sehr Einfluss auf die Auswahl der Tags nehmen würde, sondern auch die Bedenken, dass gerade die schwer zu lesenden kleinen oder sehr weit am Rand liegenden Tags den Nutzer dazu verleiten könnten, sich von seiner eigentlichen Aufgabe ablenken zu lassen.

Als weiteren Grund gegen verschiedene große Buttons und Schriftarten spricht nicht nur die Aufmerksamkeitssteuerung, sondern auch die Unterbrechung des Leseflusses und schnellere Ermüdung des Nutzers, durch eine wellen-ähnliche Leselinie, die die lese-typischen Sakkaden unterbricht, und der parallel stattfindenden wechselnden Akkommodation (Nah- und Fernakkommodation)³. Diese resultierende ungleichmäßige Anordnung bei verschiedener Größe gilt es also unbedingt zu vermeiden. Gegen die letzte Darstellungsform - die Helligkeit der Farbe spricht der, aus der Dunkelheit des Hintergrundes, nötige Wechsel der Schriftfarbe zu einer helleren, besser lesbaren Farbe. Auch hier würde der Nutzer zu sehr von der eigentlichen Aufgabe abgelenkt und beeinflusst werden.

Da die meisten der oben genannten Darstellungsformen nur in Kombination mit einer der oben genannten Relevanz-Darstellungsformen sinnvoll sind, muss nur noch die alphabetische Anordnung als Alternative zur jetzigen Kreuzverweis Anordnungsform der Tags untersucht werden. Durch alle bisher genannten Nachteile der Relevanzdarstellungen hätte man mit einer alphabetischen Sortierung, keinerlei Möglichkeiten die Relevanz auf zurückhaltende Art darzustellen. Bei einem System, das den Nutzer unterstützen, aber nicht beeinflussen möchte, darf die Relevanz dennoch nicht gänzlich außer Acht gelassen werden.

Bedenkt man des weiteren auch noch, dass es sich bei dem Suggestions-Bereich, um den Teil handelt, der immer wieder mit neuen Schlagworten aktualisiert und erweitert wird, ergibt sich hieraus ein weiteres Gegenargument, da die neuen Schlagworte alphabethisch eingesortiert werden müssten. Diese ständigen Positionswechsel der Tags würden den Nutzer zu sehr ablenken und irritieren. Tabelle 3.3 gibt noch einmal eine Übersicht über alle diskutierten Anordnungen, Ideen und Darstellungsformen.

³<http://www2.hu-berlin.de/reha/eye/Blickbewegung\%20beim\%20Lesen.pdf>

	Diskutierte Ideen	Pro (P)/Contra (C)	Entscheidung
1.	Scrollen vermeiden	Lenkt den Nutzer zu sehr von der Annotationsaufgabe ab (P)	bei Fragebögen auf Grund der Anzahl der Fragen nicht möglich, aber generell hier auch nicht nötig
2.	2-spaltiges Layout	Übersichtlichkeit Unterstützt 1. (P)	Übernommen für alle Interfaces
3.	Variation in der Anordnung rechts und links	Unwesentlicher Unterschied (C)	Nicht in die Evaluation aufgenommen
4.	Suggestion-Tags als Tag Cloud darstellen	Unübersichtlich, Verwirrend Eingewöhnungszeit (C)	Nicht übernommen
5.	Relevance Score durch Größe der Tags darstellen	Aufmerksamkeitssteuernd, anstrengend zu Lesen (C)	Nicht übernommen
6.	Relevance Score durch Helligkeit der Farbe des Tags darstellen	Aufmerksamkeitssteuernd, anstrengend zu Lesen, bei Farbenblindheit kritisch (C)	Nicht übernommen
7.	Alphabetische Anordnung der Tags	Es müssen immer alle Tags gelesen werden (C)	Nicht übernommen, Anordnung nach Relevance Feedback sinnvoller

Tabelle 3.3: Diskutierte und zum Teil begründet verworfene Ideen

Interfaces

Da es bisher noch keine derartige, wie die hier vorgestellte, Benutzerstudie im Bereich der Annotationsinterfaces gibt, wurden zunächst Ideen für die Gestaltung der Interfaces gesammelt. Bereits zu Beginn der Arbeit war der Einsatz von den Kreuzverweis Services klar. Es sollte nicht nur ein Interface geschaffen werden, was basierend auf den visuellen und gestalterischen Aspekten unterstützt, sondern auch auf inhaltlicher Ebene. Die Kreuzverweis Services umfassen nicht nur eine Autocompletion Funktion, sondern auch einen intelligenten Algorithmus, um basierend auf eingegebenen Schlagworten, weitere in Reihenfolge ihres Relevance Feedbacks in Bezug auf die eingegebenen Schlagworte, vorzuschlagen.

Nach Analyse der gesammelten Ideen und der Gestaltung des von Kreuzverweis bereitgestellten Webinterfaces⁴ zur testweisen Nutzung ihrer Services, zum einen durch die im Kapitel 2 beschriebenen Verwandten Arbeiten und zum anderen durch eine Befragung von Experten, im Umfeld des Semantic Web, ergaben sich letztendlich vier zu vergleichende Interfaces.

Interfaceaufbau

Basierend auf der Analyse fiel der Entschluss auf ein zweispaltiges Layout, bei dem zu aller Wahrscheinlichkeit, auch wenn Bilder im Hochformat vorliegen, nicht gescrollt werden muss. Die Breite der beiden Spalten beträgt jeweils die Hälfte der Bildschirmbreite. In der linken Spalte befindet sich als einziges das Bild, um auch bei kleinen Auflösungen eine gewisse Bildgröße garantieren zu können. In der rechten Spalte findet man zunächst das „Enter a keyword ...“-Textfeld und darunter, innerhalb einer dezent farblich abgesetzten und mit „Suggestions“ benannten Box, die möglichen Vorschläge. Unter der Suggestions-Box ist die auf gleiche Weise farblich dezent hervorgehobene und beschriftete Selections-Box zu finden.

Bildbereich

Da die Bilder unterschiedliche Formate und Größen haben dürfen, das menschliche Auge aber die kleinsten Sprünge zwischen zwei Bildpositionen in zwei hintereinander folgenden Sequenzen wahrnehmen kann, befindet sich die linke obere Bildecke immer an der gleichen Bildschirmkoordinate. Sollte ein Bild breiter als die linke Spalte sein, wird es die Proportionen erhaltend skaliert.

⁴<http://kreuzverweis.herokuapp.com/>

Keyword-Textfeld

Das Webinterface von Kreuzverweis bietet bereits eine Autocompletion. Sobald ein Nutzer hier Buchstaben in das Keyword-Textfeld eingibt, erscheint über ein Dropdown Fenster unterhalb des Textfeldes eine passende Auswahl an möglichen Wörtern. Diese Funktion wird im Prototyp übernommen, da so Rechtschreibfehler zum Großteil ausgeschlossen und der Nutzer bei der Wortfindung unterstützt werden kann.

Selections-Box

Ein Tag wird in die Selections-Box übernommen, indem der Nutzer entweder Enter nach Eingabe des Keywords drückt, Enter nach Auswahl aus der Autocompletion-liste drückt oder aber aus der Suggestions-Box einen Tag auswählt, indem er es anklickt. Sobald ein Keyword in der Selections-Box erscheint, ändert sich dessen Hintergrundfarbe zu Blau, um die getroffene Auswahl deutlicher von den anderen Tags zu trennen.

Suggestions-Box

Die Abwägungen aus dem Abschnitt sec:Analyse führen zu folgendem Fazit für die Suggestions-Box: Die Tags erhalten alle die gleiche Schriftgröße und -farbe. Sie sind als anklickbare Buttons mit gleicher Höhe und variabler Breite dargestellt. Ihre Sortierung erfolgt nach Relevance Score, sodass die relevanten Tags an den ersten Stellen beginnend bei der linken oberen Ecke der Suggestions-Box zu finden sind. Abbildung 3.7 zeigt ein Interface mit bereits gesetzten Tags und Vorschlägen.

Schlagworte

Die Schlagworte oder auch Tags erscheinen in der Selections-Box mit blauem und in der Suggestions-Box mit hellem Hintergrund. Beide Designs sind mit Farbverlauf und dezentem Rahmen, sodass die Button-Funktionalität, welche sich hinter den Tags verbirgt, auch optisch hervorgehoben wird. Klickt man auf die Buttons in der Suggestions-Box werden diese in die Auswahl, also die Selections-Box, übernommen und tauchen hier mit entsprechend geänderter Optik an erster Stelle beziehungsweise nach dem zuletzt ausgewählten Tag auf. Möchte man ein bestimmtes Tag aus seiner Auswahl entfernen, kann man dieses mit einfachem Klick auf den Button des betreffenden Schlagwortes herauslöschen. Wurde ein Schlagwort in die Auswahl übernommen und anschließend doch wieder aus der Selection herausgelöscht, so taucht es nicht sofort wieder in der Suggestions-Box auf. Sollte man aber weitere Keywords in das Keyword-Feld eingeben, welche basierend auf der

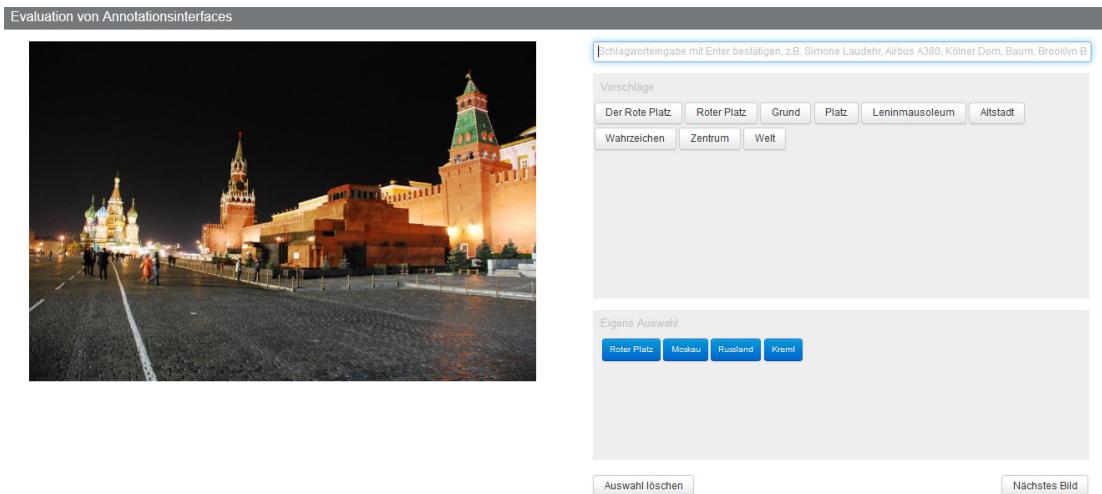


Abbildung 3.7: Interface mit 3 User Interface Elementen und Vorschlägen

Relevanz zu dem eben herausgelöschten Keyword passen, taucht dieses wieder als mögliche Option in der Suggestions-Box auf. Es wird also kein Keyword für immer als mögliches Tag aus der Datenbank oder auch für einen Annotationsvorgang eliminiert. Zur Übersicht seien hier noch einmal die Basiseigenschaften der Interfaces genannt (Abbildung 3.8):

1. Zweispläntiges Layout, links Bildbereich, rechts für Annotation
2. Keyword-Textfeld
3. Suggestions-Box
4. Selections-Box
5. Button zum Löschen aller gesetzten Tags
6. Button für das nächste Bild

Variationen

Die Basiseigenschaften der Interfaces werden durch die im Absatz sec:Hypothesen beschriebenen unabhängigen Variablen erweitert. Der Evaluations-Wizard bestimmt, ob ein Nutzer die Between-Gruppen Variable Merger, die festlegt, ob das Interface aus zwei oder drei User Interface Elementen besteht, zu bearbeiten hat und mit welchem Zustand der zweiten Variable Farbe, er die Evaluation beginnt. Die zweite

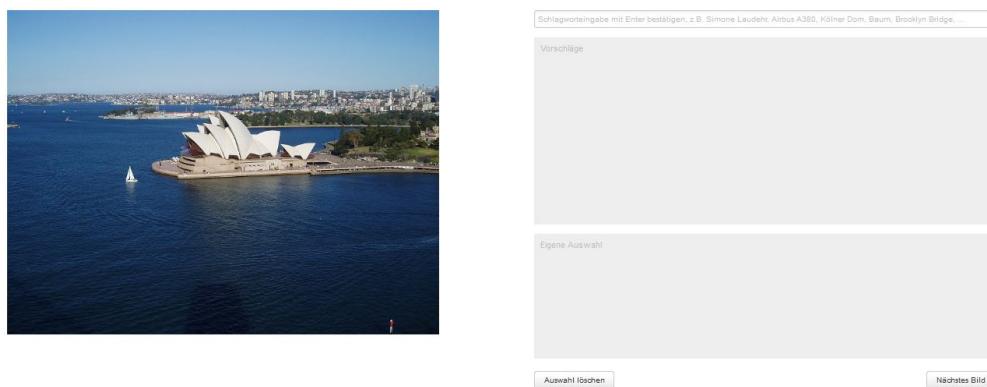


Abbildung 3.8: Interface mit 3 User Interface Elementen, Beschriftung auf Deutsch

Variable Farbe ist als Within-Gruppen Test festgelegt, daher hat jeder Nutzer ein Interface welches ohne und welches mit Einfärbung der Tags zu absolvieren ist. Der Wizard bestimmt lediglich die Reihenfolge, in der diese Zustände durchlaufen werden.

Erste Variation - Merger

Diese Variation ist inspiriert von Google+. Tags können direkt im Textfeld als solche durch das Hinzufügen eines vorangestellten „#“-Zeichens definiert werden und verwandeln sich als solche sichtbar, sobald der Nutzer ein Leerzeichen eintippt [Coh12]. Um die Konsistenz zu wahren, gibt es eine Autocompletion Funktion. Es existieren keine freizugänglichen Nutzerstudien hierzu, aber die Vermutung, dass ein Nutzer auf diese Weise schneller tippen und annotieren kann, liegt nahe. Um diese Vermutung näher zu untersuchen, belegte man die erste Variable mit dieser Eigenschaft. Diese Eigenschaft wird im Folgenden Merger oder 2- und 3-elementiges User Interface genannt.

Die erste Variation und damit Vergleichsmöglichkeit wird aus dem Verschmelzen vom Keyword-Feld und der Selections-Box gebildet. In der linken Spalte befindet sich nur das Bild. In der rechten Spalte liegt nun das Keyword-Textfeld auf der Selections-Box. Die Suggestions-Box wird auf der y-Achse nach unten verschoben, sodass diese unterhalb der Selections-Box positioniert wird. Oberhalb vom Keyword-Textfeld steht nun „Selections“ und das Textfeld wird von der farblich abgehobenen Selections-Box umsäumt. Im Textfeld steht anfangs weiterhin „Enter a keyword ...“ (Abbildung 3.9).

Bestätigt der Nutzer ein Keyword im Textfeld, verwandelt es sich hier zu einer Auswahl, indem das Tag zu einem Button wird und die gleiche Farbe annimmt, wie in der Selections-Box der Grundkonstellation. Der Cursor wandert nun hinter den



Abbildung 3.9: Interface mit 2 User Interface Elementen, Beschriftung auf Deutsch

im vorherigen Schritt festgelegten Selection-Button, um auf eine weitere Eingabe zu warten. Um weitere Tags hinzuzufügen oder seine eigenen Eingaben korrigieren zu lassen, kann der Nutzer auch hier wieder auf die gleiche Funktionalität der Suggestions-Box und des Autocompletion- Dropdownfensters zurückgreifen.

Zweite Variation - Farbe

Die zweite Variation ist inspiriert von dem im Kapitel 2 beschriebenen Voxalead⁵ [LTG11]. Die Schlagworte werden im Interface nach ihrer Relevanz angeordnet. Da diese Anordnung zu jedem neuen Bild und jedem neuen Keyword eine andere ist, ist diese Anordnung für den Nutzer weniger verständlich. Generell werden basierend auf dem Relevance Feedback Schlagworte zur Verfügung gestellt, die zum betreffenden Bild passen. Es besteht dennoch die Gefahr, dass der Nutzer sehr lange benötigt, um ein spezifisches Tag zu finden, da hierfür alle Tags durchgelesen oder betrachtet werden müssen. Um dieser Gefahr entgegen zu wirken, sollen die Tags gruppiert werden. Es gibt verschiedene Möglichkeiten Tags zu gruppieren, sei es für jede Schlagwortgruppe eine eigene Suggestions-Box zu präsentieren oder die Tags den Gruppen nach, ähnlich der alphabetischen Sortierung, anzugeordnen. Keine dieser möglichen Lösungen bietet eine Anordnung nach Relevance Feedback, die es zeitgleich auch ermöglicht die Inhalte der Tags schnell und problemlos zu erfassen. Einzig die auf Voxlead basierende farbliche Kennzeichnung der Gruppenzugehörigkeit, kann gleichzeitig die Anordnung nach Relevanz in einer einzigen Suggestions-Box bereitstellen.

⁵<http://voxaleadnews.labs.exalead.com/>

Der Evaluations-Wizard

Der Wizard organisiert auf der einen Seite die Nutzeranfragen und verteilt an die Nutzer die Interfaces, auf der anderen Seite handhabt er auch die Speicherung der Daten, die ein Nutzer eingibt, und die, die über die Dauer der Evaluation mitgeloggt werden. Außerdem stellt er die Kommunikationsschnittstelle zu den Kreuzverweis Services dar. Da man erreichen möchte, dass alle Interface Kombinationen (Tabelle 3.1) von gleich vielen Nutzern absolviert werden, gibt der Wizard die Kombinationen nicht in einer zufälligen Reihenfolge an die Nutzer aus, sondern in einer kontrollierten.

Anforderungsliste

Hier sollen noch einmal alle Anforderungen, die die Interfaces und der Evaluations-Wizard zu erfüllen haben, aufgeführt werden, um bei der Implementation berücksichtigt werden zu können.

1. Die Evaluation kann vom Nutzer in einem Internetbrowser seiner Wahl durchgeführt werden.
2. Der Evaluations-Wizard verteilt die Interface Kombinationen gleichmäßig an die Probanden.
3. Der Evaluations-Wizard steuert, dass jeder Proband zehn Bilder pro Interface annotiert und anschließend jeweils die Fragenbögen ausfüllt.
4. Es sollen über die ganze Dauer der Evaluation die Zeiten, die ein Nutzer auf einem Interface oder einem Fragebogen verbringt gespeichert werden.
5. Es sollen alle Daten die ein Nutzer auf den Interfaces und den Fragebögen eingibt gespeichert werden.
6. Es sollen alle Daten, die ein Nutzer über die Kreuzverweis Services für seine Annotation gewinnt gespeichert werden.
7. Das System soll überprüfen, ob ein Nutzer alle Fragen der Fragebögen beantwortet hat und den Nutzer gegebenenfalls auf fehlende Angaben hinweisen.
8. Die Daten sollen in einem Format gespeichert werden, sodass sie für eine anschließende statistische Analyse verwendbar sind.

Interfaces

9. Der Nutzer soll die Möglichkeit haben, Tags definieren und Tags löschen zu können

10. Der Nutzer soll die Möglichkeit haben, so viele Tags zu setzen, wie er möchte.
11. Der Nutzer soll die Möglichkeit haben, alle Tags mit einem Klick löschen zu können.
12. Das Interface soll bevor Tags gelöscht werden, eine Bestätigung des Nutzers zur Löschung einfordern.
13. Das Interface stellt dem Nutzer die Kreuzverweis Autocompletion und Suggestions Services zur Verfügung.

Wizard

14. Der Wizard soll die Kommunikation über OAuth2 mit den Kreuzverweis Servern für die Nutzung der Kreuzverweis Services sicherstellen.

3.8 Implementation

Die Implementation des Systems umfasst zu viele einzelne Teilbereiche, sodass eine Beschreibung an dieser Stelle zu sehr von der Evaluation und der Analyse der Ergebnisse ablenkt. Eine tiefergehendere Beschreibung findet sich im Anhang A. Im nachfolgenden Teil, soll nur so viel erläutert werden, dass der Ablauf während der Evaluation und die stattfindende Kommunikation zwischen den einzelnen Komponenten klar wird.

Interfaces - Clientseitige Implementation

Die Interfaces bauen auf dem von Kreuzverweis zur Verfügung gestellten Basissystem auf⁶. Das HTML und CSS Grundgerüst wurde um weitere Tags und Attribute erweitert, um die im Kapitel Prototyp beschriebenen Anforderungen an die neuen User Interfaces erfüllen zu können. Auch hier wurde auf JavaScript und jQuery zurückgegriffen, um eine erweiterte Funktionalität zu erhalten. Neben den bereits vorhandenen JavaScript Funktionen wurden weitere erstellt, um den Evaluations-Ablauf und alle die Evaluation betreffenden Elemente, wie beispielsweise die randomisiert wechselnden Bilder, die vom Testsystem definierte Probanden ID oder die vom Probanden eingegebenen Tags, in Kommunikation mit dem Server steuern zu können. Neben den vier Interfaces, die es zu untersuchen gilt, basieren der Welcome-Screen, der den Nutzer begrüßt und ihm einiges zum Evaluations-Ablauf mitteilt, ebenso wie die Fragebögen zur Evaluation, auch auf HTML, CSS und JavaScript.

⁶https://github.com/kreuzverweis/smartkeywroding_web.git

Webserver - Serverseitige Implementation

Der Webserver wurde auf einer linuxbasierten virtuellen Maschine auf den Universitätsservern installiert. Ubuntu LTS 10.04 stellt das Betriebssystem, welches um Apache 2.4.2 und PHP5 erweitert wurde. Darüberhinaus benötigt das implementierte System die PECL Erweiterungen `pecl_http`⁷, welches ermöglicht mit PHP HTTP Requests zu stellen, und `pecl_oauth`⁸, welches die vom Kreuzverweis Backoffice⁹ vorgeschriebene Authentifizierung mittels Oauth 2 unterstützt. Die in PHP erstellten Klassen kontrollieren von hier aus nicht nur die Kommunikation der Interfaces respektive der Probanden mit dem Evaluations-Wizard, sondern garantiert auch die vom Nutzer weiterhin unbemerkte Authentifizierung mit dem Kreuzverweis Webservice Server. Hierzu wurden die im Anhang A Abschnitt Oauth 2 Authentifizierung und Kommunikation beschriebenen Methoden und Authentifizierungsabläufe in PHP implementiert und auf den Ablauf des Evaluationssystems abgestimmt.

Evaluations-Wizard - Kommunikation

Sobald ein Nutzer auf der Willkommens-Seite des Evaluationssystems den Start-Button drückt, bekommt er eine ID und eine Interfacekombination, wie sie Tabelle 3.1 zu entnehmen ist, zugewiesen. Sobald das erste Bild erscheint wird ein Counter aktiviert, sodass ein Nutzer nach den zehn annotierten Bildern aus dem betreffenden Bildset, welches ebenfalls Tabelle 3.1 zu entnehmen ist, den ersten Fragebogen ausfüllen muss. Anschließend muss der Nutzer die nächsten zehn Bilder mit dem zweiten Interface annotieren. Eines der beiden Interfaces besitzt die Typinformation der Tags codiert mit Farbe. Da diese Funktion nicht intuitiv ist, wird dem Nutzer, sobald ein Interface mit Farbe zum ersten Mal vom Wizard angesteuert wird, eine Meldung gezeigt, die die Typkategorien und die zugehörige Farbe zeigt (Abbildung 3.10). Wenn der Nutzer alle 20 Bilder annotiert und die beiden Fragebögen bezüglich der Interfaces beantwortet hat, muss er nur noch einen abschließenden Fragebogen beantworten, der demographische Daten abfragt. Der Wizard überprüft, ob alle Fragen beantwortet wurden. Die Freitextfelder sind hierbei optional auszufüllen, fließen aber in die spätere Betrachtung der Daten mit ein.

Während des Annotationsprozesses speichert der Wizard die Zeit, die ein Nutzer für ein Bild benötigt. Darüberhinaus werden nicht nur die Tags, die ein Nutzer manuell eingibt, sondern auch die Tags die er aus der Autocompletion und der Suggestions-Box auswählt, in die Daten der Session mit aufgenommen. Löscht ein

⁷http://pecl.php.net/package/pecl_http

⁸<http://pecl.php.net/package/oauth>

⁹<https://backoffice.kreuzverweis.com/>



Abbildung 3.10: Informationsmeldung zu den Interfaces mit Typinformation

Nutzer Tags, so wird auch das bei dem betreffenden Tag vermerkt. Der Evaluations-Wizard organisiert auch die Kommunikation zwischen dem Client und dem Kreuzverweis Webservice Server. Jedes Tag, was der Nutzer eingibt, wird an den Kreuzverweis Webservice Server weitergegeben. Die Antwort mit Autocompletion-Inhalten oder Tag-Vorschlägen vom Kreuzverweis Server wird vom Evaluations-Server verarbeitet und an das Interface übergeben. Sobald ein Nutzer neue Tags eingibt, einen Eintrag aus der Autocompletion oder aus der Suggestions-Box auswählt, wird mit diesen hinzugefügten Inhalten eine neue Anfrage gestartet und bereits vorhandene Tag-Vorschläge in der Suggestions-Box gegebenenfalls anders eingesortiert, da sich mit jeder Anfrage auch das Relevance Feedback verändern kann.

Kapitel 4

Ergebnisse

Die Ergebnisse wurden mit OpenOffice Version 3.4 und SPSS 19, einem IBM Statistik Programm ausgewertet. Tabelle 4.1 sind die Variablen zu entnehmen, die sich aus den Daten der Fragebögen und der gespeicherten Sessions ergeben. Diese Daten werden im Folgenden statistischen Analysen unterzogen, um mit den Ergebnissen die in 3.1 aufgestellten Hypothesen zu untersuchen.

SPSS Variablen und ihre Bedeutungen		
Interface ohne Farbe	Interface mit Farbe	Bedeutung
Reaktionszeit	Farbe_Reaktionszeit	Das Interface reagiert ...
Bedienbarkeit	Farbe_Bedienbarkeit	Das Interface ist einfach zu bedienen.
Funktionalität	Farbe_Funktionalität	Das Interface bietet alle Funktionen, ... (Tabelle 3.2)
Hilfe	Farbe_Hilfe	Das Interface bietet Hilfe oder Erklärungen ... (Tabelle 3.2)
Überblick	Farbe_Überblick	Das Interface bietet einen guten Überblick ... (Tabelle 3.2)
Eingaben	Farbe_Eingaben	..., welche Eingaben nötig sind. (Tabelle 3.2)
Erfolgreich	Farbe_Erfolgreich	..., ob eine Eingabe erfolgreich war. (Tabelle 3.2)
Bedienung	Farbe_Bedienung	... lässt sich durchgehend nach einem Prinzip bedienen. (Tabelle 3.2)
Lernen	Farbe_Lernen	... erfordert wenig Zeit zum Lernen. (Tabelle 3.2)
Fehlerkorrektur	Farbe_Fehlerkorrektur	Korrekturen von Fehler sind ... (Tabelle 3.2)
Bildset	Farbe_Bildset	Wert: B1 oder B2. Ein Nutzer hat beide Bildsets zu bearbeiten
D_Tags	Farbe_D_Tags	Wieviele Tags hat ein Nutzer im Durchschnitt gesetzt
D_Suggestions	Farbe_D_Suggestions	Wie oft wurde durchschnittlich ein Tag aus den Vorschlägen übernommen
D_AutoComp	Farbe_D_AutoComp	Wie oft wurde durchschnittlich ein Tag aus der Auto Completion übernommen
D_Zeit	Farbe_D_Zeit	Wieviel Zeit benötigte ein Nutzer im Durchschnitt für ein Interface
Merger		Between-Group Variable. Hat ein Nutzer 2 oder 3-elementige Interfaces bearbeitet

Tabelle 4.1: Übersicht über die SPSS Variablen und ihre Bedeutungen

4.1 Teilnehmer

Der Anwendungszweck der Interfaces zielt auf den semi- bis professionellen Bereich der Bildannotation ab. Daher werden die ersten Probanden über die InetBib¹ gewonnen. Die InetBib ist eine E-Mailliste, die von der Universitätsbibliothek Dortmund 1994 ins Leben gerufen wurde und mittlerweile 7000 Mitglieder umfasst². Sie diente ursprünglich dem Zweck das Internet in die Bibliotheksarbeit einzugliedern. Dadurch sind gerade hier sehr viele Dokumentare und Archivare zu finden. Mittlerweile diskutiert diese Liste auch über neue Technologien und aktuelle Themen, die sich mit der Arbeit rund um das Bibliothekswesen, also auch der Verschlagwortung von Medien, befassen.

Weitere Probanden für die Evaluation werden über die Universität gewonnen. Zunächst werden hier diejenigen, die bereits Erfahrung im Gebiet der Annotation gesammelt haben, und anschließend Probanden, die keinerlei Vorerfahrung in der Verschlagwortung haben, angesprochen. Die Ergebnisse der Probanden mit und ohne Vorerfahrung werden separat ausgewertet, sofern sich hier elementare Unterschiede in den ausgewerteten Fragebögen ergeben sollten. Besonderes Augenmerk wird hierbei auf die Selbstbeschreibungsfähigkeit und die Aufgabenangemessenheit gelegt, da ein Evaluationsteilnehmer, der beruflich mit Annotationsinterfaces arbeitet, eine realistischere Einschätzung hierzu abgeben kann, als ein Laie.

An der Evaluation haben 36 Probanden teilgenommen. Drei Probanden haben intuitiv kein Enter nach Eingabe von einem Schlagwort gedrückt, sodass keine Suchanfrage generiert und an den Kreuzverweis Service geschickt werden konnte. Drei weitere Probanden haben die Evaluation nicht bis zum Ende durchgeführt. Diese sechs Probanden wurden bei der Auswertung nicht berücksichtigt. Von den 30 Probanden, deren Werte analysiert werden, sind 12 weiblich und 18 männlich. Vier der Probanden haben angegeben, dass sie bereits mit anderen Annotationsinterfaces gearbeitet haben. Das Durchschnittsalter der Probanden beträgt 28 Jahre. Im Durchschnitt haben die Evaluationsteilnehmer eine 14-jährige Erfahrung (Wert: 2,57) mit der Arbeit am Computer. Hierbei beträgt die Standartabweichung 1,1, die ungefähr 5,5 Jahren entspricht.

4.2 Effektivität

Eine der Forschungshypothesen, die untersucht werden, lautet F4: "Das Verwenden von Farbe als Hinweis zur Typinformation beeinflusst die Effektivität positiv". Darüber hinaus gilt es zu klären, ob die Anzahl der gesetzten Schlagworte von der Art der Bilder abhängen könnte, da gewisse Bilder bessere Vorschläge der

¹<http://www.inetbib.de/>

²<http://www.inetbib.de/wasistinetbib.html>

Kreuzverweis Services zur Folge haben, als andere. Außerdem könnte eine höhere Nutzerzufriedenheit ebenfalls dazu führen, dass mehr Schlagworte gesetzt werden. Die Mittelwerte der gesetzten Schlagworte, werden also auch in Relation zu dem Merger untersucht, da hier F2 besagt: „Ein Interface Design mit 2 User Interface Elementen erhöht die Benutzerzufriedenheit“. Es wurden im Mittel 3,75 Schlagworte gesetzt. Dabei erzielte ein Interface ohne Farbe einen Mittelwert von 3,58 Schlagworte pro Bild mit einer Standardabweichung von 1,56 (Abbildung 4.1) und ein Interface mit Farbe 3,92 Schlagworte pro Bild mit einer Standardabweichung von 1,58 (Abbildung 4.2).

D_Tags

Bildset	Mittelwert	N	Standardabweichung	Maximum	Minimum
Bildset 1	3,4889	12	1,93024	8,00	,00
Bildset 2	3,6383	18	1,30883	6,20	1,89
Insgesamt	3,5786	30	1,55659	8,00	,00

Abbildung 4.1: Mittelwert der gesetzten Tags eines Interfaces ohne Farbe

Farbe_D_Tags

Farbe_Bildset	Mittelwert	N	Standardabweichung	Maximum	Minimum
Bildset 1	3,6872	18	1,25230	6,63	2,30
Bildset 2	4,2685	12	1,99345	8,70	1,60
Insgesamt	3,9197	30	1,58447	8,70	1,60

Abbildung 4.2: Mittelwert der gesetzten Tags eines Interfaces mit Farbe

Setzt man nun die abhängigen Variablen D_Tags und Farbe_D_Tags, die als Platzhalter für die durchschnittliche Anzahl der gesetzten Tags bei den Interfaces ohne und mit Farbe stehen, in Relation zu den Variablen für die Bildsets, die sowohl für die Interfaces ohne und mit Farbe existieren (Bildset und Farbe_Bildset), ergeben sich nachfolgende Mittelwerte:

Für ein Interface ohne Farbe unterscheiden sich die Mittelwerte für Bildset 1 (Mittelwert: 3,49) unwesentlich von Bildset 2 (Mittelwert 3,64). Zur Erinnerung: Bildset 1 besteht aus Bildern die Sehenswürdigkeiten zeigen, Bildset 2 aus Nachrichtenbildern des Monats September 2012. Einen etwas höheren Unterschied haben die im Durchschnitt gesetzten Tags bei Interfaces mit Farbe. Bildset 1 erzielt hier im Mittel 3,69 und Bildset 2 4,27 Schlagworte. Betrachtet man nun die Werte in Bezug auf die 2- beziehungsweise 3-elementige Interfacegestaltung, kann man Abbildung 4.3

entnehmen, dass die 2-elementige Gestaltung marginal mehr Schlagworte erzielt. Das Interface ohne Farbe hat 0,21 und das Interface mit Farbe 0,36 Schlagworte mehr.

Merger		D_Tags	Farbe_D_Tags
3 UI Elemente	Mittelwert	3,4734	3,6415
	N	15	15
	Standardabweichung	1,72013	1,68882
	Maximum	8,00	8,70
	Minimum	1,78	1,60
2 UI Elemente	Mittelwert	3,6837	4,1979
	N	15	15
	Standardabweichung	1,42706	1,47732
	Maximum	5,50	6,90
	Minimum	,00	2,33
Insgesamt	Mittelwert	3,5786	3,9197
	N	30	30
	Standardabweichung	1,55659	1,58447
	Maximum	8,00	8,70
	Minimum	,00	1,60

Abbildung 4.3: Mittelwert der gesetzten Tags eines Interfaces in Bezug auf den Merger

Setzt man nun die Schlagworte in Relation zu dem Merger, die Variable, die Aufschluss darüber gibt, ob es sich bei dem Interface um ein 2- oder 3-elementiges User-Interface handelt, und in Relation zu den Bildsets, erhält man für ein Interface ohne Farbe mit 2 User Interface Elementen und den Nachrichtenbildern (Bildset 2) und für ein Interface mit Farbe ebenfalls mit den Nachrichtenbildern, allerdings mit 3 User Interface Elementen die besten Ergebnisse (Abbildung 4.4, Abbildung 4.5).

Da die Mittelwerte alleine relativ wenig Aussagekraft haben und man den vorher genannten Tabellen nicht nur die Mittelwerte, sondern auch die Minima und Maxima entnehmen kann, die eine Differenz von bis zu 8 Schlagworten aufweisen, werden die Variablen auf Normalverteilung überprüft, um weitere statistische Untersuchungen durchführen zu können. Es werden zwei Testarten für die Überprüfung, ob es sich bei der vorliegenden Verteilung um eine Normalverteilung handelt, zur Untersuchung herangezogen. Sowohl der Shapiro-Wilk, als auch der Kolmogorov-Smirnov Signifikanztest, haben den Vorteil, dass sie für geringere Stichprobenumfänge geeignet sind [JL07]. Der Shapiro-Wilk Test weist bei kleinen Stichproben eine hohe Teststärke auf, reagiert allerdings anfälliger auf Ausreißer.

D_Tags

Merger	Bildset	Mittelwert	N	Standardabweichung	Maximum	Minimum
3 UI Elemente	Bildset 1	3,8873	7	2,05437	8,00	1,78
	Bildset 2	3,1113	8	1,40677	6,20	1,89
	Insgesamt	3,4734	15	1,72013	8,00	1,78
2 UI Elemente	Bildset 1	2,9311	5	1,80242	4,90	,00
	Bildset 2	4,0600	10	1,11893	5,50	2,50
	Insgesamt	3,6837	15	1,42706	5,50	,00
Insgesamt	Bildset 1	3,4889	12	1,93024	8,00	,00
	Bildset 2	3,6383	18	1,30883	6,20	1,89
	Insgesamt	3,5786	30	1,55659	8,00	,00

Abbildung 4.4: Mittelwert der gesetzten Tags eines Interfaces ohne Farbe in Bezug zum Merger und den Bildsets

Farbe_D_Tags

Merger	Farbe_Bildset	Mittelwert	N	Standardabweichung	Maximum	Minimum
3 UI Elemente	Bildset 1	3,0403	8	,50976	3,87	2,30
	Bildset 2	4,3286	7	2,30631	8,70	1,60
	Insgesamt	3,6415	15	1,68882	8,70	1,60
2 UI Elemente	Bildset 1	4,2047	10	1,44590	6,63	2,33
	Bildset 2	4,1844	5	1,71300	6,90	2,56
	Insgesamt	4,1979	15	1,47732	6,90	2,33
Insgesamt	Bildset 1	3,6872	18	1,25230	6,63	2,30
	Bildset 2	4,2685	12	1,99345	8,70	1,60
	Insgesamt	3,9197	30	1,58447	8,70	1,60

Abbildung 4.5: Mittelwert der gesetzten Tags eines Interfaces mit Farbe in Bezug zum Merger und den Bildsets

Der Kolmogorow-Smirnow Test hat dagegen eine geringere Teststärke, ist aber robuster gegen Ausreißer [CF04]. Beide Tests arbeiten mit einem Signifikanzniveau von $\alpha = 5\%$ bzw. 0,05.

Für den Shapiro-Wilk Test kann man einer Tabelle [WH87] den kritischen Wert $W_{n,\alpha}$ für $n = 30$ und $\alpha = 0,05$ entnehmen:

$$W_{30,0,05} = 0,927$$

Abbildung 4.6 ist das Ergebnis dieser beiden Tests zu entnehmen. Für den Shapiro-Wilk Test sind $W_{D_Tags} = 0,912$ und $W_{Farbe_D_Tags} = 0,874$ und damit $W_{D_Tags} < W_{30,0,05}$ und $W_{Farbe_D_Tags} < W_{30,0,05}$. Womit beide Variablen in den Ablehnungsbereich von $W_{30,0,05} = 0,927$ fallen.

Außerdem sind für beide Variablen (mit und ohne Farbe) die Ergebnisse beider Tests mit 0,01 und 0,04 bei dem Komogorov-Smirnov und mit 0,016 und 0,002 bei

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
D_Tags	,195	30	,005	,943	30	,112
Farbe_D_Tags	,197	30	,004	,874	30	,002

a. Signifikanzkorrektur nach Lilliefors

Abbildung 4.6: Test auf Normalverteilung mit Hilfe von Kolmogorov-Smirnov und Shapiro-Wilk

dem Shapiro-Wilk Test sehr signifikant ($\rho < 0,05$) und damit ist die Normalverteilungsannahme abzulehnen. Für die anderen Werte gilt die Normalverteilungsannahme. Dieses Resultat hat zur Folge, dass keine Varianzanalyse durchgeführt werden kann, sondern auf nicht-parametrische Verfahren zurückgegriffen werden muss. Da es sich bei dem Versuchsaufbau, allerdings um einen Split-Plot Aufbau mit Within-Group und Between-Group Anteil handelt, ist die Analyse nicht trivial. In [ZVS07] werden diverse Rangordnungsstatistiken für verschiedene Versuchsaufbauten beschrieben. Für zweifaktorielle Split-Plot Versuchspläne wird der bifaktorielle D^{pq} -Test, welcher 1974 bei Bredenkamp beschrieben wurde [Bre74], vorgeschlagen. Dieser Test basiert auf dem H-Test nach Kruskal-Wallis [Con98] und beschreibt die Auswirkungen eines p-fach gestuften Blockfaktors A und eines q-fach gestuften Messwiederholungsfaktors B. Mit diesem Verfahren werden verschiedene Prüfgrößen berechnet, um damit die zwei aufgestellten Nullhypotesen zu überprüfen [ZVS07].

1. Nullhypotesen:

$$H_0^B : \mu_1 = \mu_2 = \dots = \mu_q \quad (4.1)$$

$$H_0^{AxB} : \mu_{11} = \mu_{12} = \dots = \mu_{pq} \quad (4.2)$$

2. Prüfgrößen:

$$D_{B|A}^{pq} = \frac{12}{n_q q(q+1)} \sum_i T_i^2 - 3n_p p(q+1) \quad (4.3)$$

$$D_B^{pq} = \frac{12}{N q(q+1)} \sum_i T_{i(b)}^2 - 3N(q+1) \quad (4.4)$$

3. Ablehnungsbereich bei adjustiertem Signifikanzniveau $\alpha^* = \alpha/2$:

H_0^B wird abgelehnt, falls

$$D_B^{pq} > \chi_{(\alpha^*; q-1)}^2 \quad (4.5)$$

$$(4.6)$$

$H_0^{Ax B}$ wird abgelehnt, falls

$$D_{Ax B}^{pq} > \chi_{(\alpha^*; (p-1)(q-1))}^2 \quad (4.7)$$

Wendet man dieses Testverfahren auf die unter dem Punkt D_Tags und Farbe_D_Tags erzielten Ergebnisse an, so kann man Abbildung 4.7 entnehmen, dass die Prüfgrößen nicht signifikant sind. Es werden nicht mehr Tags bei einem Interface mit Farbe gesetzt, als bei einem Interface ohne Farbe, da die Prüfgröße $D_B^{pq} = 0,133[0,71 < p(\chi_{(1)}^2) < 0,72]$ bei adjustiertem $\alpha^* = 0,025$ nicht signifikant ist. Die gleiche Erkenntnis kann man auch für die Merger Eigenschaft ablesen. Es macht keinen Unterschied, ob es sich um ein Interface mit zwei oder drei User Interface Elementen handelt. Hier ist die Prüfgröße $D_{Ax B} = 0,533[0,46 < p(\chi_{(1)}^2) < 0,47]$, die ebenfalls bei adjustiertem $\alpha^* = 0,025$ nicht signifikant ist.

Zusammenfassung

Es ist festzuhalten, dass Nutzer im Durchschnitt bei Nachrichtenbildern mehr Schlagworte setzen, als bei Sehenswürdigkeiten. Darüber hinaus kann eine Tendenz zu mehr Schlagworten bei einem farbigen Interface wahrgenommen werden. 2-elementige Interfaces schneiden ebenfalls etwas besser als 3-elementige ab. Kombiniert man den Merger mit der Farbeigenschaft, werden es mehr Schlagworte bei einem Interface mit 3 Elementen. Allerdings ist mit Hilfe des Signifikanztest keine eindeutige Tendenz für oder gegen ein Interface feststellbar.

	D_Tags	Farbe_D_Tags	D_Sugg	Farbe_D_Sugg
Rangsumme:	44	46	42	48
QuadratSumme:	4052		4068	
Multipliziert:	270,1333333333		271,2	
D_B	0,1333333333		1,2	
Chi Quadrat	0,7150006547		0,2733216783	
QuadratSumme:	2030		2070	
Multipliziert:	270,6666666667		276	
D_BxA	0,6666666667		6	
D_AxB	0,5333333333		4,8	
Chi Quadrat	0,4652088185		0,0284597369	
	D_AutoComp	Farbe_D_AutoComp	D_Zeit	Farbe_D_Zeit
Rangsumme:	47	43	42	48
QuadratSumme:	4058		4068	
Multipliziert:	270,5333333333		271,2	
D_B	0,5333333333		1,2	
Chi Quadrat	0,4652088185		0,2733216783	
QuadratSumme:	2054		2038	
Multipliziert:	273,8666666667		271,7333333333	
D_BxA	3,8666666667		1,7333333333	
D_AxB	3,3333333333		0,5333333333	
Chi Quadrat	0,0678891549		0,4652088185	

Abbildung 4.7: Signifikanzanalyse nach [ZVS07] für die Tag- und Kreuzverweis-Service-Variablen

Kreuzverweis Services

In Bezug auf Effektivität ist nicht nur die Frage interessant, ob gewisse Interface-Gestaltungen einen positiven Einfluss auf die Anzahl der gesetzten Schlagworte haben, sondern auch die Intensität der Nutzung der Services, da diese von Kreuzverweis bereit gestellten Services eine Besonderheit sind. Es wurden verschiedene Daten, während ein Nutzer die Evaluation durchgearbeitet hat, aufgezeichnet. Zu diesen Daten gehören neben den Schlagworten und der Anzahl der Schlagworte, die ein Nutzer aus der Suggestions-Box aus- beziehungsweise wieder abgewählt hat, auch die Worte, die er aus der Autocompletionliste ausgewählt hat. Für jeden Nutzer ist auch hier wieder nur das statistische Mittel für diese beiden - beziehungsweise in Hinsicht auf die Datenstruktur in SPSS vier Variablen eingetragen.

Suggestions

Die Anzahlen der aus der Suggestions-Box ausgewählten Tags ist codiert mit den Variablen D_Sugg und D_Farbe_Sugg (Tabelle 4.1), um die Werte von Interfaces ohne und mit Farbe einzeln betrachten und auf ihre Unterschiede hin prüfen zu

können. Im Gegensatz zu den vorherigen Daten, liegen bei den aus der Suggestions-Box ausgewählten Tags Normalverteilungen vor Abbildung 4.8, sodass eine parametrische Analyse möglich wäre.

Kolmogorov-Smirnov-Anpassungstest			
		D_Sugg	Farbe_D_Sugg
N		30	30
Parameter der Normalverteilung ^{a,b}	Mittelwert	,8523	,8694
	Standardabweichung	1,07496	1,01188
Extremste Differenzen	Absolut	,214	,238
	Positiv	,201	,238
	Negativ	-,214	-,195
Kolmogorov-Smirnov-Z		1,172	1,305
Asymptotische Signifikanz (2-seitig)		,128	,066

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Abbildung 4.8: Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei D_Sugg und Farbe_D_Sugg

Untersucht man aber zunächst die Mittelwerte, ist kaum ein Unterschied zwischen der Anzahl der ausgewählten Tags in Bezug auf die Anzahl der User Interface Elemente festzustellen (Abbildung 4.9). Allerdings besitzt ein Interface mit Typinformation und zwei User Interface Elementen mit 0,77 die geringste Standardabweichung dieser vier Werte.

Merger		D_Sugg	Farbe_D_Sugg
3 UI Elemente	Mittelwert	,8432	1,0304
	N	15	15
	Standardabweichung	1,32605	1,21255
2 UI Elemente	Mittelwert	,8615	,7084
	N	15	15
	Standardabweichung	,79689	,77144
Insgesamt	Mittelwert	,8523	,8694
	N	30	30
	Standardabweichung	1,07496	1,01188

Abbildung 4.9: Mittelwerte von D_Sugg und Farbe_D_Sugg in Bezug auf den Merger

Ähnlich verhält es sich, wenn man die Mittelwerte der ausgewählten Tags bei Interfaces mit und ohne Farbe im Hinblick auf die Art der Bilder analysiert. Der Unterschied zwischen Bildset 1 und Bildset 2 beträgt hier noch nicht einmal 0,3 Tags (Abbildung 4.10 und Abbildung 4.11).

D_Sugg

Bildset	Mittelwert	N	Standardabweichung
Bildset 1	,6731	12	,90983
Bildset 2	,9718	18	1,18229
Insgesamt	,8523	30	1,07496

Abbildung 4.10: Mittelwerte von D_Sugg in Bezug auf die Bildsets

Farbe_D_Sugg

Farbe_Bildset	Mittelwert	N	Standardabweichung
Bildset 1	,7236	18	,63304
Bildset 2	1,0880	12	1,41179
Insgesamt	,8694	30	1,01188

Abbildung 4.11: Mittelwerte Farbe_D_Sugg in Bezug auf die Bildsets

Auch auf die Ergebnisse von D_Sugg und Farbe_D_Sugg in Bezug auf den Merger wurde das Verfahren, welches im Abschnitt 4.2 beschrieben wurde angewendet. Der Übersicht (Abbildung 4.7) ist zu entnehmen, dass die Unterschiede in Bezug auf die Farbe nicht signifikant sind. Es ergibt sich für die Prüfgröße $D_B^{pq} = 1,2[0,27 < p(\chi^2_{(1)}) < 0,28]$, welche bei adjustiertem $\alpha^* = 0,025$ nicht signifikant ist. Betrachtet man die Prüfgröße für die Untersuchung der Nutzung des Suggestion-Services in Bezug auf den Merger, stellt man fest, dass die Prüfgröße nur 0,003 vom adjustierten $\alpha^* = 0,025$ entfernt liegt. Bei einer größeren Probandenanzahl, die Ausreißer in ihrer Mächtigkeit verringern würde, ist die Wahrscheinlichkeit für eine Signifikanz gegeben. Mit der vorliegenden Prüfgröße $D_{AxB} = 4,8$ wird die Signifikanzannahme allerdings abgelehnt. Es kann also keine Aussage bezüglich einer intensiveren Nutzung bei zwei-elementigen Interfaces in Bezug auf

die Typinformation gemacht werden, die mögliche Tendenz dahingehend, sollte dennoch in die Endauswertung mit einfließen.

Zusammenfassung

Zwischen den einzelnen Kombinationsmöglichkeiten ist kein Unterschied in Bezug auf die Intensität der Nutzung des Suggestions-Service festzustellen. Dennoch sollte die Tendenz zu einer intensiveren Nutzung bei Interfaces mit zwei Elementen in Bezug auf die Typinformation berücksichtigt werden.

Autocompletion

Die Normalverteilungsannahme kann auch bei dieser Verteilung mit Hilfe des Kolmogorov-Smirnov Tests bestätigt werden (Abbildung 4.12).

Kolmogorov-Smirnov-Anpassungstest			
		D_AutoComp	Farbe_D_AutoComp
N		30	30
Parameter der Normalverteilung ^{a,b}	Mittelwert	,7215	,5632
	Standardabweichung	,48386	,38535
Extremste Differenzen	Absolut	,151	,147
	Positiv	,120	,147
	Negativ	-,151	-,106
Kolmogorov-Smirnov-Z		,827	,804
Asymptotische Signifikanz (2-seitig)		,502	,538

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Abbildung 4.12: Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei D_AutoComp und Farbe_D_AutoComp

Untersucht man die Mittelwerte in Relation zu den 2- und 3-elementigen Interfaces, stellt man den höchsten Unterschied mit 0,276 mehr Wörtern bei den 2-elementigen und den Interfaces ohne Farbe, zu den Interfaces mit Farbe fest (Abbildung 4.13).

Die Variablen D_AutoComp und Farbe_D_AutoComp wurden ebenfalls in Bezug auf die Bildsets untersucht. Zwischen einem farbigen und nicht farbigen Layout bei dem gleichen Bildset 1 wird der höchste Unterschied mit 0,264 erreicht (Abbildung 4.14 und Abbildung 4.15).

Auch die Werte für die Verwendung der Autocompletion hat man einer Signifikanzanalyse unterzogen. Die Ergebnisse aus Abbildung 4.7 zeigen auch hier keine

Merger		D_AutoComp	Farbe_D_AutoComp
3 UI Elemente	Mittelwert	,6312	,5901
	N	15	15
	Standardabweichung	,50582	,42403
2 UI Elemente	Mittelwert	,8118	,5362
	N	15	15
	Standardabweichung	,46004	,35529
Insgesamt	Mittelwert	,7215	,5632
	N	30	30
	Standardabweichung	,48386	,38535

Abbildung 4.13: Mittelwerte von D_AutoComp und Farbe_D_AutoComp in Bezug auf den Merger

D_AutoComp

Bildset	Mittelwert	N	Standardabweichung
Bildset 1	,8074	12	,51287
Bildset 2	,6642	18	,46956
Insgesamt	,7215	30	,48386

Abbildung 4.14: Mittelwerte von D_AutoComp in Bezug auf die Bildsets

signifikante Unterschiede, aber wiederholt Tendenzen zur Signifikanz bei Interfaces mit der Variablen Farbe in Bezug auf die Merger-Eigenschaft. Die Wahrscheinlichkeit der Prüfgröße $D_{Ax_B} = 3,33$ ergibt 0,068 und liegt damit 0,043 vom adjustierten α^* -Wert.

In Bezug auf die Typinformation ist weder die Signifikanz, noch eine Tendenz festzustellen:

Prüfgröße $D_B^{pq} = 0,533[0,46 < p(\chi^2_{(1)}) < 0,47]$ und damit nicht signifikant bei adjustiertem $\alpha^* = 0,025$.

Zusammenfassung

Die Autocompletion wird bei einem 2-elementigen Interface häufiger genutzt. Diese Annahme kann auch durch die Tendenz, welche dem Test auf Signifikanz zu

Farbe_D_AutoComp

Farbe_Bildset	Mittelwert	N	Standardabweichung
Bildset 1	,5430	18	,31833
Bildset 2	,5935	12	,48292
Insgesamt	,5632	30	,38535

Abbildung 4.15: Mittelwerte von Farbe_D_AutoComp in Bezug auf die Bildsets

entnehmen ist, bestätigt werden. Bedenkt man, dass die Autocompletion bei jedem der 4 Interfaces gleich aussieht und somit keine Unterschiede, ob das Interface mit oder ohne Farbe gestaltet wurde, bestehen sollten, lässt sich dieses Ergebnis mit der höheren Standardabweichung beim höheren Wert des Interfaces ohne Farbe und der geringeren Anzahl Probanden für die Kombination aus Bildset 1 und ohne Farbe begründen. Diesen Unterschied in der Anzahl der Probanden für die Bildset-Kombinationen ließ sich, durch die Probanden, die die Evaluation nicht korrekt ausgefüllt haben, nicht vermeiden.

4.3 Effizienz

Die erste und die dritte Forschungshypothese beschäftigt sich mit der Effizienz. Es sollen Antworten auf die Frage gefunden werden, ob die Gestaltung eines Interfaces einen positiven Einfluss auf die zeitliche Dauer eines Annotationsprozesses haben kann. Bei den 20 zu annotierenden Bildern der Evaluation ist diese Frage vielleicht noch nicht von elementarer Bedeutung, wenn man aber den Verwendungszweck und die Zielgruppe des Systems betrachtet, können einige Sekunden weniger pro Bild, zu einigen eingesparten Minuten oder sogar Stunden werden. Zunächst wird die Effizienz in Bezug auf die Anzahl der Interface Elemente und damit Forschungshypothese 1, die besagt: „Der Einsatz von nur 2 User Interface Elementen erhöht die Effizienz“, untersucht. Danach werden die Zeiten gruppiert nach dem Zustand, den die Variable Farbe hat, ausgewertet. Mit dieser Analyse kann eine Aussage zu Forschungshypothese 3: „Das Verwenden von Farbe als Hinweis zur Typinformation beeinflusst die Effizienz positiv“, getroffen werden. Sowohl bei der Variablen D_Zeit als auch bei der Variablen Farbe_D_Zeit, kann die Annahme, dass eine Normalverteilung vorliegt bestätigt werden (Abbildung 4.16).

		Kolmogorov-Smirnov-Anpassungstest	
		D_Zeit	Farbe_D_Zeit
N		30	30
Parameter der Normalverteilung ^{a,b}	Mittelwert	00:01:04,552	00:01:05,338
	Standardabweichung	00:00: 32,10030	00:00: 39,65036
Extremste Differenzen	Absolut	,133	,158
	Positiv	,133	,158
	Negativ	-,079	-,131
Kolmogorov-Smirnov-Z		,726	,864
Asymptotische Signifikanz (2-seitig)		,667	,444

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Abbildung 4.16: Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei D_Zeit und Farbe_D_Zeit

Somit wäre es auch hier möglich, auf eine Varianzanalyse, nach Auswertung der Mittelwerte, zurückzugreifen. Die Mittelwerte der Interfaces ohne Farbe sind nahezu identisch (Abbildung 4.17). Bei den Interfaces mit farbigen Schlagworten, sind die Nutzer im Schnitt 10 Sekunden schneller, bei einem 2-elementigen Interface, als bei einem 3-elementigen Interface (Abbildung 4.17). Allerdings hat ein 3-elementiges Interface eine geringe Standardabweichung, daher kann man diesen zeitlichen Vorteil des 2-elementigen Interfaces, wahrscheinlich erst mit einer höheren Anzahl an Probanden oder größeren Bildsets nachweisen.

Merger		D_Zeit	Farbe_D_Zeit
3 UI Elemente	Mittelwert	00:01:04,533	00:01:10,103
	N	15	15
	Standardabweichung	00:00: 27,65313	00:00: 38,98310
2 UI Elemente	Mittelwert	00:01:04,571	00:01:00,573
	N	15	15
	Standardabweichung	00:00: 37,01032	00:00: 41,08848
Insgesamt	Mittelwert	00:01:04,552	00:01:05,338
	N	30	30
	Standardabweichung	00:00: 32,10030	00:00: 39,65036

Abbildung 4.17: Mittelwerte von D_Zeit und Farbe_D_Zeit in Bezug auf den Merger

Betrachtet man die Werte in Bezug auf Forschungshypothese 3, kann man feststellen, dass die mittlere Zeit für ein Interface mit Farbe bei Bildset 1 und 2 weniger stark variiert und sich am vorher gemessenen Mittelwert des 2- und 3-elementigen Interfaces (1:05 Minuten Abbildung 4.16) mit 1:02 Minuten für Bildset 1 und 1:09 Minuten für Bildset 2 orientiert (Abbildung 4.19). Abbildung 4.18 zeigt die Werte für ein Interface ohne Farbe. Hier existiert zwischen Bildset 1 mit 1:12,7 Minuten und Bildset 2 mit 0:59 Minuten eine Differenz von fast 14 Sekunden.

D_Zeit			
Bildset	Mittelwert	N	Standardabweichung
Bildset 1	00:01:12,748	12	00:00: 30,31230
Bildset 2	00:00:59,088	18	00:00: 32,92766
Insgesamt	00:01:04,552	30	00:00: 32,10030

Abbildung 4.18: Mittelwerte von D_Zeit in Bezug auf die Bildsets

In Abbildung 4.7 sind die Signifikanzwerte der Zeit-Variablen abzulesen. Weder Prüfgröße $D_B^{pq} = 1,2 [0,27 < p(\chi^2_{(1)}) < 0,28]$, noch Prüfgröße $D_{AxB} = 0,533 [0,46 < p(\chi^2_{(1)}) < 0,47]$ weisen eine Signifikanz bei adjustierten $\alpha^* = 0,025$ auf.

Farbe_D_Zeit			
Farbe_Bildset	Mittelwert	N	Standardabweichung
Bildset 1	00:01:02,573	18	00:00: 45,07403
Bildset 2	00:01:09,487	12	00:00: 31,20293
Insgesamt	00:01:05,338	30	00:00: 39,65036

Abbildung 4.19: Mittelwerte von Farbe_D_Zeit in Bezug auf die Bildsets

Zusammenfassung

Festzustellen ist, dass die mittlere Zeit bei dem einen Interface ohne Farbe bei Bildset 2, im Verhältnis zum Bildset 1, und bei dem Interface mit Farbe genau bei dem anderen Bildset 1 besser ist. Die Anzahl der Probanden ist zu gering, um eine verbindliche Aussage zu machen, aber die Hypothese, dass die Einfärbung der Schlagworte einen zeitlichen Vorteil, bei der Verwendung von Nachrichtenbildern, bietet, kann ebenso nicht gestützt werden. Es ist allerdings die Tendenz der zeitlichen Verbesserung durch die Kombination von Farbinformation und einem 2-elementigen Layout abzulesen. Mit Hilfe des Signifikanz-Tests, kann diese Theorie der Verbesserung nicht gestützt werden, allerdings zeigt der Test, dass es auch keine Verschlechterung der Effizienz durch ein zwei-elementiges Interface mit Farbe gibt.

4.4 Benutzerzufriedenheit

Die Zufriedenheit eines Nutzers ließe sich am besten innerhalb eines Interviews nach der Evaluation klären. Anfangs hatte man sich, auf Grund der späteren Zielgruppe und des Anwendungszweckes, auf die Online-Evaluation festgelegt, da es nur auf diese Weise möglich war, geeignete Probanden zu finden. Zumal die meisten der Probanden, die über den InetBib E-Mailverteiler gewonnen werden konnten, sich physisch nicht in unmittelbarer Umgebung zur Universität Koblenz aufhalten. Daher musste die Aufgabe des persönlichen Interviews von Fragebögen übernommen werden. Ein Proband hat mehrere Interfaces zu bearbeiten, um diese untereinander vergleichen zu können, wurde nicht nur ein abschließender Fragebogen, bei dem der Nutzer die Fragebögen miteinander vergleichen muss, benutzt, sondern nach jedem Interface ein Fragebogen, der sich nur auf das bearbeitete Interface bezieht. Auf diese Weise kann ein Nutzer sich auf die Bewertung des betreffenden Interfaces konzentrieren und der Vergleich wird später einfach anhand der statistischen Auswertung vorgenommen. Darüber hinaus ermöglichen Fragebögen die Gegenüberstellung von subjektiver Einschätzung und beobachteten Werten, der aufgezeichneten Sessions. Alle Fragen werden in diesem Abschnitt ausgewertet, weil es sich hierbei um die Einschätzung der Nutzer handelt. Allerdings können die Ergebnisse einiger Fragen auch andere, die Effizienz und Effektivität betreffende Forschungshypothesen stützen oder widerlegen. Daher werden die Ergebnisse aus diesem Abschnitt in das Endresümee am Ende dieses Kapitels mit einfließen. Es gilt während der Analyse der Fragebögen Forschungshypothesen 2 und 5, die sich mit der Auswirkung des Layouts auf die Benutzerzufriedenheit beschäftigen, zu klären. Je nach Inhalt der Freitextfelder, welche im nächsten Abschnitt betrachtet werden, fließen diese ebenfalls in die Einschätzung der Benutzerzufriedenheit mit ein.

Es gibt verschiedene Kategorien von Fragen, aus denen man einen Fragebogen zusammenstellen kann. Dem Abschnitt 3.6 sind die Fragen, die den Nutzern gestellt wurden zu entnehmen. Zwar behandelt keine der Fragen direkt die Zufriedenheit, aber alle Fragen in ihrer Gesamtheit lassen Rückschlüsse auf die Zufriedenheit zu. Zur Übersichtlichkeit werden die Fragen nach ihrer Reihenfolge im Fragebogen, allerdings die Ergebnisse der Interfaces mit und ohne Farbe gemeinsam beziehungsweise die 2-elementigen in Relation zu den 3-elementigen, ausgewertet. Auch hier wird mit der Prüfung auf die Erfüllung der Normalverteilungsannahme und der Betrachtung der Mittelwerte begonnen. Theoretisch sollte es kaum Unterschiede in den Ergebnissen zwischen Bildset 1 und 2 geben, da die Fragen weniger die Inhalte der Interfaces als viel mehr die Funktionalität und Gestaltung betreffen. Dennoch ist es durchaus denkbar, dass gerade in Bezug auf die angebotenen Services, ein gewisses Interface eine bessere Interaktion bei bestimmten Inhalten zulässt. Daher werden die Fragen, wie bei den Abschnitten zuvor, auch in Hinblick auf die Bildsets

untersucht. Abbildung 4.20 ist zu entnehmen, dass bei vier Fragen die Normalverteilungsannahme nicht erfüllt ist. Es handelt sich hierbei um die Fragen 3, 8 und 9 (Tabelle 3.2). Drei der vier Fragen betreffen die Fragebögen der Interfaces ohne Farbe. Bei der vierten Frage, ist sowohl die Verteilung des Interfaces ohne Farbe, als auch des Interfaces mit Farbe nicht normal verteilt. Darüberhinaus haben die Daten ein ordinales Messniveau. Folglich sollten parametrische Auswertungen vermieden werden. Ebenso, wie auch schon die Daten zuvor, werden die Ergebnisse der zehn Fragen der Fragebögen durch den in [ZVS07] beschriebenen nichtparametrischen SPF_{pq} -Test untersucht (Abbildung 4.21). Die Reaktionszeit der Interfaces wird von den Nutzern als am schnellsten bei der Kombination aus Farbe und 2 User Interface Elementen eingeschätzt. Die Varianz bei dieser Kombination ist im Vergleich zu den Varianzen der anderen Werte mit bis zur Hälfte am geringsten (Abbildung 4.22).

Der Test auf Signifikanz (Abbildung 4.21) kann weder für den Within-Group Anteil, noch für den Between-Group Anteil in Bezug auf die Typinformation eine Signifikanz feststellen. $D_{Ax_B} = 0,533$ hat eine Wahrscheinlichkeit von 0,465.

Für Prüfgröße, die Aussagen über die Signifikanz in Bezug auf die Typinformation macht, beträgt $D_B^{pq} = 1,2[0,27 < p(\chi^2_{(1)}) < 0,28]$ und ist damit nicht signifikant bei adjustiertem $\alpha^* = 0,025$. In Bezug auf die Verbindung des Zustandes der Variablen Farbe und der Bildsets besteht ein Ungleichgewicht. Es haben weniger Probanden (12 Probanden) Bildset 1 mit einem Interface ohne Farbe bearbeitet, als Bildset 2 (18 Probanden). Wie beim Abschnitt zuvor, tritt auch hier der Effekt auf, dass jeweils das Bildset, bei dem weniger Probanden teilgenommen haben, im Durchschnitt schlechter abschneidet, als das Bildset an dem die größere Gruppe teilgenommen hat. Bei dieser kleineren Probandengruppe ist bei beiden Interfaces eine höhere Varianz festzustellen.

Die zweite Frage, welche sich mit der Bedienbarkeit beschäftigt, weist keinen Unterschied zwischen Interfaces ohne und mit Farbe auf. Prüfgröße $D_B^{pq} = 0,133[0,71 < p(\chi^2_{(1)}) < 0,72]$. Es existiert keine Signifikanz für die Interfaces mit Typinformation in Bezug auf den Merger. Der vorher beschriebene Effekt mit der kleineren Probandengruppe ist hier nur bei den Interfaces mit Farbe zu beobachten. Bei den Interfaces ohne Farbe wird die Vermutung bestätigt, dass bezüglich der Beantwortung der Fragebögen eine Unterscheidung von Bildset 1 und 2 nicht relevant ist, da die Mittelwerte gleich sind. Allerdings zeigt die Varianz der kleineren Gruppe, dass die Meinungen der Nutzer stärker variieren.

Bei der dritten Frage wird ein Unterschied der beiden Within-Faktoren von 0 gemessen. Betrachtet man aber die Prüfgröße $D_{Ax_B} = 1,2[0,27 < p(\chi^2_{(1)}) < 0,28]$, kann man ablesen, dass die Werte der Within-Faktoren in Bezug auf die Between-Group Eigenschaft nicht gleich sind. Es besteht also eine Tendenz zu einem Unterschied von Interfaces mit oder ohne Farbe in Bezug auf den Merger (Abbil-

Kolmogorov-Smirnov-Anpassungstest

		Reaktionszeit	Bedienbarkeit	Funktionalität	Hilfe	Überblick
N		30	30	30	30	30
Parameter der Normalverteilung ^{a,b}	Mittelwert	79,2000	78,1667	72,6333	56,2667	55,4667
	Standardabweichung	23,19988	26,45893	32,88911	31,44446	31,25615
Extremste Differenzen	Absolut	,224	,218	,274	,130	,123
	Positiv	,185	,205	,203	,107	,103
	Negativ	-,224	-,218	-,274	-,130	-,123
Kolmogorov-Smirnov-Z		1,227	1,193	1,501	,710	,673
Asymptotische Signifikanz (2-seitig)		,098	,116	,022	,694	,755
		Eingaben	Erfolgreich	Bedienung	Lernen	Fehlerkorrektur
N		30	30	30	30	30
Parameter der Normalverteilung ^{a,b}	Mittelwert	57,7667	70,8000	83,4000	84,1667	71,8333
	Standardabweichung	28,95381	34,09864	26,30668	21,74790	29,04394
Extremste Differenzen	Absolut	,118	,205	,290	,252	,164
	Positiv	,072	,196	,284	,233	,164
	Negativ	-,118	-,205	-,290	-,252	-,158
Kolmogorov-Smirnov-Z		,644	1,125	1,589	1,382	,900
Asymptotische Signifikanz (2-seitig)		,801	,159	,013	,044	,392
		Farbe_Reaktionszeit	Farbe_Bedienbarkeit	Farbe_Funktionalität	Farbe_Hilfe	Farbe_Überblick
N		30	30	30	30	30
Parameter der Normalverteilung ^{a,b}	Mittelwert	77,0667	82,0667	71,6000	54,8667	59,1667
	Standardabweichung	26,23230	23,71953	26,44004	30,26861	27,62568
Extremste Differenzen	Absolut	,191	,225	,232	,130	,125
	Positiv	,191	,225	,182	,130	,122
	Negativ	-,165	-,202	-,232	-,110	-,125
Kolmogorov-Smirnov-Z		1,046	1,231	1,269	,711	,686
Asymptotische Signifikanz (2-seitig)		,224	,096	,080	,692	,734
		Farbe_Eingaben	Farbe_Erfolgreich	Farbe_Bedienung	Farbe_Lernen	Farbe_Fehlerkorr
N		30	30	30	30	30
Parameter der Normalverteilung ^{a,b}	Mittelwert	61,7333	74,7000	83,8667	81,4667	77,2333
	Standardabweichung	28,05528	29,37528	22,04030	19,52140	19,35456
Extremste Differenzen	Absolut	,129	,200	,287	,223	,147
	Positiv	,102	,195	,232	,171	,120
	Negativ	-,129	-,200	-,287	-,223	-,147
Kolmogorov-Smirnov-Z		,709	1,096	1,573	1,219	,805
Asymptotische Signifikanz (2-seitig)		,696	,181	,014	,102	,536

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Abbildung 4.20: Kolmogorov-Smirnov Test zur Überprüfung der Normalverteilungsannahme bei den Ergebnissen der Fragebögen

	Reaktionszeit	FarbeReaktionszeit	Bedienbarkeit	FarbeBedienbarkeit	Funktionalität	FarbeFunktionalität
Rangsumme:	48	42	44	46	45	45
QuadratSumme:	4068		4052		4050	
Multipliziert:	271,2		270,1333333333		270	
D_B	1,2		0,1333333333		0	
Chi Quadrat	0,2733216783		0,7150006547		1	
QuadratSumme:	2038		2030		2034	
Multipliziert:	271,7333333333		270,6666666667		271,2	
D_B A	1,7333333333		0,6666666667		1,2	
D_AxB	0,5333333333		0,5333333333		1,2	
Chi Quadrat	0,4652088185		0,4652088185		0,2733216783	
	Hilfe	FarbeHilfe	Überblick	FarbeÜberblick	Eingaben	FarbeEingaben
Rangsumme:	46	44	43	47	41	49
QuadratSumme:	4052		4058		4082	
Multipliziert:	270,1333333333		270,5333333333		272,1333333333	
D_B	0,1333333333		0,5333333333		2,1333333333	
Chi Quadrat	0,7150006547		0,4652088185		0,1441270348	
QuadratSumme:	2030		2030		2042	
Multipliziert:	270,6666666667		270,6666666667		272,2666666667	
D_B A	0,6666666667		0,6666666667		2,2666666667	
D_AxB	0,5333333333		0,1333333333		0,1333333333	
Chi Quadrat	0,4652088185		0,7150006547		0,7150006547	
	Erfolgreich	FarbeErfolgreich	Bedienung	FarbeBedienung	Lernen	FarbeLernen
Rangsumme:	46	44	45	45	48	42
QuadratSumme:	4052		4050		4068	
Multipliziert:	270,1333333333		270		271,2	
D_B	0,1333333333		0		1,2	
Chi Quadrat	0,7150006547		1		0,2733216783	
QuadratSumme:	2030		2034		2038	
Multipliziert:	270,6666666667		271,2		271,7333333333	
D_B A	0,6666666667		1,2		1,7333333333	
D_AxB	0,5333333333		1,2		0,5333333333	
Chi Quadrat	0,4652088185		0,2733216783		0,4652088185	
	Fehlerkorrektur	FarbeFehlerkorrektur				
Rangsumme:	44	45				
QuadratSumme:	3961					
Multipliziert:	264,0666666667					
D_B	-5,9333333333					
Chi Quadrat	1					
QuadratSumme:	1983					
Multipliziert:	264,4					
D_B A	-5,6					
D_AxB	0,3333333333					
Chi Quadrat	0,5637028617					

Abbildung 4.21: Signifikanzanalyse der Ergebnisse der Fragebögen

Merger		Reaktionszeit	Farbe_Reaktionszeit
3 UI Elemente	Mittelwert	74,8667	69,4667
	N	15	15
	Standardabweichung	26,04081	32,20662
2 UI Elemente	Mittelwert	83,5333	84,6667
	N	15	15
	Standardabweichung	19,91362	16,25980
Insgesamt	Mittelwert	79,2000	77,0667
	N	30	30
	Standardabweichung	23,19988	26,23230

Abbildung 4.22: Mittelwerte von Reaktionszeit und Farbe_Reaktionszeit in Bezug auf den Merger

Merger		Bedienbarkeit	Farbe_Bedienbarkeit
3 UI Elemente	Mittelwert	75,8667	77,6667
	N	15	15
	Standardabweichung	30,38530	29,98492
2 UI Elemente	Mittelwert	80,4667	86,4667
	N	15	15
	Standardabweichung	22,70578	14,99460
Insgesamt	Mittelwert	78,1667	82,0667
	N	30	30
	Standardabweichung	26,45893	23,71953

Abbildung 4.23: Mittelwerte von Bedienbarkeit und Farbe_Bedienbarkeit in Bezug auf den Merger

Merger		Funktionalität	Farbe_Funktionalität
3 UI Elemente	Mittelwert	65,6667	68,7333
	N	15	15
	Standardabweichung	35,88208	28,60686
2 UI Elemente	Mittelwert	79,6000	74,4667
	N	15	15
	Standardabweichung	29,09295	24,74114
Insgesamt	Mittelwert	72,6333	71,6000
	N	30	30
	Standardabweichung	32,86911	26,44004

Abbildung 4.24: Mittelwerte von Funktionalität und Farbe_Funktionalität in Bezug auf den Merger

dung 4.21). Die Ergebnisse des Signifikanz-Tests der nächsten beiden Fragen, die zur Kategorie Selbstbeschreibungsfähigkeit gehören, befinden sich auf der Skala relativ mittig. Nur bei Frage sechs „Das Interface liefert genügend Informationen darüber, welche Eingaben nötig sind“, besteht wieder eine Tendenz zu einem Unterschied zwischen einem Interface mit und ohne Farbe. Die Prüfgröße $D_B^{pq} = 2,133[0,14 < p(\chi^2_{(1)}) < 0,15]$ ist dennoch nicht signifikant.

Bei diesen drei Fragen existiert die höchste Irrtumswahrscheinlichkeit innerhalb der Ergebnisse der Fragebögen. Da das bedeutet, dass die Wahrscheinlichkeit sich zu irren, wenn man die Normalverteilungshypothese ablehnt, so hoch ist, wie die Zahl der asymptotischen Signifikanz, ist die Wahrscheinlichkeit, dass eine Normalverteilung vorliegt, hier eher gegeben (Abbildung 4.25, Abbildung 4.26 und Abbildung 4.27). Bei der ersten Frage aus der Kategorie „Erwartungskonformität“, kann mit $D_B^{pq} = 0,133[0,71 < p(\chi^2_{(1)}) < 0,72]$ und $D_{AxB} = 0,533[0,46 < p(\chi^2_{(1)}) < 0,47]$ bei adjustiertem α^* ebenfalls keine Signifikanz festgestellt werden. Es macht somit keinen Unterschied, ob ein Interface mit oder ohne Typinformation und zwei oder drei Elementen vorliegen. Damit erfüllen alle Interfaces den Punkt Erwartungskonformität. Für alle Interfaces liegt die gleiche Funktionsweise für die Eingabe, Selektion und Deselektion von Tags vor. Dieser Wert bestätigt die bereits schon erwähnte Annahme, dass Typinformationen in Form von Farbe und zwei User Interface Elementen nicht unerwartet agieren. Basierend auf diesem Ergebnis, darf ein

Merger		Hilfe	Farbe_Hilfe
3 UI Elemente	Mittelwert	52,9333	55,9333
	N	15	15
	Standardabweichung	33,87470	35,16993
2 UI Elemente	Mittelwert	59,6000	53,8000
	N	15	15
	Standardabweichung	29,61129	25,66042
Insgesamt	Mittelwert	56,2667	54,8667
	N	30	30
	Standardabweichung	31,44446	30,26861

Abbildung 4.25: Mittelwerte von Hilfe und Farbe_Hilfe in Bezug auf den Merger

Merger		Überblick	Farbe_Überblick
3 UI Elemente	Mittelwert	53,9333	60,7333
	N	15	15
	Standardabweichung	31,79952	30,70706
2 UI Elemente	Mittelwert	57,0000	57,6000
	N	15	15
	Standardabweichung	31,74002	25,15324
Insgesamt	Mittelwert	55,4667	59,1667
	N	30	30
	Standardabweichung	31,25615	27,62568

Abbildung 4.26: Mittelwerte von Überblick und Farbe_Überblick in Bezug auf den Merger

Merger		Eingaben	Farbe_Eingaben
3 UI Elemente	Mittelwert	53,1333	57,4667
	N	15	15
	Standardabweichung	32,68435	29,70778
2 UI Elemente	Mittelwert	62,4000	66,0000
	N	15	15
	Standardabweichung	24,94508	26,62437
Insgesamt	Mittelwert	57,7667	61,7333
	N	30	30
	Standardabweichung	28,95381	28,05528

Abbildung 4.27: Mittelwerte von Eingabe und Farbe_Eingabe in Bezug auf den Merger

Interface durchaus solche Eigenschaften besitzen.

Frage acht, die ebenfalls zur Kategorie „Erwartungskonformität“ gehört, bestätigt diese Annahme ebenfalls ($D_B^{pq} = 0$). Es muss allerdings erwähnt werden, dass die Prüfgröße D_{AxB} mit 1,2 und $[0,27 < p(\chi^2_{(1)}) < 0,28]$ wieder gegen den adjustierten α^* -Wert tendiert. Es gibt also Unterschiede in der Bedienung nach einem Prinzip zwischen Interfaces mit und ohne Typinformationen in Bezug auf den Merger.

Frage Neun, die sich mit der Lernförderlich beschäftigt, gibt gleichzeitig auch Aufschluss über das Nutzerempfinden zum Thema Erlernbarkeit. Zwar gibt es keine signifikanten Unterschiede zwischen Interfaces mit und ohne Farbe (Within-Group Anteil), da die Prüfgröße $D_B^{pq} = 1,2 [0,27 < p(\chi^2_{(1)}) < 0,28]$ beträgt, dennoch kann man diesen Wert wieder als Tendenz interpretieren. Es fiel den Nutzern also schwerer Interfaces ohne Farbe zu erlernen, als Interfaces mit Farbe.

Bei der vorletzten Frage wird auch hier die Normalverteilungshypothese abgelehnt (Abbildung 4.30).

Bei der letzten Frage gibt es kaum Unterschiede in der Möglichkeit Fehler zu korrigieren bei Interfaces mit und ohne Farbe ($D_B^{pq} = -5,93$). Somit bestätigt auch dieser Wert die Annahme, dass die gleiche Funktionalität gewährleistet werden kann, obwohl Informationen zu Tag-Typen angegeben werden.

Von einer Normalverteilung kann bei der letzten Frage wieder eher ausgegangen werden. Die Tendenz zum 2-elementigen Interface mit Farbe als beliebtestes In-

terface nach Meinung der Probanden, kann auch hier wieder bestätigt werden (Abbildung 4.31). Allgemein stützen diese Ergebnisse die Vermutung, dass in

Merger		Erfolgreich	Farbe_Erfolgreich
3 UI Elemente	Mittelwert	66,7333	71,7333
	N	15	15
	Standardabweichung	35,41079	30,04156
2 UI Elemente	Mittelwert	74,8667	77,6667
	N	15	15
	Standardabweichung	33,45331	29,42950
Insgesamt	Mittelwert	70,8000	74,7000
	N	30	30
	Standardabweichung	34,09864	29,37528

Abbildung 4.28: Mittelwerte von Erfolgreich und Farbe_Erfolgreich in Bezug auf den Merger

den Themenebreichen, in denen gleiche Ergebnisse erzielt werden sollen, wie beispielsweise bei der Fehlerkorrektur, keine Unterschiede zwischen den Interfaces existieren. Es sind teilweise positive Tendenzen zu einem farbigen 2-elemtigen Interface zu verzeichnen.

Kommentare von Probanden

Manche Probleme oder weitere Gestaltungsmöglichkeiten gehen erst aus einem persönlichen Gespräch hervor. Da es sich hierbei um eine anonyme Online-Evaluation handelt, die es nicht ermöglicht, den Nutzer persönlich zu interviewen, und die darauf vertraut, dass ein Proband die Fragen wahrheitsgemäß beantwortet, wurden die Freitextfelder eingefügt. Diese ermöglichen dem Nutzer Anmerkungen zu hinterlassen oder auch weitere Erklärungen für eine bestimmte Bewertung innerhalb der Fragebögen Skalen.

Aus diesen Gründen fließen die Inhalte der Freitextfelder ebenfalls in die Betrachtung der Ergebnisse und Analyse mit ein. Ungefähr zwei Drittel der Nutzer haben gänzlich auf die Nutzung der Freitextfelder verzichtet. Aus den Kommentaren der

Merger		Bedienung	Farbe_Bedienung
3 UI Elemente	Mittelwert	74,3333	77,0667
	N	15	15
	Standardabweichung	33,55947	27,18578
2 UI Elemente	Mittelwert	92,4667	90,6667
	N	15	15
	Standardabweichung	11,45093	12,96515
Insgesamt	Mittelwert	83,4000	83,8667
	N	30	30
	Standardabweichung	26,30668	22,04030

Abbildung 4.29: Mittelwerte von Bedienung und Farbe_Bedienung in Bezug auf den Merger

Merger		Lernen	Farbe_Lernen
3 UI Elemente	Mittelwert	77,8000	80,0000
	N	15	15
	Standardabweichung	25,36646	18,91334
2 UI Elemente	Mittelwert	90,5333	82,9333
	N	15	15
	Standardabweichung	15,79270	20,66559
Insgesamt	Mittelwert	84,1667	81,4667
	N	30	30
	Standardabweichung	21,74790	19,52140

Abbildung 4.30: Mittelwerte von Lernen und Farbe_Lernen in Bezug auf den Merger

Merger		Fehlerkorrektur	Farbe_Fehlerkorr
3 UI Elemente	Mittelwert	73,2667	75,2667
	N	15	15
	Standardabweichung	27,79123	20,45366
2 UI Elemente	Mittelwert	70,0000	79,2000
	N	15	15
	Standardabweichung	31,13335	18,68995
Insgesamt	Mittelwert	71,6333	77,2333
	N	30	30
	Standardabweichung	29,04394	19,35456

Abbildung 4.31: Mittelwerte von Fehlerkorrektur und Farbe_Fehlerkorrektur in Bezug auf den Merger

anderen Nutzer geht unter anderem hervor, dass einigen die volle Funktionalität der Button-Lösung verschlossen geblieben ist. Es haben zwei Nutzer angegeben, dass es ihnen nicht möglich war, einzelne Tags aus ihrer Auswahl wieder herauszulöschen.

Ebenfalls zwei Nutzer haben den Wunsch nach einer automatischen Rechtschreibkorrektur geäußert. Ein weiterer Nutzer, der die Möglichkeit des Herauslöschen eines Buttons intuitiv angewendet hat, hat angemerkt, dass gerade bei längeren Buchstabenketten und möglicher auftretender Rechtschreibfehler das Herauslöschen nicht ideal ist. Daher wäre zu überdenken, eine weitere Funktion einzubauen, um Buttons, also die bereits gesetzten Schlagworte, zu erhalten, aber ihre Texte korrigieren zu können. Ein Nutzer hat angegeben, dass die Einfärbung der Tags seine Aufmerksamkeit auf sich gezogen haben und er diese daher häufiger ausgewählt hat. Er empfand die Einfärbung als hilfreicher, als das Interface bei dem alle Tags die gleiche Farbe habe. Dahingegen hat ein anderer Nutzer angegeben, dass ihm die Einfärbung überhaupt nicht geholfen hat. Insgesamt drei Probanden haben angemerkt, dass sie in den Vorschlägen keinerlei brauchbare Schlagworte gefunden haben. Gerade bei diesen Nutzern wurden zur Gewichtung dieser Aussage, die Daten der Sessions, also das Nutzungsverhalten der Kreuzverweis Services, hinzugezogen. Wie sich herausstelle, hat einer dieser Probanden die Services wirklich nicht in Anspruch genommen. Der zweite Proband beschränkte seine Nutzung

auf die Autocompletion und der Dritte nutzte beide Services sehr wenig (maximal einmal pro Bild).

Zwei Probanden haben ihre Verwirrung bezüglich der Anordnung der Vorschläge und wie die Vorschläge generiert werden zum Ausdruck gebracht. Eine mögliche Erklärung hierfür wäre eine mangelnde Erfahrung im Bereich des Semantic Web beziehungsweise im Umgang mit Ontologien.

Zusammenfassung

Obwohl die Bewertungen der Probanden für ein gewisses Interface bei der ein oder anderen Frage erheblich differieren, lässt sich doch feststellen, dass ein Interface mit Farbe und 2 User Interface Elementen von den Probanden als hilfreicher und angenehmer angesehen werden, als die anderen Interface-Kombinationen. Jedoch muss auch erkannt werden, dass ein Großteil der Probanden die Interfaces nicht zu 100% als intuitiv empfunden haben. Zwar gibt es eine Nutzergruppe, die viele Funktionen intuitiv gefunden und benutzt hat, aber auf der anderen Seite gibt es wiederum Probanden, die trotz kleiner Hilfen diese überhaupt nicht erkannt haben, was zur Folge hatte, dass diese sich allein gelassen gefühlt haben. Dieser Unterschied zwischen den Nutzern und ihrem Empfinden für intuitiv, wäre eine mögliche Begründung für die vielen Werte um die 50% bei den Fragen der Kategorie „Selbstbeschreibungsfähigkeit“, welche letztendlich auch zu der Normalverteilungsannahme führten.

4.5 Diskussion

Gerade bei einer Online-Evaluation muss man auf die Zuverlässigkeit der Nutzer vertrauen. Die Probleme beginnen aber im Grunde schon bevor der Nutzer die Evaluation durchführt, bei den technischen Schwierigkeiten: Man kann für die verschiedenen Systeme und Konfigurationen oder favorisierten Internetbrowsern der Nutzer nicht die gleiche Darstellung garantieren. Einem Nutzer kann bei eventuellen Verbindungsproblemen mit dem Kreuzverweis-Server, obwohl sich diese authentifizierte Verbindung automatisch aufbauen soll, nicht geholfen werden, außer der Nutzer würde von dem Angebot des E-Mailkontaktes Gebrauch machen. Zwar wurde nur von einem Probanden berichtet, initial ein Verbindungsproblem gehabt zu haben, aber man muss dennoch bedenken, dass man durch Probleme jedweder Art Probanden verlieren kann.

Manche Probanden haben teilweise nur ein bis zwei Schlagworte gesetzt, was sich für eine spätere repräsentative Auswertung als schwierig gestaltet. Eine Spekulation wäre, dass man diese Verhaltensweise während einer Evaluation mit dem Eye-Tracker positiv hätte beeinflussen können, da eine solche Evaluation nicht in

einem vollkommen anonymen Rahmen stattgefunden hätte.

Wenn man Funktionen dahingehend testen möchte, wie intuitiv sie sind, kann es passieren, dass diese Funktion einigen Nutzern verborgen bleiben. Die Ergebnisse deuten darauf hin, dass die Selektion eines Tags zwar nicht beschrieben werden muss, da jeder der Probanden, herausgefunden hat, wie man die Tags aus der Suggestions-Box auswählen kann, dass der umgekehrte Fall, also das Deselektieren oder im übertragenen Sinne Löschen eines Tags, aber durchaus eines Hinweises bedarf.

Ob und wie ein Schlagwort eingefärbt wird, hängt von der Qualität der DBpedia ab. Diese ist für die deutsche Sprache noch nicht so ausgereift, dass sofort bei der Eingabe eines Schlagwortes ausschließlich farbige Tags präsentiert wurden. Einige Nutzer haben teilweise auch nur eine Farbe gesehen. Diese fehlenden Informationen in der Dbpedia führen letztendlich im getesteten System dazu, dass die Aussagen bezüglich des Vergleichs von Interfaces mit und ohne Farbe sich nur in Tendenzen ausdrücken lassen. Eine schlechte Bewertung oder ein schlechteres Annotationsverhalten kann auch in der Frustration und Demotivation eines Nutzers durch so einen Effekt begründet sein. Eine nächste Nutzerstudie in diesem Bereich sollte entweder im englischen Sprachraum oder nach Erweiterung der Dbpedia durchgeführt werden. Als letztes Problem sei noch die Struktur der Ontologien genannt: Bei einigen Nutzern war die Erwartungshaltung bezüglich der Vorschläge, dass der Algorithmus genau das Bild sieht, das auch der Nutzer sieht und basierend auf diesen und den Nutzereingaben Vorschläge unterbreitet. Der Algorithmus arbeitet aber einzig mit den Eingaben des Nutzers und nicht mit Bildverarbeitungsalgorithmen. Daher wurden weitere Länder vorgeschlagen, wenn ein Nutzer einen Ländernamen eingegeben hat. Der Nutzer hat aber erwartet, dass der Algothmus nicht in der Breite der Baumstrutur sucht, sondern die Struktur in die Tiefe traversiert und damit beispielsweise die Unterkategorie Stadt respektive einen Stadtnamen vorschlägt.

4.6 Schlussfolgerung

Es lassen sich aus den vorliegenden und analysierten Daten folgende Erkenntnisse zusammenfassen:

Im Schnitt werden mehr Schlagworte bei Nachrichtenbildern und Interfaces mit Typinformation gesetzt. Was sich zum einen mit den Daten, welche durch die Fragebögen gewonnen werden konnten begründen lässt, da bei Interfaces mit Farbe eine Tendenz zur höheren Benutzerzufriedenheit zu vermerken ist, zum anderen bestanden teilweise Schwierigkeiten die Sehenswürdigkeiten zu erkennen. Den Probanden fehlten hierzu Assoziationen, da es sich bei den Bildern nicht um eigene handelte. Nachrichten hingegen begegnen uns mehrfach am Tag, sodass in diesen

Bereichen viele Probanden oft eine Fülle an Informationen besitzen, ohne sich der wirklich bewusst zu sein. Die Autocompletion wurde oft in Verbindung mit Bildset 1 benutzt. Da es auf Grund der Speicherung gewisser Daten möglich war, nachzuvollziehen, welche Tags aus der Autocompletion stammten, lässt sich festhalten, dass diese oft dann zum Einsatz kam, wenn schwierige oder besonders lange Namen und Begriffe, wie beispielsweise Taj Mahal, Sydney Opera House oder Freiheitsstatue, als Schlagwort definiert werden sollten.

Bezüglich der Vorschläge hätten die Meinungen der Probanden und die Intensität der Nutzung dieses Services nicht unterschiedlicher sein können. Während die einen nicht ein Schlagwort aus den Vorschlägen entnommen haben, haben andere diesen Services fast ausschließlich genutzt. Einige Probanden haben angemerkt, dass sie nichts mit den Vorschlägen anfangen konnten. Eine mögliche Begründung hierfür liegt vielleicht in der Diskrepanz zwischen Erwartungshaltung des Nutzers, dass die Vorschläge basierend auf Bildverarbeitungsalgorithmen generiert werden oder, dass eine ähnlich mächtige bereits annotierte Bilddatenbank, wie bei LabelMe [RTMF08], zu Grunde liegt und damit die Annotation zu großen Teilen automatisiert, und der tatsächlichen Funktionsweise des Services. Andere Nutzer dagegen haben mit Hilfe der Vorschläge Sehenswürdigkeiten herausfinden können, deren Namen sie nicht kannten.

Die Zeit, welche ein Nutzer im Durchschnitt für die Bearbeitung mit einem Interface benötigte variierte nur minimal, mit der Tendenz beim 2-elementigen Interface mit Typinformation schneller zu sein.

Der Service Vorschläge zu generieren, ist im Ganzen als positiv und unterstützend für den Nutzer zu bewerten. Probanden, die mit einem 2-elementigen Interfaces mit Typinformation gearbeitet haben, waren im Durchschnitt um fünf Sekunden schneller, bei geringerer Standartabweichung, als bei allen anderen Interfaces. Dieses Ergebnis deutet zwar nur auf ein mögliches Potential zur Zeitersparnis hin, es besagt aber auch, dass kein negativer Effekt durch den Einsatz dieser Gestaltungsmerkmale nachgewiesen werden kann. Dieses Fazit wird durch die Ergebnisse der Analyse der Fragebogen gestützt. Bei den meisten Fragen lässt sich bei einem Interface mit 2 Elementen und Typinformation eine höhere durchschnittliche Punktzahl und eine niedrigere Varianz feststellen.

Kapitel 5

Fazit

5.1 Zusammenfassendes Fazit

Im Rahmen dieser Arbeit wurden verschiedene Interfaces hinsichtlich ihrer Usability Eigenschaften untersucht. Auf Grund des Evaluationsdesigns, des ordinalen Messniveaus und fehlender Normalverteilungen eines großen Teils der Daten, musste auf ein nicht-parametrisches Verfahren zur Analyse zurückgegriffen werden. Nicht-parametrische Verfahren haben generell den Nachteil, dass sie weniger mächtig, als parametrische Verfahren sind [CF04]. Die 30 Probanden die an der Evaluation teilgenommen haben, konnten mit ihren Anmerkungen Begründungen für bestimmte statistische Werte geben, aber auch Anregungen für weitere Ideen und das Fazit, dass einige Funktionen für bestimmte Personengruppen durchaus intuitiv sind. Dennoch lassen sich mit den Ergebnissen der 30 Probanden nur Tendenzen zusammenfassen und keine allgemeingültigen Aussagen treffen. Der Einsatz von unterstützenden Funktionen, wie die Autocompletion oder den Service, Vorschläge für Tags zu unterbreiten, sollte bei der Gestaltung von Annotationsinterfaces berücksichtigt werden. Ebenso kann die Codierung von Typinformationen mit Hilfe von Farbe, sofern sie in Maßen geschieht, für einen Nutzer hilfreich sein und somit zur Steigerung der Zufriedenheit im Umgang mit dem Interface beitragen. Es bleibt weiterhin festzuhalten, dass Nutzerstudien unerlässlich sind, wenn man möglichen Problemen vorbeugen und keinen Nutzer im voraus ausschließen möchte. Es werden immer unterschiedliche Meinungen darüber existieren, was intuitiv ist und was nicht. Es wird auch immer die verschiedensten Nutzer, mit unterschiedlichen Erfahrungen im Umgang mit Computern oder physischen Grenzen, wie zum Beispiel der Farbenblindheit, geben, aber es gilt im Rahmen einer Nutzerstudie herauszufinden, welche Hilfen man einbauen muss, um möglichst alle Nutzer anzusprechen, ohne dabei durch zu viel Hilfe zu verärgern.

5.2 Ausblick

Die Evaluation sollte noch einmal, allerdings diesmal im Labor mit einem Eye Tracker, stattfinden. Daten eines Eye Trackers würden die Möglichkeit bieten, subjektive Angaben der Probanden innerhalb der Fragebögen und der Freitextfelder, zu bestätigen oder zu widerlegen. Außerdem ließen sich beispielsweise auch design-betreffende Probleme mit diesen Daten erkennen, wenn zum Beispiel Nutzer einen gewissen Button auf Grund seiner Position gar nicht bemerken, da dieser außerhalb ihres Wahrnehmungsfeldes liegt. 30 Ergebnisse von Probanden, wie sie in der Evaluation analysiert wurden, sind gut um Tendenzen festzustellen, allerdings könnte man mit einer noch größeren Anzahl an Probanden gewisse Ausreißer minimieren. Daher wäre eine Versuchsreihe, wie man sie auf Amazon Mechanical Turk implementieren kann, denkbar. Im Rahmen dieser Arbeit konnten zwei der vier unabhängigen Variablen noch nicht evaluiert werden. In Folge dessen, sollte eine Erweiterung des Evaluations-Wizard stattfinden, um den Lernalgorithmus einer Benutzerstudie zu unterziehen. Eine zweite Erweiterung sollte der Untersuchung einer Synonym-Funktion dienen: Oft gibt es für einen Gegenstand, ein Land, Ereignisse und vieles mehr verschiedene Begriffe. In der Basisversion muss der Nutzer, wenn er für beispielsweise einen Gegenstand auf einem Bild mehrere Begrifflichkeiten annotieren möchte, diese einzeln eingeben oder falls diese schon als Vorschläge in der Suggestions-Box aufgeführt werden über das Anklicken der zugehörigen Buttons auswählen. Die Idee für eine weitere Version wäre, dass die Synonyme eines Begriffes automatisch ebenfalls zum Bild hinzugefügt werden, sobald dieser Begriff vom Nutzer als Schlagwort für das Bild definiert wurde. Aus dem Web- und Softwaredesign kennt man einige selbsterklärende Icons, wie zum Beispiel das Kreuz zum Schließen einer Anwendung. Da dem Nutzer nicht die Freiheit genommen werden soll, selber festlegen zu können, welche Begriffe er letztendlich wirklich dem betreffenden Bild zuordnen möchte, versteckt man die Synonyme eines Begriffes hinter einem ebensolchen Symbol. Klickt der Nutzer auf das Symbol zum Expandieren der Synonyme, erscheinen diese hell im Vordergrund, während alles andere verdunkelt im Hintergrund verschwindet. Da die Synonyme ebenfalls Tags sind, werden diese auch dargestellt wie Tags: Als einzelne Buttons mit der gleichen Funktionalität. Möchte der Nutzer ein bestimmtes Synonym nicht in der Auswahl lassen, kann er auch hier mit einem Klick auf den betreffenden Button, dieses Schlüsselwort aus der Auswahl löschen. Um diese Ansicht wieder zu schließen genügt ein Klick außerhalb des hellen Bereiches, indem die Synonym-Tags dargestellt werden, oder auf den Button mit dem Kreuz, welcher sich rechts oben im hellen Bereich befindet. Einige Eigenschaften, wie zum Beispiel das Klicken auf ein Button in der Suggestions-Box als Möglichkeit diesen Button in die Auswahl zu übernehmen oder den umgekehrten Fall, dass ein Klicken auf ein Schlagwort in der Selections-Box zu einer Deselektierung führt, hat man in der Theorie als intuitiv

verständlich eingeschätzt. Um diese Theorie zu stützen hat man auf weitere Hilfen zu dieser Funktion verzichtet. Einige wenige Probanden haben nicht alle Funktionen herausfinden können, daher sollte über eine ähnliche optische Erweiterung bei den Tags nachgedacht werden. Eine Möglichkeit wäre, dass sobald man mit der Computer-Maus über die Tags in der Selection-Box fährt, ein kleines Kreuz, wie man es vom Schließen von Programmen kennt eingeblendet wird.

Kapitel 6

Anhang A

6.1 Implementation

6.1.1 Basissystem Kreuzverweis

Kreuzverweis stellt die Webapplikation SSmart Keywording for the Web^{“1}, unter der Apache Lizenz Version 2.0, als GitHub² Repository zum Download³ bereit. Möchte man direkt auf den vollen Funktionsumfang des Verzeichnisses zugreifen können, empfiehlt sich ein linuxbasiertes Betriebssystem.

Interface

Das Basissystem, Smart Keywording for the Web, von Kreuzverweis bietet drei untereinander angeordnete User-Interface Elemente an: Ein Textfeld, eine Suggestions- und eine Selections-Box. Es baut in der Grundstruktur auf HTML und CSS auf. Um eine erweiterte Funktionalität zu gewährleisten, wurden einige Funktionen in JavaScript implementiert. Die Bibliothek jQuery diente hierbei als Hilfe, da diese unter anderem eine zuverlässige Kommunikation über Ajax, bei einem geringeren Codezeilenaufwand, gewährleistet. Darüberhinaus griff man auf JavaScript Klassen von Twitter und jQuery, welche unter der Apache Lizenz Version 2.0 verfügbar sind, zurück.

¹<http://kreuzverweis.herokuapp.com/>

²<https://github.com/>

³https://github.com/kreuzverweis/smartkeywroding_web.git

6.1.2 Lokaler Webserver

Um das heruntergeladene System Smart Keywording for the Web in einer realitätsnahen Szenerie, so wie man sie auf der Webseite von Kreuzverweis⁴ vorfindet, testen zu können, liegt dem ausgecheckte GitHub Verzeichnis ein bereits vorkonfigurierter Localhost bei. Diesen kann man allerdings erst verwenden, wenn man die Pakete Yum (Yellow dog Updater, Modifier), zum Installieren der im folgenden benötigten Pakete, und sbt, ein Build-Tool für Java und Scala Projekte, herunter geladen und die nötigen Schritte zur Installation durchgeführt hat. Der lokale Webserver (Server.scala) und die Schnittstelle (Proxy.scala) zur Kommunikation mit dem Kreuzverweis Keyword Webservice Server sind in Scala implementiert und um das Toolkit Unfiltered⁵, welches HTTP Requests in Scala bereitstellt, und die Bibliothek Dispatch⁶, welche die asynchrone Interaktion mit HTTP unterstützt, erweitert. Eine Datei namens demo.conf legt weitere Eigenschaften, wie zum Beispiel die Portnummer des Servers oder die Client ID und das Client Secret, fest.

6.1.3 Kreuzverweis Server

Generell werden alle Verbindungen, die zum Server aufgebaut werden, über SSL, also eine sichere HTTPS Verbindung (Port 443), hergestellt. Diese Regel gilt also sowohl für die initiale Authentifizierung, als auch für die gesamte Nutzungsdauer der Services. Alle Daten, die der Server als Antwort sendet, liegen im XML Format vor. Dies gilt ebenfalls sowohl für die Kommunikation während des Authentifizierungsprozesses, als auch für die Autocompletion- und Suggestion-Vorschläge.

OAuth 2 Authentifizierung

Die Kommunikation mit dem Kreuzverweis Keyword Webservice verläuft gesichert über OAuth 2. Man benötigt für die Kommunikation die ein Nutzer mit einer Anfrage über ein Interface, welches die Kreuzverweis Funktionalität der Autocompletion und Proposals nutzt, an den Keyword Webserver stellt, eine Clientid und ein Clientsecret für das jeweilige Interface. Sowohl Client ID als auch Client Secret sind nur dem Entwickler oder dem Betreiber des eigenen Webservers bekannt. Darüber hinaus bekommt jeder Nutzer des Interfaces eine User ID und einen User Namen zugewiesen. Der User Name kann nicht zu Authentifizierungszwecken benutzt werden einzig mit der User ID, kann dann ein Token generiert werden,

⁴<http://kreuzverweis.herokuapp.com/>

⁵<http://unfiltered.databinder.net/Unfiltered.html>

⁶<http://dispatch.databinder.net/Dispatch.html>

welches 24 Stunden gültig ist. Diese Daten werden in Browser Cookies hinterlegt, damit der Nutzer vom eigentlichen Authentifizierungsprozess nicht behelligt wird.

Backoffice

Ist man Entwickler oder möchte das Smartkeywording for the Web auf dem eigenen Localhost testen, muss man sich zunächst auf der Seite des Kreuzverweis Backoffices⁷ registrieren. Hier hat man die Möglichkeit in einer GUI eine Organisation zu erstellen. Diese benötigt man, um darunter einen Client zu registrieren, welchem nach Registrierung eine ID und ein Secret zugewiesen wird. Übersichtshalber ist es möglich einen Namen zu definieren. Auch hier kann der Name nicht zur Authentifizierung benutzt werden, er ist der übersichtlichkeitshalber frei wählbar. Unter diesem Client befindet sich eine Übersicht über alle unter diesem Client registrierten User mit ihren IDs, Namen und Tokens.

Keyword Webservice

Als Webserver dient der in Java geschriebene Jetty, welcher ebenfalls unter der Apache Lizenz Version 2.0 zur Verfügung steht. Die Daten werden in NoSQL-Datenbanken verwaltet. NosQL-Datenbanken garantieren eine kurze Antwortzeit und kurze Entwicklungszyklen⁸. Implementiert wurde in Scala und Akka⁹

⁷<https://backoffice.kreuzverweis.com/>

⁸community.oreilly.de/blog/2012/07/25/smart-keywording-die-durchdringung-der-multimedialen-flut

Kapitel 7

Anhang B

7.1 Quellennachweise zu den Bildsets

Bildset B1

Sydney Opera House

<http://www.geo.de/reisen/community/bild/bigImage/89833/Sydney-Opera-House.jpg>

Taj Mahal

<http://www.flickr.com/photos/bestarns/7500747498/>

Berlin Brandenburgertor

(Eigenes Foto)

Berlin Fernsehturm

(Eigenes Foto)

Köln

<http://www.wo-war-ich-schon.de/photo/normal/27d5bad76c34dbf6.jpg>

Moskau

<http://www.flickr.com/photos/ricci-armani/map/>

Paris

<http://www.flickr.com/photos/ricci-armani/7670961666/in/photostream>

NewYork

http://www.holidaycheck.de/vollbild-Freiheitsstatue%2BFreiheitsstatue%2Bvom%2BHelikopter%2Baus-ch_ub-id_1156062482.html

Shanghai

http://www.flickr.com/photos/further_to_fly/7101106115/

Rom

<http://www.flickr.com/photos/mrscurlyhead/7479079134/>

Bildset B2**Adenauer, Brandt**

<http://www.google.com/imgres?um=1&hl=de&client=firefox-a&sa=N&rls=org.mozilla:de:official&biw=1366&bih=596&tbo=isch&tbnid=jNe-oHBSzvuN3M:&imgrefurl=http://1960inberlin.wordpress.com/about/&docid=THZQQTOPqOSI3M&imgurl=http://1960inberlin.files.wordpress.com/2009/10/kennedy-willy.jpg%253Fw%253D510%2526h%253D361&w=510&h=361&ei=PJiPUOWPH0ng4QTMxYCgAQ&zoom=1&iact=hc&vpx=134&vpy=294&dur=1959&hovh=189&hovw=267&tx=208&ty=92&sig=103364757428647631&tbnh=138&tbnw=214&start=0&ndsp=17&ved=1t:429,i:87>

Steve Jobs

http://www.google.com/imgres?start=115&um=1&hl=de&client=firefox-a&rls=org.mozilla:de:official&biw=1366&bih=596&tbo=isch&tbnid=XSuuu6nBiaBS_M:&imgrefurl=http://www.iphonophile.fr/hier-soir-steve-jobs-a-aussi-ete-nommé-a&docid=UCqM9YbMSTcSJM&imgurl=http://www.iphonophile.fr/wp-content/uploads/2012/02/Steve-Jobs-Oscars.jpg&w=400&h=258&ei=YpiPUMmkEITH4QSgxYHABg&zoom=1&iact=rc&dur=333&sig=103364757428647639715&page=6&tbnh=131&tbnw=194&ndsp=24&ved=1t:429,i:73&tx=93&ty=88

London Olympia

<http://www.faz.net/aktuell/bilder-des-tages-bilder-des-tages-1135321.html>

Gedenken an den 11. September

<http://m.faz.net/;bild=56/aktuell/gesellschaft/menschen/bilder-des-tages-bilder.html>

Bundestag

<http://www.domradio.de/aktuell/74754/bundestag-diskutiert-organspenden.html>

Angelina Jolie

<http://www.welt.de/vermisches/bilder-des-tages/article109194902/Die-besten-Bilder-des-Tages-2012.html>

ESM Börse

<http://www.mz-web.de/servlet/ContentServer?pagename=ksta/page&atype=ksArtikel&aid=1347465970195>

ESM Bundesverfassungsgericht

<http://www.sueddeutsche.de/politik/entscheidung-ueber-fiskalpakt-und-esm-karlsruhe-1395420>

Prinz William

<http://www.n-tv.de/leute/William-waere-gern-unsichtbar-article7201131.html>

Barack Obama

<http://www.spiegel.de/politik/ausland/bill-clinton-kaempft-fuer-obama-kaempferische-r.html>

Literaturverzeichnis

- [AD04] AHN, Luis von ; DABBISH, Laura: Labeling images with a computer game. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2004 (CHI '04). – ISBN 1-58113-702-8, S. 319–326
- [AHO⁺09] AMIN, Alia ; HILDEBRAND, Michiel ; OSSENBRUGGEN, Jacco V. ; EVERS, Vanessa ; HARDMAN, Lynda: *Organizing Suggestions in Autocompletion Interfaces*. Berlin, Heidelberg, 2009 (LNCS 5478)
- [ALB06] AHN, Luis von ; LIU, Ruoran ; BLUM, Manuel: Peekaboom: a game for locating objects in images. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2006 (CHI '06). – ISBN 1-59593-372-7, S. 55–64
- [ASSB05] APPAN, Preetha ; SHEVADE, Bageshree ; SUNDARAM, Hari ; BIRCHFIELD, David: Interfaces for networked media exploration and collaborative annotation. In: *Proceedings of the 10th international conference on Intelligent user interfaces*. New York, NY, USA : ACM, 2005 (IUI '05). – ISBN 1-58113-894-6, S. 106–113
- [Bre74] BREDENKAMP, Jürgen: Nonparametrische Prüfung von Wechselwirkungen. In: *Psychologische Beiträge*. Berlin : VEB Verlag Technik, 1974, S. 398–416
- [CF04] CANN, A.J. ; FEUERBACHER, B.: *Mathe für Biologen*. Wiley-VCH, 2004. – ISBN 9783527311835
- [Coh12] COHEN, James N.: Professional Resource: The Potential of Google+ as a Media Literacy Tool. In: *The National Association for Media Literacy Education's*, 2012 (Journal of Media Literacy Education 4:1), 93–96

- [Con98] CONOVER, W. J.: *Practical Nonparametric Statistics*. John Wiley & Sons, 1998 <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471160687>. – ISBN 0471160687
- [Fun10] FUNKE, Frederik: *Internet-Based Measurement with Visual Analogue Scales: An Experimental Investigation*, Universität Tübingen, Diss., 2010. http://frederikfunke.net/dissertation/Funke_%282010%29_Internet-based_measurement_with_visual_analogue_scales-Dissertation.pdf
- [GH99] GEDIGA, Günther ; HAMBORG, Kai-Christoph: IsoMetrics: An usability inventory supporting summative and formative evaluation of software systems. In: *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I - Volume I*. Hillsdale, NJ, USA : L. Erlbaum Associates Inc., 1999. – ISBN 0-8058-3391-9, S. 1018–1022
- [HSWW03] HOLLINK, Laura ; SCHREIBER, Guus ; WIELEMAKER, Jan ; WIELINGA, Bob: Semantic Annotation of Image Collections. In: *Proceedings of the Second International Conference on Knowledge Capture K-CAP*. Sanibel, Florida, USA, 10 2003
- [IS11] IAKOVIDIS, Dimitris K. ; SMAILIS, Christos V.: Efficient Semantically-Aware Annotation of Images. In: *In Proc. IEEE International Conference of Imaging Systems and Techniques (IST)*, 2011, S. 146–149
- [JL07] JANSSEN, Jürgen ; LAATZ, Wilfried: *Statistische Datenanalyse mit SPSS für Windows*. 6., neu bearb. und erw. Aufl. Berlin : Springer, 2007 <http://www.hwp-hamburg.de/janssenJ/spss.html>. – ISBN 9783540729778
- [LS04] LIU, H. ; SINGH, P.: ConceptNet A Practical Commonsense Reasoning Tool-Kit. In: *BT Technology Journal* 22 (2004), Oktober, Nr. 4, S. 211–226. – ISSN 1358–3948
- [LTG11] LAW-TO, Julien ; GREFENSTETTE, Gregory: VOVALEAD: a scalable video search engine based on content. In: *Proceedings of the 19th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2011 (MM '11). – ISBN 978-1-4503-0616-4, S. 747–748
- [Lux09] LUX, Mathias: Caliph & Emir: MPEG-7 photo annotation and retrieval. In: *Proceedings of the 17th ACM international conference on*

- Multimedia.* New York, NY, USA : ACM, 2009 (MM '09). – ISBN 978-1-60558-608-3, S. 925–926
- [LZT09] LOHmann, Steffen ; ZIEGLER, Jürgen ; TETZLAFF, Lena: Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In: *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I.* Berlin, Heidelberg : Springer-Verlag, 2009 (INTERACT '09). – ISBN 978-3-642-03654-5, S. 392–404
- [MC01] MATEOS, Gins G. ; CHICOTE, Cristina V.: *Face Detection on Still Images Using HIT Maps.* 2001
- [RGMM07] RIVADENEIRA, A. W. ; GRUEN, Daniel M. ; MULLER, Michael J. ; MILLEN, David R.: Getting our head in the clouds: toward evaluation studies of tagclouds. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York, NY, USA : ACM, 2007 (CHI '07). – ISBN 978-1-59593-593-9, S. 995–998
- [RTMF08] RUSSELL, Bryan C. ; TORRALBA, Antonio ; MURPHY, Kevin P. ; FREEMAN, William T.: LabelMe: A Database and Web-Based Tool for Image Annotation. In: *Int. J. Comput. Vision* 77 (2008), Mai, Nr. 1-3, S. 157–173. – ISSN 0920-5691
- [SJ08] SINHA, Pinaki ; JAIN, Ramesh: Classification and annotation of digital photos using optical context data. In: *Proceedings of the 2008 international conference on Content-based image and video retrieval.* New York, NY, USA : ACM, 2008 (CIVR '08). – ISBN 978-1-60558-070-8, S. 309–318
- [THB⁺06] TUFFIELD, Mischa M. ; HARRIS, Stephen ; BREWSTER, Christopher ; GIBBINS, Nicholas ; CIRAVEGNA, Fabio ; SLEEMAN, Derek ; SHADBOLT, Nigel R. ; WILKS, Yorick: Image annotation with photoco-pain. In: *In Proceedings of Semantic Web Annotation of Multimedia (SWAMM-06) Workshop at the World Wide Web Conference 06.* WWW, 2006, S. 22–26
- [WDS⁺01] WENYIN, Liu ; DUMAIS, Susan ; SUN, Yanfeng ; ZHANG, Hongjiang ; CZERWINSKI, Mary ; FIELD, Brent: Semi-automatic image annotation. In: *In Proc. of Interact 2001: Conference on Human-Computer Interaction,* 2001, S. 326–333

- [Wer78] WERSIG, Gernot: *Thesaurus-Leitfaden: eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis.* Pullach bei München : Dokumentation, 1978
- [WH87] WILRICH, Peter-Theodor ; HENNING, Hans-Joachim ; GRAF, Ulrich (Hrsg.): *Formeln und Tabellen der angewandten mathematischen Statistik.* 3., völlig neu bearb. Aufl. Berlin [u.a.] : Springer, 1987. – ISBN 3540169016
- [ZVS07] ZENDLER, Andreas ; VOGEL, Markus ; SPANNAGEL, Christian: Experimentelle Unterrichtsforschung: Nützliche Versuchspläne und Rank Order Statistics. In: *Notes on Educational Informatics - Section A: Concepts and Techniques* 3, 2007, S. 41–66