Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

DEPARTMENT INFORMATION

Bachelorarbeit

Evaluierung von Suchmaschinen – Qualitätsvergleich von Google- und Bing-Suchergebnissen unter besonderer Berücksichtigung von Universal-Search-Resultaten

vorgelegt von

Markus Günther

Studiengang Bibliotheks- und Informationsmanagement

erster Prüfer: Prof. Dr. Dirk Lewandowski

zweiter Prüfer: Prof. Dr. Franziskus Geeb

Hamburg, Juli 2012

Abstract

Diese Arbeit geht der Frage nach, ob und wie sich die Relevanz des Gesamtergebnisses einer Suche mit einer Suchmaschine ändert, wenn in die Ergebnisseite Resultate spezieller Ergebniskollektionen eingebunden werden (so das Prinzip der „Universal Search“). Die untersuchten Kollektionen hierbei sind Bilder, Produkte, Videos, Nachrichten und lokale Resultate. Weiter wird untersucht, ob Google relevantere reguläre Treffer (Dokumente) liefert als Bing, und ob die auf der Suchergebnisseite sichtbaren Trefferbeschreibungen den Inhalt und damit die Relevanz der dahinterliegenden Treffer korrekt widerspiegeln. Die theoretischen Hintergründe dieser Themen werden in einem Literaturteil behandelt. Für die Beantwortung der Forschungsfragen wurde im Mai/Juni 2012 ein Retrievaltest für Google und Bing durchgeführt, um die Relevanz der Suchergebnisse, die beide Suchmaschinen auf jeweils 50 Suchanfragen lieferten, erheben und vergleichen zu können. Dazu wurden Juroren ca. 2.400 Items vorgelegt – nämlich jeweils die ersten zehn regulären Ergebnisse (Beschreibungen und Treffer) sowie alle auf der ersten Ergebnisseite angezeigten Resultate der berücksichtigten speziellen Kollektionen (ebenfalls Beschreibungen und Treffer). Diese wurden nach ihrer Relevanz bewertet. Die Auswertung des Tests ergibt, dass Trefferbeschreibungen die Relevanz der Treffer i.d.R. korrekt widerspiegeln, jedoch mit leichter Tendenz zu geringerer Relevanz als die Treffer. Google gibt tatsächlich relevantere reguläre Treffer aus als Bing, jedoch ist der Unterschied nicht gravierend. Eingebundene Resultate spezieller Kollektionen sind zu einem relativ hohen Teil irrelevant, und wenn sie als relevant eingeschätzt wurden, war die Relevanz wiederholt so niedrig, dass als Folge das Gesamtergebnis abgewertet wurde. Diese Arbeit bietet eine erste Einschätzung zum praktischen Nutzen von in die Ergebnisseite eingestreuten Resultaten spezieller Kollektionen, die als Ausgangsbasis für weitere Forschungsarbeiten in diesem Bereich dienen kann, z.B. um einzelne Kollektionen auf spezifische Merkmale hin zu untersuchen, die Nutzern helfen.

Keywords: Web Information Retrieval, Suchmaschinen, Retrievaltests, Retrievaleffektivität, Trefferbeschreibungen, Google, Bing, Universal Search, Blended Search

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

IR	Information Retrieval
US	Universal Search
BS	Blended Search

1. Einleitung

In diesem Kapitel werden die Untersuchungsbereiche dieser Arbeit durch eine thematische Einführung vorgestellt, sowie ihre Motivation und ihr Aufbau erläutert.

Suchmaschinen stellen heutzutage ein substanzielles Mittel zur Informationsbeschaffung dar, dem *Information Retrieval* (IR). Sie sind neben E-Mails die am häufigsten verwendete Anwendung im Internet (vgl. VAN EIMEREN & FREES 2009, S. 340f): In Deutschland werden pro Monat mehrere Milliarden Suchanfragen abgeschickt (vgl. COMSCORE 2010).

Die mit Abstand populärste Suchmaschine in Deutschland ist *Google* (vgl. WEBHITS 2012). In *Retrievaltests*, mit denen die Fähigkeit von Informationssystemen gemessen wird, zum Informationsbedürfnis des Nutzers passende Dokumente auszugeben, schnitt Google in diversen Fällen (z.B. GRIESBAUM 2004, TAWILEH & GRIESBAUM & MANDL 2010) am besten ab und gilt daher auch als Referenzsuchmaschine (vgl. LEWANDOWSKI 2011A, S. 208). Anfang dieses Jahres ließ *Bing* ihren Betastatus in Deutschland hinter sich (vgl. MICROSOFT 2012) und ist damit als vollwertiger Gegenspieler zu sehen. Es gilt zu untersuchen, ob Microsoft Google so ernstlich Konkurrenz machen kann.

Bei ihren Suchen sind Suchmaschinennutzer i.d.R. weder bereit, viel Mühe in die Formulierung ihrer Anfragen zu investieren (was sich in deren Knappheit niederschlägt), noch viele Ergebnisse auf vielen Ergebnisseiten zu betrachten (vgl. CUTRELL & GUAN 2007, S. 5, HÖCHSTÖTTER & KOCH 2008, pdf S. 9). Die Suchmaschine soll das Informationsbedürfnis hinter der Anfrage erkennen und das passende Ergebnis möglichst an erster Stelle liefern. Suchmaschinenbetreiber passen ihre Produkte diesem Verhalten an.

So helfen grundsätzlich die *Trefferbeschreibungen*, kleine Ausschnitte der Treffer, die heute bei bestimmten Ergebnistypen noch zusätzliche Informationen enthalten (vgl. LEWANDOWSKI 2012, pdf S. 4f), die Relevanz eines Ergebnisses einzuschätzen, ohne es anklicken und direkt betrachten zu müssen. Doch spiegeln die Beschreibungen tatsächlich den Inhalt und damit die Relevanz der zugehörigen Treffer wider? LEWANDOWSKI etwa kam zu dem Ergebnis, dass die Beschreibungen im Durchschnitt relevanter als die Treffer sind (vgl. LEWANDOWSKI 2008C, pdf S. 10f). Zu untersuchen ist also, ob Nutzer so tatsächlich zu vermeintlich relevanten Treffern gleitet werden, und ob ihnen evtl. auch relevante Treffer durch irrelevante Beschreibungen verborgen bleiben.

Ein großer Entwicklungsschritt von Suchmaschinen, der den Nutzern ebenfalls den Suchprozess erleichtern soll, spielt in Deutschland erst seit wenigen Jahren eine bedeutende Rolle: die *Universal Search* (US, „universelle Suche“). Ihr Prinzip ist es, ausgewählte Ergebnisse aus speziellen Kollektionen, wie bspw. Bilder, Produkte und Videos, in die Ergebnisseite der Standard-Websuche einzustreuen. Klar ist, dass dadurch das Design der

Ergebnisseite und damit das Blickverhalten der Nutzer verändert wird (vgl. QUIRMBACH 2009, S. 231ff) - doch es gilt zu untersuchen, ob diese zusätzlichen Resultate für die Nutzer grundsätzlich überhaupt relevant sind.

Für diese Untersuchungsziele wurde im Rahmen dieser Arbeit ein Retrievaltest mit Fokus auf dem Thema Universal Search durchgeführt. Dabei wurde Bing an Google gemessen, und bei alldem die Qualität der Trefferbeschreibungen berücksichtigt.

1.1 Motivation

In diesem Abschnitt wird die Motivation der vorliegenden Arbeit erläutert.

Für die Durchführung des Retrievaltests und die Anfertigung dieser Arbeit gibt es mehrere Motive. Generell ist es sinnvoll, die *Pertinenz*, also die subjektive Relevanz (vgl. STOCK 2007, S. 68f), von Suchergebnissen aus Nutzersicht zu ergründen und daraus Rückschlüsse sowohl für die Wissenschaft als auch für die Suchmaschinenentwicklung zu ziehen. Darüber hinaus ist es reizvoll, die erste Arbeit über die Relevanz von Resultaten spezieller Kollektionen anzufertigen und dadurch einen Eckstein für die weitere Forschung zu legen.

Die Methode des Retrievaltests wurde gewählt, weil sie eine lange Tradition im Information Retrieval hat (die ersten Retrievaltests fanden 1953 statt (vgl. GRIESBAUM 2000, S. 13)). Mit der Entstehung des *Web Information Retrieval* wurde sie für Suchmaschinen adaptiert und stellt heute auch in diesem Teilbereich das Standardverfahren zur Evaluierung von Informationssystemen dar. Weil die Suchergebnisse in solchen Tests meist von realen Suchmaschinennutzern bewertet werden, weisen die so gewonnenen Erkenntnisse i.d.R. eine hohe Relevanz auf.

1.2 Aufbau der Arbeit

In diesem Kapitel wird der Aufbau dieser Arbeit dargelegt.

Die vorliegende Arbeit gliedert sich in zwei grundlegende Teilbereiche: den Stand der Forschung und den Retrievaltest.

Im ersten Teilbereich wird zunächst kurz der Bereich des Web Information Retrieval vom klassischen IR abgegrenzt. Danach werden die einzelnen Aspekte von Retrievaltests erläutert, um im Anschluss auf Kritik, neuere Impulse und einige bedeutende Tests einzugehen. Zuletzt wird das Prinzip der Universal Search beleuchtet.

Im zweiten Teilbereich werden zunächst die Forschungsfragen und –hypothesen präsentiert. Danach wird die angewandte Methodik des Retrievaltests in korrespondierender Struktur zu den im ersten Teilbereich erläuterten Aspekten von Retrievaltests dargelegt. Abschließend werden die Ergebnisse des Tests im Hinblick auf die Forschungsfragen und –hypothesen präsentiert und diskutiert.

In dieser Arbeit bezeichnet der Begriff „Trefferbeschreibung“ das, was von einem Ergebnis direkt auf der Suchergebnisseite zu sehen ist. Ein „Treffer“ ist das Ergebnisdokument selbst, das durch einen Klick erreicht wird. „Ergebnis“ meint die Einheit aus Beschreibung und Treffer - „Resultat“ wird dazu synonym verwendet (meist im Zusammenhang mit speziellen Ergebniskollektionen).

2. Stand der Forschung

Durch dieses Kapitel wird das grundlegende Verständnis vom Bereich Web Information Retrieval vermittelt, bevor ausführlich auf Retrievaltests eingegangen wird. Daran knüpfen Kritik an dieser Evaluierungsmethode, neuere Entwicklungen und die Vorstellung einiger bedeutender Tests an. Durch die Untersuchung des Prinzips der Universal Search wird die Verortung dieser Arbeit und die Erklärung ihrer Notwendigkeit abgeschlossen.

2.1 (Web) Information Retrieval

Web Information Retrieval ist ein Bereich des Information Retrieval. Deshalb werden im Folgenden zuerst die Grundlagen des IR vermittelt, bevor im Anschluss das Web IR vorgestellt wird. Es folgt eine Darstellung wichtiger Unterschiede der beiden Bereiche sowie ein kurzer Einstieg in die Evaluierung von Suchmaschinen.

2.1.1 Information Retrieval

In diesem Abschnitt wird der Bereich des Information Retrieval in Grundzügen vorgestellt.

Unter dem Begriff des Information Retrieval, auf Deutsch etwa „Informationsrückgewinnung“, werden Verfahren zum Auffinden von Inhalten in einer Computerumgebung,

klassischerweise in Datenbanken, subsummiert. Es geht darum, dem Nutzer die richtige Information zur richtigen Zeit zu beschaffen (vgl. CHU 2010, S. 18).

Die gesuchten Inhalte können die Form von Texten, Bildern oder Sound haben oder auch multimedial sein, weswegen sie im Folgenden allgemein als *Dokumente* bezeichnet werden.

Die grundlegende Problematik des Auffindens liegt darin, dass das Informationsbedürfnis des Suchenden nur über Repräsentationen auf Dokumente abgebildet werden kann. Dieser Umweg ist in *Abbildung 1* dargestellt.

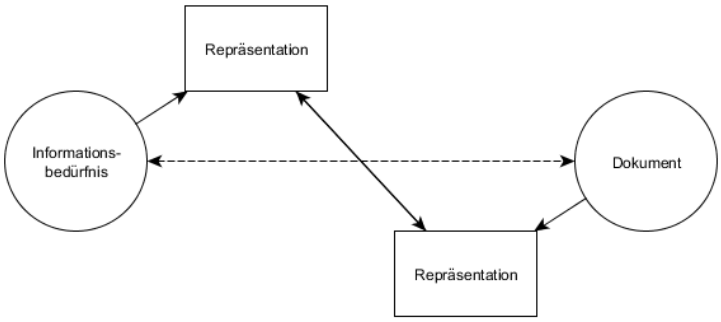


Abbildung 1: Suchvorgang über Repräsentationen

Die Repräsentation des Informationsbedürfnisses stellt die Suchanfrage dar, während die des Dokuments ein *Indexat* ist. Nach SCHULZ ist dies „der Oberbegriff für alle Benennungen, die zur Indexierung verwendet werden, z.B. Stichwörter, Schlagwörter, Deskriptoren oder Notationen“ (vgl. SCHULZ 2012). Durch Übereinstimmungen dieser beiden wird die Verbindung hergestellt. Folglich sind die Repräsentationen als Unsicherheitsfaktoren im Suchprozess zu sehen. Bei der Zuordnung gibt es zwei grundlegende Probleme:

- **Inhaltlich**

  Oft ist es schwierig, ein Informationsbedürfnis oder den Inhalt eines Dokuments in allen Facetten in einer Repräsentation wiederzugeben.

- **Formal**

  Das Informationsbedürfnis wird (zuerst) in natürlicher Sprache formuliert, wohingegen für die Artikulation des Indexats oft ein *kontrolliertes Vokabular* (z.B. ein Thesaurus) verwendet wird (um Uneinheitlichkeit zu vermeiden, die sonst durch Synonyme, Homonyme und Akronyme auftreten und zu Informationsballast bzw. –verlust führen würde).

So kann es passieren, dass sich Informationsbedürfnis und Dokument verfehlen, obwohl sie zueinander gepasst hätten.

Um den Suchvorgang möglichst effizient und effektiv zu gestalten, wurden verschiedene

Verfahren und Modelle entwickelt. Ihnen allen ist das Prinzip des Abgleichs („matching“) gemein (vgl. CHU 2010, S. 19f, 109f), wie es in *Abbildung 1* veranschaulicht wurde.

Die gängigen Verfahren, mit denen Dokumente untersucht werden, sind die Bestimmung der

- ***Term frequency*** (wie häufig ein bestimmter *Term* (eine Bezeichnung) vorkommt)
- ***Term proximity*** (der Abstand zwischen Termen)
- ***Term location*** (wo der Termen auftaucht) und
- ***Inverse document frequency*** (Relation der Dokumente mit dem Term zu allen Dokumenten in der Datenbank) (vgl. CHU 2010, S. 76).

Die wichtigsten IR-Modelle sollen jeweils kurz in ihrer Grundform vorgestellt werden:

- Im ***Booleschen Modell*** (benannt nach GEORGE BOOLE) werden die Terme der Suchanfrage mit den logischen Operatoren AND, OR und NOT verknüpft und mit den Indextermen des Dokuments abgeglichen (vgl. FERBER 2003, S. 61).
  GANTERT und HACKER erklären die Operatoren (GANTERT & HACKER 2008, S. 209f): AND kombiniert die Terme und liefert nur Dokumente, in denen alle vorkommen. Bei OR werden nur Dokumente ausgegeben, in denen einer, mehrere oder alle der verknüpften Terme vorkommen (sinnvoll für Synonyme). NOT schließt nachfolgende Terme aus und liefert nur Dokumente, in denen diese *nicht* vorkommen.

- Nach LEWANDOWSKI (vgl. LEWANDOWSKI 2005B, S. 84f) wird im ***Vektorraum-modell*** (nach GERARD SALTON) jeder Term durch eine Dimension dargestellt, die so gemeinsam den Vektorraum aufspannen. Die Suchanfrage und das Dokument werden durch Vektoren repräsentiert, die aus den Such- bzw. Indextermen bestehen. Je kleiner der Cosinus des Winkels zwischen Suchanfrage und Dokument ist, desto ähnlicher sind sich die beiden, also desto relevanter ist das Dokument.

- Nach CHU (vgl. CHU 2010, S. 106f) geht es im ***probalistischen Modell*** (nach MELVIN MARON und JOHN KUHNS) um die Wahrscheinlichkeit, mit der das Dokument für die Suchanfrage relevant ist. Diese wird anhand der Ähnlichkeit der beiden berechnet; Grundlage dabei ist die Term frequency.

Während die Zuordnung von Dokument zu Suchanfrage im Booleschen Modell lediglich binär erfolgt (stimmt (nicht) überein), finden beim Vektorraum- und probalistischen Modell differenzierte Einschätzungen statt, die ein *Ranking* der Ergebnisdokumente ermöglichen, also ein Ordnen nach Relevanz (vgl. LEWANDOWSKI 2005B, S. 89).

Diese (und weitere) Maßnahmen zum effizienteren und effektiveren Auffinden von Inhalten brachten komplexe Retrieval- oder Abfragesprachen mit sich, die Laien nicht geläufig waren. Außerdem gab es früher technische Einschränkungen, als Computer noch selten und sehr teuer waren.

Die Situation änderte sich ab Anfang der 1990er Jahre grundlegend, als das *World Wide Web*, das im Folgenden als *Web* bezeichnet und vorgestellt wird, eingeführt wurde.

2.1.2 Web Information Retrieval

In diesem Kapitel wird der Bereich des Web Information Retrieval vorgestellt.

Beim Web Information Retrieval spielt sich das Suchen und Finden von Inhalten in der Umgebung des Webs ab.

Nach LACKES ET AL. (vgl. LACKES ET AL. 2011) ist das Web ein System von *Hypertext-dokumenten* (Webseiten), die über *Hyperlinks* miteinander verknüpft sind, denen die Nutzer folgen können. Mittels eines *Browsers* können diese Dokumente aufgerufen werden, wobei die Daten über das *Hypertext-Transferprotokoll* (HTTP) übertragen werden.

Der Grundgedanke des Webs war es, wissenschaftliche Informationen für jedermann umsonst zugänglich zu machen (vgl. BERNERS-LEE 1991).

Das Erfassen und Indexieren von Webseiten wird durch Suchmaschinen vorgenommen; diese stellen für die Nutzer das Tor zu den Inhalten des Webs dar (sofern sich die Nutzer nicht bloß durch *Browsen*, also „Entlanghangeln“ an den Links (vgl. CHU 2010, S. 179f), bewegen).

Nach GANTERT und HACKER (vgl. GANTERT & HACKER 2008, S. 363ff) ist eine Suchmaschine ein System aus Soft- und Hardware, das aus drei Hauptkomponenten besteht:

- einer Software zum Sammeln von Informationen (*Crawler*)
- einer Indizierungssoftware (*Indexer*)
- einer Software zur Bearbeitung von Suchanfragen (*Searcher*)

Hinzu kommen Systeme zum Speichern der gesammelten Informationen.

Das Zusammenspiel dieser Komponenten ist in *Abbildung 2* dargestellt. Im Prinzip sammelt der Crawler Webseiten, der Indexer indexiert sie, und der Searcher liefert dem Nutzer aus diesem Index zu seiner Suchanfrage passende Ergebnisse.

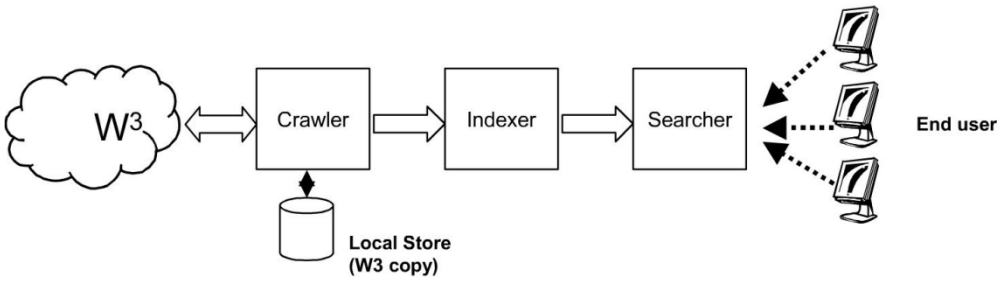


Abbildung 2: Referenzmodell einer Suchmaschine (RISVIK & MICHELSEN 2002, pdf S. 2)

Nach LEWANDOWSKI (vgl. LEWANDOWSKI 2005B, S. 48ff) besucht der **Crawler** (auch *Spider* oder *Robot* genannt) bestimmte, möglichst weit verteilte Webseiten und speichert diese („oft in konzentrierter Form“ (GANTERT UND HACKER 2008, S. 363)) in einer Datenbank ab. Diesen Vorgang wiederholt er periodisch und achtet dabei auf neue, veränderte, gelöschte und verschobene Seiten. Neue Seiten werden ggf. In die Datenbank aufgenommen, veränderte werden aktualisiert, gelöschte entfernt, und verschobene entweder entfernt oder aktualisiert.

Weiter wird nach LEWANDOWSKI (vgl. LEWANDOWSKI 2005B, S. 78) i.d.R. bereits bei diesem Sammeln versucht, unerwünschte Inhalte auszuschließen, nämlich Spam, Dubletten und inhaltsarme Seiten. Da das Ergebnis dieser Bemühungen die Qualität der Ergebnislisten erheblich beeinflusst, halten Suchmaschinenbetreiber ihre hierfür verwendeten Verfahren geheim.

Durch diese Traversierung der Linkstruktur werden große Bereiche des Webs erfasst. Jedoch gibt es Einschränkungen – bspw. fanden BRODER ET AL. heraus, dass das Web aus verschiedenen Bereichen besteht, die für Suchmaschinen unterschiedlich schwierig zu erfassen sind (vgl. BRODER ET AL. 2000). Und FERBER nennt Einzelfaktoren wie etwa Einträge in den Seiten-Metadaten, die die Erfassung verhindern, langsame Verbindungen, Passwortschutz, Domainbeschränkungen und spezifische Dateiformate, z.B. *Adobe Flash*, die zum Teil nicht indexiert werden (vgl. FERBER 2003, S. 300f).

Nach GANTERT UND HACKER (vgl. GANTERT & HACKER 2008, S. 364) strukturiert der **Indexer** die gesammelten Daten und macht sie durchsuchbar. Im Prinzip werden dabei alle Wörter, die in den Webseiten vorkommen, in einem Index abgelegt. Die meisten Suchmaschinen verwenden dabei das Vektorraummodell (vgl. DOPICHAJ 2009, S. 103) (s. Kapitel *2.1.1 Information Retrieval*).

Bei Suchanfragen wird dann auf den Index zugegriffen, nicht auf die Volltexte.

Der **Searcher** ist das vermittelnde Element zwischen Nutzer und Suchmaschinen-Index. Er wertet Suchanfragen aus, leitet sie an den Index weiter und stellt eine Kollektion aus Suchergebnissen zusammen (vgl. GANTERT & HACKER 2008, S. 364). Von besonderer Bedeutung ist dabei das Ranking. Nach CHU kommen dabei zu den den klassischen Methoden der Dokumentauswertung (Term frequency usw., s. Kapitel *2.1.1 Information Retrieval*) zwei grundlegende Arten von Verfahren hinzu, nämlich *linktopologische* und *nutzungsstatistische* (vgl. CHU 2010, S. 146ff):

- **Linktopologische Verfahren** basieren auf der Linkstruktur des Webs (s.o.). Mit ihnen wird versucht, die Linkpopularität von Seiten festzulegen. Das Prinzip dabei ist, dass eine Seite als umso relevanter eingeschätzt wird, je mehr andere Seiten auf sie verweisen (da sie dann wahrscheinlich wichtige Informationen enthält). Dabei spielt auch die Reputation der verweisenden Seiten eine Rolle (vgl. LEWANDOWSKI 2004, pdf S. 1).

  Der wahrscheinlich bekannteste Vertreter ist Googles *PageRank*-Algorithmus (benannt nach seinem Erfinder LAWRENCE PAGE), dem dieses Verfahren zugrunde liegt. (Allerdings fließen in diesen noch mehr als 200 weitere Faktoren mit ein (vgl. GOOGLE 2012B).)

  Ebenfalls bekannt ist das von KLEINBERG entwickelte Verfahren *HITS* („Hyperlink Induced Topic Search“) (KLEINBERG 1999), das klassische Methoden mit dem linktopologischen Ansatz kombiniert.
- **Nutzungsstatistische Verfahren** werten das Klickverhalten von Nutzern aus. Das Prinzip dabei ist, dass eine Seite als umso relevanter eingeschätzt wird, je mehr Nutzer sie besuchen. Damit sind es in diesem Verfahren quasi die Nutzer, die die Relevanz bestimmen.

Zwei weitere Verfahren stellen nach LEWANDOWSKI die Aktualität und der Standort des Nutzers dar (vgl. LEWANDOWSKI 2008A, pdf S. 5f):

- Bei der **Aktualität** werden neue oder kürzlich aktualisierte Dokumente höher bewertet als alte und lange unveränderte.
- Beim **Standort des Nutzers** werden Dokumente bevorzugt, die besser zur geografischen Lage des Nutzers passen. LEWANDOWSKI nennt hier das Beispiel, dass derselben Suchanfrage in derselben Sprache unterschiedliche Intentionen

zugrunde liegen können; so zielten ein Deutscher und ein Schweizer bei der Anfrage „bundesrat“ jeweils auf die Institution ihres Landes ab.

Der Rankingverfahrenkomplex stellt das Herzstück einer Suchmaschine dar und bestimmt ihre individuelle Qualität – deshalb schweigen sich Suchmaschinenbetreiber i.d.R. über diesen Bereich aus (vgl. LEWANDOWSKI 2005A, S. 10).

Über die Größe des Webs liegen keine verlässlichen Angaben vor - KUNDER gibt aktuell knapp sieben Milliarden Webseiten an, was sich aus dem Vergleich von Suchmaschinen-indizes ergibt (vgl. KUNDER 2012), jedoch findet sich bei LEWANDOWSKI in diesem Zusammenhang bereits 2009 die Zahl 20 Milliarden (vgl. LEWANDOWSKI 2008B, S. 54). Hinzu kommt außerdem der Teil des Webs, der nicht durch Suchmaschinen erfasst ist; das *invisible* oder *deep web*. Dieses besteht größtenteils aus Datenbankinhalten, die von Suchmaschinen nicht erreicht werden, da diese in Datenbankformularen keine sinnvollen Abfragen einzugeben vermögen (vgl. LEWANDOWSKI 2008B, S. 54). Das invisible web ist wahrscheinlich größer, auf jeden Fall aber genauso groß wie der erfasste Teil des Webs (vgl. LEWANDOWSKI 2008B, S. 54). Trotz dieser vagen Angaben kann davon ausgegangen werden, dass den Suchmaschinen eine erreichbare Seitenmenge von mehreren Milliarden gegenübersteht.

2.1.3 Unterschiede

In diesem Abschnitt wird kurz auf die Unterschiede zwischen dem klassischen Information Retrieval und dem Web Information Retrieval eingegangen.

LEWANDOWSKI ordnet die Unterschiede zwischen klassischem IR und Web IR in die vier Klassen „Merkmale des Dokumentenkorpus“, „Inhalte“, „Nutzer“ und „IR-System“ ein und listet sie in einer Tabelle auf, die hier leicht verkürzt als *Tabelle 1* wiedergegeben ist.

Tabelle 1: Unterschiede klassisches und Web IR (vgl. LEWANDOWSKI 2005A, S. 8)

Unterscheidungsmerkmal	Web	Klassische Datenbanken
Merkmale des Dokumentenkorpus Sprachen	Dokumente liegen in einer Vielzahl von Sprachen vor; aufgrund der Volltext-erschließung keine einheitliche Erschließung über Sprachgrenzen hinweg	Einzelne Sprache oder Dokumente liegen in vorher definierten Sprachen vor; Erschließung von Dokumenten verschiedener Sprachen mittels einer einheitlichen Indexierungssprache.
Medienarten	Dokumente in unterschiedlichen Formaten	Dokumente liegen in der Regel in nur einem Format vor.
Spam	Problem der von den Suchmaschinen unerwünschten Inhalte	Beim Aufbau der Datenbank wird vorab definiert, welche Dokumente erschlossen werden.
Hyperlink-Struktur	Dokumente sind miteinander verbunden.	Dokumente sind in der Regel nicht miteinander verknüpft; keine Notwendigkeit, aus Verlinkungsstrukturen auf die Qualität der Dokumente zu schließen.
Inhalte Datenmenge/Größe des Datenbestands	genaue Datenmenge nicht bestimmbar; keine vollständige Indexierung möglich	genaue Datenmenge aufgrund formaler Kriterien bestimmbar
Abdeckung des Datenbestands	Abdeckung der Zielmenge unklar	Abdeckung gemäß dem bei der Planung der Datenbank gesteckten Ziel in der Regel vollständig
Dubletten	Dokumente können mehrfach/vielfach vorhanden sein; teils auch in unterschied-lichen Versionen	Dublettenkontrolle bei der Erfassung der Dokumente. Versionskontrolle in der Regel nicht notwendig,da jeweils eine endgültige Fassung existiert und diese in die Datenbank eingestellt wird
Nutzer unterschiedliche Interessen Art der Anfragen	aufgrund heterogener Informationsbedürfnisse der Nutzer sehr unterschiedlich	genaue Zielgruppe mit klarem Informationsbedürfnis
Ill-formed queries	geringe Kenntnis der Nutzer über angebotene Suchfunktio-nen/Recherche allgemein	Nutzer sind mit der jeweiligen Abfragesprache vertraut
IR-System Interface	einfache, intuitiv bedienbare Interfaces für Laien-Nutzer	oft komplexe Interfaces; Einarbeitung notwendig
Ranking	Relevance Ranking aufgrund der großen Treffermengen notwendig	Relevance Ranking aufgrund genau formulierter Suchan-fragen und dadurch geringerer Treffermengen meist nicht nötig
Suchfunktionen	beschränkte Suchfunktionen	komplexe Abfragesprachen

Hinsichtlich des **Dokumentenkorpus** wird deutlich, dass sich die Dokumente im Web durch Heterogenität auszeichnen (z.B. in Format, Sprache und Nützlichkeit), während sie in der Umgebung klassischer Datenbanken weitgehend vereinheitlicht sind. Außerdem sind die Dokumente im Web miteinander verknüpft.

Bzgl. der **Inhalte** zeigt sich, dass diese im Web quantitativ unbestimmbar sind und Elemente mehrmals vorkommen können. Das macht die Erfassung schwierig. Beim klassischen IR ist die Datenmenge bestimmbar, was bei der Gestaltung der Datenbank hilft. Dubletten werden bei der Erfassung vermieden.

Die **Nutzer** zeichnen sich im Web durch Heterogenität bzgl. ihrer Informationsbedürfnisse und Recherchekenntnisse aus, während sie im klassischen IR eine eindeutigere Zielgruppe mit genauen Informationsbedürfnissen und ausgeprägten Recherchefähigkeiten darstellen.

Die **IR-Systeme** des Webs sind simpel, intuitiv und in ihren Funktionen eingeschränkt, während sie im klassischen IR komplex, stark strukturiert und mit umfangreichen Funktionen ausgestattet sind.

Zusammenfassend lässt sich sagen, dass der Heterogenität des Webs die Homogenität des klassischen IR gegenübersteht. Dies findet in den jeweiligen IR-Systemen seinen Niederschlag, was in *Tabelle 2* als zusammenfassendes Fazit dargestellt ist.

Tabelle 2: Fazit Unterschiede klassisches und Web IR

	Web	**Klassische Datenbanken**
Dokumente	heterogen	homogen
Nutzer	heterogen	homogen
IR-Systeme	simpel	komplex

2.1.4 Evaluierung von Suchmaschinen

Mittels dieses Abschnitts findet eine kleine Einführung in das Thema Evaluierung von Suchmaschinen statt.

Welche Suchmaschine gibt die besten Suchergebnisse aus? Welche bietet die nützlichsten Funktionen? Bevorzugen manche Suchmaschinen kommerzielle Treffer? Inwiefern geben Suchmaschinen unterschiedliche Ergebnisse aus? ...

Es gibt unzählige Fragen, die durch die Evaluierung von Suchmaschinen beleuchtet werden können.

Dieses Feld beschäftigt Forscher und Praktiker schon lange. So gibt es seit einigen Jahren Bestrebungen, Suchmaschinen aus einer möglichst ganzheitlichen Perspektive zu sehen, siehe z.B. SPINK 2002 und LEWANDOWSKI 2004, also bspw. besondere Funktionen oder das Nutzerverhalten mit einzubeziehen. Die gängigen Mittel der Suchmaschinenevaluierung sind immer noch bzw. waren Retrievaltests, auf die im nachfolgenden Kapitel eingegangen wird, und der Vergleich der Indexgrößen (vgl. LEWANDOWSKI 2005A, S. 10), das **„**einzige direkt vergleichbare objektive Qualitätsmerkmal**“** (LEWANDOWSKI 2004, pdf S. 3).

Bis vor einigen Jahren veröffentlichten viele Suchmaschinenbetreiber (auch zu Werbezwecken) Zahlen über den Umfang ihrer Suchmaschinenindizes. Heute werden Indexgrößen zwar geschätzt, doch sollte sowieso bedacht werden, dass Quantität nicht gleich Qualität ist und dieses Merkmal ebenfalls nichts über die Aktualität der erfassten Dokumente aussagt (vgl. LEWANDOWSKI & HÖCHSTÖTTER 2007, pdf S. 4f).

2.2 Retrievaltests

In diesem Kapitel werden der Bereich Retrievaltests mit Fokus auf Suchmaschinen vorgestellt und seine einzelnen Teilbereiche untersucht. Daran knüpfen Kritik und neuere Impulse in diesem Feld an. Den Abschluss bilden Betrachtungen bedeutender Tests.

Mit Retrievaltests wird die Retrievaleffektivität von Informationssystemen gemessen. Im Laufe der Jahre hat sich mit Blick auf Suchmaschinen aus zwei Faktoren ein diesbzgl. Standardaufbau herausgebildet. Diese Faktoren sind:

- **Die IR-Literatur**

  Hier sind besonders drei Arbeiten hervorzuheben: In „The pragmatics of information retrieval experimentation, revisited“ von TAGUE-SUTCLIFFE (TAGUE-SUTCLIFFE 1992) werden zehn Schritte zur Planung eines Retrievaltests aufgestellt, an denen sich meistens orientiert wird (vgl. LEWANDOWSKI 2011A, S. 208):

  1. Need for testing (To test or not to test?)
  2. Type of test (What kind of test?)
  3. Definition of variables (How to operationalize the variables)
  4. Database development (What database to use?)
  5. Finding queries (Where to get queries?)
  6. Retrieval software (How to process queries)
  7. Experimental design (How will treatments be assigned to experimental units?)
  8. Data collection (How to collect data?)
  9. Data analysis (How to analyze the data?)
  10. Presenting results (How to present results)

Die beiden anderen Arbeiten beschäftigen sich mit Besonderheiten von Suchmaschinen: GORDON und PATHAK entwickeln in „Finding information on the World Wide Web: the retrieval effectiveness of search engines“ (GORDON & PATHAK 1999) sieben Kriterien für einen Suchmaschinen-Retrievaltest, die von HAWKING ET AL. in „Measuring Search Engine Quality“ (HAWKING ET AL. 2001) auf fünf reduziert werden. Diese beziehen sich auf:

1. die Abbildung echter Informationsbedürfnisse durch die Suchanfragen
2. die Mitteilung des Informationsbedürfnisses, sofern nicht der Urheber als Juror fungiert
3. die ausreichende Menge von von Suchanfragen
4. die Einbeziehung der wichtigsten Suchmaschinen
5. den überlegten Versuchsaufbau und eine sorgfältige Durchführung

- **Evaluierungsinitiativen**

  Hier ist vor allem *TREC* (Text REtrieval Conference) zu nennen, eine seit 1992 bestehende Reihe groß angelegter Retrievaltests, hervorgegangen aus US-amerikanischer Initiative. Bei diesen Tests lassen Informationssystem-Betreiber ihre Produkte verschiedene Aufgaben lösen. Die Ergebnisse werden ausgewertet und verglichen, und so werden Forschungserkenntnisse gewonnen.

  Auf TREC und eine weitere Evaluierungsinitiative, *CLEF* (Conference and Labs of the Evaluation Forum), die sich mit der sprachlichen Dimension des IR befasst, wird im Kapitel *2.2.8 Bedeutende Tests* eingegangen.

So entstand der Standardaufbau für Retrievaltests. In Hinsicht auf Suchmaschinen sieht er nach LEWANDOWSKI wie folgt aus (vgl. LEWANDOWSKI 2011A, S. 206):

Eine bestimmte Anzahl Suchanfragen wird unterschiedlichen Suchmaschinen gestellt. Die Ergebnisse (i.d.R. eine bestimmte Anzahl von Trefferpositionen, da es beliebig viele Treffer geben kann) werden anonymisiert, gemischt und daraufhin Juroren zur Bewertung vorgelegt. Danach werden die Ergebnisse wieder ihren Suchmaschinen und Trefferpositionen zugeordnet. Gemessen wird i.d.R. die *Precision*, also der Anteil relevanter Ergebnisse.

Nachem der Bereich Retrievaltests in Hinsicht auf Suchmaschinen eingeführt wurde, erfolgt nun die Vorstellung der einzelnen Aspekte solcher Tests.

2.2.1 Suchmaschinen

In diesem Abschnitt wird auf die Wahl der zu testenden Suchmaschinen eingegangen.

Die Wahl der zu testenden Suchmaschinen hängt freilich vom Zweck des Tests ab. Z.B. kann eine (neue) Suchmaschine an einer Referenzsuchmaschine getestet werden, oder verschiedene Suchmaschinen untereinander.

Für einen Vergleich gängiger Suchmaschinen sollten diejenigen mit den größten Marktanteilen ausgewählt werden, wobei darauf zu achten ist, dass diese eigene Ergebnisse liefern und nicht etwa auf den Pool einer der bekannten Suchmaschinen zurückgreifen (vgl. LEWANDOWSKI 2011A, S. 208). Außerdem sollte besonders bei länderübergreifenden Tests auf die Verteilung der Marktanteile geachtet werden, die sich teils erheblich unterscheidet (vgl. LEWANDOWSKI 2011A, S. 208).

2.2.2 Retrievaleffektivität

In diesem Kapitel werden Methoden zur Messung der Retrievaleffektivität vorgestellt.

Seit einer Reihe von Jahren werden Maße zur Messung der Retrievaleffektivität aus dem allgemeinen IR-Bereich für Suchmaschinen adaptiert (vgl. LEWANDOWSKI & HÖCHSTÖTTER 2007, pdf S. 6).

Grundsätzlich können Ergebnisse relevant oder irrelevant sein. Daraus ergeben sich nach CHU (vgl. CHU 2010, S. 210) vier unterschiedliche Ergebnisarten („hits“, „noises“, „misses“ und „rejects“), die in *Tabelle 3* dargestellt sind. Dabei können die Ergebnisse danach unterteilt werden, ob sie ausgegeben wurden oder nicht.

Tabelle 3: Arten von Ergebnissen (CHU 2010, S. 210)

	Relevant	**Not Relevant**	**Total**
Retrieved	a (hits)	b (noise)	a + b (all retrieved)
Not retrieved	c (misses)	d (rejects)	c + d (all nonretrieved)
Total	a + c (all relevant)	b + d (all nonrelevant)	a + b + c + d (total in the system)

Im Optimalfall werden alle relevanten und keine irrelevanten Ergebnisse ausgegeben.

Bzgl. des Begriffs Relevanz sei angemerkt, dass es schon seit etwa 1950 Diskussionen über eine Definition gibt, bis heute aber kein Konsens gefunden wurde - „Relevance, like love, is, in a sense, in the eye of the beholder. It is not a concept that can be explicitly described and easily quantified“ (CHU 2010, S. 211). Ein Knackpunkt ist die Abgrenzung von objektiver und subjektiver Relevanz (Pertinenz). Doch so schwierig es auch sein mag, ein akkurates und objektives Urteil über Relevanz zu fällen, sind Einschätzungen trotzdem möglich und bis zu einem gewissen Grad auch aussagekräftig und vergleichbar (vgl. CHU 2010, S. 211).

Für eine Übersicht der Debatte siehe z.B. MIZZARO 1997 und CHU 2010.

Die beiden klassischen Retrievalmaße sind die bereits genannte **Precision** und der ***Recall***.

Die **Precision** ist der Anteil der relevanten ausgegeben Treffer an allen ausgegebenen Treffern (vgl. FERBER 2003, S. 86f). Auf *Tabelle 3* bezogen lautet die Formel also:

$$\frac{a}{a+b}$$

Das Ergebnis liegt im Bereich 0 (schlechtester Fall: kein ausgegebener Treffer ist relevant) bis 1 (bester Fall: alle ausgegebenen Treffer sind relevant).

Die Precision ist das bedeutendste und am häufigsten verwendete Maß in Retrievaltests (vgl. LEWANDOWSKI 2011A, S. 213).

Der **Recall** ist der Anteil der relevanten ausgegebenen Treffer an allen relevanten Treffern (vgl. FERBER 2003, S. 86f). Auf *Tabelle 3* bezogen lautet die Formel also:

$$\frac{a}{a+c}$$

Das Ergebnis liegt im Bereich 0 (schlechtester Fall: keiner der relevanten Treffer wurde ausgegeben) bis 1 (bester Fall: alle relevanten Treffer wurden ausgegeben).

Im Webkontext ist die Messung des Recalls unmöglich, da die Menge aller relevanten Dokumente nicht bestimmt werden kann (vgl. LEWANDOWSKI 2011A, S. 213). Eine Variation ist der ***Relative recall***, bei dem die Gesamtzahl der relevanten Dokumente geschätzt wird – allerdings wird die eben genannte Problematik dabei lediglich verlagert (vgl. CHU 2010, S. 215). Eine Alternative zum Recall stellt das ***Pooling*** dar, wobei „die Gesamtzahl *der von allen getesteten Systemen* zu einer Suchanfrage ausgegebenen relevanten Dokumente als Basis genommen werden [sic]“ (LEWANDOWSKI 2011A, S. 213).

Die Beziehung zwischen Precision und Recall tendiert dazu, umgekehrt proportional zu sein (vgl. FERBER 2003, S. 87), jedoch besteht keine mathematische Abhängigkeit (vgl. LEWANDOWSKI 2012, pdf S. 13), und z.B. FUGMANN zeigt, dass eine Verbesserung der Precision nicht unbedingt zu einer Verschlechterung des Recalls führt und umgekehrt (vgl. FUGMANN 1993, S. 203ff). Trotzdem gilt die umgekehrt proportionale Beziehung als Faustregel.

Von diesen klassischen Precision- und Recall-Formen gibt es moderne Varianten. So definiert etwa SU (SU 1998, S. 557f) die ***Importance of precision of the search to user*** und die ***Importance of completeness of search***, indem sie eine mögliche Gewichtung von Precision vs. Recall mit einbezieht.

Für weitere Variationen siehe z.B. CHU 2010.

Es ist zu beachten, dass mittels Precision und Recall nur die Retrievaleffektivität in Hinsicht auf *informationsorientierte* Anfragen, bei denen also mehrere Ergebnisse gewünscht und auch vorhanden sind (vgl. BRODER 2002), gemessen werden kann (LEWANDOWSKI 2011A, S. 214). Dies ist der Standardfall in Suchmaschinen-Retrievaltests. Für andere Arten von Anfragen (s. *Kapitel 2.2.3 Suchanfragen*) sind andere Kennzahlen bzw. Messmethoden zu wählen.

LEWANDOWSKI gibt zu bedenken, dass sowohl die klassischen Precision- und Recall-Formen als auch ihre modernen Varianten die Treffer auf Dokumentenebene betrachten, also als voneinander unabhängig - „So könnte eine Suchmaschine, die zehnmal hintereinander dasselbe relevante Dokument ausgibt, eine perfekte Precision erreichen, auch wenn die Ergebnismenge als Ganzes für einen Nutzer natürlich nicht relevant wäre“ (LEWANDOWSKI 2011A, S. 214).

Der ***Fallout*** stellt die Umkehrung des Recalls dar; er ist „der Anteil der ausgegebenen, aber nicht relevanten Treffer an der Gesamtzahl der nicht relevanten Treffer im Datenbestand“ (LEWANDOWSKI & HÖCHSTÖTTER 2007, pdf S. 6). Auf *Tabelle 3* bezogen lautet die Formel also:

$$\frac{b}{b+d}$$

Das Ergebnis liegt wieder im Bereich 0 (bester Fall: keiner der irrelevanten Treffer wurde ausgegeben) bis 1 (schlechtester Fall: alle irrelevanten Treffer wurden ausgegeben).

Im Webkontext besteht dieselbe grundlegende Problematik wie beim Recall; die Menge aller irrelevanten Dokumente im Web ist unbestimmbar, sodass sich hier ebenfalls auf andere Weise beholfen werden muss.

Zwei weitere, speziell für Suchmaschinen interessante Kennzahlen sind ***Salience*** nach DING und MARCHIONINI (vgl. DING & MARCHIONINI 1996), und ***Ability to retrieve top ranked pages*** nach VAUGHAN (vgl. VAUGHAN 2004). Diese stellen das Abschneiden der anderen am Test beteiligten Suchmaschinen in Relation zum erreichten Ergebnis einer Suchmaschine. Hintergrund ist die Annahme, dass die Schwierigkeit, relevante Ergebnisse auszugeben, maßgeblich von den Suchanfragen abhängt und daher stark variieren kann (vgl. LEWANDOWSKI 2011A, S. 215).

Es gibt eine Fülle weiterer Retrievalmaße (KORFHAGE 1997 bietet einen Überblick, in DEMARTINI & MIZZARO 2006 findet sich eine Vollständigkeit anstrebende Auflistung, und in LEWANDOWSKI 2012 gibt es eine Auswahl weiterer in Bezug auf Suchmaschinen interessanter (s. *Kapitel 2.2.7.3 Rahmenwerk zur Messung von Suchmaschinen-Retrievaleffektivität nach LEWANDOWSKI)*, auf die an dieser Stelle jedoch nicht weiter eingegangen werden soll.

Es bleibt festzuhalten, dass sich im Web-Bereich noch kein allgemeines Maß durchgesetzt hat und dass besondere Maße benötigt werden, um den spezifischen Anforderungen dieses Umfelds gerecht zu werden (vgl. LEWANDOWSKI 2011A, S. 215).

2.2.3 Suchanfragen

In diesem Abschnitt wird auf den Umgang mit unterschiedlichen Arten von Suchanfragen eingegangen.

BRODER (vgl. BRODER 2002) teilt Suchanfragen in drei Typen ein; die bereits erwähnten „informationsorientierten“, die „navigationsorientierten“ und die „transaktionsorientierten“.

Informationsorientierte Anfragen sind dem klassischen IR am nächsten; mit ihnen werden vielfältige Informationen zu bestimmten Themen gesucht. Es kann also nicht davon ausgegangen werden, dass ein derartiges Informationsbedürfnis mit dem Betrachten lediglich eines Ergebnisses befriedigt ist.

Navigationsorientierte Anfragen dagegen zielen auf eine bestimmte Webseite ab, z.B. „gmx“ auf http://www.gmx.net/. Hier ist also ein Treffer (der richtige) ausreichend.

Transaktionorientierte Anfragen verfolgen nach dem Aufrufen einer Webseite eine weitere Aktion, z.B. die Suche in einer Datenbank, den Kauf eines Produkts oder den Download einer Datei.

Aufgrund der spezifischen Informationsbedürfnisse und den damit verbundenen unterschiedlichen Anforderungen an Ergebnisse sollten diese Anfragetypen in Retrievaltests nicht gemischt werden (vgl. LEWANDOWSKI 2012, pdf S. 3).

Es gibt einige Faktoren bzgl. Suchanfragen, die direkt vom Zweck des Tests abhängen, z.B. Anzahl, Thema, Struktur und Herkunft.

Allgemein hat sich ein Minimum von 50 Anfragen etabliert – soll ein breites Themenspektrum oder verschiedenartiges Suchverhalten abgedeckt werden, ist es ratsam, die Anzahl entsprechend zu erhöhen (vgl. LEWANDOWSKI 2011A, S. 208).

Soll eine allgemeine Aussage über die Trefferqualität getroffen werden, sind die Suchanfragen möglichst breit zu wählen (bei bestimmten zu untersuchenden Themenfeldern freilich dementsprechend) (vgl. LEWANDOWSKI 2011A, S. 209).

Die Anfragen können z.B. aus Logfiles, Jahresendveröffentlichungen (z.B. Google: www.googlezeitgeist.com/de) oder Tools wie *GoogleAdWords*, die solche Daten bereitstellen, gewählt werden (vgl. LEWANDOWSKI 2011A, S. 209). Es ist allerdings zu beachten, dass diese Methoden keine zugehörigen Kontextinformationen liefern – die aber für die Juroren abhängig vom Zweck des Tests in unterschiedlichem Maße wichtig zum Bewerten sind. Besser ist es daher, Suchanfragen samt Kontextinformationen von echten Nutzern abzufragen (vgl. LEWANDOWSKI 2011A, S. 209f), die im optimalen Fall auch als Juroren dienen und die zugehörigen Ergebnisse bewerten (vgl. LEWANDOWSKI 2012, pdf S. 11f). Bei der Auswahl sollte auch die Länge und ggf. die Verteilung nach Wortanzahl berücksichtigt werden (vgl. LEWANDOWSKI 2011A, S. 209) - die diesbzgl. Durchschnittswerte können z.B. HÖCHSTÖTTER & KOCH 2008 entnommen werden.

2.2.4 Bewertung der Ergebnisse

In diesem Kapitel wird auf die Bewertung der Suchergebnisse eingegangen.

Gemäß dem Retrievaltest-Standardaufbau werden die Ergebnisse den Juroren anonymisiert und gemischt vorgelegt. Anonymisierung meint hier die Verschleierung der Herkunft, also der zugehörigen Suchmaschine. Findet diese nicht statt, sind Markeneffekte zu beobachten, die teils erhebliche Auswirkungen auf das Urteil der Juroren haben (vgl. JANSEN & ZHANG & ZHANG 2007, S. 2474ff). Die Durchmischung verhindert Lerneffekte beim Bewerten (vgl. LEWANDOWSKI 2011A, S. 212).

Dubletten sollten den Juroren nur einmal (pro Suchaufgabe) vorgelegt werden, um einheitliche Urteile zu erhalten (vgl. LEWANDOWSKI 2011A, S. 212).

Die Bewertung kann grundsätzlich binär (relevant ja/nein) oder mittels einer differenzierenderen Skala erfolgen. In der Praxis hat sich gezeigt, dass bei der binären

Variante kaum Diskriminierung stattfindet – oft werden auch Ergebnisse von relativ geringer Relevanz von den Juroren als relevant bewertet (vgl. LEWANDOWSKI 2011A, S. 212, 219). Dagegen treten bei Einschätzungen mittels einer differenzierenderen Skala größere Unterschiede zutage. Es hat sich bewährt, die Juroren die Ergebnisse sowohl binär, als auch mittels einer Fünfer-Skala bewerten zu lassen (vgl. LEWANDOWSKI 2011A, S. 212).

I.d.R. wird jeder Treffer nur von einer Person bewertet, jedoch ist es sinnvoll, die Bewertung von zwei Personen vornehmen zu lassen – ob weitere Bewertungen von wieder anderen Juroren die Zuverlässigkeit von Tests signifikant erhöhen, wurde bislang nicht untersucht (vgl. LEWANDOWSKI 2011A, S. 211); Erkenntnisse etwa in Bezug auf die *Interrater-Reliabilität*, also dem Übereinstimmungsausmaß von Einschätzungen unterschiedlicher Personen (vgl. PSYCHOLOGY48 2010), sind noch wenig vorhanden (s. z.B. SCHAER & MAYR & MUTSCHKE 2010).

2.2.5 Juroren

In diesem Abschnitt wird auf die Juroren eingegangen, die die Ergebnisse bewerten.

Die Anzahl der Juroren richtet sich freilich nach dem Umfang des Tests. Sofern möglich, ist es üblich, dass ein Juror alle Ergebnisse der beteiligten Suchmaschinen zu einer Anfrage bewertet (vgl. LEWANDOWSKI 2011A, S. 211).

Wenn der Retrievaltest ein bestimmtes Thema behandelt, sollten freilich Juroren gewählt werden, die sich in diesem Bereich auskennen. Bei allgemeinen Tests können Laien als Juroren dienen – oft werden Studierende ausgewählt, da sie einfach zu rekrutieren sind (vgl. LEWANDOWSKI 2012, pdf S. 11). Wegen der Verwendung von Studenten als Juroren werden Tests oft kritisiert (vgl. LEWANDOWSKI 2011A, S. 211).

2.2.6 Kritik

In Hinblick auf den Standardaufbau von Suchmaschinen-Retrievaltests lassen sich einige Kritikpunkte identifizieren, von denen hier eine Auswahl genannt werden soll:

1. Es wird davon ausgegangen, dass der Nutzer alle (im Test einbezogenen) Ergebnisse nacheinander durchgeht. Untersuchungen haben jedoch gezeigt, dass Nutzer großes Interesse lediglich für die ersten Treffer zeigen (die weit oben und somit sichtbar sind, ohne dass gescrollt oder gar geblättert werden muss), deren Reihenfolge ebenfalls einen erheblichen Einfluss auf das Klickverhalten hat (s. z.B. KEANE & O'BRIEN & SMYTH 2008).

2. Trefferbeschreibungen werden im Test nicht berücksichtigt. Diese beeinflussen das Klickverhalten jedoch maßgeblich (s. z.B. LEWANDOWSKI 2008C, pdf S. 4ff).
3. Die Vielfalt der Ergebnisse findet keinen Eingang in die Tests. Da nur die ersten Treffer breite Beachtung finden, wäre es wichtig zu untersuchen, ob durch diese verschiedene Aspekte des Themas abgebildet werden (vgl. LEWANDOWSKI 2011A, S. 220).
4. Auf der Suchergebnisseite finden sich neben den regulären Treffern oft weitere Elemente wie Textanzeigen (Werbung), Treffer aus weiteren Kollektionen (Bilder, Videos usw.) und Shortcuts (Hinweise auf Treffer anderer Kollektionen), die nicht berücksichtigt werden (vgl. LEWANDOWSKI 2011A, S. 217f, 220).
5. Der tatsächliche Rechercheprozess, der oft interaktiv ist und aus mehreren Schritten besteht (z.B. durch eine Reformulierung der Suchanfrage oder die Nutzung von Filtern) und mitnichten dem simplen Anfrage-Ergebnis-Paradigma entspricht (s. z.B. MARCHIONINI 2006), wird nicht einmal annähernd abgebildet (vgl. LEWANDOWSKI 2011A, S. 216, 220).
6. Da in den Tests i.d.R. lediglich informationsorientierte Anfragen behandelt werden, lässt sich aus ihnen keine allgemeine Aussage über die Qualität der Ergebnisse treffen (vgl. LEWANDOWSKI 2011A, 216f).
7. Es wird davon ausgegangen, dass der Nutzer gleichermaßen an einer hohen Precision und einem hohen Recall interessiert ist - dabei sind Nutzer eher auf der Suche nach *ein paar sehr relevanten* Dokumenten (Precision), als nach *allen relevanten* Dokumenten (Recall) (vgl. LEWANDOWSKI 2008A, pdf S. 4).
8. Oft werden die Ergebnisse lediglich binär bewertet, was weit weniger aussagekräftig ist als mittels einer differenzierenderen Skala (vgl. LEWANDOWSKI 2011A, 219) (s. Kapitel *2.2.4 Bewertung der Ergebnisse*).

2.2.7 Neuere Impulse

In diesem Kapitel werden einige neuere Entwicklungen im Bereich Suchmaschinen-Retrivaltests vorgestellt.

2.2.7.1 Einbeziehung von Trefferbeschreibungen

Dieser Abschnitt befasst sich mit der Einbeziehung von Trefferbeschreibungen in Retrievaltests.

Beim Standardaufbau von Retrievaltests wird davon ausgegangen, dass der Nutzer bei einer Suche alle Treffer einer bestimmten Anzahl der ersten Positionen durchgeht. Tatsächlich pickt er sich aber einige heraus, die ihm relevant erscheinen, und ignoriert andere. Diese Entscheidung trifft er u.a. aufgrund von Trefferbeschreibungen, die Metadaten enthalten. *Abbildung 3* zeigt eine solche Trefferbeschreibung exemplarisch.

Abbildung 3: Trefferbeschreibung (Google)

Bei bestimmten Ergebnistypen (s. *Kapitel 2.3.1 Hintergrund und Prinzip*) werden zusätzliche Informationen ausgegeben, wie bspw. ein Bild oder navigatorische Links.

Es ist für den Suchprozess also entscheidend, dass die Trefferbeschreibungen den Inhalt der zugehörigen Dokumente adäquat wiedergeben, und dabei prägnant sind, sodass der Nutzer möglichst auf einen Blick sieht, worum es sich handelt. Dies ist das Spannungsfeld, in dem Trefferbeschreibungen realisiert werden müssen. So lauten auch nach CRYSTAL und GREENBERG die beiden entscheidenden Fragen (vgl. CRYSTAL & GREENBERG 2006, pdf S. 5f):

- *Welcher Umfang* der Metadaten ist sinnvoll?
- *Was für* Metadaten sind besonders hilfreich?

Beim Umfang der Trefferbeschreibungen hat sich bei den gängigen Suchmaschinen in Deutschland Google, Bing, *Yahoo*, *T-Online AOL* der maximale Standard von zwei vollen Zeilen mit bis zu ca. 25 Wörtern etabliert (bei Standard-Beschreibungen).

Bzgl. der zweiten von CRYSTAL und GREENBERG gestellten Frage nennt LEWANDOWSKI drei Herangehensweisen, die die Komposition von Trefferbeschreibungen betreffen (vgl. LEWANDOWSKI 2011A, S. 5):

- ***KWIC*** (KeyWord In Context)

  Bei dieser Methode werden Schlüsselwörter des Dokuments identifiziert und innerhalb ihres Kontexts (jeweils ein paar Wörter) in der Trefferbeschreibung

angezeigt. Diese Methode wird am meisten von Suchmaschinen genutzt (auch von Google in *Abb. 3*). Vorteilhaft ist, dass sie immer angewendet werden kann (da ja jedes Dokument Inhalt hat) - nachteilig ist, dass als Ergebnis oft unvollständige Sätze herauskommen, mit denen der Nutzer evtl. nur wenig anfangen kann.

- **Daten aus externen Verzeichnissen**

  Hierbei werden Daten aus externen Web-Verzeichnissen wie dem *Open Directory Project* (ODP) verwendet. Im optimalen Fall entsteht so eine prägnante Trefferbeschreibung, doch Nachteile sind, dass Webverzeichnisse nur Daten über komplette Webseiten und keine Einzelseiten führen, und außerdem nicht vollständig sind. Diese Methode ist also nur für populäre Webseiten geeignet.

- **Metadaten der Webseite selbst**

  In diesem Fall werden Trefferbeschreibungen aus den seiteneigenen Metadaten generiert. So können nützliche Beschreibungen entstehen, doch sind solche Daten nicht immer vorhanden.

Genau wie die Treffer selbst können auch die zugehörigen Beschreibungen entweder relevant oder irrelevant sein. Hieraus ergeben sich folgende Kombinationsmöglichkeiten:

- Relevante Beschreibung – relevanter Treffer: Der Nutzer wird den Treffer anklicken und ein relevantes Dokument erhalten. Die Trefferbeschreibung hat ihren Zweck erfüllt.
- Relevante Beschreibung – irrelevanter Treffer: Der Nutzer wird den Treffer anklicken und ein irreleventes Dokument erhalten. Die Trefferbeschreibung hat ihren Zweck verfehlt.
- Irrelevante Beschreibung – relevanter Treffer: Der Nutzer wird den Treffer nicht anklicken und verpasst somit ein relevantes Dokument. Die Trefferbeschreibung hat ihren Zweck verfehlt.
- Irrelevante Beschreibung – irrelevanter Treffer: Der Nutzer wird den Treffer nicht anklicken und überspringt so ein irrelevantes Dokument. Die Trefferbeschreibung hat ihren Zweck erfüllt.

Es wird deutlich, dass Trefferbeschreibungen in demselben Maße, wie sie helfen, auch schaden können. Daher ist auf eine geeignete Form und eine sorgfältig durchdachte Generierungsmethode zu achten.

Aufgrund ihrer entscheidenden Rolle, die Trefferbeschreibungen bei der Selektion von Ergebnissen durch den Nutzer spielen, sollten sie in Retrievaltests einbezogen werden.

Tatsächlich gibt es ein paar Tests, die dies tun - mehr dazu im Kapitel *2.2.8 Bedeutende Tests*.

2.2.7.2 Ganzheitliche Betrachtung

In diesem Kapitel wird der Ansatz vorgestellt, das System Suchmaschine in Retrievaltests ganzheitlich zu betrachten und nicht auf die bloße Retrievaleffektivität zu reduzieren.

Durch den Standardaufbau von Retrievaltests entsteht der Eindruck, dass die Qualität einer Suchmaschine lediglich auf ihrer Retrievaleffektivität basiert. Vernachlässigt wird dabei jedoch der Nutzer, der eine Suchmaschine nur dann fortwährend nutzt, wenn er zufrieden ist und so den Erfolg der Suchmaschine definiert.

Deshalb plädieren LEWANDOWSKI und HÖCHSTÖTTER für eine „integrierte Betrachtungsweise von Technik und Nutzer“ (LEWANDOWSKI & HÖCHSTÖTTER 2007, pdf S. 2), indem sie die vier Evaluationsbereiche „Qualität des Index“, „Qualität der Suchresultate“, „Qualität der Suchfunktionen“ und „Nutzerfreundlichkeit von Suchmaschinen (Usability)“ definieren (vgl. LEWANDOWSKI & HÖCHSTÖTTER 2007, pdf S. 3-9):

- **Qualität des Index**

  Der Index hat entscheidenden Einfluss auf die Qualität einer Suchmaschine. Was nicht im Index ist, kann auch nicht ausgegeben werden - daher ist es für Suchmaschinenbetreiber wichtig, eine möglichst hohe Abdeckung des Webs zu bieten (von unerwünschten Inhalten wie Spam abgesehen). Die ist jedoch schwierig zu messen, da die Größe des Webs unbekannt ist (s. Kapitel *2.1.2 Web Information Retrieval*). Oft werden Indizes verschiedener Suchmaschinen miteinander verglichen und über das Maß der Überschneidung die Zahl der insgesamt indizierten Seiten bestimmt. Dieser kann dann die eigene Indexgröße gegenübergestellt werden.

  Studien (z.B. VAUGHAN & THELWALL 2004) untersuchen mit unterschiedlichen Ergebnissen die Abdeckung in verschiedenen Ländern, denn natürlich ist es besonders wünschenswert, dass das „nationale“ Web erfasst ist.

  Ein weiterer wichtiger Faktor des Index‘ ist die Aktualität; häufig wird nach aktuellen Inhalten gesucht, und auch beim Ranking spielt sie eine Rolle (s. Kapitel *2.1.2 Web Information Retrieval*).

- **Qualität der Suchresultate**

Dies ist der Bereich, auf den sich der Standardaufbau von Suchmaschinen-Retrievaltests mit der Relevanzbewertung von Suchergebnissen bezieht (s. Kapitel *2.2 Retrievaltests* und *2.2.2 Retrievaleffektivität*).

Ein weiterer Faktor ist hier die Einzigartigkeit der Ergebnisse; basieren Suchmaschinen auf demselben Index, können sie aufgrund ihrer individuellen Rankingverfahren trotzdem unterschiedliche Ergebnisse zur selben Suchanfrage ausgeben. Studien zeigen, dass sich dabei (im Gegensatz zu früher) die Sets der ersten zehn Ergebnisse erheblich unterscheiden (s. z.B. SPINK ET AL. 2006, pdf S. 11).

- **Qualität der Suchfunktionen**

Spezielle Suchfunktionen, wie etwa Boolesche Operatoren (s. Kapitel *2.1.1 Information Retrieval*), Feldbeschränkungen (z.B. auf den Titel von Dokumenten), Einschränkung der Dokumentherkunft (z.B. über die Sprache), zeitliche Eingrenzung, Bestimmung des Dokumenttyps usw. können Suchergebnisse erheblich verbessern. Ein Problem ist jedoch, dass die Wirkung solcher Funktionen oft unterschätzt wird, oder manche nicht richtig funktionieren. Bei einzelnen Funktionen besteht noch Evaluierungsbedarf. Eine Übersicht angebotener Funktionen einiger Suchmaschinen bietet NOTESS 2007.

- **Nutzerfreundlichkeit von Suchmaschinen (Usability)**

Dieser Bereich beschäftigt sich mit der Frage, ob Suchmaschinen ihre Nutzer effizient und effektiv vorgehen lassen.

Das Interface von Suchmaschinen sieht i.d.R. ähnlich aus - ein Gegensatzpaar bilden jedoch z.B. Google, deren Design sich durch Schlichtheit auszeichnet, und Yahoo, das als Portal auftritt und eine vollgepackte Startseite präsentiert.

Anfragen mit Füllwörtern (ohne Phrasensuchfunktion) oder selten und falsch benutzte Operatoren zeigen, dass die Akzeptanz von speziellen Suchfunktionen bei Nutzern generell sehr niedrig ist, zumal diese i.d.R. auch ohne diese Funktionen finden, was sie suchen.

Hinsichtlich der Ergebnisse beklagen viele Nutzer, dass Werbung oft nicht ausreichend kenntlich gemacht ist. Außerdem werden Seiten, die nur auf ein besseres Ranking hin optimiert wurden, bemängelt.

2.2.7.3 Rahmenwerk zur Messung von Suchmaschinen-Retrievaleffektivität nach LEWANDOWSKI

In diesem Abschnitt wird ein neues, umfassendes Modell für Retrievaltests vorgestellt.

In seiner Arbeit „A Framework for Evaluating the Retrieval Effectiveness of Search Engines“ definiert LEWANDOWSKI ein komplettes Gerüst für einen Suchmaschinen-Retrievaltest, das auch einige Kritikpunkte am Standardaufbau von Suchmaschinen-Retrievaltests (s. Kapitel *2.2.6 Kritik*) aufgreift und Lösungen bietet.

Das Rahmenwerk besteht aus den fünf Bereichen „Query selection“, „Results collection“, „Results presentation“, „Results judgement“ und „Data analysis“. Es ist in *Abbildung 4* dargestellt.

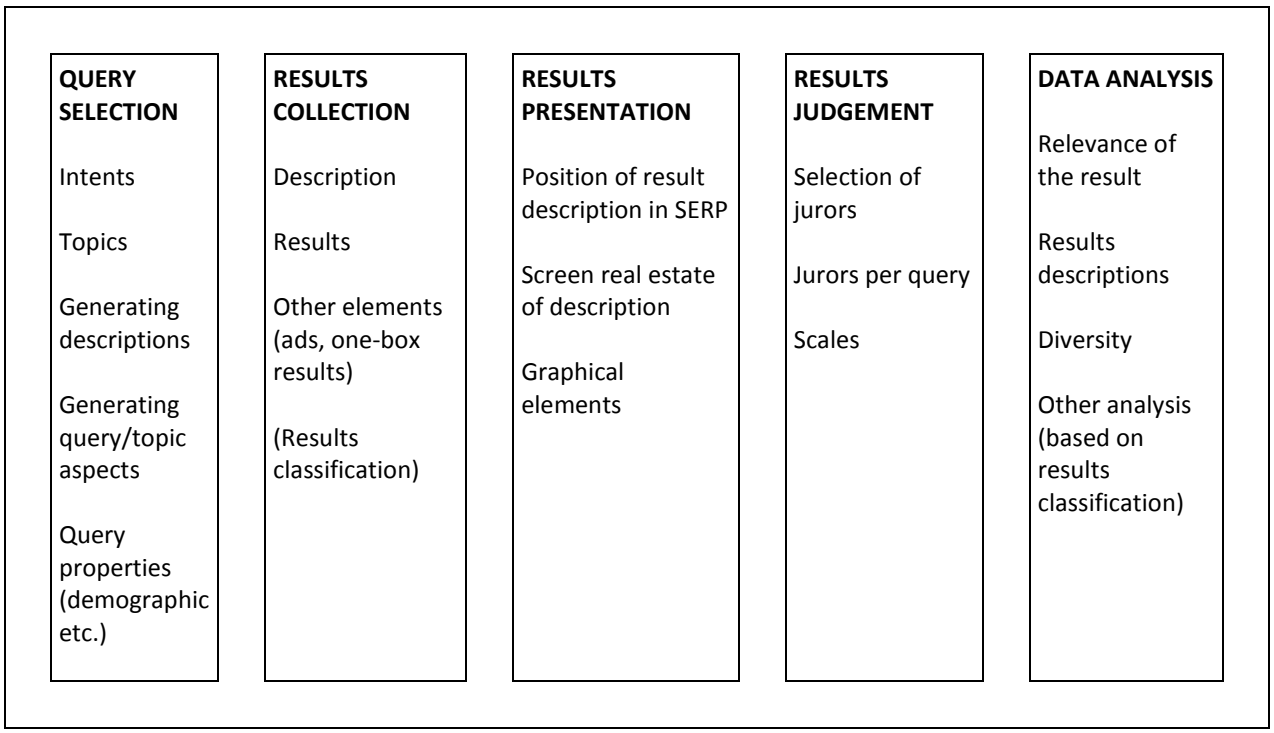


Abbildung 4: Rahmenwerk LEWANDOWSKI (LEWANDOWSKI 2012, pdf S. 8)

Im Folgenden werden die einzelnen Bereiche kurz nach LEWANDOWSKI vorgestellt (vgl. LEWANDOWSKI 2012, pdf S. 8-14):

Query selection

Dieser erste Bereich beschäftigt sich mit der Frage, welche Art von Suchanfragen für den Test ausgewählt werden sollte.

- **Query intent**

  Aufgrund ihrer unterschiedlichen Ansprüche sollte sich auf einen der drei Anfragetypen nach BRODER (s. Kapitel *2.2.3 Suchanfragen*) konzentriert werden. Sollen verschiedene Anfragetypen in einem Test behandelt werden, müssen die Anfrage-Sets individuell sein, und bei der Auswertung (Bereich „Data analysis“) sind verschiedene Kennzahlen zu verwenden (s. Kapitel *2.2.6 Kritik*, Punkt 6).

- **Query topics**

  Die Suchanfragen sollten verschiedene Themen durch echte Informationsbedürfnisse von Nutzern behandeln.

- **Generating descriptions**

  Es ist notwendig, dass Beschreibungen der Informationsbedürfnisse zu den Anfragen generiert werden. Am besten tut dies die Person, von der die jeweilige Anfrage stammt.

  Wurden Anfragen aus Logfiles o.Ä. gewonnen, sollten mehrere Personen ihre Vermutungen zu den zugehörigen Informationsbedürfnissen äußern, und die Beschreibungen aus den Übereinstimmungen formulieren.

- **Generating query aspects**

  Anfragen haben oft unterschiedliche Aspekte; z.B. kann sich „James Bond“ auf Filme, Bücher, Charakterbiografien, Schauspieler usw. beziehen. Diese Aspekte sollten generiert werden, was auf ähnliche Weise wie bei den Beschreibungen der Informationsbedürfnisse geschehen kann (s.o.), und dann sollte bei der Auswertung darauf geachtet werden, ob die Suchergebnisse wenigstens die Hauptaspekte der Anfragen abdecken (s. Kapitel *2.2.6 Kritik*, Punkt 3).

- **Query properties**

  Es gibt weit mehr Anfrageeigenschaften als bloß Informationsbedürfnis und Aspekte. CALDERON-BENAVIDES, GONZALES-CARO und BAEZA-YATES etwa definieren neun Facetten: „genre“, „topic“, „task“, „objective“, „specificity“, „scope“, „authority sensitivity“, „spatial sensitivity“ und „time sensitivity“ (CALDERON-BENAVIDES & GONZALES-CARO & BAEZA-YATES 2010, pdf S. 2). Sogar demografische Daten können für Anfragen beschafft werden (etwa durch Tools wie Microsofts *Demographics Prediction*).

  Alle diese Eigenschaften können in einem kombinierten Schritt gewonnen werden. (S. Kapitel *2.2.6 Kritik*, Punkt 3)

Results collection

Die zweite Säule beschäftigt sich mit Informationen, die zu den Trefferlinks ausgegeben werden.

- **Results descriptions**

  Trefferbeschreibungen sind von großer Wichtigkeit, da sie dem Nutzer bei der Entscheidung helfen, einen Treffer (nicht) anzuklicken – sie sollten daher einbezogen werden (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen* und Kapitel *2.2.6 Kritik*, Punkt 2).

- **Capturing organic results**

  In welcher Form die *organischen*, also die regulären, Ergebnisse gespeichert werden sollten, hängt von der Form des Tests ab. Handelt es sich um eine Testumgebung unter der Kontrolle der Testleitung, mag es genügen, die Links der Treffer zu speichern. Andernfalls sollten die Ergebnisse selbst und ihre Positionen gespeichert werden. Werden unterschiedliche Suchmaschinen untersucht, sollte außerdem der Name der Suchmaschine, die ein individuelles Ergebnis ausgegeben hat, festgehalten werden.

- **Capturing other elements of the results page**

  Wenn Ergebnisse nicht in derselben Art und Weise präsentiert werden, sollte der Typ des Ergebnisses festgehalten (regulärer Treffer, Werbung etc., s. Kapitel *2.3.1 Hintergrund und Prinzip*) werden (da verschiedene Ergebnistypen von Nutzern unterschiedlich wahrgenommen werden können) (s. Kapitel *2.2.6 Kritik*, Punkt 4).

- **Results classifications**

  Auch Suchergebnisse können weiter klassifiziert werden, etwa als von Blogs, Nachrichtenseiten, Regierungsbehördenseiten usw. stammend. So kann die Vielfalt untersucht werden. (S. Kapitel *2.2.6 Kritik*, Punkt 3)

Results weighting

Suchergebnisse werden von Nutzern nicht gleichwertig wahrgenommen. In diesem Bereich geht es daher um das Gewichten von Treffern nach ihrer optischen Aufmachung, z.B. ihre Position und ihr Design.

- **Position of the result descriptions within the *SERP*** (Search Engine Result Page)

  In traditionellen Suchmaschinen-Retrievaltests wird zwar die Position der Ergebnisse innerhalb der Liste berücksichtigt, doch reicht das nicht mehr aus, wenn die Ergebnispräsentation eine Liste mit mehr als einer Spalte oder überhaupt nicht listenbasiert ist. Es sollte daher zusätzlich überhaupt die Position innerhalb der Suchergebnisseite untersucht werden.

- ***Screen real estate***

  Ursprünglich steht der Term „Screen real estate“ für das Verhältnis von durch Inhalt belegtem Bereich zum leeren Teil einer Webseite (vgl. NIELSEN & TAHIR 2002, z.B. S. 57). Nach NICHOLSON ET AL. meint er das Verhältnis des zur Verfügung gestellten Raums auf der Suchergebnisseite zwischen bezahlten und regulären Ergebnissen (vgl. NICHOLSON ET AL. 2006). Es lohnt sich, den Screen real estate von Treffern unterschiedlichen Typs zu messen. Auch in einer

Listenform kann das der Fall sein, sofern die Treffer nicht in derselben Art und Weise präsentiert werden.

- **Graphic elements**

  Suchergebnisse können durch Farben oder Bilder hervorgehoben werden, was eine erhöhte Klickwahrscheinlichkeit darstellt (s. Kapitel *2.3.2 Usability*). Wegen dieses Einflusses auf das Nutzerverhalten sollten sie noch stärker als nur durch den Screen real estate gewichtet werden.

Results judgment

Die vierte Säule beschäftigt sich mit den Fragen, wer die Bewertung anhand welcher Skalen vornehmen sollte und wie Klickverhalten genutzt werden kann.

- **Selection of jurors**

  Juroren sollten immer nach dem Thema und der Zielgruppe der Suchmaschine ausgewählt werden. Ein Ansatz ist, Nutzer der Suchmaschine als Juroren einzusetzen.

  Juroren sollten stets sorgfältig ausgewählt werden.

- **Jurors per query**

  Idealerweise sollten mehrere Juroren eine Suchanfrage bearbeiten. Bewertungsunterschiede bei Ergebnissen allgemeiner Tests sind gering, daher ist hierbei auch eine geringe Anzahl Juroren möglich. Bei speziellen Tests jedoch sollten es tatsächlich mehrere sein.

  Wenn die Suchmaschine mehrere Zielgruppen hat, sollten Juroren aus allen Zielgruppen vertreten sein.

- **Scales**

  Die grundsätzliche Frage lautet, ob die Bewertungen binär oder mittels einer Skala erfolgen sollen. Die binäre Methode ist simpler, jedoch liegt die Schwierigkeit für Suchmaschinen weniger darin, *irgendwie* relevante Ergebnisse auszugeben, als vielmehr *hoch* relevante. Daher sollten differenzierende Skalen verwendet werden. (S. Kapitel *2.2.6 Kritik*, Punkt 8)

Data analysis

Dieser Bereich beschäftigt sich mit der Analyse der gewonnenen Daten, etwa, welche Kennzahlen zu verwenden sind.

- **Relevance of the results**

  Neben den klassischen Kennzahlen Precision und Recall und weiteren gängigen wie Fallout (s. Kapitel *2.2.2 Retrievaleffektivität*), führt LEWANDOWSKI hier neuere Ansätze an:

- *Median Measure* (GREISDORF & SPINK 2001) berücksichtigt die Gesamtzahl aller gefunden Ergebnisse und misst nicht nur, wie gut die Ergebnisse sind, sondern auch, wie sehr im Verhältnis zu allen schlechten.
- *Importance of completeness of search results* und *Importance of precision of the search to user* (s. Kapitel *2.2.2 Retrievaleffektivität*).
- *Value of Search Results as a Whole* (SU 1998) korreliert offenbar gut mit anderen wichtigen Maßen und kann den Evaluationsprozess so effizienter und effektiver machen.
- *Salience* (s. Kapitel *2.2.2 Retrievaleffektivität*).
- *Relevance concentration* (DING & MARCHIONINI 1996) misst die Anzahl der Treffer mit der Wertung vier oder fünf (bei einer Fünf-Punkte-Skala) innerhalb der ersten zehn oder 20 Treffer.
- *CBC ratio* (MACCALL & CLEVELAND 1999) misst die Zahl der *content-bearing clicks* (CBC) im Verhältnis zu anderen Klicks (die nicht ausgeführt werden, um direkt möglicherweise relevante Informationen aufzufinden, wie etwa navigationsbedingte) im Suchprozess.
- *Quality of result ranking* (VAUGHAN 2004) beachtet die Korrelation zwischen Suchmaschinenranking und menschlichem Ranking.
- Bei der *Ability to retrieve top ranked pages* (VAUGHAN 2004) werden die Ergebnisse aller am Test beteiligten Suchmaschinen kombiniert und von Menschen gerankt - die Ability to retrieve top ranked pages misst dann den Anteil dieser Top-75-Prozent-Ergebnisse in der Ergebnisliste einer Suchmaschine.

Da die Kennzahlen starken Einfluss auf die Ergebnisse haben können, sollten sie mit Bedacht und immer in Hinsicht auf die Ziele des Tests gewählt werden.

- **Results descriptions**

Trefferbeschreibungen haben großen Einfluss auf das Klickverhalten der Nutzer (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*).

LEWANDOWSKI führt hier folgende Kennzahlen zur Auswertung von Trefferbeschreibungen an (vgl. LEWANDOWSKI 2008C, pdf S. 17-19): *Description-result precision, Description-result conformance, Description fallout* und *Description deception*. Auf diese Kennzahlen wird im Kapitel *3.2.2 Retrievaleffektivität* eingegangen. (S. Kapitel *2.2.6 Kritik*, Punkt 2)

- **Diversitiy**

Bereits auf der ersten Ergebnisseite sollte eine Vielfalt an Quellen abgebildet werden, um dem Nutzer einen Überblick des Themas zu geben.

Alleine vielfältige Quellen müssen jedoch noch nichts über die Abdeckung der Informationsbedürfnis-Aspekte aussagen. Bei der results classification (s.o.) können allerdings die Aspekte der Dokumente extrahiert und miteinander verglichen werden. Im Anschluss kann gemessen werden, wieviele Dokumente der Nutzer anschauen müsste, bis er Informationen über alle Aspekte eines Themas hat. Und auch wenn der Nutzer nicht zu allen Aspekten Informationen benötigt, können so Suchmaschinen identifiziert werden, die relevante, aber zu ähnliche Ergebnisse ausgeben. (S. Kapitel *2.2.6 Kritik*, Punkt 3)

- **Other analysis (based on results classification)**

Weitere Analysen der results classifications können eröffnen, ob bestimmte Ergebnistypen nützlicher sind als andere. So können Weblogs für allgemeine Websuchen relevant sein, während sie es im Kontext seriöser Nachrichtensuche weniger sind. Separat berechnete Kennzahlen für die unterschiedlichen Ergebnistypen können bestimmte Typen als nützlicher als andere enthüllen.

2.2.8 Bedeutende Tests

In diesem Kapitel werden zwei Evaluierungsinitiativen eingeführt, bevor ein kleiner Vergleich von älteren und neueren Suchmaschinen-Retrievaltests angestellt wird und abschließend einige markante Tests vorgestellt werden.

TREC

TREC (Text REtrieval Conference) ist eine Reihe jährlich stattfindender Evaluierungsexperimente, die 1992 aus dem *TIPSTER Text Program* hervorging und vom US-amerikanischen *National Institute of Standards and Technology* (NIST) durchgeführt wird (vgl. NIST 2010). Der Grundgedanke war, eine Plattform für umfangreiche Retrievaltests zu bieten, um den Erkenntnisaustausch zwischen Industrie, Forschung und Regierung zu ermöglichen, und so die IR-Forschung voranzutreiben (vgl. NIST 2010).

Der Ablauf dieser Tests sieht so aus, dass die teilnehmenden Teams die Kollektion der Testdokumente (hauptsächlich Zeitungsartikel) und verschiedene Aufgaben vom NIST erhalten, diese mittels ihrer Informationssysteme bearbeiten, und die Ergebnisse dem NIST zur Auswertung übermitteln. Danach findet immer ein Workshop statt, bei dem sich die Teams treffen, die Ergebnisse diskutieren und sich austauschen (vgl. CHU 2010, S. 236f).

Es gibt zwei grundsätzliche Aufgabenstellungen; den *ad hoc* und den *routing task*. Der ad hoc task entspricht dem klassischen Retrieval: Es müssen rückwirkend Informationen zu bestimmten Themen gefunden werden. Der routing task ist mit einem Pressespiegel

vergleichbar: Es müssen selektiv neu produzierte Informationen gefunden werden (vgl. FERBER 2003, S. 94f). Seit TREC-4 gibt es zusätzliche Aufgaben, *tracks* genannt, von denen jede auf ein bestimmtes Teilproblem des IR eingeht (vgl. CHU 2010, S. 239). Z.B. musste beim *Confusion track* mit schadhaften Daten umgegangen werden, beim *Cross-language (CLIR) track* deckten die Dokumente verschiedene Sprachen ab (auch Deutsch), und beim *Web track* stellte die Dokumentkollektion einen Ausschnitt des Webs dar (vgl. NIST 2012).

Die eingereichten Ergebnisse der Teams werden mittels Pooling und gemittelter Precision (s. Kapitel *2.2.2 Retrievaleffektivität*) verglichen (vgl. FERBER 2003, S. 95).

Die Ergebnisse von TREC sind freilich zu zahlreich und vielfältig, um hier vollständig aufgelistet zu werden. Wichtige Erkenntnisse sind z.B., dass mit vielen verschiedenen Ansätzen ähnliche Ergebnisse erzielt werden, dass automatisches Retrieval ebenso gut ist wie manuelles (abgesehen vielleicht bei schwachen Anfragen), und dass auf Statistiken basierende Indexierungs- und Sucharten günstig und wettbewerbstauglich sind, insbesondere für umfangreiche Dokumentsammlungen (vgl. CHU 2010, S. 246).

Kritik an den TREC-Tests bezieht sich zum großen Teil auf ihre Künstlichkeit. Nach CHU (vgl. CHU 2010, S. 247) hat diese drei Dimensionen: Erstens seien die Testdokumente vielmehr für die Tests geschaffen, als dass sie einem natürlich gewachsenem Informationsbestand entsprächen. Zweitens gebe es bei TREC keine echten Nutzer. Das Retrieval würde von Systementwicklern durchgeführt und die Retrievaleffektivität von Assessoren vorgenommen, die auch die Themen ausgewählt hätten. Drittens würden die Tests in einer Laborumgebung mit einer gewissen Kontrolle und Manipulation stattfinden, was nicht der Realität entspräche.

Nach CHU (vgl. CHU 2010, S. 248) ist TREC von großer Wichtigkeit für das IR. Weil es eine öffentliche und anregende Plattform ist, hat es viele Institutionen und Organisationen aus den USA sowie dem Ausland angezogen und seine Partizipanten ihre Systeme weiterentwickeln lassen. Mittels TREC wurden zahlreiche IR-Ansätze getestet, die Testkollektionsgröße bewegt sich mittlerweile im Gigabyte-Bereich, und durch ständige Optimierungen ist TREC realistischer und verallgemeinerbarer als viele andere Evaluierungsinitiativen.

CLEF

Die CLEF Initiative (früher *Cross-Language Evaluation Forum*, heute Conference and Labs of the Evaluation Forum) ist ein seit 2000 bestehendes europäisches Projekt, das sich mit multilingualem Retrieval befasst. Es widmet sich also dem Thema, das bei TREC durch den *Cross-language (CLIR) track* behandelt wurde; dem Finden von Dokumenten in anderen Sprachen als der Anfrage. Dabei liegt der Schwerpunkt auf europäischen Sprachen

(Deutsch, Englisch, Französisch, Italienisch und Spanisch), sodass diese 2002 aus TREC ausgegliedert wurden (vgl. GEY & OARD 2002, pdf S. 2).

Der Zweck von CLEF ist es, die Forschung und Weiterentwicklung von schwerpunktmäßig multilingualen Informationssystemen im europäischen Raum zu fördern und eine entsprechende Plattform zu bieten (vgl. CLEF 2012A).

Das Projekt besteht aus zwei Hauptbereichen (vgl. CLEF 2012A):

- Evaluierung von Informationssystemen und Workshops
- von Experten begleitete, jährlich stattfindende Konferenzen zu den Ergebnissen der Evaluierungen und Workshops, Experimenten mit multilingualen Daten, sowie Forschung und Herausforderungen im IR-Bereich

Bei der Evaluierung von Informationssystemen gibt es verschiedene Aufgaben (vgl. CLEF 2012B): Die standardmäßigen *Adhoc tracks* entsprechen dem o.g. Schwerpunkt des mehrsprachigen Retrievals. Bei den *Domain-specific tracks* wird nach wissenschaftlichen Informationen in speziellen Dokumentsammlungen gesucht (wie bei TREC besteht der Datenhauptbestand aus Zeitungsartikeln). Darüber hinaus gibt es *tracks*, die auf spezifische IR-Bereiche eingehen, z.B. *QA@CLEF*, mit dem Frage-Antwort-Systeme evaluiert werden, *ImageCLEF*, bei dem Bilder in *Cross-language*-Manier gesucht werden, und *MusiClef*, bei dem Musiksuchmaschinen, die auf auditiven und textlich mehrsprachigen Inhalten basieren, getestet werden.

Die Evaluierung findet wie bei TREC mittels Pooling und gemittelter Precision statt (vgl. GEY & OARD 2002, pdf S. 5) (s. Kapitel *2.2.2 Retrievaleffektivität*), jedoch sind die Verfahren aufgrund der unterschiedlichen Sprachen komplexer.

CLEF spielt eine bedeutende Rolle im europäischen IR und hat sich durch vielfältige Untersuchungen auch international etabliert (vgl. CLEF 2012A).

Nachdem zwei bedeutende Evaluierungsinitiativen vorgestellt wurden, wird nun ein kurzer Vergleich zwischen älteren und neueren Suchmaschinen-Retrievaltests angestellt.

LEWANDOWSKI nennt hierzu einige Punkte die Testeigenschaften betreffend, die in *Tabelle 4* als Übersicht dargestellt sind.

Tabelle 4: Testeigenschaften früher und heute (vgl. LEWANDOWSKI 2012, pdf S. 3f)

Früher	Heute
wenige Anfragen (5-15)	mehr Anfragen (mind. 25, i.d.R. 50+)
Anfragen aus Referenzfragen oder kommerziellen Online-Systemen	meist Anfragen aus allgemeinem Interesse (manchmal auch Kombinationen)
Untersuchung der ersten 10 oder 20 Ergebnisse	insgesamt mehr Ergebnisse, da zusätzliche Treffer
wenig Anonymisierung der Ergebnisse	mehr Anonymsierung der Ergebnisse
wenig Mischung der Ergebnisse	wenig Mischung der Ergebnisse
Bewertung oft durch Wissenschaftler selbst	Bewertung i.d.R. durch Studierende

Aus dieser Gegenüberstellung wird deutlich, dass sich Suchmaschinen-Retrievaltests aufgrund technischer Fortschritte sowie dem Bemühen der Durchführenden, die Realität abzubilden, weiterentwickelt haben. Durch die heute gegebenen technischen Möglichkeiten werden umfangreichere Tests durchgeführt, und durch bspw. die Wahl echter Suchanfragen und die Mischung der Ergebnisse wird versucht, die Künstlichkeit zu reduzieren. Auch ökonomische Aspekte, wie das Verwenden von Studierenden als Juroren, spielen heute eine größere Rolle.

Zu diesem kleinen Überblick werden nun einige Suchmaschinen-Retrievaltests mit besonderen, korrespondierenden Merkmalen oder anderen Schwerpunkten vorgestellt.

LEWANDOWSKI - The retrieval effectiveness of search engines on navigational queries (LEWANDOWSKI 2011C)

Der geschätzte Anteil von navigationsorientierten Anfragen (s. Kapitel *2.2.3 Suchanfragen*) an allen Suchanfragen differiert in Studien, jedoch handelt es sich sicher um einen nenneswerten Anteil (vgl. LEWANDOWSKI 2011C, pdf S. 3f) und allgemein um einen wichtigen Bereich des Web IR, der evaluiert werden sollte. Denn es ist für Nutzer hier sehr einfach, die Ergebnisse zu beurteilen, was sich in ihrer Meinung zur Suchmaschine niederschlägt (vgl. LEWANDOWSKI 2011C, pdf S. 2).

Aus Sicht von Suchmaschinenbetreibern ist es also wichtig, das eine richtige Ergebnis überhaupt und möglichst auf der ersten Trefferposition zu liefern.

Genau das untersucht LEWANDOWSKI in seiner Arbeit. Anhand 100 echter deutscher navigationsorientierter Suchanfragen wurden die sechs Suchmaschinen Google, Yahoo, *MSN*, *Ask*, *Seekport*, und *Exalead* mittels der Kennzahlen *Success N* und *Mean reciprocal rank* evaluiert. Success N misst, ob sich das richtige Ergebnis unter den ersten N Treffern befindet (vgl. HAWKING & CRASWELL 2005, pdf S. 4), und beim Mean reciprocal rank (einem TREC-Maß) werden Treffer auf einer absteigenden Skala nach ihrer Position bewertet (vgl. LEWANDOWSKI 2011C, pdf S. 4).

Die Ziele der Arbeit waren, herauszufinden, welche Suchmaschine die meisten richtigen Ergebnisse als erste Treffer ausgibt, wieviele Anfragen von den Suchmaschinen jeweils unbeantwortet bleiben, und in welchem Ausmaß richtige Ergebnisse auf hinteren Positionen ausgegeben werden (berücksichtigt wurden die ersten zehn Treffer).

Die Success-N-Berechnungen ergaben, dass Google mit 84 von 92 am meisten richtige Treffer auf der ersten Position ausgab.

Vier Anfragen konnten von keiner Suchmaschine beantwortet werden, lediglich ein Drittel der Anfragen wurde korrekt von allen Suchmaschinen beantwortet, und die höchste Quote unbeantworteter Anfragen erzielte Seekport mit 57 Prozent.

Bei den Mean-reciprocal-rank-Berechnungen erreichten Google und Yahoo eine *Gain ratio*, also eine Verbesserung, wenn bis zu alle einbezogenen Trefferpositionen berücksichtigt wurden, von knapp zehn Prozent. Entsprechend höher war sie bei den anderen Suchmaschinen, die die richtigen Treffer seltener als erste ausgaben; Seekport erreichte hier 22 Prozent.

Generell ergab die Untersuchung, dass Google, Yahoo und MSN mit ca. 90 Prozent korrekt beantworteter Anfragen besser abschnitten als die anderen drei Suchmaschinen, die trotzdem akzeptable Ergebnisse lieferten.

GRIESBAUM – Evaluation of three German search engines: Altavista.de, Google.de and lycos.de (GRIESBAUM 2004)

In dieser Arbeit wurde die Retrievaleffektivität von Google, *Altavista* und *Lycos* verglichen; es sollte herausgefunden werden, ob und inwiefern Google tatsächlich überlegen ist.

Das Mittel dieser Untersuchung waren die Treffer und Trefferbeschreibungen der ersten 20 Positionen. Für die Bewertung wurden binäre Skalen verwendet; für die Treffer standen die Möglichkeiten „relevant", „not relevant" und „links to relevant page(s)" zur Auswahl, für die Trefferbeschreibungen „seems to be relevant, I would click this link" und „seems not to be relevant, I would not click this link". Gemessen wurde die Precision, außerdem wurden die *Number of retrieved items* und *Number of answered questions* erfasst.

Die 50 Anfragen stellten eine von prekären (z.B. pornografischen) Inhalten bereinigte Auswahl echter Websuchanfragen der Pay-per-Click-Suchmaschine *QualiGO* und ihren Partnersuchseiten dar. Die Informationsbedürfnisse wurden generiert.

Die 27 Juroren setzten sich aus Studierenden, wissenschaftlichem Personal und Angestellten von QualiGO zusammen.

Die Untersuchung ergab, dass Google die beste Retrievaleffektivität erreichte, also mit 591 von ca. 1000 am meisten „relevant"- und „links to relevant page(s)"-Ergebnisse lieferte. Lycos belegte mit 530 solcher Treffer den zweiten Platz, gefolgt von Altavista mit 510.

Diese Reihenfolge setzte sich bzgl. der Trefferbeschreibungen fort; Google lieferte mit 590 am meisten relevante, gefolgt von Lycos mit 517 und Altavista mit 458.

Die Konsistenz zwischen Trefferbeschreibungen und Treffern war bei Google zu fast 100 Prozent gegeben, wich bei Lycos etwas ab (mehr Treffer als Trefferbeschreibungen waren relevant) und unterschied sich bei Altavista auf dieselbe Weise um mehr als fünf Prozent.

In Hinsicht auf *Micro-* und *Macro-Precision*, also auf die Verteilung der relevanten Treffer auf alle berücksichtigten Trefferpositionen, bzw. auf die Retrievalperformance in Hinsicht auf die einzelnen Anfragen (vgl.GRIESBAUM 2004), schnitt Google am besten ab, wieder gefolgt von Lycos und Altavista.

Generell ergab die Untersuchung, dass Google die besten Werte erreichte, gefolgt von erst Lycos und dann Altavista, die Unterschiede zwischen den Suchmaschinen jedoch relativ gering waren. Es gab beträchtliche Abweichungen bei den Bewertungen von Trefferbeschreibungen und Treffern, die sich bei Google und Lycos zwar fast gänzlich aufhoben, bei Altavista jedoch dergestalt in Erscheinung traten, dass viele Trefferbeschreibungen schlechter bewertet wurden als die Treffer selbst.

Die Arbeit von GRIESBAUM weist gleich mehrere interessante Merkmale auf, wegen denen sie hier Erwähnung findet: So wurden die drei Anfragetypen nach BRODER gemischt, was es eigentlich zu vermeiden gilt (s. Kapitel *2.2.3 Suchanfragen*), es wurden sinnvollerweise die Trefferbeschreibungen einbezogen (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*), für die Bewertung wurden binäre Skalen verwendet, wobei differenzierendere effektiver sind (s. Kapitel *2.2.4 Bewertung der Ergebnisse*), und die Bewertung wurde mitunter von den Durchführenden vorgenommen.

LEWANDOWSKI – The Retrieval Effectiveness of Web Search Engines: Considering Results Description (LEWANDOWSKI 2008C)

In dieser Arbeit wurde die Retrievaleffektivität der Suchmaschinen Google, Yahoo, MSN, Ask.com und Seekport unter Berücksichtigung der Trefferbeschreibungen untersucht.

Die Kernfragen waren, welche Suchmaschine die beste Precision erreicht, ob die Trefferbeschreibungen die Relevanz der Treffer widerspiegeln, und welchen Einfluss die Einbeziehung der Trefferbeschreibungen auf die allgemeine Relevanz hat.

Die 40 echten informationsorientierten Suchanfragen wurden von Studierenden abgefragt, die auch das zugehörige Informationsbedürfnis beschrieben. Sie waren es auch, die die ersten 20 Ergebnisse samt Trefferbeschreibungen jeweils „ihrer“ Anfrage sowohl binär als auch mittels zweier differenzierenderen Skalen bewerteten, jedoch wurden in dieser Arbeit lediglich die binären Bewertungen ausgewertet.

Die Untersuchung ergab, dass Google und Yahoo in Hinsicht auf die Treffer ähnliche Ergebnisse von knapp 50 Prozent Relevanz erzielten (Yahoo sogar geringfügig bessere).

Ask.com befand sich im Mittelfeld, während MSN und Seekport mit nur 34 Prozent die Schlusslichter darstellten.

Die Bewertung der Trefferbeschreibungen fiel bei allen Suchmaschinen besser aus als die der Treffer selbst; am stärksten bei Google und MSN mit jeweils gut zehn Prozent Differenz, am schwächsten bei Yahoo mit vier.

In Hinsicht auf die Micro-Precision der Treffer schnitten Google und Yahoo auf den vorderen Plätzen am besten ab. Bei der Betrachtung aller 20 Positionen sticht Ask.com etwas mit einer gleichmäßigeren Verteilung hervor (was für eine Verbesserung des Rankings spricht).

Bzgl. der Macro-Precision der Treffer beantwortete keine der Suchmaschinen *alle* Anfragen am besten, jedoch stellte mit 16 Yahoo als Sieger heraus.

Generell ergab die Untersuchung, dass die Precisionwerte aller Suchmaschinen der ersten 20 Trefferpositionen relativ schlecht waren.

Die Gründe für die Erwähnung dieser Arbeit sind der Schwerpunkt auf Trefferbeschreibungen (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*), der Umstand, dass die Anfragen von Nutzern stammten, die dann auch die zugehörigen Ergebnisse bewerteten, was den Optimalfall darstellt (s. Kapitel *2.2.3 Suchanfragen*), und dass die Ergebnisse sowohl anonymisiert als auch gemischt wurden.

JOACHIMS ET AL. – Accurately Interpreting Clickthrough Data as Implicit Feedback (JOACHIMS ET AL. 2005)

Einen alternativen Ansatz zur aufwändigen Suchmaschinenevaluierung mittels Juroren stellt die Nutzung von Klickdaten dar. Nach LEWANDOWSKI (vgl. LEWANDOWSKI 2012, pdf S. 4) wird dabei die Qualität eines Ergebnisses dadurch festgestellt, wie oft es angeklickt und wie lange es gelesen wurde („dwell time“), und wie oft ein Nutzer direkt zur Ergebisseite zurückkehrt („bounce rate“). Auf diese Weise können beträchtliche Mengen an Anfragen und Ergebnissen bewältigt werden, jedoch finden nur jene Ergebnisse, die tatsächlich angeklickt wurden, Eingang in die Untersuchung.

In der Arbeit von JOACHIMS ET AL. wurden als Teil des Klickverhaltens *Eyetracking*-Daten gemessen. Eyetracking meint das Aufzeichnen der Blickbewegungen von Nutzern mit einer Kamera, meist durch einen von der Hornhaut reflektierten Infrarotstrahl („Pupil Centre Corneal Reflection-Technik“) (vgl. TOBII 2010, S. 6). Unterschieden wird dabei zwischen *Fixationen*, während denen das Auge bewegungslos auf einen kleinen Bereich des Bildschirms ausgerichtet ist, und *Sakkaden*, die Sprünge zwischen Fixationen (vgl. FLOTHOW 2009, S. 3f).

Außerdem wurden in dieser Arbeit die Suchergebnisse zusätzlich klassisch „manuell“ bewertet, um das „implizite Feedback“ mit diesem „expliziten“ vergleichen zu können.

Die Studie fand in zwei Phasen statt. In Phase eins wurden Klick- und Eyetracking-Daten von 29 Studierenden bei der Beantwortung von zehn Fragen (navigationsorientierte und informationsorientierte) anhand von Google gesammelt. Die zweite Phase stellte eine Wiederholung der ersten dar, doch wurden nun die Google-Ergebnisse hinsichtlich ihrer Reihenfolge manipuliert. Das „explizite“ Feedback sah so aus, dass Juroren die Suchergebnisse nach ihrer Relevanz ordneten.

Die Untersuchung ergab, dass „implizites“ und „explizites Feedback“ im Großen und Ganzen miteinander übereinstimmten. Die Klickentscheidung von Nutzern hing von den Treffern selbst ab, wurde aber vom Ranking und der allgemeinen Qualität des Ergebnissets beeinflusst, was es schwierig macht, Klicks als „absolutes Feedback“ anzusehen.

Einen Überblick über bedeutende Suchmaschinen-Retrievaltests zwischen 1996 und 2007 bietet LEWANDOWSKI 2008C, für eine Übersicht älterer Tests siehe GORDON & PATHAK 1999.

2.3 Universal Search

In diesem Kapitel wird die Entstehung und das Prinzip der Universal Search aufgearbeitet, bevor einige Usability-Aspekte vermittelt werden, der Frage nachgegangen wird, wie die US gesteuert werden kann, und abschließend einige Fakten zum aktuellen Stand genannt werden.

2.3.1 Hintergrund und Prinzip

In diesem Abschnitt wird auf die Entstehung und das Prinzip der Universal Search eingegangen.

Mit den Zeiten haben sich die Ergebnisseiten von Suchmaschinen gewandelt. Bestanden sie vor einigen Jahren lediglich aus den berühmt-berüchtigten „zehn blauen Links“, tummelt sich heute eine Vielfalt um Aufmerksamkeit heischender Elemente vor den Augen der Nutzer. Einen Überblick verschafft die folgende Auflistung (vgl. LEWANDOWSKI 2008a, pdf S. 7f; HÖCHSTÖTTER & LEWANDOWSKI 2009, S. 3-8; LEWANDOWSKI 2012, pdf S. 4f):

Grundsätzliche Elemente

- ***Reguläre/organische Treffer***: Die konventionellen, „objektiven“ Web-Ergebnisse, unbeeinflusst von monetären Interessen.
- ***Sponsored links***: Werbe-Ergebnisse, von regulären abgegrenzt.
- ***Anzeigen***: Werbe(-Text-)Anzeigen, von der tatsächlichen Ergebnisliste abgegrenzt.
- ***Shortcuts***: Hervorgehobene Verweise auf Treffer anderer Kollektionen (z.B. Bilder, Videos etc.) oder besonders vertrauenswürdige Quellen.
    - ***Ergebnisse aus anderen Kollektionen***: Ein Bereich der Shortcuts; z.B. Bilder und Videos. Einige Beschreibungen/Vorschauen werden angezeigt, für weitere Resultate ist ein Klick nötig.
- ***Vorschläge zur Einschränkung/Erweiterung/Korrektur der Anfrage***: Hinweise zur Optimierung der Suchanfrage (z.B. bei Ein-Wort-Anfragen, „Meinten Sie ...“).

Hervorhebungen bestimmter Elemente

- ***Primary search result***: Erweitertes Ergebnis, das verschiedene Kollektionen abdeckt und dessen Beschreibung mit einem Bild sowie weiteren Informationen angereichert ist.
- ***Prefetch***: Ergebnis einer bevorzugten Quelle.
- ***Snippet***: Reguläres Ergebnis, dessen Beschreibung zusätzliche navigatorische Links enthält.
- ***Child***: Zweites Ergebnis eines Servers, dessen Beschreibung einen Link zu weiteren Treffern des Servers enthält.

Diese Elemente sind freilich nicht bei jeder Suchmaschine zu finden und werden i.d.R. auch nicht alle zusammen ausgelöst. Dennoch stellen sich Ergebnisseiten heute als vielfach komplexer dar als früher, was *Abbildung 5* demonstriert. Die Ziffern stellen eine Zählung der Links dar.

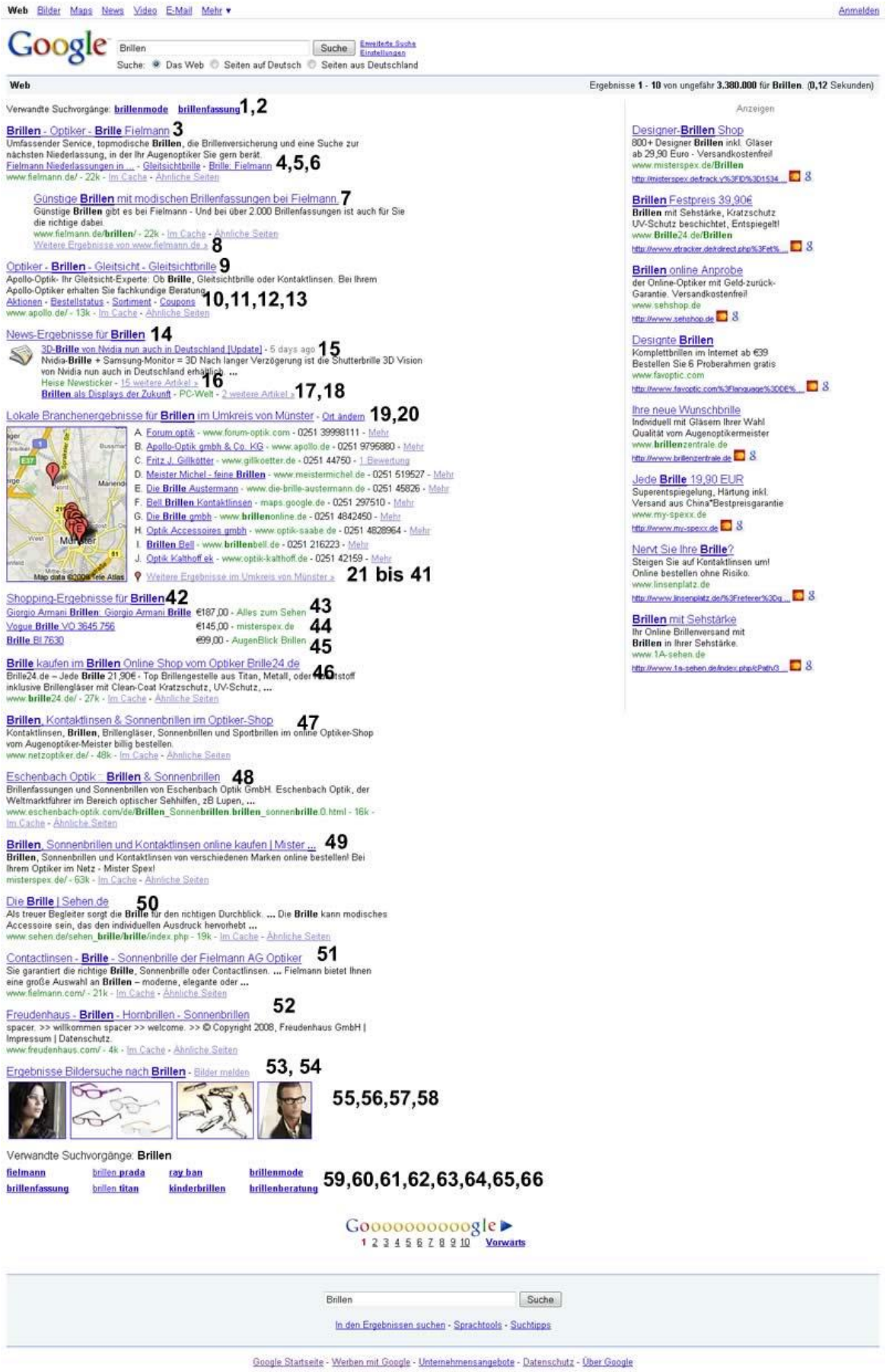


Abbildung 5: Google-Suchergebnisseite (SCHRÄPLER 2009)

Elemente dieser Ergebnisseite sind:

Reguläre/organische Ergebnisse: Links 46-52
Anzeigen: rechte Seite
Shortcuts: 14, 16, 18, 19, 42, 53
Ergebnisse aus anderen Kollektionen: 15, 17, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 43-45, 55-58
Vorschläge zur Einschränkung/Erweiterung/Korrektur der Anfrage: 1, 2, 59-66
Snippet: 3
Child: 7

Da Suchmaschinen mit den Inhalten des Webs umgehen, ist ein Grund für diese Entwicklung sicher deren Veränderung. Stellten früher einfache Webseiten den Löwenanteil der Web-Inhalte dar, entwickelten sich mit der Zeit bestehende Medienformen und es kamen neue hinzu (z.B. Bilder bzw. Videos), sodass Suchmaschinenbetreiber begannen, diese Themenstränge eigens zu crawlen (s. Kapitel *2.1.2 Web Information Retrieval*) und den Nutzern über Tabs (meist oberhalb der Sucheingabebox) zugänglich zu machen, was in *Abbildung 6* dargestellt ist. Die „vertikale“ Suche wurde etabliert. Im Gegensatz zur gewöhnlichen „horizontalen“ Suche, d.h. im gesamten Web, wird bei dieser lediglich in bestimmten Bereichen gesucht (vgl. SULLIVAN 2007), etwa in Bildern, Produkten oder Videos.

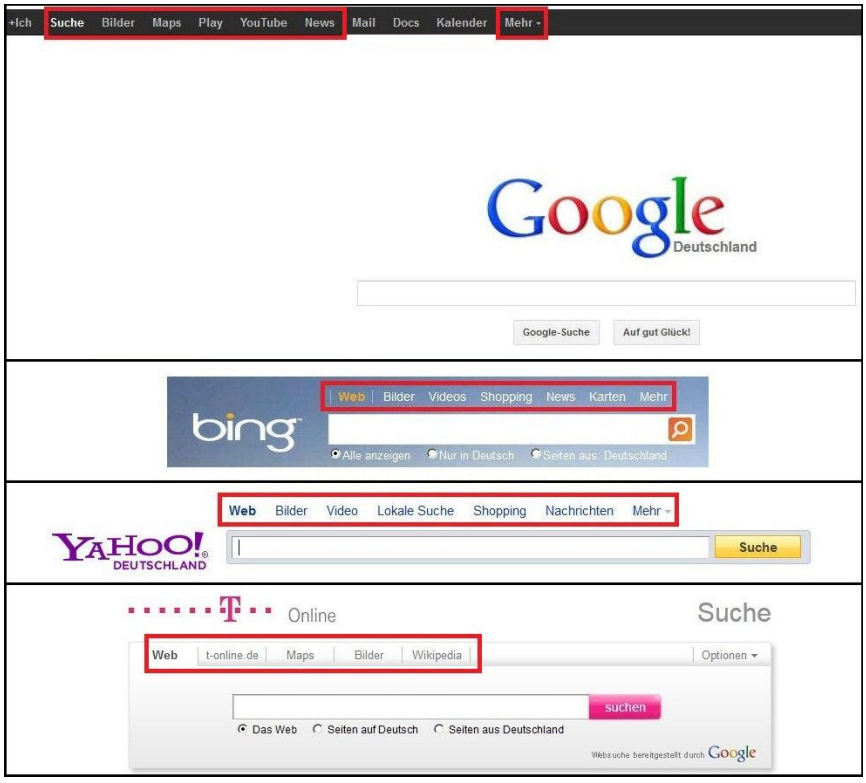


Abbildung 6: Tabs für vertikale Suchen

Der Nutzer kann so aus verschiedenen Kollektionen jene auswählen, die am besten zu seinem Informationsbedürfnis passt.

Diese Tab-Metapher wurde von den Nutzern jedoch nicht richtig wahrgenommen und verstanden, sodass in diesem Zusammenhang der Begriff „Tab-blindness“ geprägt wurde (vgl. SULLIVAN 2003).

Der nächste Entwicklungsschritt trug diesem Umstand Rechnung und wurde als Universal Search bekannt (auch *Blended Search* genannt).

Das Prinzip dabei ist es, dem Nutzer das Auswählen der Kollektionen zu ersparen und stattdessen ein paar ergänzende Top-Ergebnisse aus ihnen in das gewöhnliche Suchergebnis zu integrieren. So werden dem Nutzer verschiedenartige Ergebnisse angeboten, die entweder sein Suchbedürfnis direkt befriedigen, oder denen er ggf. zu den Kollektionen folgen kann. Es handelt sich also sozusagen um eine Kombination aus horizontaler und vertikaler Suche (in gleichzeitig allen Kollektionen). Dies ist in *Abbildung 7* veranschaulicht.

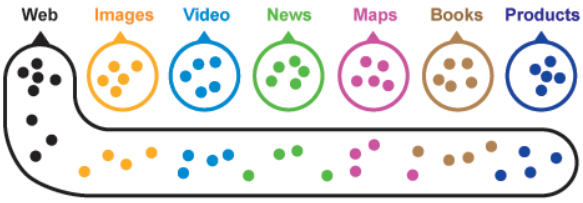


Abbildung 7: Prinzip der Universal Search (SULLIVAN 2007)

Spezielle Kollektionen gab es bei Google angefangen mit der Bildersuche bereits seit 2001 (vgl. GOOGLE 2012A), doch die Universal Search wurde erst 2007 in den USA offiziell unter der Kernaussage „We want to help you find the very best answer, even if you don't know where to look“ (MAYER 2007) eingeführt. Die damaligen Kollektionen waren Bilder, Karten, Bücher, Videos und Nachrichten (vgl. MIßFELDT 2011). 2008 ließ Google sich das Prinzip der Universal Search patentieren (s. USTPO 2008).

In Deutschland ist die US seit etwa 2010 ein wichtiges Thema (vgl. EMAGNETIX 2012).

2.3.2 Usability

Dieses Kapitel beschäftigt sich mit dem Prinzip der Universal Search in Hinsicht auf *Usability*-Aspekte.

Nach NIELSEN (vgl. NIELSEN 2012) ist Usability (auf Deutsch etwa „Gebrauchstauglichkeit“) ein Qualitätsmerkmal, das Aussage darüber gibt, wie einfach und angenehm Benutzeroberflächen zu bedienen sind. Sie setzt sich aus folgenden Elementen zusammen:

- **Erlernbarkeit**

  Wie einfach ist es, ohne weitere Kenntnis der Oberfläche grundlegende Aufgaben zu erledigen?
- **Effizienz**

  Wie schnell können nach dem Kennenlernen Aufgaben erledigt werden?
- **Einprägsamkeit**

  Wie gut können Nutzer ihre Kenntnisse nach längerer Nichtbenutzung rekonstruieren?
- **Fehler**

  Wieviele werden gemacht, wie schlimm sind sie und wie können sie behoben werden?
- **Zufriedenheit**

  Wie angenehm ist die Nutzung?

Ein weiteres entscheidendes Element der Usability ist die ***Utility***; erreicht der Nutzer sein Ziel?

In Bezug auf Suchmaschinen sind Nutzer i.d.R. nicht bereit, sich viele Suchergebnisse anzusehen; in den meisten Fällen werden lediglich die ersten Ergebnisseiten (vgl. HÖCHSTÖTTER & KOCH 2008, pdf S. 9) und dort jeweils nur die ersten Treffer betrachtet (vgl. CUTRELL & GUAN 2007, S. 5).

Für Suchen mit traditionell listenförmig angeordneten, organischen Ergebnissen gilt folgendes schematische Scanverhalten der Nutzer: Der Hauptfokus liegt nach Erscheinen der Ergebnisseite oben links (auf dem ersten Treffer). Von dort aus wird einmal kurz nach unten und zweimal nach rechts gescannt. Dieses Verhalten ist in der Literatur als *Golden Triangle* bzw. als *F-Scanning* bekannt (vgl. ENQUIRO 2005, S. 7ff) und in *Abbildung 8* dargestellt.

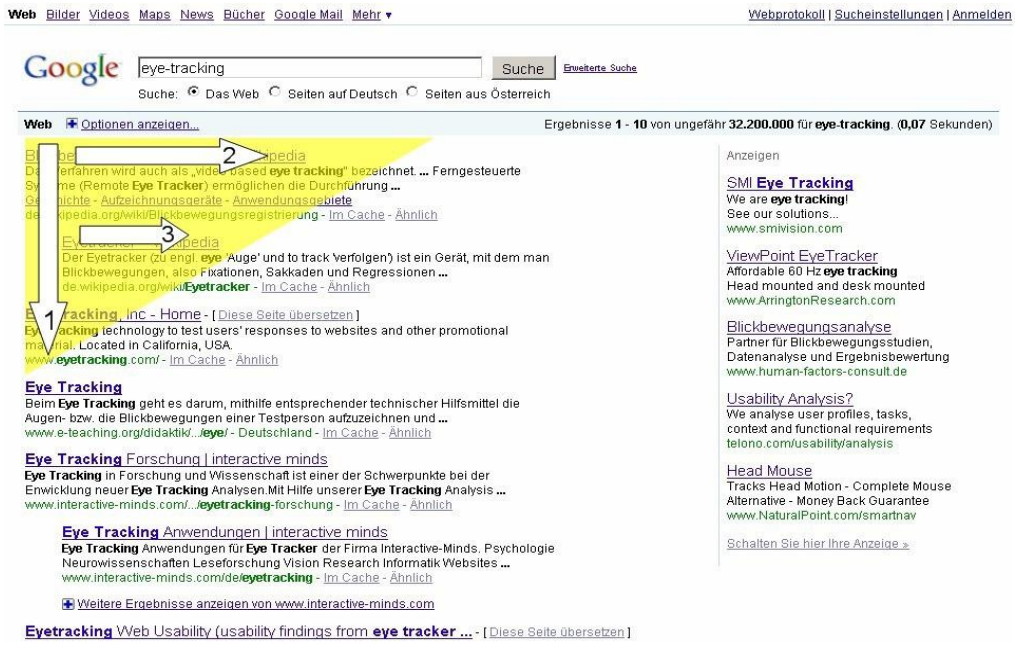


Abbildung 8: Golden Triangle bzw. F-Scan-Muster (STEINER 2010, S. 3)

Dieses Verhalten ergibt sich aus unserer Schriftrichtung (primär von links nach rechts, sekundär von oben nach unten), die beim Ranking aufgegriffen wird.

Unsere Aufmerksamkeit wird im Allgemeinen von Intensität und im Speziellen u.a. von Farben angezogen (vgl. WIRTH 2009). Da US-Resultate oft solche Reize darstellen (etwa bei Bilder-, Video- und Karten-Vorschauen) wird das Golden-Triangle-Scanverhalten dadurch aufgebrochen (vgl. QUIRMBACH 2009, S. 231). Dies zeigen *Heatmaps* (grafisch dargestellte Eyetracking-Daten, die zeigen, wo und wie lange der Blick von Nutzern verweilte; je länger ein Bereich betrachtet wurde, desto heller ist er dargestellt) in *Abbildung 9*.

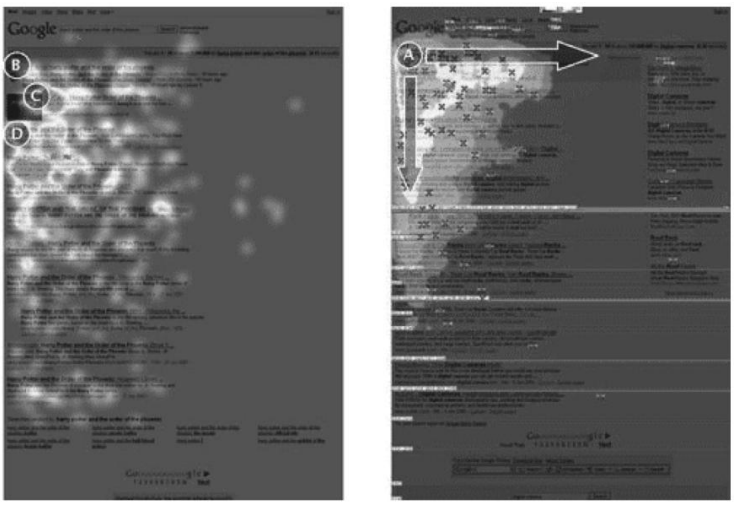


Abbildung 9: Heatmaps Standard-Suche (rechts) und US (links) (QUIRMBACH 2009, S. 232)

Die Aufmerksamkeit liegt bei der US also nicht mehr so konzentriert im Golden-Triangle-Bereich, sondern ist gleichmäßiger auf die Ergebnisseite verteilt. Das ist wichtig, um den Fokus des Nutzers auf die unterschiedlichen Ergebniskollektionen zu lenken. Denn was nützt es, solche Treffer auszugeben, wenn der Nutzer sie nicht wahrnimmt, weil er auf die ersten paar (regulären) Ergebnisse fixiert ist? „Wo nicht hingesehen wird, wird auch nicht geklickt!“ (QUIRMBACH 2009, S. 233)

2.3.3 Steuerung

In diesem Abschnitt wird auf die gezielte Verwendung des Prinzips der Universal Search eingegangen.

Die beiden Kernfragen bzgl. der Steuerung der Universal Search lauten (vgl. LEWANDOWSKI 2011B, S. 61):

- Wann sollen Ergebnisse aus einer bestimmten Kollektion eingeblendet werden?
- Wie sollen diese Ergebnisse platziert werden?

Generell ist die Einblendung von Ergebnissen spezieller Kollektionen freilich dann sinnvoll, wenn Suchanfrage (und damit Informationsbedürfnis) und Kollektion(en) gut zusammenpassen.

Wie dies aussehen kann, ist in *Tabelle 5* an einigen Kollektionen und Typen von Suchanfragen beispielhaft dargestellt.

Tabelle 5: Wann US-Resultate ausgegeben werden sollten

Ergebnisse	**Suchanfragen**
Bilder	nach bekannten Persönlichkeiten und Objekten („stephen king“, „eiffelturm“)
Produkte	die wahrscheinlich eine Kaufabsicht darstellen („sneakers 45“, „digitalkamera“)
Videos	nach Filmen, Serien, TV-Shows und Persönlichkeiten aus dem Fernsehen („game of thrones“, „jon stewart“)
Nachrichten	nach aktuellen Geschehnissen („regierung griechenland“, „breivik prozess“)
Lokale Ergebnisse	mit lokalem Bezug (auch angenommen) („rathaus hamburg“, „zahnarzt“)

Passen Suchanfragen zu mehreren Kollektionen, sollten diese Facetten durch die Einbindung von entsprechenden Resultaten möglichst abgedeckt werden.

Bei der Frage, *wie* solche Ergebnisse platziert werden sollen, sind mehrere Aspekte zu berücksichtigen.

Nach QUIRMBACH (vgl. QUIRMBACH 2009, S. 220, 235, 241) verändert die Universal Search sowohl das Suchinterface und die Nutzerführung im Ganzen (Makroebene), als auch die Darstellung der Ergebnisse im Detail (Mikroebene). Beide Aspekte haben direkte Auswirkungen auf die Wahrnehmung des Nutzers (s. Kapitel *2.3.2 Usability*). Die Ausgabe von US-Resultaten muss wohlüberlegt sein, denn für die Akzeptanz des Nutzers muss dieser nachvollziehen können, wann welches US-Modul ausgegeben wird. D.h. die Module müssen zu seiner Suchanfrage (und dem dahinterliegenden Informationsbedürfnis) passen. Jedes eingebundene Modul stellt dabei auf der anderen Seite eine Konkurrenz zu Anzeigen dar (über die der Umsatz der Suchmaschine generiert wird), denn der sichtbare Bereich auf dem Bildschirm des Nutzers ist begrenzt. Es gilt daher bei der Konzeption der Universal Search, die Suchergebnisseite sowohl inhaltlich als auch in ihrer Darstellung emotional ansprechend zu gestalten, dabei aber nicht zu überladen (wie es z.B. in *Abb. 5* passiert).

Prinzipiell gibt es zwei Möglichkeiten, US-Module zu steuern:

- **Die Anforderungen an das Produkt Suche ändern sich**

  Die Aussteuerung der Module wird an die Anzeigen- und Filterlogiken angepasst (Top-down).
- **Eine neue Datenquelle soll eingebunden werden**

  Die Datenquelle wird in ein Modul und die zugehörigen Aussteuerungsregeln überführt (diese müssen in Gesamtkonzept und Strategie überführt werden) (Bottom-up).

Diese beiden Wege sind in *Abbildung 10* dargestellt.

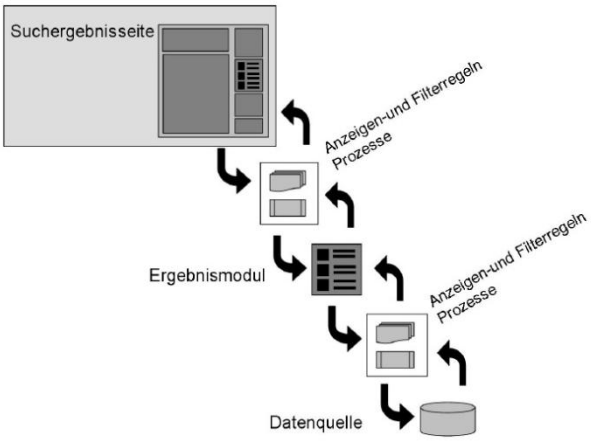


Abbildung 10: Top-down- und Bottom-up-Steuerung von US-Modulen (QUIRMBACH 2009, S. 236)

Generell ist die Erstellung von Universal-Search-Modulen ein iterativer Prozess des Messens, Beobachtens und Anpassens.

2.3.4 Fakten

In diesem Kapitel werden einige Fakten zur Universal Search präsentiert.

Die fünf wichtigsten der heute für die Universal Search genutzten Kollektionen sind (vgl. ALPAR 2012):

- **Bilder** (Images)
- **Produkte** (Shopping)
- **Videos**
- **Nachrichten** (News)
- **Lokale Ergebnisse** (*Google Maps* bzw. *Places*)

Die prozentualen Häufigkeiten, mit denen US-Module auftauchen, wurden im November 2011 für die ersten zehn Suchergebnisseiten anhand einer Datenbasis von 100 Millionen Keywords ausgewertet (vgl. BACHOR 2011). Die Ergebnisse sind in *Tabelle 6* dargestellt.

Tabelle 6: Häufigkeit US-Module (prozentual) (BACHOR 2011)

	Seite 1	**Seite 2**	**Seite 3**	**Seite 4**	**Seite 5**	**Seite 6**	**Seite 7**	**Seite 8**	**Seite 9**	**Seite 10**
organisch	100,00	100,00	200,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
HP Adwords	50,13	48,36	48,36	48,32	48,36	48,38	48,35	48,34	48,35	48,35
Adwords	42,91	37,35	39,72	35,95	40,96	34,37	41,33	35,66	39,42	36,08
News	11,01									
Maps	8,71									
Videos	28,75	30,09	26,50	22,91	18,65	15,21	13,12	11,57	10,75	10,05
Images	33,59	21,64	1,67	0,22	0,03	0,00	0,02		0,00	
Shopping	29,92									
Books	0,00	0,02	0,02	0,03	0,02	0,02	0,14	0,02	0,02	0,02
iGoogle	0,15									
Translate	0,50									
Scholar	0,01									
Profiles	0,04	0,15	0,17	0,18	0,19	0,16	7,59	0,14	0,15	0,14
SiteLinks	3,14									

Von den o.g. wichtigsten US-Kollektionen befinden sich also Bilder mit einem Vorkommen in 34 Prozent der Suchen auf Platz eins der ersten Ergebnisseite. Produkte und Videos

folgen mit jeweils knapp 30 Prozent. Nachrichten kommen nur noch in elf Prozent der Suchen vor und das Schlusslicht bilden lokale Ergebnisse mit weniger als zehn Prozent.

In Hinsicht auf die weiteren Ergebnisseiten ist festzustellen, dass lediglich Bilder und Videos überhaupt noch ausgegeben werden, wobei Bilder mit Seite sieben enden. Während sich ihre Häufigkeit stetig verringert, auf der zweiten Seite um etwa ein Drittel und der dritten um ca. 95 Prozent, erhöht sich die Häufigkeit von Videos in Relation zur ersten Seite auf der zweiten, und ist erst auf der sechsten Seite ungefähr halbiert.

Wieviele Resultate durschnittlich pro US-Modul vertreten sind, ist in *Tabelle 7* ebenfalls für die ersten zehn Ergebnisseiten dargestellt. Die Datenbasis ist dieselbe wie bei *Tabelle 6*.

Tabelle 7: Anzahl Resultate pro US-Kollektion (BACHOR 2011)

	Seite 1	**Seite 2**	**Seite 3**	**Seite 4**	**Seite 5**	**Seite 6**	**Seite 7**	**Seite 8**	**Seite 9**	**Seite 10**
organisch	9,56	9,54	9,63	9,69	9,76	9,81	9,84	9,86	9,87	9,88
HP Adwords	2,39	2,34	2,34	2,34	2,34	2,34	2,35	2,35	2,34	2,35
Adwords	7,24	6,79	6,96	6,96	7,09	6,35	7,15	6,51	6,90	6,60
News	1,18									
Maps	3,51									
Videos	2,45	2,05	1,82	1,66	1,45	1,37	1,36	1,35	1,35	1,33
Images	4,93	5,12	5,08	5,02	4,91	4,66				
Shopping	3,05									
Books	1,14	1,09	1,10	1,34	1,34	1,82	0,15	1,63	1.47	1,17
iGoogle	1,00									
Translate	1,02									
Scholar	2,83									
Profiles	1,04	1,04	1,18	1,18	1,15	1,10	0,02	1,07	1,07	1,08
SiteLinks	5,03									

Spitzenreiter sind deutlich Bilder mit durchschnittlich fast fünf Exemplaren auf der ersten Ergebnisseite. Es folgen lokale Ergebnisse mit dreieinhalb, und Produkte mit drei Resultaten. Videos werden im Mittel zweieinhalb ausgegeben, und Nachrichtenartikel bilden mit nur gut einem Resultat das Schlusslicht.

In Hinsicht auf die weiteren Ergebnisseiten nimmt die Anzahl der Bilder auf den Seiten zwei bis vier sogar noch zu, bis sie wieder leicht abfällt, und schießlich keine mehr ausgegeben werden. Videos werden auf den Seiten abnehmend ausgegeben, bis sich die Zahl ab Seite fünf etwa bei eineinhalb bis eins einpendelt.

In *Abbildung 11* ist die Entwicklung des prozentualen Vorkommens einiger US-Kollektionen von Februar 2011 bis Februar 2012 dargestellt. Unterschiede zu den *Tabellen 6* und *7* ergeben sich aufgrund der Datenbasis: Es wurden Wochendaten mit vielen *Shorttail*

bzw. Monatsdaten mit vielen *Longtail Keywords* untersucht (vgl. BACHOR 2011). (Shorttail Keywords sind allgemeine Suchanfragen, bei denen die Ergebniskonkurrenz groß ist und die daher eher US-Resultate auslösen als Longtail Keywords, die spezifischere Suchanfragen darstellen (vgl. GATES 2012).)

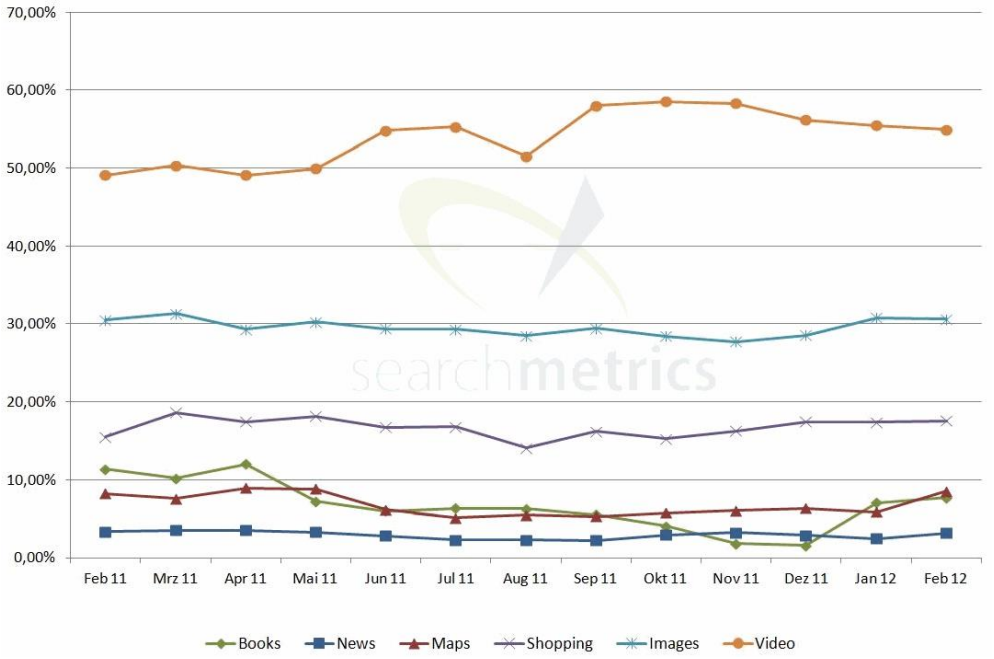


Abbildung 11: Entwicklung Vorkommen von US-Kollektionen (BACHOR 2012)

Den ersten Platz stellen im Bereich zwischen 50 und 60 Prozent mit Abstand Videos dar, gefolgt von Bildern bei ca. 30 Prozent. Produkte kommen in knapp 20 Prozent der Suchen vor, und Bücher, Karten und Nachrichten liegen durchschnittlich bei ca. fünf Prozent nah beieinander.

Es ist festzustellen, dass alle Kollektionen auf den gesamten Zeitraum betrachtet eine Tendenz zum leicht vermehrten oder gleichbleibenden Vorkommen zeigen – bis auf Bücher, die zum Jahreswechsel fast gar nicht mehr angezeigt, dann aber wieder vermehrt ausgegeben wurden.

In *Abbildung 12* sind die Quellen der wichtigsten US-Kollektionen nach Häufigkeit dargestellt. Datenbasis sind erneut die „Millionen von Keywords“ (BACHOR 2012).

Video	Maps	News	Shopping	Images
youtube.com	google.de	focus.de	google.de	wikipedia.org
dailymotion.com	accorhotels.com	welt.de	amazon.de	blogspot.com
myvideo.de	arbeitsagentur.de	bild.de	ebay.de	amazon.de
yatego.com	sixt.de	abendblatt.de	otto.de	preisroboter.de
vimeo.com	berlin.de	sueddeutsche.de	yatego.com	wordpress.com
clipfish.de	mercedes-benz.de	rp-online.de	neckermann.de	ebay.de
metacafe.com	marriott.com	morgenpost.de	mercateo.com	yatego.com
netzwelt.de	bestwestern.de	derwesten.de	buecher.de	wikia.com
bild.de	europcar.de	stern.de	hood.de	fotocommunity.de
sevenload.com	motel-one.com	zeit.de	rakuten.de	motor-talk.de

Abbildung 12: Quellen US-Kollektionen (BACHOR 2012)

Es zeigt sich, dass Google, sofern vorhanden, die eigenen Quellen präferiert. Bei Karten und Produkten, aber auch bei Videos, denn *YouTube* wurde 2006 von Google übernommen (vgl. BERCHER 2006).

Auch Bing (vgl. INTERNETWORLD 2009) verwendet das Prinzip der Universal Search, jedoch unter dem Namen Blended Search. Dies zeigt *Abbildung 13*.

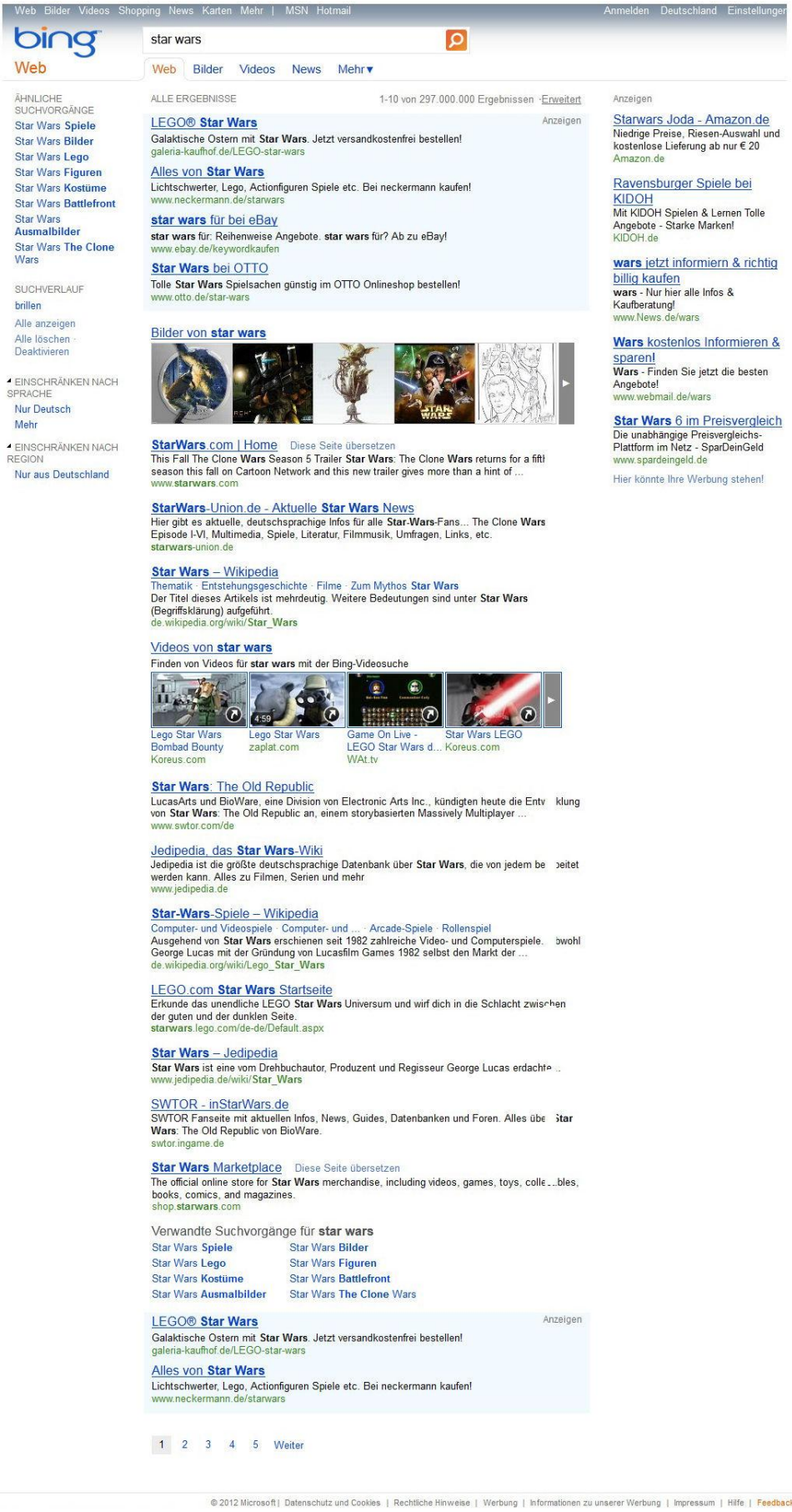


Abbildung 13: Blended Search

Die Elemente der Ergebnisseite ähneln denen von Google, sind lediglich etwas anders angeordnet. Eingebundene Kollektionen sind in diesem Fall Bilder und Videos.

3. Retrievaltest

Dieses Kapitel ist dem Retrievaltest gewidmet, der im Rahmen dieser Arbeit durchgeführt wurde.

Es werden zunächst die Forschungsfragen samt Hypothesen dargelegt, bevor die Methodik des Tests erläutert und anschließend die Ergebnisse präsentiert sowie diskutiert werden.

3.1 Forschungsfragen und -hypothesen

In diesem Abschnitt werden die Forschungsfragen und –hypothesen dargelegt.

Der eingangs dargestellte Stand der Forschung zeigt Bereiche der Suchmaschinenevaluierung auf, die mithilfe empirischer Forschungsmethoden zu untersuchen sind.

Das primäre Ziel des vorliegenden Retrievaltests ist es, Resultate spezieller Kollektionen auf ihre Relevanz und ihren Einfluss auf das Gesamtergebnis von Suchmaschinen hin zu untersuchen. Weiter wird die Relevanz von Ergebnissen verschiedener Suchmaschinen untersucht. Außerdem wird die Beziehung von Trefferbeschreibungen und Treffern betrachtet. Auf diese Weise wird der Schwerpunkt Universal Search in den Suchmaschinenergebnis-Kontext eingebettet.

Die resultierenden Forschungsfragen mit ihren zu prüfenden Hypothesen werden basierend auf dem gegenwärtigen Forschungsstand wie folgt formuliert:

F1: Spiegelt sich die Relevanz von Treffern in deren Beschreibungen wider?

Nutzer entscheiden in hohem Maße aufgrund der Trefferbeschreibung, ob sie einen Treffer anklicken (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*).

H1.1: Werden die regulären Beschreibungen und Treffer in Hinsicht auf ihre Relevanz miteinander verglichen, schneiden im Durchschnitt die Beschreibungen besser ab.

Dies war z.B. bei LEWANDOWSKI 2008C der Fall, weswegen auch hier davon ausgegangen wird.

H1.2: Werden die Beschreibungen/Vorschauen der speziellen Kollektionen und die Treffer in Hinsicht auf ihre Relevanz miteinander verglichen, ergibt sich Folgendes: Bei Produkten und lokalen Ergebnissen ähneln sich Beschreibungen und Treffer sehr, bei Videos und Nachrichten schneiden im Durchschnitt die Beschreibungen besser ab.

Da die „Beschreibung" von Bildern eine bloße Vorschau ist, müssten sich diese und das Ergebnis hinsichtlich der Relevanz gleichen (wovon in diesem Test ausgegangen wird). Dies müsste in leicht abgeschwächter Form auch auf Produkte und lokale Ergebnisse zutreffen, da die wichtigen Aspekte dieser Treffer (Bild und Preis bzw. Adresse und Kartenausschnitt) bereits in der Beschreibung enthalten sind. Bei Videos und Nachrichten hingegen zeigen die Beschreibungen tatsächlich nur kleine Ausschnitte der Treffer, weswegen hier von demselben Verhalten wie bei H1.1 ausgegangen wird.

F2: Wer liefert die relevanteren Treffer – Google oder Bing?

Siehe hierzu Kapitel *3.2.1 Suchmaschinen*.

H2.1: Werden die regulären Treffer von Google und Bing in Hinsicht auf ihre Relevanz miteinander verglichen, schneiden die von Google besser ab.

Da Google generell als Referenzsuchmaschine („Gold Standard") gilt, in Hinsicht auf Relevanz in zahlreichen Studien (z.B. GRIESBAUM 2004, TAWILEH & GRIESBAUM & MANDL 2010) am besten abschnitt und Bing als fertiges Produkt noch neu am deutschen Suchmaschinenmarkt ist, wird hier von einer Überlegenheit Googles ausgegangen.

H2.2: Werden alle Treffer (inkl. speziellen Kollektionen) von Google und Bing in Hinsicht auf ihre Relevanz miteinander verglichen, schneiden die von Google besser ab.

Da Google in Hinsicht auf reguläre Treffer bisher i.d.R. am besten abschnitt und Erfinder der US ist, wird auch bzgl. solcher Treffer von diesem Verhalten ausgegangen, was auch im Gesamtergebnis seinen Niederschlag finden würde.

F3: Werten die Treffer der speziellen Kollektionen das Trefferergebnis insgesamt auf?

Die Integration von Resultaten spezieller Kollektionen stellt einen Entwicklungsschritt von Suchmaschinen dar. Es gilt zu untersuchen, ob dieser Schritt nach vorne, auf der Stelle oder nach hinten gemacht wird.

H3: Werden Treffer spezieller Kollektionen ausgegeben, sind sie im Durchschnitt relevant und werten damit das Suchtrefferergebnis auf.

Da sich die Suchmaschinenbetreiber für dieses Feature entschieden haben, wird davon ausgegangen, dass es nützlich ist und das Gesamtergebnis verbessert.

3.2 Methodik

In diesem Kapitel werden die Umstände des Retrievaltests erläutert. Dabei wird die Struktur des Kapitels *2.2 Retrievaltests* von den Punkten *2.2.1 Suchmaschinen* bis *2.2.5 Juroren* gespiegelt. Für die jeweiligen theoretischen Hintergründe siehe also diese Entsprechungen.

3.2.1 Suchmaschinen

In diesem Abschnitt wird die Wahl der verwendeten Suchmaschinen begründet.

Bei der Wahl der zu untersuchenden Suchmaschinen spielten drei Faktoren entscheidende Rollen:

- Verwendung des Prinzips der Universal Search
- Unabhängigkeit von anderen Suchmaschinen
- möglichst hoher Marktanteil in Deutschland

Tabelle 8 zeigt eine diesbzgl. Übersicht der fünf populärsten Suchmaschinen in Deutschland:

Tabelle 8: Potenzielle Test-Suchmaschinen (vgl. KRONENBERG 2012, W&V 2011, WEBHITS 2012)

	Google	Bing	Yahoo	T-Online	AOL
US-Prinzip	ja	ja	ja	nein	nein
Unabhängigkeit	ja	ja	nein (Bing)	nein (Google)	nein (Google)
Marktanteil	ca. 90 %	ca. 4 %	ca. 3 %	ca. 2 %	ca. 0,3 %

Da T-Online und AOL aufgrund ihrer Nichtverwendung des Prinzips der US und ihrer Abhängigkeit von Google, und Yahoo wegen seiner Abhängigkeit von Bing nicht in Frage

kommen, stellten sich Google und Bing als für den vorliegenden Test am geeignetsten heraus.

Google ist von anderen Suchmaschinen unabhängig, dominiert die Suchmaschinenlandschaft deutlich mit ungefähr 90 Prozent Marktanteil und hat die Universal Search erfunden.

Bing ist die Suchmaschine mit dem nächstgrößeren Marktanteil von ca. vier Prozent - zählt man Yahoo dazu, die Bing-Ergebnisse ausgibt, sind es etwa sieben - sie verwendet das US-Prinzip als Blended Search und ist ebenfalls unabhängig. Davon abgesehen hat Bing Anfang 2012 nach über zwei Jahren ihren Betastatus in Deutschland hinter sich gelassen (vgl. MICROSOFT 2012), sodass es Zeit ist, sie am „Gold Standard“ Google zu messen.

3.2.2 Retrievaleffektivität

In diesem Kapitel wird beschrieben, mithilfe welcher Kennzahlen die Auswertung durchgeführt wurde.

Im vorliegenden Test wurde die Retrievaleffektivität der Suchmaschinen anhand regulärer und US/BS-Ergebnissen grundsätzlich durch das klassische Maß Precision berechnet.

Eine besondere Form war dabei die *Micro-Precision*, die misst, wie sich die relevanten Ergebnisse auf die Trefferpositionen verteilen (vgl. GRIESBAUM 2004, LEWANDOWSKI 2008C, pdf S. 11): Dabei werden die Ergebnisse den Trefferpositionen zugeordnet, es werden die Precisionwerte berechnet, und dann werden diese für die einzelnen Trefferpositionen verrechnet, sodass bspw. der Wert von Trefferposition fünf die Precision der ersten fünf Trefferpositionen zusammen anzeigt. Das Nutzermodell des Tests sieht nämlich so aus, dass der Nutzer die Ergebnisse nacheinander durchgeht.

Für die Auswertung der Beziehungen zwischen Beschreibungen/Vorschauen und Treffern wurden einige hierfür adaptierte Kennzahlen verwendet (vgl. LEWANDOWSKI 2008C, pdf S. 17ff):

- **Description-result precision**

  Sie misst den Anteil jener Ergebnisse, bei denen sowohl Beschreibung als auch Treffer relevant sind, an allen ausgegeben Ergebnissen. Es werden also die beidseitig relevanten Beschreibung-Treffer-Paare durch alle ausgegeben Paare geteilt. Der anzustrebende Wert ist 1, dann wären alle Beschreibungen und Treffer relevant.

- **Description-result conformance**

  Sie misst den Anteil jener Ergebnisse, bei denen Beschreibung und Treffer in binärer Relevanzbewertung übereinstimmen, an allen ausgegebenen Ergebnissen. Es werden also die übereinstimmenden Beschreibung-Treffer-Paare durch alle ausgegebenen Paare geteilt. Der anzustrebende Wert ist 1, dann wären Beschreibungen und Treffer immer konsistent.

- **Description fallout**

  Er misst den Anteil jener Ergebnisse, bei denen die Beschreibung irrelevant, doch der Treffer relevant ist. Es werden also die Paare bestehend aus irrelevanter Beschreibung und relevantem Treffer durch alle ausgegebenen Paare geteilt. Der anzustrebende Wert ist 0, dann würde dieser Fall nicht auftreten.

- **Description desception**

  Sie misst den Anteil jener Ergebnisse, bei denen die Beschreibung relevant, doch der Treffer irrelevant ist. Es werden also die Paare bestehend aus relevanter Beschreibung und irrelevantem Treffer durch alle ausgegebenen Paare geteilt. Der anzustrebende Wert ist 0, dann würde dieser Fall nicht auftreten.

Außerdem wurde, um den Durchschnitt bei Häufigkeitsverteilungen herauszufinden, der *Median* berechnet – dieser Wert ist die Mitte dieser Verteilungen, teilt sie also in gleich große Hälften und stellt so den Durchschnitt dar (vgl. STATISTA 2012).

3.2.3 Suchanfragen

In diesem Abschnitt wird die Wahl der im Test verwendeten Suchanfragen dargelegt.

Es wurden 50 deutsche informationsorientierte Anfragen samt zugehörigem Informationsbedürfnis ausgewählt. Der Großteil stammt aus anderen Studien, während einige indirekt aus solchen abgeleitet wurden – Prämisse war jeweils die Auslösung möglichst vieler US/BS-Resultate. Für eine vollständige Auflistung der Suchanfragen siehe Anhang *B Suchanfragen*.

In *Tabelle 9* ist die Länge der verwendeten Anfragen dargestellt.

Tabelle 9: Suchanfragenlänge

Länge	Anzahl an Suchanfragen	Prozent
ein Wort	17	34
zwei Wörter	24	48
drei Wörter	8	16
vier Wörter	1	2

Die durchschnittliche Länge deutscher Suchanfragen (aufgrund der häufigen Kompositabildung geringer als bei englischen) liegt zwischen 1,6 und 1,8 Wörtern (vgl. HÖCHSTÖTTER & KOCH 2008, pdf S. 12). In diesem Test beträgt sie 1,4, liegt also etwas unter diesem Durchschnitt.

Generell gibt Bing weniger Ergebnisse spezieller Kollektionen aus als Google, sodass von 100 gestellten Anfragen 69 solche Resultate auslösten; davon 42 bei Google und 27 bei Bing. Welche das im Detail waren, ist in Anhang *B Suchanfragen* dargestellt.

Tabelle 10 zeigt die Einordnung der Suchanfragen nach Themen.

Tabelle 10: Themen der Suchanfragen (Klassifikation nach CALDERON-BENAVIDES & GONZALES-CARO & BAEZA-YATES 2010, pdf S. 3)

Thema	Anzahl der Suchanfragen	Prozent
Nicht Jugendfreies & Sex	0	0
Kunst & Kultur	9	18
Schönheit & Stil	3	6
Autos und Verkehr	0	0
Computer & Internet	0	0
Bildung	1	2
Unterhaltung	9	18
Musik & Spiele	4	8
Finanzielles	0	0
Essen & Trinken	0	0
Gesundheit	0	0
Haus & Garten	1	2
Industrielle Produkte & Services	0	0
Politik & Regierung	5	10
Religion	0	0
Wissenschaft & Mathematik	0	0
Sozialwissenschaften	0	0
Sport	0	0
Technologie & Elektronik	7	14
Reisen	3	6
Undefiniert	7	14
Arbeit	1	2

Die Schwerpunkte der gewählten Anfragen liegen also auf den Bereichen „Kunst & Kultur“ sowie „Unterhaltung“ und „Technologie & Elektronik“. Auch „Politik & Regierung“ ist stärker vertreten.

Für die Einordnung der Suchanfragen im Detail siehe Anhang *C Suchanfragen – thematische Einordnung*.

3.2.4 Bewertung der Ergebnisse

In diesem Kapitel wird auf die Bewertung der Ergebnisse durch die Juroren eingegangen.

Der Retrievaltest wurde mit dem *Relevance Assessment Tool* (RAT) der *HAW Hamburg* durchgeführt (vgl. LEWANDOWSKI & SÜNKLER 2012, S. 237-245): Das RAT ist ein Web-Browser-Tool, mit dem Suchmaschinen-Retrievaltests gestaltet und die betreffenden Daten automatisiert erfasst, aufbereitet und ausgewertet werden können.

Es besteht aus folgenden vier Komponenten:

- **Suchmaschinenscraper**
  Der Scraper erfasst automatisiert Suchmaschinenergebnisse.
- **Backend zur Verwaltung und zum Testdesign**
  Im Backend werden Projekte, Projektbenutzer, Projektadministratoren und Suchaufgaben administriert.
- **Frontend zur Durchführung der Tests**
  Im Frontend findet die Relevanzbewertung durch die Juroren statt.
- **Auswertung**
  Diese Komponente befindet sich noch in der Entwicklung.

Die Vorbereitung des Tests sah folgendermaßen aus: Nach dem Einspeisen der Suchanfragen samt Informationsbedürfnissen und dem Verfassen einer Projektbeschreibung für die Juroren (s. beiliegende CD) wurden die Scraping- und weitere Testeinstellungen vorgenommen:

Suchmaschinen: Google, Bing

Art der Ergebnisse:

- die ersten zehn regulären; jeweils Beschreibung und Treffer
- alle US/BS-Resultate der ersten Ergebnisseite; jeweils die Übersichten

(US/BS-Resultate werden oft als Gruppe in die Ergebnisseite eingebunden) sowie Beschreibung/Vorschau und Treffer

Eine Ausnahme stellten Bilder dar; da die Vorschauen verkleinerte Versionen der Treffer sind, wurde aus ökonomischen Gründen darauf verzichtet, diese Treffer selbst noch einmal bewerten zu lassen. Diesen wurde bei der Auswertung die jeweilige Vorschaubewertung zugeordnet.

Aufgaben pro Juror: eine (d.h. eine komplette Suchanfrage; die Ergebnisse beider Suchmaschinen)

Abfrage demografischer Daten: ja (Geschlecht, Alter und Suchmaschinenerfahrung in Jahren)

Für die Bewertung zu verwendende Skalen: binäre und Fünfer-Skala

Daraufhin fand das Scraping statt; die dabei erfassten Daten wurden in einer Datenbank gespeichert (sicherheitshalber wurden parallel die Ergebnisseiten der Suchmaschinen abgespeichert, s. beiliegende CD).

Nun war der Test fertig vorbereitet, also wurde der zugehörige Link samt Zugangscode über *Facebook* verbreitet.

Der Test lief im Zeitraum vom 30. Mai bis zum 13. Juni 2012.

Nach dem Aufrufen wurden die demografischen Daten abgefragt und die Projektbeschreibung angezeigt, danach fand die Bewertung statt. Das Interface, das die Juroren dabei vor Augen hatten, ist in *Abbildung 14* abgebildet.

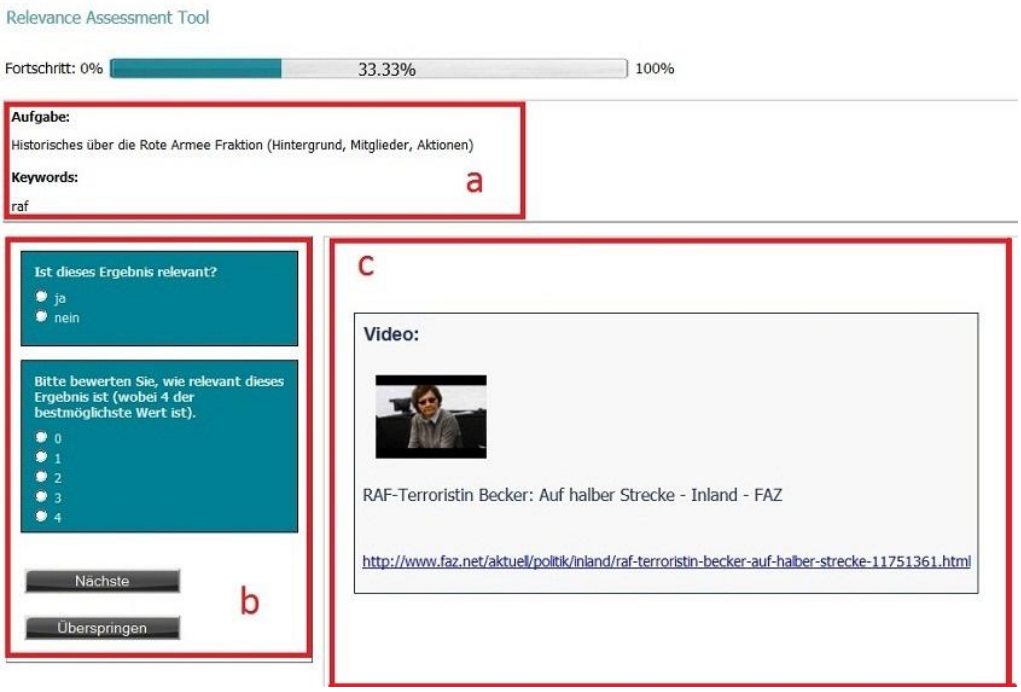


Abbildung 14: RAT - Bewertungsumgebung

Die Bewertungsumgebung war also in drei Bereiche aufgeteilt: Im Bereich a war die jeweilige Suchanfrage mit dem zugehörigen Informationsbedürfnis zu sehen. Im Bereich c wurden nacheinander die Beschreibungen/Vorschauen und Treffer angezeigt (zu sehen ist hier eine Videobeschreibung), und zwar anonymsiert und gemischt (wobei die Beschreibung/ Vorschau-Treffer-Reihenfolge beibehalten wurde), und im Bereich b fand die Bewertung statt; binär und mittels Fünfer-Skala.

3.2.5 Juroren

In diesem Abschnitt wird auf die Juroren des Tests eingegangen.

Der Test wurde anonym durchgeführt und umfasste 50 Aufgaben. Jedoch wurden diese nicht zwangsläufig von verschiedenen Personen bearbeitet, denn eine Person konnte beliebig viele Aufgaben erledigen. Eine vollständige Rekonstruktion dessen hätte (bei korrekter Angabe) anhand der demografischen Daten angestellt werden können, jedoch wurden diese aufgrund eines Skriptfehlers nicht korrekt in der Datenbank abgespeichert, weswegen sie nicht ausgewertet werden konnten.

Obwohl die Identität der Juroren also unbekannt ist, kann davon ausgegangen werden, dass der Großteil der Aufgaben von Studierenden bearbeitet wurde.

3.3 Ergebnisse

In diesem Kapitel werden die Ergebnisse des Retrievaltests in Hinsicht auf die Forschungsfragen und -hypothesen präsentiert. Dabei wird zuerst auf die Datenbasis eingegangen.

3.3.1 Datenbasis

In diesem Abschnitt wird die Datenbasis des Tests erläutert, die bei der Auswertung zur Verfügung stand.

Insgesamt wurden den Juroren 2.600 zu bewertende Items (Beschreibungen/ Vorschauen, Treffer und Übersichten) vorgelegt.

211 davon stellten sich später als unbrauchbar heraus, weil sie übersprungen worden waren oder es Probleme beim Scraping gegeben hatte. Diese Fehlschläge verteilen sich auf die verschiedenen Itemarten.

Die Übersichten der Ergebnisse spezieller Kollektionen bewerten zu lassen, hatte den Zweck zu prüfen, ob sich Ergebnisse evtl. sehr ähneln. Dies war lediglich bei den Suchanfragen „21 gramm dvd“ und „esp ltd-alexi-200“ der Fall, die sehr ähnliche Bilder auslösten, wie in *Abbildung 15* zu sehen ist.

Abbildung 15: Sehr ähnliche Ergebnisse spezieller Kollektionen

Die Bewertung dieser Bilder fiel dementsprechend (fast) gleich aus.

Insgesamt wurden 101 dieser Übersichten zum Bewerten vorgelegt - sie wurden jedoch für die tatsächliche Auswertung nicht weiter verwendet.

Des Weiteren wurden 133 Bildervorschauen bewertet - diese Wertungen wurden dupliziert, um die Bewertungen der Bilder selbst zu simulieren, die ja lediglich die Vorschauen in höherer Auflösung darstellen.

Nach Abzug der Fehlschläge und Übersichten, sowie dem Hinzurechnen der simulierten Bilderbewertungen ergab sich letztlich eine Datenbasis von 2.425 Items.

Den Löwenanteil davon stellen mit ca. 75 Prozent die regulären Ergebnisse dar (Beschreibungen und Treffer). Die nächstgrößere Menge stellen Bilder mit ca. zehn Prozent Anteil. Die restlichen speziellen Kollektionen sind mit jeweils knapp fünf Prozent vertreten. Diese Zusammensetzung ist in *Abbildung 16* im Detail dargestellt.

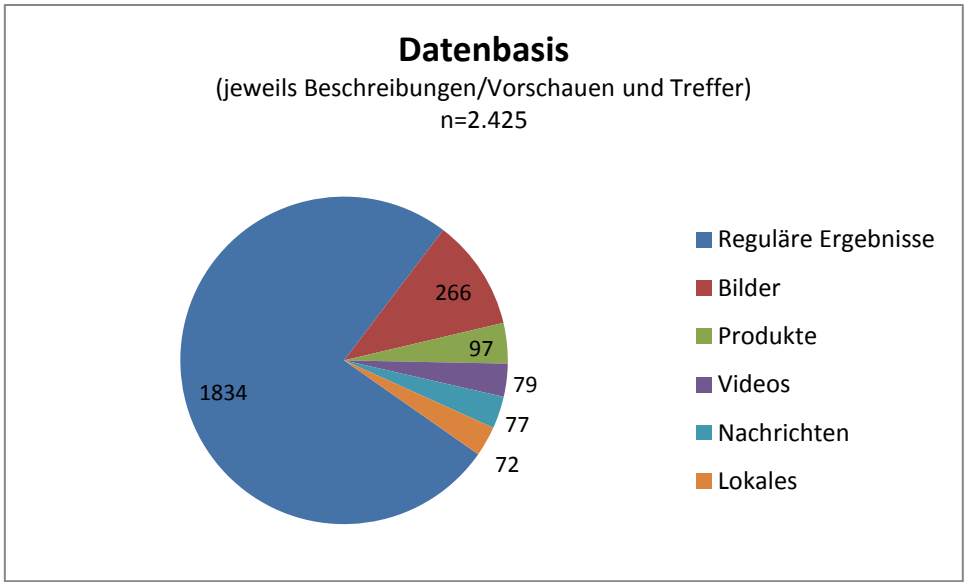


Abbildung 16: Datenbasis

Dieses Bild lässt sich freilich aufgrund des erheblich geringeren Testumfangs und der Einbeziehung von Bing nicht direkt mit dem von BACHOR in *Tabelle 6* gezeichneten vergleichen, doch stimmt immerhin die Reihenfolge der Häufigkeit des Auftretens überein; am meisten sind auf der ersten Ergebnisseite Bilder vertreten, dann folgen Produkte, Videos, Nachrichten und zum Schluss lokale Resultate.

Beim Betrachten der einzelnen Auswertungen ist zu beachten, dass Bing generell weniger Resultate spezieller Kollektionen ausgibt als Google.

Abweichungen der Grundgesamtheiten sind den Fehlschlägen zu schulden.

Nachdem die Entstehung und die Zusammensetzung der Datenbasis erläutert wurde, wird in den folgenden Kapiteln auf die Einzelauswertungen eingegangen, die mit den Forschungsfragen und –hypothesen korrespondieren. Dabei werden Ergebnisse herausgestellt, die direkt mit ihnen korrespondieren. Am Ende der jeweiligen Auswertungen werden diese Ergebnisse für eine zusammenfassende Beantwortung der jeweiligen Forschungshypothese resp. –frage verwendet.

Der Diskussionsteil (Kapitel *3.3.6 Diskussion*) spiegelt die Struktur der einzelnen Auswertungskapitel, und dort werden die gewonnenen Erkenntnisse jeweils noch einmal tabellarisch vorangestellt.

3.3.2 Reguläre Ergebnisse

Dieses Kapitel ist das erste von zweien, das der Beantwortung der ersten Forschungsfrage gewidmet ist (s. Kapitel *3.1 Forschungsfragen und -hypothesen*). Es werden jene Untersuchungen angestellt, die nötig sind, um die erste Forschungshypothese zu beantworten. Dafür werden die regulären Ergebnisse in Hinsicht auf ihre Relevanz ausgewertet. Im Verlauf des Kapitels werden Ergebnisse herausgestellt, die direkt mit der Hypothese korrespondieren.

Die Forschungshypothese wird am Ende des Kapitels beantwortet.

Abbildung 17 zeigt die binäre Relevanzbewertung der regulären Beschreibungen und Treffer, *Abbildung 18* die mittels der Fünfer-Skala.

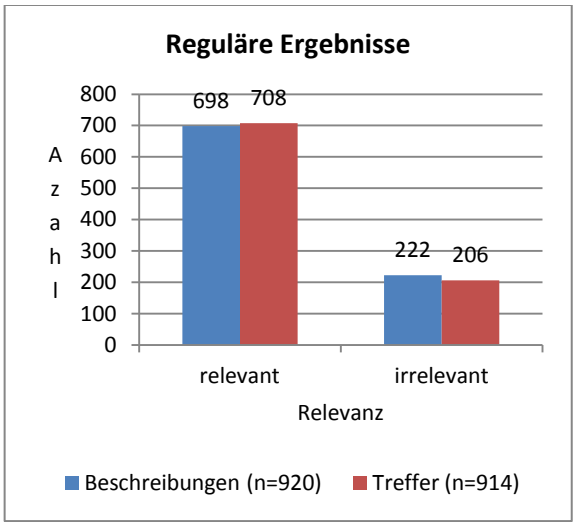


Abbildung 17: Relevanz reguläre Ergebnisse binär

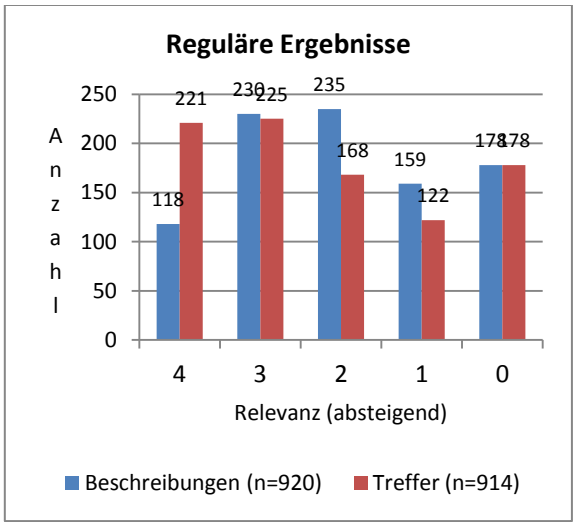


Abbildung 18: Relevanz reguläre Ergebnisse differenziert

Die binäre Sicht zeigt, dass das Verhältnis von relevanten Beschreibungen und Treffern zu irrelevanten ungefähr drei zu eins beträgt. Es gibt knapp zwei Prozent mehr relevante Treffer als Beschreibungen und ebenfalls knapp zwei Prozent weniger irrelevante Treffer als Beschreibungen.

Im Detail zeigt sich, dass deutlich weniger Beschreibungen als Treffer mit der besten Bewertung (4) versehen wurden. In den anderen Fällen sind die Beschreibungen relevanter als die Treffer, und bei der niedrigsten Bewertung (0) stimmen sie überein.

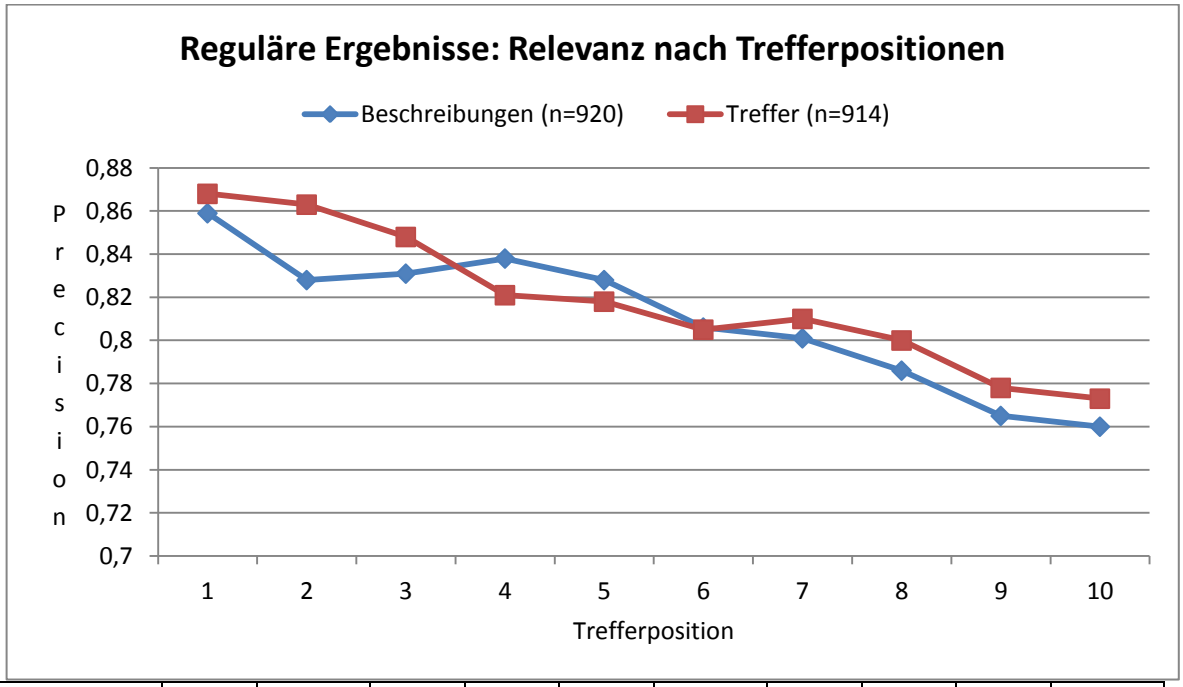
Sowohl für die Beschreibungen als auch für die Treffer ist der Durchschnitt die Relevanzbewertung 2 (≈ „neutral“) (Median = 460 bzw. 457).

Ergebnis 1: Binär: Es gibt knapp 2% mehr relevante reguläre Treffer als Beschreibungen.

Ergebnis 2: Differenziert: Es wurden ca. doppelt so viele reguläre Treffer wie Beschreibungen als „relevant" bewertet.

Ergebnis 3: Differenziert: Es wurden verschwindend wenig mehr reguläre Beschreibungen als Treffer als „eher relevant" bewertet.

In *Abbildung 19* ist das Verhältnis der relevanten zu allen erfassten regulären Ergebnissen der berücksichtigten Trefferpositionen dargestellt.



Beschreibungen	0,859	0,828	0,831	0,838	0,828	0,806	0,801	0,786	0,765	0,76
Treffer	0,868	0,863	0,848	0,821	0,818	0,805	0,81	0,8	0,778	0,773

Abbildung 19: Reguläre Ergebnisse nach Trefferpositionen

Die Trefferpositionen lassen sich in drei Bereiche einteilen: Auf den ersten dreien sind die Treffer relevanter als die Beschreibungen. Bis Trefferposition vier hat sich das umgekehrt und hält bis zur sechs. Dann sind bis zur zehnten Trefferposition wieder die Treffer relevanter.

Das Verhältnis ist (beinahe) optimal, also gleich, zwischen der dritten und vierten, und auf der sechsten Trefferposition. Die größte Lücke (0,035 Prozent) klafft bei der zweiten. Auf den ersten zehn Trefferpositionen büßen die Beschreibungen etwas mehr Relevanz ein als die Treffer (0,099 bzw. 0,095 Prozent).

<u>Ergebnis 4</u>: Bzgl. der Precision hinsichtlich der Trefferpositionen schneiden meistens die regulären Treffer, manchmal die Beschreibungen besser ab.

Für *Tabelle 11* wurden die Bewertungen der regulären Ergebnisse in Hinsicht auf die vier Beschreibung-Treffer-Fälle (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*) ausgewertet.

Tabelle 11: Binäre Relevanzmatrix reguläre Ergebnisse

	Beschreibung	**Treffer**	**Google & Bing**
a	relevant	relevant	621
b	relevant	irrelevant	65
c	irrelevant	relevant	79
d	irrelevant	irrelevant	129
e	alle ausgegeben Paare		894

Es zeigt sich, dass beidseitig relevante Ergebnispaare mit ca. 70 Prozent den größten Anteil ausmachen. Den nächstgrößeren stellen mit ungefähr 15 Prozent beidseitig irrelevante Paare. In diesen beiden Fällen erfüllen die Beschreibungen ihren Zweck, indem sie die Relevanz des Treffers korrekt abbilden. In den restlichen beiden Fällen, die gemeinsam ca. 15 Prozent Anteil ausmachen, ist dies nicht so, und der Nutzer wird zu irrelevanten Informationen geleitet bzw. es bleiben ihm relevante verborgen.

Anhand der Daten aus *Tabelle 11* wurden die Maße Description-result precision, Description-result conformance, Description fallout und Description desception (s. Kapitel *3.2.2 Retrievaleffektivität*) berechnet. Die Ergebnisse sind in *Tabelle 12* dargestellt.

Tabelle 12: Kennzahlberechnungen reguläre Ergebnisse

Maß	**Formel**	**Google & Bing**
Description-result precision	a/e	0,695
Description-result conformance	(a+d)/e	0,839
Description fallout	c/e	0,088
Description desception	b/e	0,073

Die Description-result precision zeigt, dass gut zwei Drittel der ausgegeben Trefferpaare beidseitig relevant sind.

Die Description-result conformance zeigt, dass ca. 84 Prozent der ausgegebenen Treffer-paare konsistent sind.

Der Description fallout zeigt, dass ungefähr neun Prozent der ausgegebenen Trefferpaare aus irrelevanter Beschreibung und relevantem Treffer bestehen.

Die Description desception zeigt, dass ca. sieben Prozent der ausgegebenen Trefferpaare aus relevanter Beschreibung und irrelevantem Treffer bestehen.

Ergebnis 5: Bei ca. 15% der regulären Ergebnispaare stimmt die Relevanzeinschätzung von Beschreibung und Treffer nicht überein. Dabei gibt es etwas mehr Fälle aus irrelevanter Beschreibung und relevantem Treffer als umgekehrt.

Zusammenfassend kann gesagt werden, dass die Relevanz regulärer Treffer durch ihre Beschreibungen in den meisten Fällen (ca. 84 Prozent) korrekt widergespiegelt wird.

Der der ersten Forschungshypothese zur ersten –frage zugrundeliegende Fall, dass die regulären Beschreibungen relevanter als die Treffer sind (s. Kapitel *3.1 Forschungsfragen und -hypothesen*), ist im Detail zwar vertreten, jedoch kann die These als solche nicht bestätigt werden, da in diesem Test im Durchschnitt die Treffer (geringfügig) relevanter als ihre Beschreibungen sind.

3.3.3 Ergebnisse nach Kollektionen

Dieses Kapitel ist das zweite, das der Beantwortung der ersten Forschungsfrage gewidmet ist (s. Kapitel *3.1 Forschungsfragen und -hypothesen*). Es werden jene Untersuchungen angestellt, die nötig sind, um die zweite Forschungshypothese zu beantworten. Dafür werden die Ergebnisse der speziellen Kollektionen in Hinsicht auf ihre Relevanz ausgewertet. Im Verlauf des Kapitels werden Ergebnisse herausgestellt, die direkt mit der Hypothese korrespondieren.

Die Forschungshypothese sowie die -frage werden am Ende des Kapitels beantwortet.

Abbildung 20 zeigt die binäre Relevanzbewertung der Bildervorschauen und -treffer, *Abbildung 21* die mittels der Fünfer-Skala. Zu beachten ist, dass die Bewertungen der Treffer von den Vorschauen gespiegel wurden (s. Kapitel *3.3.1 Datenbasis*).

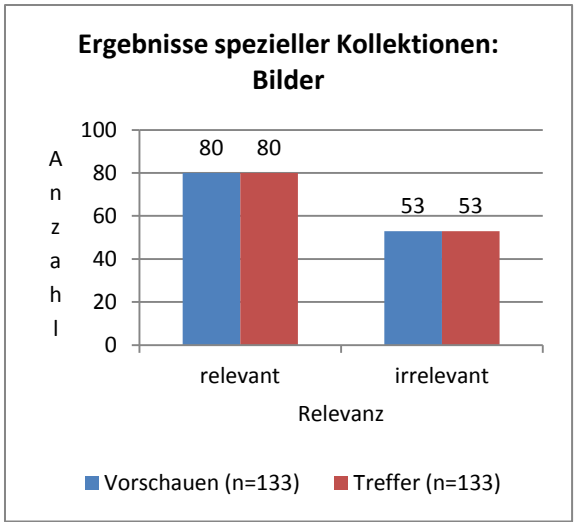


Abbildung 20: Ergebnisse spezieller Kollektionen: Bilder binär

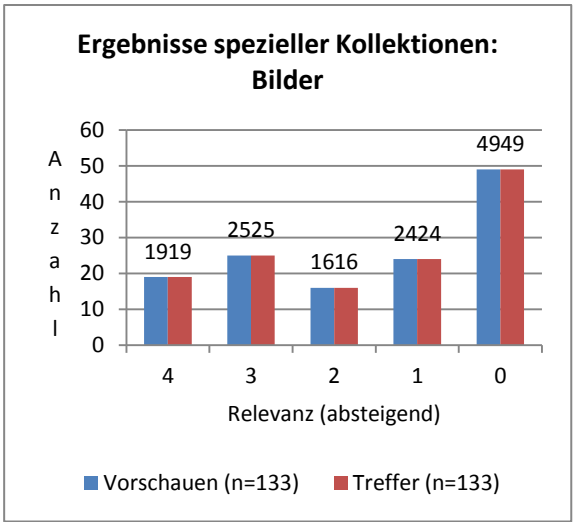


Abbildung 21: Ergebnisse spezieller Kollektionen: Bilder differenziert

Die binäre Sicht zeigt, dass fast 40 Prozent der Bilder als irrelevant eingestuft wurden.

Im Detail zeigt sich ebenfalls, dass die niedrigste Relevanzbewertung stark vertreten ist und fast doppelt so oft gewählt wurde wie die des nächstgrößeren Anteils (3 ≈ „eher relevant“). In Hinsicht auf die anderen Bewertungen wurden die „gemäßigten“, 3 und 1 (≈ „eher irrelevant“) öfter gewählt als 4 und 2. Der Durchschnitt der Relevanzbewertungen ist die 1 (Median = 66,5).

Durch die gespiegelten Vorschaubewertungen entfallen die Fälle, in denen die Bewertungen von Vorschauen und Treffern divergieren.

Die Description-result precision zeigt, dass 60 Prozent der ausgegebenen Trefferpaare beidseitig relevant sind.

Dementsprechend zeigt die Description desception, dass 40 Prozent der ausgegebenen Trefferpaare beidseitig irrelevant sind.

Abbildung 22 zeigt die binäre Relevanzbewertung der Produktbeschreibungen und -treffer, *Abbildung 23* die mittels der Fünfer-Skala.

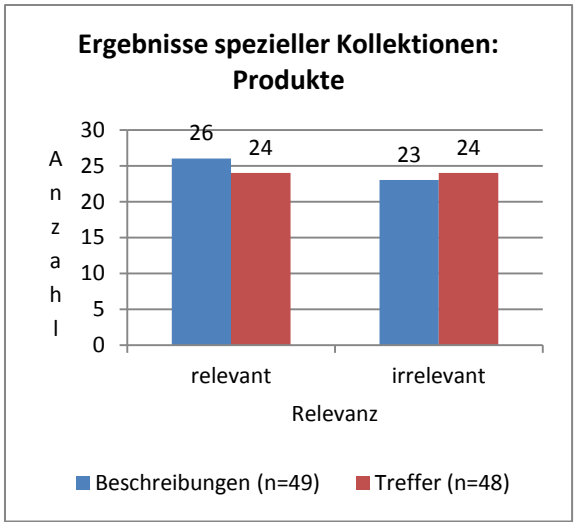


Abbildung 22: Ergebnisse spezieller Kollektionen: Produkte binär

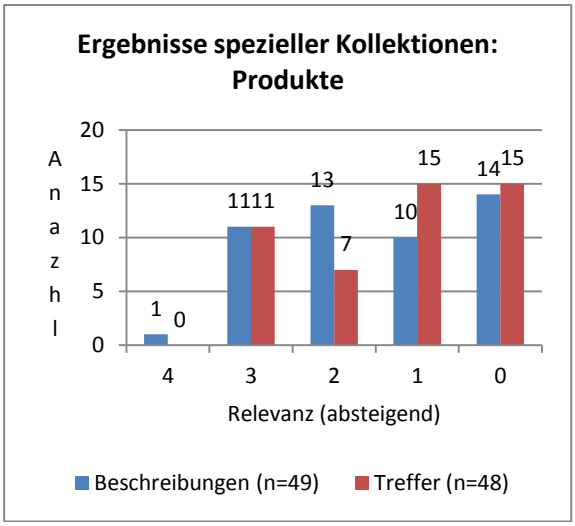


Abbildung 23: Ergebnisse spezieller Kollektionen: Produkte differenziert

Die binäre Sicht zeigt, dass es ungefähr so viele irrelevante wie relevante Resultate gibt und sich die Anteile ähneln. Es gibt lediglich etwas mehr relevante Beschreibungen als irrelevante.

Im Detail fällt auf, dass kein Treffer und nur eine Beschreibung die höchste Relevanzbewertung erhielten. Als „eher relevant“ wurden ungefähr genauso viele Beschreibungen wie Treffer bewertet.

Mit der neutralen Bewertung wurden deutlich mehr Beschreibungen als Treffer versehen. Dies kehrt sich bei den „(eher) irrelevant“-Wertungen um, indem dort mehr Treffer als Beschreibungen zu verzeichnen sind.

Der Durchschnitt der Relevanzbewertungen der Beschreibungen ist die 2, der der Treffer die 1 (Median = 24,5 bzw. 24).

<u>Ergebnis 1</u>: Binär: Produktbeschreibungen und –treffer ähneln sich in puncto Relevanz in hohem Maße (Differenz: 3%).

<u>Ergebnis 2</u>: Differenziert: Es wurden geringfügig mehr Produktbeschreibungen als „relevant“ und etwas mehr als „neutral“ als –treffer bewertet.

Für *Tabelle 13* wurden die Bewertungen der Produkte in Hinsicht auf die vier Beschreibung-Treffer-Fälle (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*) ausgewertet.

Tabelle 13: Binäre Relevanzmatrix Produkte

	Beschreibung	**Treffer**	**Google & Bing**
a	relevant	relevant	24
b	relevant	irrelevant	1
c	irrelevant	relevant	0
d	irrelevant	irrelevant	23
e	alle ausgegeben Paare		48

Es zeigt sich, dass sich die Paare in etwa gleichmäßig auf die beiden konsistenten Fälle verteilen.

Anhand der Daten aus *Tabelle 13* wurden einige Beschreibungsmaße (s. Kapitel *3.2.2 Retrievaleffektivität*) für Produkte berechnet. Die Ergebnisse sind in *Tabelle 14* dargestellt.

Tabelle 14: Kennzahlberechnungen Produkt-Ergebnisse

Maß	**Formel**	**Google & Bing**
Description-result precision	a/e	0,5
Description-result conformance	(a+d)/e	0,979
Description fallout	c/e	0
Description desception	b/e	0,021

Die Description-result precision zeigt, dass die Hälfte der ausgegebenen Trefferpaare beidseitig relevant ist.

Die Description-result conformance zeigt, dass ca. 98 Prozent der ausgegebenen Trefferpaare konsistent sind.

Der Description fallout zeigt, dass der Fall irrelevante Beschreibung und relevanter Treffer nicht vorkommt.

Die Description desception zeigt, dass ca. zwei Prozent der ausgegebenen Trefferpaare aus relevanter Beschreibung und irrelevantem Treffer bestehen.

Ergebnis 3: Fast alle Produktpaare sind konsistent (ca. 98 %).

Bzgl. der Hypothese 1.2 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) kann bestätigt werden, dass sich Produktbeschreibungen und -treffer hinsichtlich ihrer Relevanz in hohem Maße ähneln (Differenz: drei Prozent). Außerdem sind fast alle Produktpaare konsistent (ca. 98 Prozent).

Abbildung 24 zeigt die binäre Relevanzbewertung der Videobeschreibungen und -treffer, *Abbildung 25* die mittels der Fünfer-Skala.

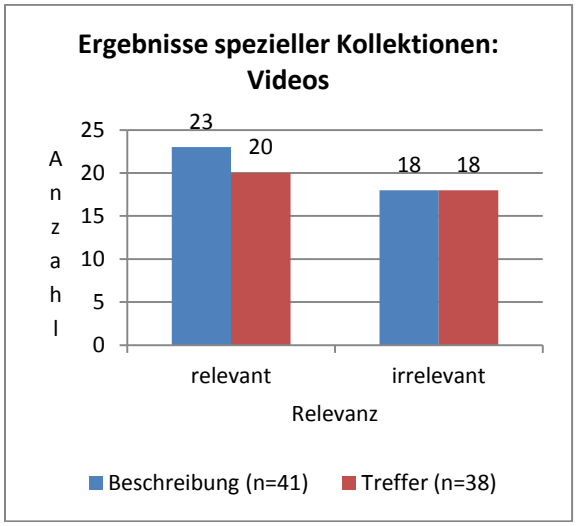


Abbildung 24: Ergebnisse spezieller Kollektionen: Videos binär

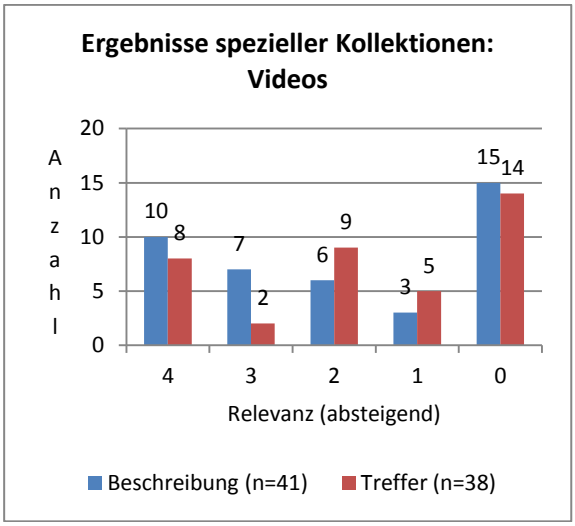


Abbildung 25: Ergebnisse spezieller Kollektionen: Videos differenziert

Die binäre Sicht zeigt, dass es etwas mehr relevante Beschreibungen als Treffer gibt (ca. 56 bzw. 53 Prozent), während etwas weniger Beschreibungen als Treffer als irrelevant eingestuft wurden (etwa 44 bzw. 47 Prozent).

In der detaillierten Sicht sind auf den beiden relevantesten Wertungsplätzen mehr Beschreibungen als Treffer vertreten, bei den Wertungen 2 und 1 ist es umgekehrt, und bei der „irrelevant"-Wertung sind wieder mehr Beschreibungen vorhanden. Auffällig ist, dass die höchsten Anzahlen von Beschreibungen und Treffern bei der negativsten Bewertung zu finden sind.

Der Durchschnitt der Relevanzbewertungen sowohl der Beschreibungen als auch der Treffer ist die 2 (Median = 20,5 bzw. 19).

Ergebnis 4: Binär: Videobeschreibungen und –treffer ähneln sich in puncto Relevanz hohem Maße (Differenz: 3%).

Ergebnis 5: Differenziert: Es wurden etwas mehr Videobeschreibungen als „(eher) relevant" und „irrelevant" als –treffer bewertet.

Für *Tabelle 15* wurden die Bewertungen der Videos in Hinsicht auf die vier Beschreibung-Treffer-Fälle (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*) ausgewertet.

Tabelle 15: Binäre Relevanzmatrix Videos

	Beschreibung	**Treffer**	**Google & Bing**
a	relevant	relevant	17
b	relevant	irrelevant	1
c	irrelevant	relevant	0
d	irrelevant	irrelevant	17
e	alle ausgegeben Paare		35

Es zeigt sich, dass die Paare sich in etwa gleichmäßig auf die beiden konsistenten Fälle verteilen.

Anhand der Daten aus *Tabelle 15* wurden einige Beschreibungsmaße (s. Kapitel *3.2.2 Retrievaleffektivität*) für Videos berechnet. Die Ergebnisse sind in *Tabelle 16* dargestellt.

Tabelle 16: Kennzahlberechnungen Video-Ergebnisse

Maß	**Formel**	**Google & Bing**
Description-result precision	a/e	0,486
Description-result conformance	(a+d)/e	0,971
Description fallout	c/e	0
Description desception	b/e	0,029

Die Description-result precision zeigt, dass knapp die Hälfte der ausgegebenen Trefferpaare beidseitig relevant ist.

Die Description-result conformance zeigt, dass fast alle ausgegebenen Trefferpaare konsistent sind.

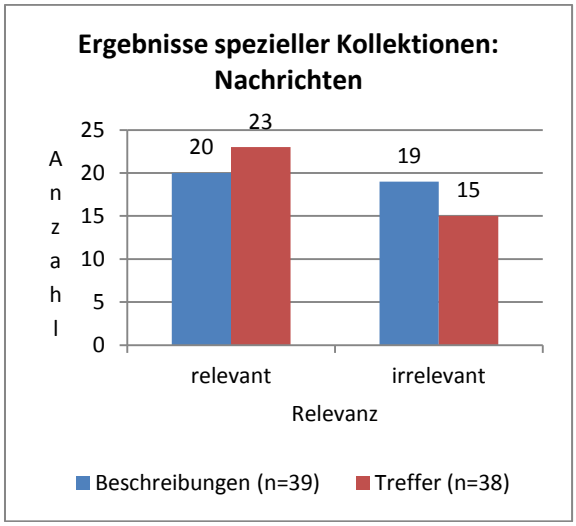
Der Description fallout zeigt, dass der Fall irrelevante Beschreibung und relevanter Treffer nicht vorkommt.

Die Description desception zeigt, dass fast drei Prozent der ausgegebenen Trefferpaare aus relevanter Beschreibung und irrelevantem Treffer bestehen.

Ergebnis 6: Fast alle Videopaare sind konsistent (ca. 97%).

Bzgl. der Hypothese 1.2 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) kann bestätigt werden, dass die Beschreibungen von Videos im Durchschnitt relevanter als die Treffer sind, jedoch nicht viel (Differenz: gut drei Prozent). Außerdem sind fast alle Videopaare konsistent (ca. 97%).

Abbildung 26 zeigt die binäre Relevanzbewertung der Nachrichtenbeschreibungen und -treffer, *Abbildung 27* die mittels der Fünfer-Skala.



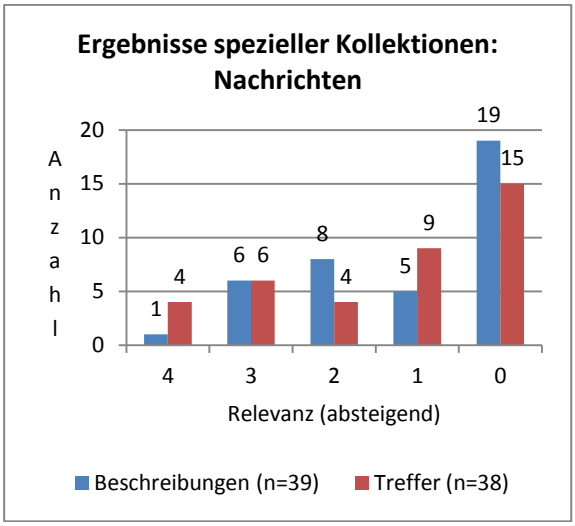


Abbildung 26: Ergebnisse spezieller Kollektionen: Nachrichten binär

Abbildung 27: Ergebnisse spezieller Kollektionen: Nachrichten differenziert

Die binäre Sicht zeigt, dass es etwas mehr relevante Treffer als Beschreibungen gibt, und etwas weniger irrelevante Treffer als Beschreibungen.

Im detaillierten Bild wurde bzgl. der Beschreibungen nur einmal die beste Bewertung vergeben. Etwa gleich viele Beschreibungen und Treffer wurden mit der „eher relevant"-Wertung versehen. Als „neutral" wurden fast doppelt soviele Beschreibungen wie Treffer eingeschätzt. Dies kehrt sich bei der Wertung „eher irrelevant" um. Als „irrelevant" wurden mehr Beschreibungen als Treffer eingestuft – von beiden Ergebnisarten sind hier die jeweils höchsten Anzahlen zu finden.

Der Durchschnitt der Relevanzbewertungen sowohl der Beschreibungen als auch der Treffer ist die 1 (Median = 19,5 bzw. 19).

Ergebnis 7: Binär: Es sind etwas weniger Nachrichtenbeschreibungen als –treffer relevant (ca. 9%).

Ergebnis 8: Differenziert: Es gibt etwas mehr als „neutral" und „irrelevant" bewertete Nachrichtenbeschreibungen als –treffer.

Für *Tabelle 17* wurden die Bewertungen der Nachrichten in Hinsicht auf die vier Beschreibung-Treffer-Fälle (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*) ausgewertet.

Tabelle 17: Binäre Relevanzmatrix Nachrichten

	Beschreibung	**Treffer**	**Google & Bing**
a	relevant	relevant	20
b	relevant	irrelevant	0
c	irrelevant	relevant	4
d	irrelevant	irrelevant	15
e	alle ausgegeben Paare		39

Es zeigt sich, dass ca. die Hälfte der Paare beidseitig relevant ist. Der Großteil der anderen Hälfte ist beidseitig irrelevant. Der Rest der Paare besteht aus irrelevanter Beschreibung und relevantem Treffer.

Anhand der Daten aus *Tabelle 17* wurden einige Beschreibungsmaße (s. Kapitel *3.2.2 Retrievaleffektivität*) für Nachrichten berechnet. Die Ergebnisse sind in *Tabelle 18* dargestellt.

Tabelle 18: Kennzahlberechnungen Nachrichten-Ergebnisse

Maß	**Formel**	**Google & Bing**
Description-result precision	a/e	0,513
Description-result conformance	(a+d)/e	0,897
Description fallout	c/e	0,103
Description desception	b/e	0

Die Description-result precision zeigt, dass ca. 50 Prozent der ausgegebenen Trefferpaare beidseitig relevant sind.

Die Description-result conformance zeigt, dass fast 90 Prozent der ausgegebenen Trefferpaare konsistent sind.

Der Description-fallout zeigt, dass ungefähr zehn Prozent der ausgegebenen Trefferpaare aus irrelevanter Beschreibung und relevantem Treffer bestehen.

Die Description desception zeigt, dass der Fall relevante Beschreibung und irrelevanter Treffer nicht vorkommt.

Ergebnis 9: Die überwiegende Mehrheit der Nachrichtenpaare ist konsistent (ca. 90%).

Bzgl. der Hypothese 1.2 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) kann nicht bestätigt werden, dass die Beschreibungen von Nachrichten im Durchschnitt relevanter als die Treffer sind; das Gegenteil ist der Fall (Differenz: ca. neun Prozent). Außerdem ist die überwiegende Mehrheit der Nachrichtenpaare konsistent (ca. 90 Prozent).

Abbildung 28 zeigt die binäre Relevanzbewertung der Beschreibungen und Treffer lokaler Ergebnisse, *Abbildung 29* die mittels der Fünfer-Skala.

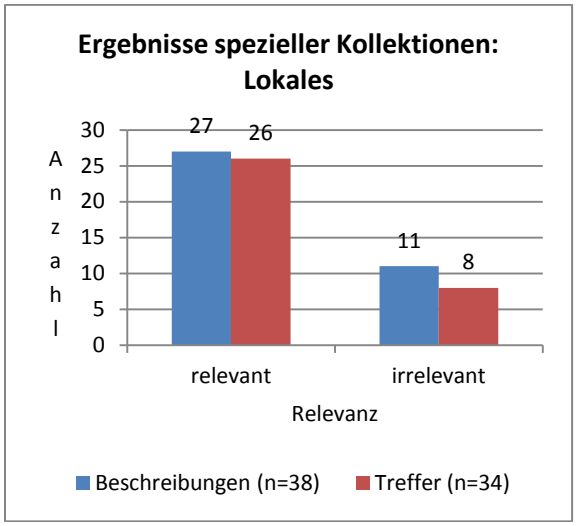


Abbildung 28: Ergebnisse spezieller Kollektionen: Lokales binär

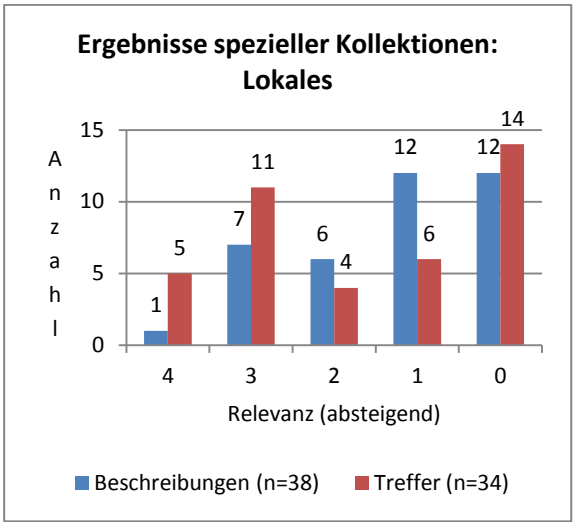


Abbildung 29: Ergebnisse spezieller Kollektionen: Lokales differenziert

Die binäre Sicht zeigt, dass deutlich mehr Beschreibungen und Treffer als relevant als irrelevant eingeschätzt wurden. Dabei sind etwas mehr Treffer als Beschreibungen relevant (ca. 77 bzw. 71 Prozent), und etwas mehr Beschreibungen als Treffer irrelevant (etwa 29 bzw. 24 Prozent).

Im detaillierten Bild ist der Anteil der als „relevant“ bewerteten Beschreibungen sehr niedrig.

Bei der „eher relevant“-Wertung tummeln sich deutlich mehr Treffer als Beschreibungen. Bei den Relevanzbewertungen 2 und 1 sind jeweils mehr Beschreibungen als Treffer verzeichnet. Dies dreht sich bei der „irrelevant“-Wertung wieder um, bei der verhältnismäßig viele Beschreibungen und Treffer anzutreffen sind.

Der Durchschnitt der Relevanzbewertungen der Beschreibungen ist die 1, der der Treffer die 2 (Median = 19 bzw. 17).

Ergebnis 10: Binär: Lokale Beschreibungen und Treffer divergieren in puncto Relevanz etwas (Differenz: ca. 6%).

Ergebnis 11: Differenziert: Es wurden im positiven Bereich etwas weniger lokale Beschreibungen als Treffer bewertet, im neutralen und negativen etwas mehr.

Für *Tabelle 19* wurden die Bewertungen der Ergebnisse in Hinsicht auf die vier Beschreibung-Treffer-Fälle (s. Kapitel *2.2.7.1 Einbeziehung von Trefferbeschreibungen*) ausgewertet.

Tabelle 19: Binäre Relevanzmatrix lokale Ergebnisse

	Beschreibung	**Treffer**	**Google & Bing**
a	relevant	relevant	22
b	relevant	irrelevant	2
c	irrelevant	relevant	3
d	irrelevant	irrelevant	5
e	alle ausgegeben Paare		32

Es zeigt sich, dass die Mehrheit der Paare (knapp 70 Prozent) beidseitig relevant ist. Es gibt ca. 15 Prozent beidseitig irrelevante Paare und vereinzelte Fälle von irrelevanter Beschreibung und relevantem Treffer, bzw. relevanter Beschreibung und irrelevantem Treffer.

Anhand der Daten aus *Tabelle 19* wurden einige Beschreibungsmaße (s. Kapitel *3.2.2 Retrievaleffektivität*) für lokale Ergebnisse berechnet. Die Ergebnisse sind in *Tabelle 20* dargestellt.

Tabelle 20: Kennzahlberechnungen lokale Beschreibungen

Maß	**Formel**	**Google & Bing**
Description-result precision	a/e	0,688
Description-result conformance	(a+d)/e	0,844
Description fallout	c/e	0,094
Description desception	b/e	0,063

Die Description-result precision zeigt, dass knapp 70 Prozent der ausgegebenen Trefferpaare beidseitig relevant sind.

Die Description-result conformance zeigt, dass ca. 85 Prozent der ausgegebenen Trefferpaare konsistent sind.

Der Description-fallout zeigt, dass ungefähr neun Prozent der ausgegebenen Trefferpaare aus irrelevanter Beschreibung und relevantem Treffer bestehen.

Die Description desception zeigt, dass ca. sechs Prozent der ausgegebenen Trefferpaare aus relevanter Beschreibung und irrelevantem Treffer bestehen.

Ergebnis 12: Der Großteil der lokalen Paare ist konsistent (ca. 85%).

Bzgl. der Hypothese 1.2 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) sollte eher gesagt werden, dass Beschreibungen und Treffer lokaler Ergebnisse etwas divergieren (Differenz: ca. sechs Prozent). Außerdem ist der Großteil der lokalen Paare konsistent (ca. 85 Prozent).

Nachdem die Beschreibungen und Treffer der einzelnen Kollektionen untersucht wurden, werden nun in *Tabelle 21* die Kollektionen in Hinsicht auf die berechneten Beschreibungsmaße miteinander verglichen. Der jeweils schlechteste Wert ist gefettet, der (die) beste(n) kursiviert.

Tabelle 21: Kennzahlberechnungen alle Kollektionen

	Bilder	Produkte	Videos	Nachrichten	Lokales
Description-result precision	0,6	0,5	**0,486**	0,513	*0,688*
Description-result conformance	-	*0,979*	0,971	0,897	**0,844**
Description fallout	-	*0*	*0*	**0,103**	0,094
Description desception	**0,4**	0,021	0,029	*0*	0,063

Bzgl. der Description-result precision schneiden Videos mit etwas weniger als der Hälfte beidseitig relevanter Paare am schlechtesten ab. Die in dieser Hinsicht beste Kollektion, lokale Ergebnisse, erreicht einen Wert von knapp 70 Prozent.

In Hinsicht auf die Description-result conformance stellen die lokalen Ergebnisse mit ca. 84 Prozent Fällen konsistenter Paare das Schlusslicht dar. Am besten stehen die Produkte mit ungefähr 98 Prozent da.

In Bezug auf den Description fallout schneiden Nachrichten mit ca. zehn Prozent Paaren mit irrelevanter Beschreibung und relevantem Treffer am schlechtesten ab. Als am besten stellen sich Produkte und Videos mit keinem solchen Fall heraus.

Hinsichtlich der Description desception stellen Bilder mit einem Anteil von 40 Prozent beidseitig irrelevanter Paare das Schlusslicht dar. Am besten stehen Nachrichten mit keinem solchen Fall da.

Abschließend ist zur Hypothese 1.2 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) Folgendes zu sagen: Die Anzahlen der als relevant bewerteten Beschreibungen und Treffer sind sich bei Produkten und lokalen Ergebnissen tatsächlich ziemlich ähnlich; die

Abweichungen betragen 3,1 bzw. 5,4 Prozent. Dass die Beschreibungen von Videos im Durchschnitt relevanter als die Treffer sind, kann bestätigt werden, jedoch beträgt die Differenz nur 3,5 Prozent, sodass sie eher als ebenfalls in hohem Maße ähnlich einzuordnen sind. Dass die Beschreibungen von Nachrichten im Durchschnitt relevanter als die Treffer sind, kann dagegen nicht bestätigt werden: Die Treffer sind relevanter als die Beschreibungen, und zwar mit einer Differenz von 9,2 Prozent.

In Zusammenfassung des vorigen und dieses Kapitels ist bzgl. der Forschungsfrage 1 zu sagen, dass sich die Relevanz von Suchergebnissen in den meisten Fällen korrekt in ihren Beschreibungen widerspiegelt. Dabei schneiden Resultate spezieller Kollektionen insgesamt (ohne Berücksichtigung von Bilderpaaren, deren Konsistenz vorausgesetzt wurde) mit durchschnittlich ca. 92 Prozent konsistenten Fällen besser ab als reguläre Ergebnisse, deren Durchschnitt diesbzgl. bei ungefähr 84 Prozent liegt.

Nachdem in diesem und dem vorangegangenen Kapitel die erste Forschungsfrage mit ihren –hypothesen untersucht wurde, wird sich im nächsten Abschnitt der zweiten zugewandt.

3.3.4 Ergebnisse nach Suchmaschinen

Dieses Kapitel ist der Beantwortung der zweiten Forschungsfrage und ihrer –hypothesen gewidmet (s. Kapitel *3.1 Forschungsfragen und -hypothesen*). Dafür werden die regulären Ergebnisse in Hinsicht auf ihre Relevanz nach ihrer jeweiligen Herkunftssuchmaschine ausgewertet, und anschließend wird der Effekt der Treffer der speziellen Kollektionen auf das Gesamtergebnis untersucht. Im Verlauf des Kapitels werden Ergebnisse herausgestellt, die direkt mit den Hypothesen korrespondieren.

Die erste Forschungshypothese wird zwischen den beiden Schritten beantwortet, die zweite zusammen mit der Forschungsfrage am Ende des Kapitels.

Abbildung 30 zeigt die binäre Relevanzbewertung der regulären Treffer nach Suchmaschinen, *Abbildung 31* die mittels der Fünfer-Skala.

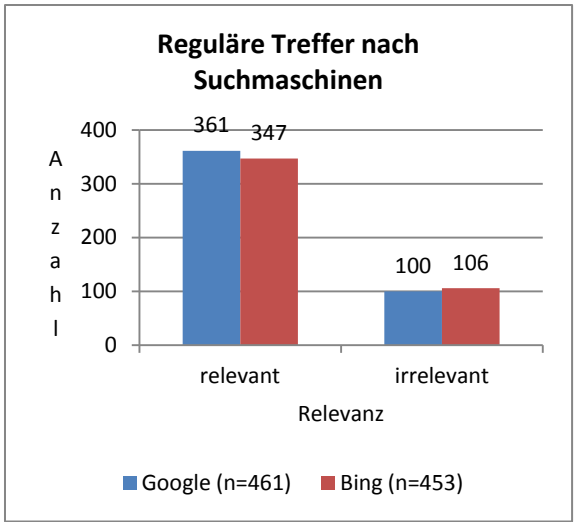


Abbildung 30: Reguläre Treffer nach Suchmaschinen binär

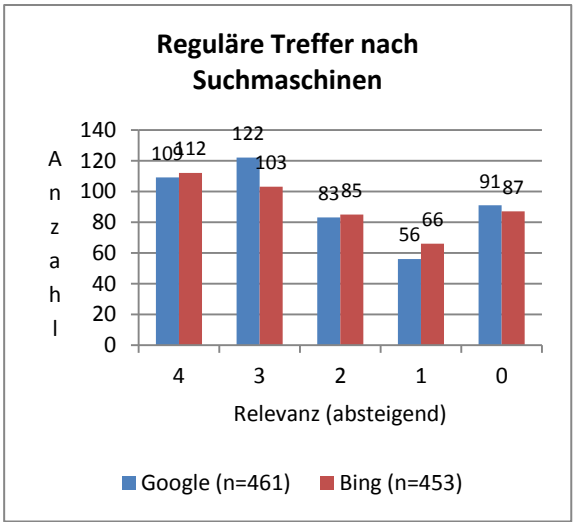


Abbildung 31: Reguläre Treffer nach Suchmaschinen differenziert

Die binäre Sicht zeigt, dass Google und Bing ähnlich abschneiden. Google liefert 1,7 Prozent mehr relevante Treffer und genauso viele irrelevante weniger. Das Verhältnis von relevanten zu irrelevanten Treffern beträgt bei beiden Suchmaschinen ungefähr 3,5 zu eins.

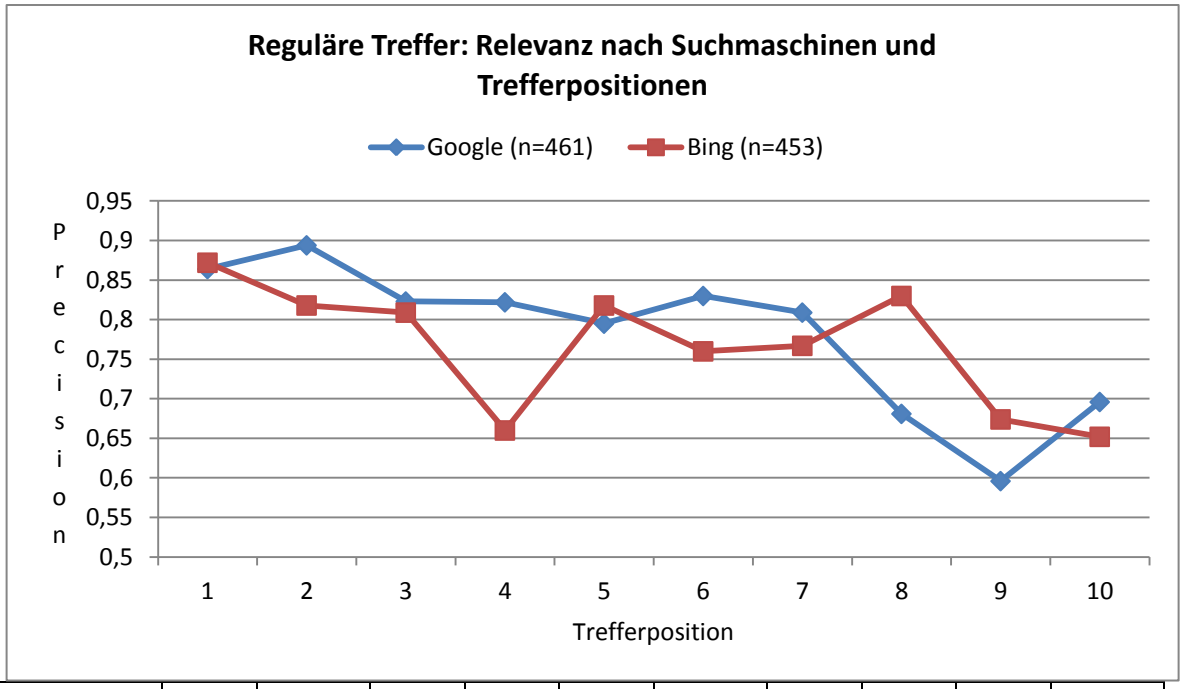
Im Detail ist interessant, dass Bings Treffer mehr beste und weniger schlechteste Wertungen als Googles erhielten. Bei der Wertung „eher relevant“ führt Google, während bei „neutral“ und „eher irrelevant“ Bing jeweils etwas mehr Treffer verzeichnet.

Sowohl für Googles als auch für Bings Treffer ist der Durchschnitt die Relevanzbewertung 3 (Median = 230,5 bzw. 226,5).

Ergebnis 1.1: Binär: Googles reguläre Treffer schneiden in puncto Relevanz besser ab als Bings, jedoch beträgt die Differenz nur 1,7%.

Ergebnis 1.2: Differenziert: Im positiven Bereich führt Google nur bei der „eher relevant“-Wertung; Differenz: 3,8%.

In *Abbildung 32* ist das Verhältnis der relevanten zu allen erfassten regulären Treffern der berücksichtigten Trefferpositionen nach Suchmaschinen dargestellt.



Google	0,864	0,894	0,823	0,822	0,795	0,83	0,809	0,681	0,596	0,696
Bing	0,872	0,818	0,809	0,66	0,818	0,76	0,767	0,83	0,674	0,652

Abbildung 32: Reguläre Treffer nach Suchmaschinen und Trefferpositionen

Auf die ersten sieben Trefferpositionen betrachtet, verlaufen die Graphen in Zweierschritten ungefähr gegenläufig: Sie starten mit einer ähnlich hohen Relevanz - bis zu Trefferposition 2 nimmt die von Google zu, die von Bing ab, bis zu Position 3 nehmen beide ab und treffen sich fast. Dieser Zyklus wiederholt sich ungefähr, indem Googles Relevanz fast gleich bleibt und Bings nach unten ausreißt. Beide Graphen treffen sich wieder bei Trefferposition 5, indem Googles Relevanz leicht abfällt und Bings stark ansteigt. Der Zyklus wiederholt sich ein weiteres Mal, indem Googles Relevanz erneut ansteigt und Bings abfällt. Die Graphen treffen sich ungefähr wieder bei der siebten Trefferposition, indem Googles Relevanz etwas ab- und Bings ansteigt. Ab hier lockert sich der Zyklus und kehrt sich um; während Google abfällt, steigt Bings Relevanz bis zur achten Trefferposition, um dann, wie Google, bis zur neunten erneut abzufallen. Bis zur zehnten Trefferposition kehrt sich der Zyklus wieder um.

Insgesamt verläuft Googles Relevanz-Graph, dessen einziger markanter Ausfall bei Trefferposition 9 festzustellen ist, gleichmäßiger als Bings, der bei den Trefferpositionen 4 und 8 stark ausreißt.

Während Google auf den ersten zehn Trefferpositionen 19,4 Prozent Relevanz einbüßt, sind es bei Bing 25,2. Die sowohl höchste als auch niedrigste Precision-Ausprägungen hat Google zu verzeichnen (Trefferposition 2 bzw. 9).

<u>Ergebnis 1.3</u>: Trefferpositionen: Google startet bei den regulären Treffern mit leicht geringerer Precision als Bing, liegt aber insgesamt öfter vorne, als zurückzuliegen. Sie schneidet um gut 4% besser ab als Bing und büßt auf den ersten 10 Positionen etwas weniger Relevanz ein (19,4 bzw. 25,2%).

Bzgl. der Hypothese 2.1 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) kann zwar bestätigt werden, dass Google in Bezug auf reguläre Treffer in puncto Relevanz besser als Bing abschneidet, jedoch beträgt die Differenz nur knapp zwei Prozent - Google und Bing schneiden hier also etwa gleich gut ab. In Trefferpositionensicht nimmt Googles Precision etwas weniger und gleichmäßiger als Bings ab, unterliegt aber auf den hinteren Positionen (7-9).

Abbildung 33 zeigt die binäre Relevanzbewertung aller Treffer (inkl. spezielle Kollektionen) nach Suchmaschinen, *Abbildung 34* die mittels der Fünfer-Skala.

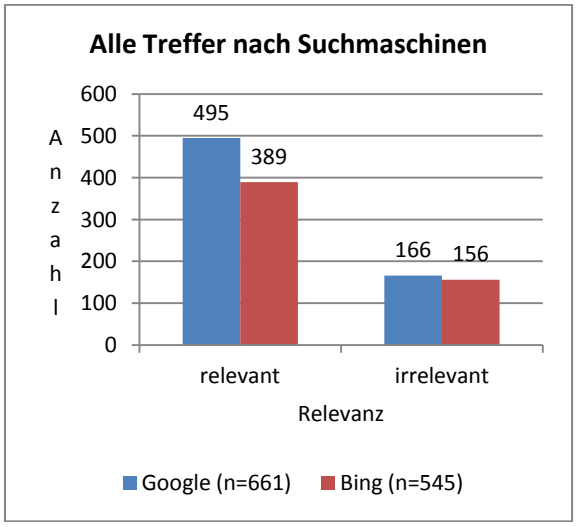


Abbildung 33: Alle Treffer nach Suchmaschinen binär

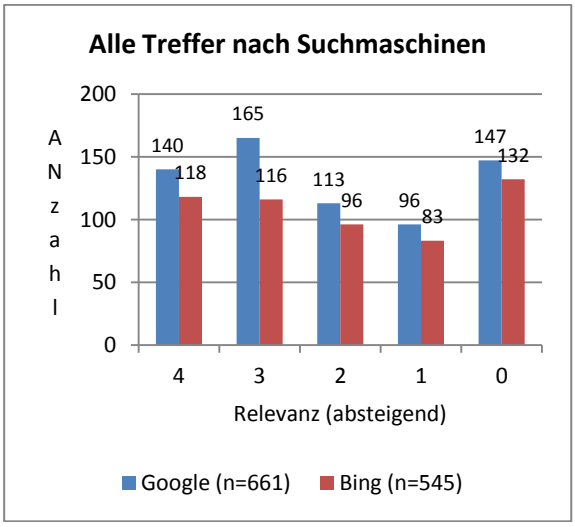


Abbildung 34: Alle Treffer nach Suchmaschinen differenziert

Unter Berücksichtigung der unterschiedlichen Grundgesamtheiten zeigt die binäre Sicht, dass Google dreieinhalb Prozent mehr relevante Treffer als Bing ausgibt und genauso viele irrelevante weniger. Das Verhältnis von relevanten zu irrelevanten Treffern liegt bei beiden Suchmaschinen in etwa bei drei zu eins.

Ebenfalls unter Berücksichtigung der unterschiedlichen Grundgesamtheiten zeigt die differenzierte Sicht, dass etwas mehr Bing-Treffer als „relevant“ bewertet wurden, wohingegen bei der „eher relevant“- und der „neutral“-Wertung Google führt. „Eher irrele-

vante“ Treffer weisen beide Suchmaschinen etwa in gleicher Höhe auf, und Bings Treffer wurden öfter als „irrelevant“ klassifiziert als Googles.

Sowohl für Googles als auch für Bings Treffer ist der Durchschnitt die Relevanzbewertung 2 (Median = 330,5 bzw. 272,5), was eine Verschlechterung von je einem Schritt zu den Durchschnitten bzgl. der nur regulären Treffer bedeutet.

Ergebnis 2.1: Binär: Googles gesamte Treffer schneiden in puncto Relevanz ein wenig besser ab als Bings, die Differenz beträgt 3,5%.

Ergebnis 2.2: Binär: Bei der Anreicherung der regulären Treffer mit Treffern spezieller Kollektionen werden bei beiden Suchmaschinen prozentual weniger relevante Treffer ausgegeben (Differenz Google: gut 3%, Bing: gut 5%)

Ergebnis 2.3: Differenziert: Googles Treffer erhalten mehr „eher relevant“-Wertungen als Bings (Differenz: knapp 4%).

Ergebnis 2.4: Differenziert: Bei der Anreicherung der regulären Treffer mit Treffern spezieller Kollektionen schneidet Google bei den Wertungen „(eher) relevant“ merkbar besser als Bing ab.

In *Abbildung 36* ist das Verhältnis der relevanten zu allen erfassten Treffern (also inkl. US/BS) der berücksichtigten Trefferpositionen nach Suchmaschinen dargestellt. Dafür wurden die Treffer spezieller Kollektionen in die Ergebnislisten eingerechnet. Dies ist in *Abbildung 35* beispielhaft dargestellt.

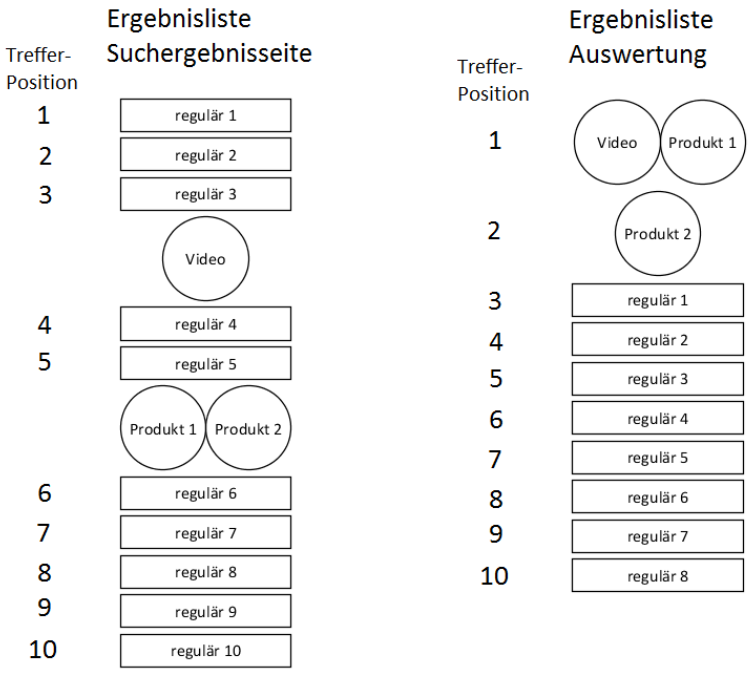
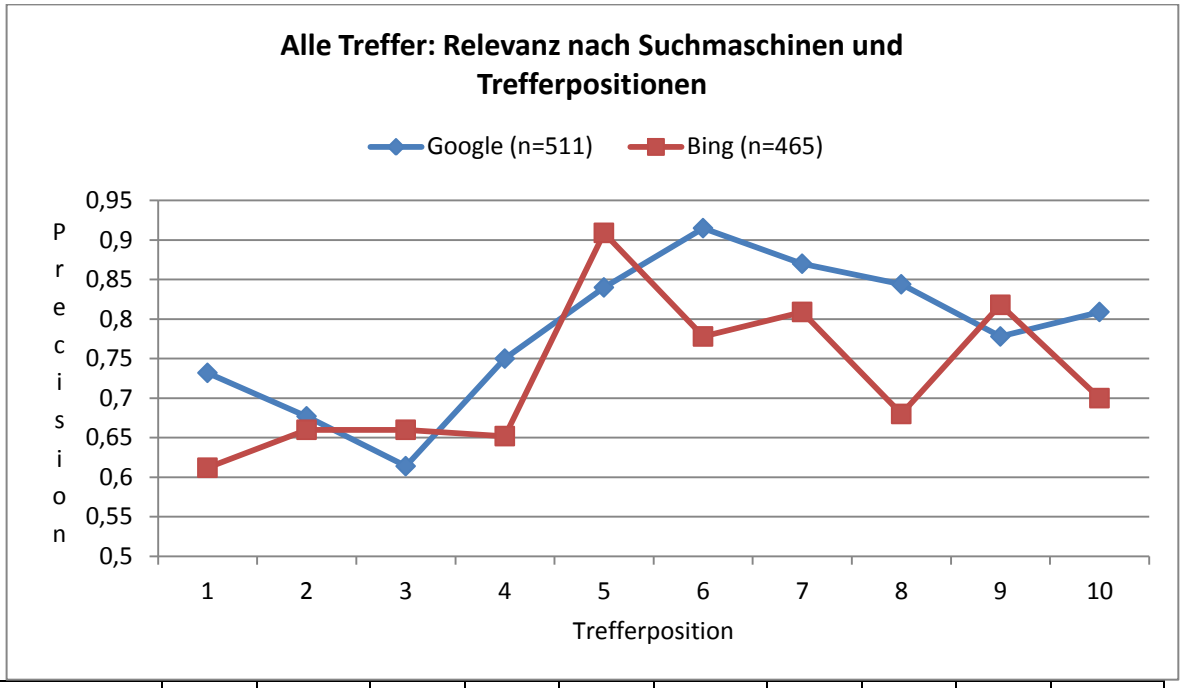


Abbildung 35: Auswertung: Einbeziehung US/BS-Treffer

In diesem Bsp. wurden also zusätzlich ein Video und zwei Produkte ausgegeben. Die Einreihung dieser Resultate in die Auswertungs-Ergebnisliste erfolgte anhand ihrer Positionen innerhalb ihres US/BS-Moduls (Ergebnisse spezieller Kollektionen werden i.d.R. als Gruppe ausgegeben). So stellen also Video (1) und Produkt 1 die Besetzung der ersten Trefferposition dar, Produkt 2 die der zweiten, und die regulären Treffer rutschen auf die folgenden Positionen, bis die zehn voll besetzt sind.

So wird versucht, der erhöhten Klickwahrscheinlichkeit solcher Ergebnisse durch oft vorhandene Hervorhebungen, z.B. Bilder, Rechnung zu tragen, wie LEWANDOWSKI es empfiehlt (s. Kapitel *2.2.7.3 Rahmenwerk zur Messung von Suchmaschinen-Retrieval-effektivität nach LEWANDOWSKI*).

Der Effekt auf die Precision der Suchmaschinen ist in *Abbildung 36* dargestellt.



Google	0,732	0,677	0,614	0,75	0,84	0,915	0,87	0,844	0,778	0,809
Bing	0,612	0,66	0,66	0,652	0,909	0,778	0,809	0,68	0,818	0,7

Abbildung 36: Alle Treffer nach Suchmaschinen und Trefferpositionen

Aufgrund der Auswertungsmethode ist die Dichte an Treffern der speziellen Kollektionen auf der ersten Trefferposition also am höchsten und nimmt im weiteren Verlauf ab. Einen deutlichen Schnitt gibt es von der vierten zur fünften Position, da Bilder (die am häufigsten unter den US/BS-Ergebnissen vertreten sind) i.d.R. zu viert ausgegeben werden. Die Kollektion mit den meisten Ergebnisexemplaren ist Produkte, deren Ergebnisse vermehrt zu siebt ausgegeben wurden. Ab der achten Trefferposition handelte es sich also um reine reguläre Treffer.

Die Graphen starten gleich mit einer schlechten Precision als bei den nur regulären Treffern (s. *Abb. 32*); Google mit ca. 13,5 Prozent weniger, Bing mit ungefähr 30. Bzgl. Bing stellt der Precisionwert der ersten Trefferposition den niedrigsten überhaupt dar.

Bis zur zweiten Trefferposition nähern sich die Precisionwerte einander an, indem Googles sich verschlechtert und Bings sich verbessert. Bis zur dritten verschlechtert sich Google weiter bis zu seinem Tiefstand, während Bings Wert konstant bleibt. Bis zur vierten Position verbessert sich Googles Precision erheblich, während Bings etwas abfällt. Bis zur fünften verbessert sich Googles Precision weiter, während Bing einen einzigartigen Anstieg auf ihren Hochstand verzeichnet. Hier befindet sich grob der Schnitt, der die Menge der durch US/BS dominierten Treffern von den eher regulär dominierten trennt; besonders für Bing eine erhebliche Verbesserung in puncto Relevanz.

Bis zur sechsten Trefferposition verzeichnet Google ihren Hochstand, während Bing sich ungefähr in der Mitte zwischen ihrem Hoch- und Tiefpunkt einpendelt. Bis zur zehnten Trefferposition büßt Google ab hier gut elf und Bing ungefähr zehn Prozent an Relevanz ein; Google ziemlich gleichmäßig mit leichtem Ausreißen bei Trefferposition 9, Bing wieder (s. *Abb. 32*) mehr im Zickzack.

Im Gegensatz zu der Reguläre-Treffer-Sicht (s. *Abb. 32*) gewinnen Google und Bing hier insgesamt Relevanz hinzu (10,5 bzw. 14,4 Prozent), anstatt welche einzubüßen, und stehen am Ende mit einer höheren Precision da als bei den regulären Treffern (Google: ca. 0,8 statt 0,7, Bing: ungefähr 0,7 statt 0,65).

Die höchste Precision-Ausprägung hat Google zu verzeichnen (Position 6; 91,5 Prozent), die niedrigste Bing (Position 1; 61,2 Prozent).

Ergebnis 2.5: Trefferpositionen: Google startet mit einer höheren Precision und liegt insgesamt deutlich öfter vorne, als zurückzuliegen. Sie schneidet um gut 10% besser ab als Bing und gewinnt auf den ersten 10 Positionen etwas weniger Relevanz hinzu als Bing (gut 10 bzw. gut 14%).

Ergebnis 2.6: Trefferpositionen: Vereinzelt können Treffer spezieller Kollektionen die Precision verbessern, doch je höher ihre Dichte ist, desto niedriger ist die Precision.

Bzgl. der Hypothese 2.2 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) kann zwar bestätigt werden, dass Google in Bezug auf alle Treffer (inkl. speziellen Kollektionen) in puncto Relevanz besser als Bing abschneidet, die Differenz beträgt jedoch lediglich dreieinhalb Prozent. In Hinsicht auf die Trefferpositionen gewinnt Google mit US-Treffern insgesamt 10,5 Prozent Relevanz hinzu (bei regulären Treffern verliert sie gut 19).

Bzgl. der Forschungsfrage 2 ist zu sagen, dass Google sowohl in Hinsicht auf die regulären als auch auf alle Treffer mehr relevante liefert, die Differenz zu Bing jedoch ziemlich gering ist (knapp zwei bzw. dreieinhalb Prozent). In Hinsicht auf die regulären Treffer der ersten zehn Trefferpositionen schneidet Google um gut vier Prozent besser ab als Bing (Google büßt dabei gut 19 Prozent Relevanz ein, Bing gut 25). Bzgl. aller Treffer auf den ersten zehn Trefferpositionen schneidet Google um gut zehn Prozent besser ab als Bing (Bing gewinnt dabei gut vierzehn Prozent Relevanz hinzu, Google gut zehn). Insgesamt ist also eine Überlegenheit Googles festzustellen, die jedoch nicht gravierend ist.

Nachdem in diesem Kapitel die zweite Forschungsfrage mit ihren –hypothesen untersucht wurde, wird sich im nächsten Abschnitt der dritten zugewandt.

3.3.5 Treffer spezieller Kollektionen

Dieses Kapitel ist der Beantwortung der dritten Forschungsfrage und ihrer –hypothese gewidmet (s. Kapitel *3.1 Forschungsfragen und -hypothesen*). Dafür werden die Treffer der speziellen Kollektionen in Hinsicht auf ihre Relevanz ausgewertet. Dabei werden Ergebnisse herausgestellt, die in direktem Bezug zur Forschungsfrage und –hypothese stehen. Die Forschungshypothese und die –frage werden am Ende des Kapitels beantwortet.

Abbildung 37 zeigt die binäre Relevanzbewertung der Treffer der speziellen Kollektionen, *Abbildung 38* die mittels der Fünfer-Skala.

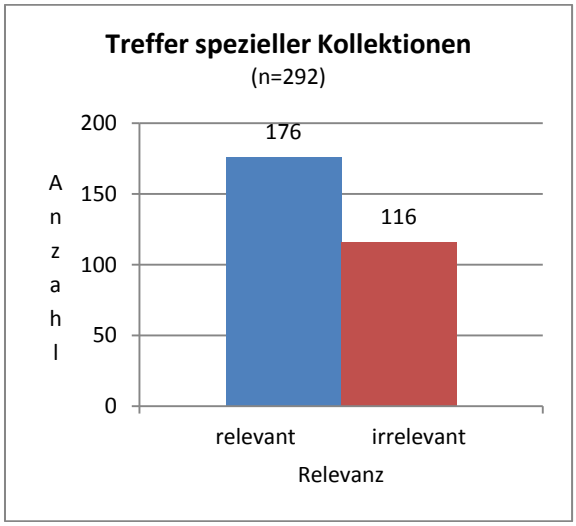


Abbildung 37: Treffer spezieller Kollektionen binär

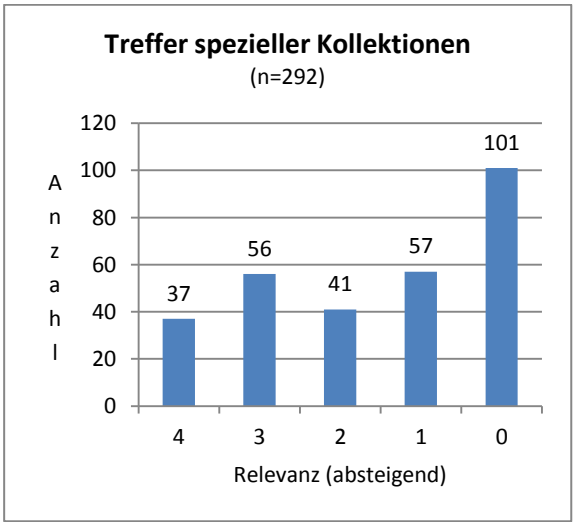


Abbildung 38: Treffer spezieller Kollektionen differenziert

Die binäre Sicht zeigt, dass ein großer Teil, ca. 40 Prozent, der Treffer der speziellen Kollektionen irrelevant ist. Das Verhältnis von relevanten zu irrelevanten Treffern beträgt also ungefähr 1,5 zu 1.

In der differenzierten Sicht ist ein Gewicht auf den „(eher) irrelevant"-Wertungen zu bemerken. Es wurden knapp dreimal soviele Treffer als „irrelevant" wie „relevant" bewertet. Selbst die „(eher) relevant"-Wertungen zusammen wiegen diesen Anteil nicht auf.

Für die Treffer der speziellen Kollektionen ist der Durchschnitt die Relevanzbewertung 1 (Median = 146).

Ergebnis 1: Binär: Ca. 40% der Treffer spezieller Kollektionen sind irrelevant.

Ergebnis 2: Differenziert: Gut die Hälfte der Treffer wurden als „(eher) irrelevant" bewertet.

Um den Einfluss der Treffer der speziellen Kollektionen auf die Gesamtperformance der Suchmaschinen zu untersuchen, werden die *Abbildungen 30* und *31*, sowie *33* und *34* des vorangegangenen Kapitels verglichen, also die Relevanzuntersuchungen der regulären und aller Treffer.

Bei der binären Sicht fällt auf auf, dass sich die Balken der relevanten Ergebnisse sowohl bei Google als auch bei Bing verkürzen (während die der irrelevanten etwas wachsen) – Google büßt 4,3 Prozent relevante Ergebnisse ein, Bing 6,8.

Ergebnis 3: Binär: Beim Vergleich der regulären mit allen Treffern nach Suchmaschinen büßt Google 4,3 Prozent relevante Treffer ein, Bing 8,6.

Im Detail zeigt sich generell, dass die langen Balken schrumpfen und die kurzen wachsen – es findet also sozusagen eine Umverteilung statt. Dies ist im Detail in *Tabelle 22* dargestellt. Dort sind die prozentualen Werte aus den Daten der *Abbildungen 30* und *31,* sowie *33* und *34* eingetragen, und zusätzlich die Abweichungen der regulären zu allen Treffern der einzelnen Relevanzbewertungen.

Tabelle 22: Einfluss Treffer spezieller Kollektionen auf die allgemeine Relevanz

Relevanz-bewertung	Google reguläre Treffer	Google alle Treffer	Google Abwei-chung	Bing reguläre Treffer	Bing alle Treffer	Bing Abwei-chung
4	23,6%	21,2%	*-10,4%*	24,7%	21,7%	*-12,4%*
3	26,5%	25%	*-5,7%*	22,7%	21,2%	*-6,5%*
2	18%	17,1%	*-5%*	18,8%	17,6%	*-6,1%*
1	12,1%	14,5%	*+19,6%*	14,6%	15,2%	*+4,5%*
0	19,7%	22,2%	*+12,7%*	19,2%	24,2%	*+26,1%*

Es wird deutlich, dass sowohl Google als auch Bing mit Treffern spezieller Kollektionen weniger „(eher) relevante“ und „neutrale“ Bewertungen erhalten. Google verzeichnet durch US-Treffer fast 20 Prozent mehr „eher irrelevant“-Wertungen (die größte Abweichung bei Google), bei Bing sind es knapp fünf. Dafür verzeichnet Bing mit BS-Treffern ca. 26 Prozent mehr „irrelevante“ Treffer (die größte Abweichung bei Bing), bei Google sind es hier knapp 13.

Bei beiden Suchmaschinen verschieben sich durch die Ausgabe von Treffern spezieller Kollektionen bzgl. der differenzierten Sicht also sozusagen ungefähr 30 Prozent der Bewertungen aus dem positiven und neutralen Bereich in den negativen.

Ergebnis 4: Differenziert: Beim Vergleich der regulären mit allen Treffern nach Suchmaschinen verlagern sich bei Google und Bing sozusagen je ca. 30% der Treffer aus dem (eher) relevanten und neutralen Bereich in den (eher) irrelevanten.

In *Tabelle 23* werden die Beschreibungsmaße der regulären Ergebnisse und die der speziellen Kollektionen einander gegenübergestellt. Die Werte bzgl. der Ergebnisse der speziellen Kollektionen sind die Durchschnitte, errechnet aus *Tabelle 21* (s. Kapitel *3.3.3 Ergebnisse nach Kollektionen*). Die Werte bzgl. der regulären Ergebnisse stammen aus *Tabelle 12* (s. Kapitel *3.3.2 Reguläre Ergebnisse*). Der schlechtere Wert ist jeweils gefettet.

Tabelle 23: Kennzahlvergleich reguläre und alle Ergebnisse

	reguläre Ergebnisse	**Ergebnisse spez. Kollektionen**	**Differenz (prozentual)**
Description-result precision	0,695	**0,557**	20
Description-result conformance	**0,839**	0,923	10
Description fallout	**0,088**	0,05	43,2
Description desception	0,073	**0,103**	41,1

Bei der Description-result precision schneiden die speziellen Kollektionen mit 20 Prozent weniger beidseitig relevanten Ergebnispaaren schlechter ab als die regulären Ergebnisse.

Bzgl. der Description-result conformance stehen die speziellen Kollektionen mit zehn Prozent mehr konsistenten Paaren besser da.

Beim Description Fallout schneiden die regulären Ergebnisse mit 43,2 Prozent mehr Paaren aus irrelevanter Beschreibung und relevantem Treffer schlechter ab.

Bzgl. der Description desception stehen die regulären Ergebnisse mit 41,1 Prozent weniger Paaren aus relevanter Beschreibung und irrelevantem Treffer besser da.

Bzgl. der Hypothese 3 (s. Kapitel *3.1 Forschungsfragen und -hypothesen*) kann bestätigt werden, dass die Treffer der speziellen Kollektionen im Durchschnitt relevant sind. Sie sind es jedoch in so geringem Maße, dass sie das Suchergebnis insgesamt bei beiden Suchmaschinen nicht auf- sondern abwerten.

Bzgl. der Forschungsfrage 3 ist zu sagen, dass die Treffer der speziellen Kollektionen das Suchergebnis sowohl von Google als auch von Bing *nicht* aufwerten. Im Gegenteil: Google büßt durch die Ausgabe dieser Treffer durchschnittlich gut drei Prozent an relevanten Treffern ein, Bing gut fünf.

3.3.6 Diskussion

In diesem Kapitel wird eine Einschätzung der Retrievaltest-Ergebnisse vorgenommen. Dafür werden zunächst übergreifende Beobachtungen dargelegt, um dann die einzelnen Auswertungskapitel in Hinsicht auf die Forschungsfragen und –hypothesen tabellarisch zusammenzufassen und zu diskutieren.

3.3.6.1 Allgemeines

In diesem Abschnitt werden einige übergreifende Beobachtungen dargelegt.

- Googles Dominanz auf dem deutschen Suchmaschinenmarkt (s. Kapitel *3.2.1 Suchmaschinen*) kann anhand des vorliegenden Retrievaltests nicht auf signifikant exponierende Retrievaleffektivität-Leistungen zurückgeführt werden. Zwar stellt sich Google gegenüber Bing in den einzelnen Bereichen i.d.R. als überlegen heraus, doch mit weniger Abstand, als man vielleicht vermuten würde – oft lieferten die beiden Suchmaschinen ähnlich gute Ergebnisse.

  Die Popularität Googles beruht folglich mehr auf anderen Faktoren, wie etwa der hohen Präsenz des Unternehmens durch weitere Produkte, wie z.B. *Gmail*. Es bleibt abzuwarten, ob Bing in Zukunft durch seine eher geringfügig schlechteren Leistungen der Suchmaschine Google Nutzer abspenstig machen kann.

- Für einen so großen Entwicklungsschritt bzgl. Suchmaschinen und der mittlerweile vergangenen Zeit der Etablierung, enttäuscht das Universal-Search-Prinzip in Anbetracht der Untersuchungsergebnisse. Im Rahmen des Retrieval-

tests wurde deutlich, dass die zusätzlichen Ergebnisse für die Nutzer weniger relevant als reguläre sind, sodass die Retrievaleffektivität verschlechtert wird.

Zu beachten ist, dass Resultate spezieller Kollektionen mit regulären Treffern um Platz auf der Ergebnisseite konkurrieren. Eine Daseinsberechtigung sollte dieses Feature folglich nur haben, wenn die zusätzlichen Resultate eine höhere Relevanz aufweisen, zumindest aber gleichauf liegen.

- Bzgl. der Auswertungsdiagramme der Fünfer-Skala-Bewertungen ist besonders bei den Trefferbeschreibungen zu bemerken, dass erheblich mehr Exemplare als „eher relevant“ als „relevant“ bewertet wurden (s. z.B. *Abb. 18*, *23*, *27* und *29*). Möglicherweise ist dies auf eine psychologische Hemmschwelle beim Bewerten zurückzuführen: einer Beschreibung, die von Haus aus nur spärliche Informationen aufweist, kann schwerlich die bestmögliche Wertung verliehen werden.

- Ebenfalls bzgl. der Auswertungsdiagramme der Fünfer-Skala-Bewertungen ist wiederholt eine Ergebnisverteilung zu bemerken, die an eine liegende Zwei (ᔓ) erinnert und typisch zu sein scheint (s. z.B. *Abb. 18*): Der Anteil der am besten bewerteten Ergebnisse ist hoch, doch der der „eher relevanten“ noch höher (evtl. aufgrund der psychologischen Hemmschwelle, s.o.), dann nehmen die Werte bis zur „eher irrelevant“-Wertung ab (es scheint für Suchmaschinen nicht schwierig zu sein, *irgendwie* relevante Treffer zu produzieren (vgl. LEWANDOWSKI 2012, pdf S. 12)), um bei „irrelevant“ noch einmal nach oben zu schnellen (dort schlagen die grundlegend irrelevanten Treffer kollektiv zu Buche).

3.3.6.2 Diskussion Reguläre Ergebnisse

In diesem Kapitel werden die Ergebnisse bzgl. der ersten Forschungshypothese der ersten –frage zusammengefasst und diskutiert.

In *Tabelle 24* sind die Untersuchungsergebnisse zur ersten Forschungshypothese und -frage sowie das resultierende Fazit (s. Kapitel *3.3.2 Reguläre Ergebnisse*) zusammengefasst.

Tabelle 24: Übersicht Ergebnisse F1, H1.1

Kapitel: *3.3.2 Reguläre Ergebnisse*	
Forschungsfrage: F1: Spiegelt sich die Relevanz von Treffern in deren Beschreibungen wider?	
Forschungshypothese: H1.1: Werden die regulären Beschreibungen und Treffer in Hinsicht auf ihre Relevanz miteinander verglichen, schneiden im Durchschnitt die Beschreibungen besser ab.	
Ergebnis	**Inhalt**
1	Binär: Es gibt knapp zwei% mehr relevante reguläre Treffer als Beschreibungen.
2	Differenziert: Es wurden ca. doppelt so viele reguläre Treffer wie Beschreibungen als „relevant“ bewertet.
3	Differenziert: Es wurden verschwindend wenig mehr reguläre Beschreibungen als Treffer als „eher relevant“ bewertet.
4	Bzgl. der Precision hinsichtlich der Trefferpositionen schneiden meistens die regulären Treffer, manchmal die Beschreibungen besser ab.
5	Bei ca. 15% der regulären Ergebnispaare stimmt die Relevanzeinschätzung von Beschreibung und Treffer nicht überein. Dabei gibt es etwas mehr Fälle aus irrelevanter Beschreibung und relevantem Treffer als umgekehrt.
Fazit zur Forschungshypothese: Der Fall, dass die regulären Beschreibungen relevanter als die Treffer sind, ist zwar im Detail vertreten, insgesamt sind jedoch die Treffer (geringfügig) relevanter als die Beschreibungen, was die These widerlegt.	

Verantwortlich für dieses unerwartete Ergebnis könnte (neben dem inhaltlichen Faktor) die im vorangegangenen Kapitel erwähnte psychologische Hemmschwelle sein, einer Beschreibung die bestmögliche Relevanzwertung zukommen zu lassen - denn dort klafft in differenzierter Sicht die mit Abstand größte Lücke. Außerdem wurden viele Beschreibungen als „neutral“ bewertet (deutlich mehr als Treffer) – hätten die Juroren sich stattdessen öfter für Relevanz entschieden, hätten Beschreibungen und Treffer in etwa gleich relevant sein, oder sogar den Befund von LEWANDOWSKI von mehr relevanten Beschreibungen als Treffern (LEWANDOWSKI 2008C) (ansatzweise) bestätigen können. Doch auch in Hinsicht auf die Paarkonstellationen tritt der Fall irrelevante Beschreibung und relevanter Treffer häufiger auf als der umgekehrte. In Hinsicht auf die Trefferpositionen sind die Beschreibungen ungefähr ab der vierten bis zur sechsten Position relevanter als die Treffer, auf den restlichen Positionen ist es umgekehrt. So ist insgesamt eine zu LEWANDOWSKI leicht gegensätzliche Tendenz das Ergebnis.

3.3.6.3 Diskussion Ergebnisse nach Kollektionen

In diesem Abschnitt werden die Ergebnisse bzgl. der zweiten Forschungshypothese der ersten –frage und dieser selbst zusammengefasst und diskutiert.

In *Tabelle 25* sind die Untersuchungsergebnisse zur zweiten Forschungshypothese der ersten -frage sowie die resultierenden Fazite (s. Kapitel *3.3.3 Ergebnisse nach Kollektionen*) zusammengefasst.

Tabelle 25: Übersicht Ergebnisse F1, H1.2

Kapitel: *3.3.3 Ergebnisse nach Kollektionen*	
Forschungsfrage: F1: Spiegelt sich die Relevanz von Treffern in deren Beschreibungen wider?	
Forschungshypothese: H1.2: Werden die Beschreibungen/Vorschauen der speziellen Kollektionen und die Treffer in Hinsicht auf ihre Relevanz miteinander verglichen, ergibt sich Folgendes: Bei Produkten und lokalen Ergebnissen ähneln sich Beschreibungen und Treffer sehr, bei Videos und Nachrichten schneiden im Durchschnitt die Beschreibungen besser ab.	
Ergebnis	**Inhalt**
1	Binär: Produktbeschreibungen und –treffer ähneln sich in puncto Relevanz in hohem Maße (Differenz: 3%).
2	Differenziert: Es wurden geringfügig mehr Produktbeschreibungen als „relevant“ und etwas mehr als „neutral“ als –treffer bewertet.
3	Fast alle Produktpaare sind konsistent (ca. 98 %).
4	Binär: Videobeschreibungen und –treffer ähneln sich in puncto Relevanz in hohem Maße (Differenz: 3%).
5	Differenziert: Es wurden etwas mehr Videobeschreibungen als „(eher) relevant“ und „irrelevant“ als –treffer bewertet.
6	Fast alle Videopaare sind konsistent (ca. 97%).
7	Binär: Es sind etwas weniger Nachrichtenbeschreibungen als –treffer relevant (ca. 9%).
8	Differenziert: Es gibt etwas mehr als „neutral“ und „irrelevant“ bewertete Nachrichtenbeschreibungen als –treffer.
9	Die überwiegende Mehrheit der Nachrichtenpaare ist konsistent (ca. 90%).
10	Binär: Lokale Beschreibungen und Treffer divergieren in puncto Relevanz etwas (Differenz: ca. 6%).
11	Differenziert: Es wurden im positiven Bereich etwas weniger lokale Beschreibungen als Treffer bewertet, im neutralen und negativen etwas mehr.
12	Der Großteil der lokalen Paare ist konsistent (ca. 85%).
Fazit zur Forschungshypothese: Die Beschreibungen und Treffer von Produkten und Videos sind sich in puncto Relevanz tatsächlich ziemlich ähnlich (Differenz: ca. 3 bzw. ca. 4%), auch sind jeweils die Produktpaare in hohem Maße konsistent (ca. 98 bzw. 97%). Lokale Ergebnisse und Nachrichten divergieren in puncto Relevanz etwas (ca. 6 bzw. ca. 9%). Dabei sind die Beschreibungen von Nachrichten *nicht* relevanter als die Treffer; es ist umgekehrt.	
Fazit zur Forschungsfrage: I.d.R. wird die Relevanz der Treffer durch ihre Beschreibungen korrekt widergespiegelt; die Rate der konsistenten Paare beträgt bei regulären Ergebnissen ca. 84% (der etwas größere Teil der restlichen Paare besteht aus irrelevanter Beschreibung und relevantem Treffer), und bei Ergebnissen spezieller Kollektionen sogar ungefähr 92%.	

Die Ergebnisse bzgl. der Hypothese 1.2 fallen unerwartet aus. Da die Relevanzunterschiede jeweils kleiner als zehn Prozent sind, wurde das „sehr ähnlich“ als Differenz kleiner oder gleich fünf Prozent definiert. Die Ergebnisse daraus sind in *Tabelle 26* als Übersicht dargestellt.

Tabelle 26: Übersicht Ergebnisse H1.2

Hypothese 1.2 (Relevanz Beschreibungen/Treffer spezieller Kollektionen)		
Sehr ähnlich (Differenz ≤ 5%)	**Unterschiedlich**	
	Beschreibungen relevanter	**Treffer relevanter**
Produkte, Lokales	Videos, Nachrichten	
Test		
Produkte, Videos	Lokales	Nachrichten

Die Hypothese zur Relevanzbeziehung von Beschreibungen und Treffern spezieller Kollektionen kann mit Blick auf Produkte bestätigt werden. Bei lokalen Ergebnissen divergieren Beschreibungen und Treffer etwas. Bei Videos ähneln sie sich. Die Beschreibungen und Treffer von Nachrichten divergieren zwar, jedoch sind dabei die Treffer relevanter. Bei allen anderen Kollektionen sind die Beschreibungen relevanter als die Treffer.

Dass die Beschreibungen lokaler Ergebnisse als relevanter als die Treffer eingeschätzt wurden, könnte daran liegen, dass in den Beschreibungen zwar die Standorte, aber kaum inhaltliche Informationen angezeigt werden. Z.B. lautete im Test das Informationsbedürfnis zu der Anfrage „anwalt hamburg“, allgemeine Informationen zu Anwälten in Hamburg zu finden (s. Anhang *B Suchanfragen*). Die ausgegebenen Ergebnisse zu dieser Anfrage sind in *Abbildung 39* abgebildet.

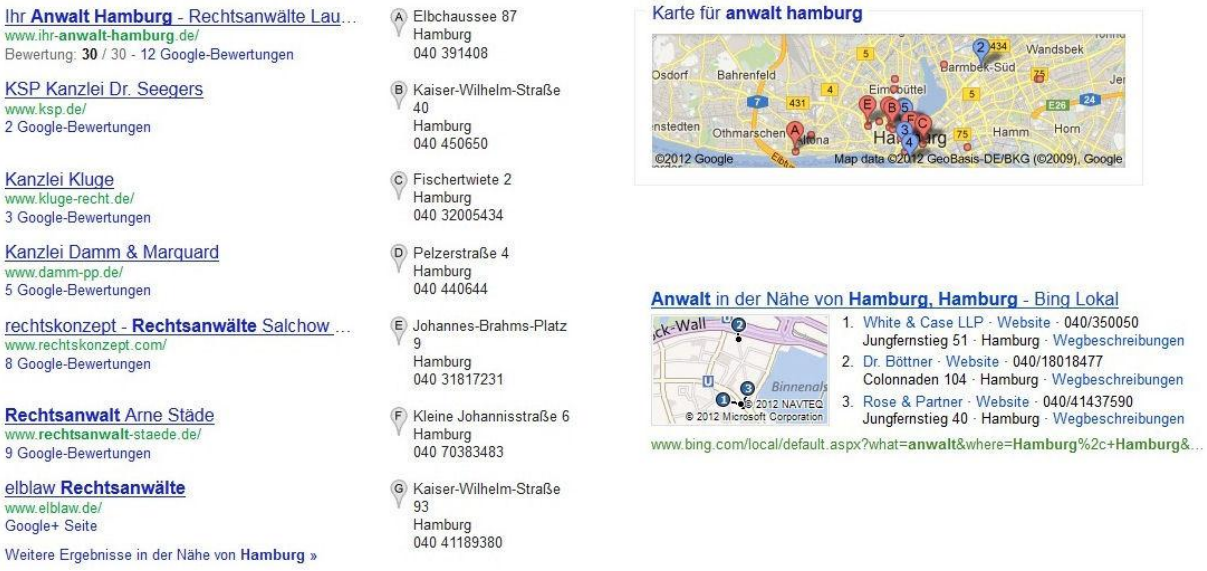


Abbildung 39: Lokale Ergebnisse Google (links und oben rechts) und Bing (unten rechts)

So sehen diese Ergebnisse erst einmal relevant aus - man erhält jedoch noch keine inhaltlichen Informationen. Nach dem Klick zeigen sich dann jedoch möglicherweise inhaltliche Einschränkungen (z.B. Anfahrt, Geschäftszeiten, Fachgebiete o.Ä. betreffend), weswegen diese Ergebnisse dann doch nicht so relevant sind.

Die Beschreibungen von Videos, die meist aus Titel, Vorschaubild, evtl. Länge und Quelle bestehen, transportieren die Relevanz der Treffer offenbar gut.

Bei Nachrichten sind die Treffer relevanter – vermutlich liegt das daran, dass bei den meisten Beschreibungen (bis auf die erste, die i.d.R. einen kleinen Textauszug beinhaltet) nur der Titel des Artikels angezeigt wird, und dieser bildet freilich selten viele inhaltliche Aspekte ab, sondern stellt i.d.R. seine Quintessenz dar.

3.3.6.4 Diskussion Ergebnisse nach Suchmaschinen

In diesem Kapitel werden die Ergebnisse bzgl. der beiden Forschungshypothesen der zweiten –frage und diese selbst zusammengefasst und diskutiert.

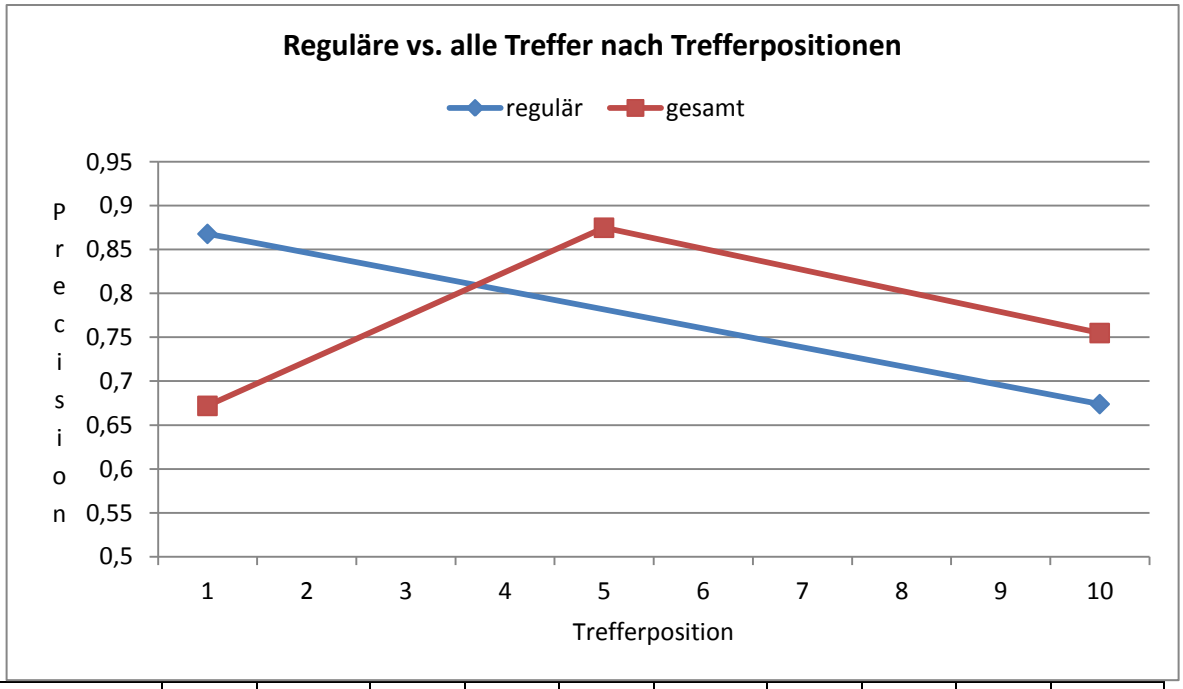
In *Tabelle 27* sind die Untersuchungsergebnisse zu beiden Forschungshypothesen der zweiten -frage sowie die resultierenden Fazite (s. Kapitel *3.3.4 Ergebnisse nach Suchmaschinen*) zusammengefasst.

Tabelle 27: Übersicht Ergebnisse F2, H2.1 & H2.2

Kapitel: *3.3.4 Ergebnisse nach Suchmaschinen*	
Forschungsfrage: F2: Wer liefert die relevanteren Treffer – Google oder Bing?	
Forschungshypothese: H2.1: Werden die regulären Treffer von Google und Bing in Hinsicht auf ihre Relevanz miteinander verglichen, schneiden die von Google besser ab.	
Forschungshypothese: H2.2: Werden alle Treffer (inkl. speziellen Kollektionen) von Google und Bing in Hinsicht auf ihre Relevanz miteinander verglichen, schneiden die von Google besser ab.	
Ergebnis	**Inhalt**
1.1	Binär: Googles reguläre Treffer schneiden in puncto Relevanz besser ab als Bings, jedoch beträgt die Differenz nur 1,7%.
1.2	Differenziert: Im positiven Bereich führt Google nur bei der „eher relevant"-Wertung; Differenz: 3,8%.
1.3	Trefferpositionen: Google startet bei den regulären Treffern mit leicht geringerer Precision als Bing, liegt aber insgesamt öfter vorne, als zurückzuliegen. Sie schneidet um gut 4% besser ab als Bing und büßt auf den ersten 10 Positionen etwas weniger Relevanz ein (19,4 bzw. 25,2%).
Fazit zur Forschungshypothese: H2.1: Google schneidet in Hinsicht auf die Relevanz regulärer Treffer tatsächlich besser ab als Bing, die Differenz beträgt jedoch nur 1,7%, sodass die beiden Suchmaschinen in etwa gleich gut abschneiden. In Trefferpositionensicht nimmt Googles Precision etwas weniger und gleichmäßiger ab als Bings, unterliegt aber auf den hinteren Positionen (7-9).	
2.1	Binär: Googles gesamte Treffer schneiden in puncto Relevanz ein wenig besser ab als Bings; die Differenz beträgt 3,5%.
2.2	Binär: Bei der Anreicherung der regulären Treffer mit Treffern spezieller Kollektionen werden bei beiden Suchmaschinen prozentual weniger relevante Treffer ausgegeben (Differenz Google: gut 3%, Bing: gut 5%)
2.3	Differenziert: Googles Treffer erhalten mehr „eher relevant"-Wertungen als Bings (Differenz: knapp 4%).
2.4	Differenziert: Bei der Anreicherung der regulären Treffer mit Treffern spezieller Kollektionen schneidet Google bei den Wertungen „(eher) relevant" merkbar besser als Bing ab.
2.5	Trefferpositionen: Google startet mit einer höheren Precision und liegt insgesamt deutlich öfter vorne, als zurückzuliegen. Sie schneidet um gut 10% besser ab als Bing und gewinnt auf den ersten 10 Positionen etwas weniger Relevanz hinzu als Bing (gut 10 bzw. gut 14%).
2.6	Trefferpositionen: Vereinzelt können Treffer spezieller Kollektionen die Precision verbessern, doch je höher ihre Dichte ist, desto niedriger ist die Precision.
Fazit zur Forschungshypothese: H2.2: Google schneidet in Hinsicht auf die Relevanz aller Treffer tatsächlich besser ab als Bing, die Differenz beträgt jedoch nur 3,5%. In Trefferpositionensicht gewinnt Google mit US-Treffern insgesamt gut 10% hinzu.	
Fazit zur Forschungsfrage: Google liefert sowohl in Hinsicht auf reguläre, als auch auf alle Treffer mehr relevante Exemplare, doch die Differenz zu Bing ist gering (knapp 2, bzw. 3,5%). In Hinsicht auf die regulären Treffer der ersten 10 Trefferpositionen schneidet Google um gut 4% besser ab als Bing (Google büßt dabei gut 19% Relevanz ein, Bing gut 25). Bzgl. aller Treffer auf den ersten zehn Trefferpositionen schneidet Google um gut 10% besser ab als Bing (Bing gewinnt dabei gut 14% Relevanz hinzu, Google gut 10). Insgesamt ist also eine Überlegenheit Googles festzustellen, die jedoch nicht gravierend ist.	

Bzgl. der Hypothese 2.1 schneidet Google etwas besser ab als Bing, jedoch weniger, als man aufgrund Googles Popularität im Suchmaschinenbereich vielleicht vermuten würde (s. Kapitel *3.3.6.1 Allgemeines*).

In Bezug auf die Hypothese 2.2 verhalten sich die Suchmaschinen ähnlich: Beide starten mit einer schlechteren Precision, erreichen am Ende aber je eine höhere als bei den nur regulären Treffern. Auf den ersten zehn Positionen nimmt die Relevanz also zu, anstatt abzunehmen. Dieser Sachverhalt ist in *Abbildung 40* vereinfacht anhand einzelner, verrechneter Werte dargestellt.



regulär	0,868									0,674
gesamt	0,672				0,875					0,755

Abbildung 40: Reguläre vs. alle Treffer nach Trefferpositionen

Der Graph der regulären Treffer verhält sich dem Ranking entsprechend: Die relevantesten Ergebnisse befinden sich auf den ersten Trefferpositionen und im weiteren Verlauf nimmt die Relevanz ab.

Um die Resultate der speziellen Kollektionen in die Ergebnislisten einzurechnen, wurden sie entsprechend ihrer Position innerhalb ihres US/BS-Moduls (Resultate spezieller Kollektionen werden i.d.R. als Gruppe ausgegeben) eingereiht (s. Kapitel *3.3.4 Ergebnisse nach Suchmaschinen*). D.h. also, dass die Dichte von zusätzlichen Resultaten auf Trefferposition 1 am höchsten ist und dann abnimmt. Eine wichtige Positionen in dieser Hinsicht ist die fünfte, da Bilder, die am häufigsten vorkamen, i.d.R. zu viert ausgegeben wurden.

Der „gesamt“-Graph zeigt nun, wie sehr die zusätzlichen Resultate die Precision verschlechtern; im Vergleich mit den nur regulären Treffern nämlich um ca. 20 Prozent auf Trefferposition 1. Diese Position ist sehr interessant, da sich hier die regulären Treffer und die der speziellen Kollektionen am „reinsten“ gegenüberstehen.

Bis zur Position fünf dünnen sich die Treffer der speziellen Kollektionen aus und es kommen mehr reguläre hinzu, bis dort die höchste Precision erreicht wird. Die Signifikanz dieser Position wurde bereits erwähnt.

Ab hier lässt die Precision im normalen Rahmen bis zur zehnten Trefferposition nach und endet auf einem ca. zehn Prozent höheren Stand als die nur regulären Treffer. Stark

vereinfacht könnte man sagen, dass beim „gesamt“-Graph, bedingt durch die Auswertungsmethode, der „normale“ Teil (der der regulären Treffer) um vier Positionen verschoben wird, ungefähr ab Trefferposition fünf anfängt und daher auch mit einer höheren Precision endet.

Das Fazit auf den durch Treffer spezieller Kollektionen dominierten Bereich lautet also, dass die Precision umso mehr leidet, je mehr Treffer spezieller Kollektionen vorhanden sind.

Freilich kann diskutiert werden, ob es „fair“ ist, die Treffer der speziellen Kollektionen den regulären in dieser Art und Weise gegenüberzustellen. Die in dieser Auswertung angewandte Methode stellt lediglich eine Möglichkeit dar, der erhöhten Klickwahrscheinlichkeit solcher Resultate (s. Kapitel *2.2.7.3 Rahmenwerk zur Messung von Suchmaschinen-Retrievaleffektivität nach LEWANDOWSKI*) und den Anforderungen an sie, z.B. bzgl. des begrenzten Platzes auf der Suchergebnisseite (s. *Kapitel 3.3.6.1 Allgemeines*), Rechnung zu tragen.

3.3.6.5 Diskussion Treffer spezieller Kollektionen

In diesem Abschnitt werden die Ergebnisse bzgl. der Forschungshypothese der dritten –hypothese und dieser selbst zusammengefasst und diskutiert.

In *Tabelle 28* sind die Untersuchungsergebnisse zur Forschungshypothese der dritten -frage sowie die resultierenden Fazite (s. Kapitel *3.3.5 Treffer spezieller Kollektionen*) zusammengefasst.

Tabelle 28: Übersicht Ergebnisse F3, H3

Kapitel: *3.3.5 Treffer spezieller Kollektionen*	
Forschungsfrage: F3: Werten die Treffer der speziellen Kollektionen das Trefferergebnis insgesamt auf?	
Forschungshypothese: H3: Werden Treffer spezieller Kollektionen ausgegeben, sind sie im Durchschnitt relevant und werten damit das Suchergebnis auf.	
Ergebnisse	**Inhalt**
1	Binär: Ca. 40% der Treffer speziellen Kollektionen sind irrelevant.
2	Differenziert: Gut die Hälfte der Treffer wurden als „(eher) irrelevant“ bewertet.
3	Binär: Beim Vergleich der regulären mit allen Treffern nach Suchmaschinen büßt Google 4,3% relevante Treffer ein, Bing 8,6.
4	Differenziert: Beim Vergleich der regulären mit allen Treffern nach Suchmaschinen verlagern sich bei Google und Bing sozusagen je 30% der Treffer aus dem relevanten und neutralen Bereich in den (eher) irrelevanten.
Fazit zur Forschungshypothese: Treffer spezieller Kollektionen sind im Durchschnitt relevant, aber in so geringem Maße, dass sie bei beiden Suchmaschinen das Suchergebnis abwerten.	
Fazit zur Forschungsfrage: Nein, Treffer spezieller Kollektionen werten das Suchergebnis insgesamt ab.	

Die Einbindung von US/BS-Treffern soll den Nutzern bei ihrer Suche helfen, indem ihnen möglichst viele thematische Aspekte angeboten werden. Andererseits werden damit das

Problem des Informationsüberflusses und die Suche selbst ein Stück weit auf sie abgewälzt. Zwar sortiert die Suchmaschine vor, doch die (evtl.) Wahl der Kollektion(en) liegt bei den Nutzern. Das kann die Suche erleichtern, sie aber auch anstrengender machen, da die Nutzer so möglicherweise mehr Mühe und Zeit aufbringen müssen, um zu finden, was sie suchen. Anstatt die „zehn blauen Links“ zu überfliegen, müssen sie nun zusätzliche und hervorgehobene Ergebnisse einbeziehen. Fraglich ist, ob dem Aufwand ein angemessener Nutzen gegenübersteht. Google hat die Ausgabe von US-Resultaten zwar bereits reduziert (*Abb. 5* zeigt einen ziemlich extremen Fall aus der Vergangenheit, der heute so nicht mehr auftritt). Doch trotzdem scheint ein erneutes Überdenken der aktuellen Antworten auf die Fragen angebracht, ob (anfragenspezifisch) überhaupt zusätzliche Resultate ausgegeben werden sollen, und wenn ja, welche und in welchem Umfang (müssen z.B. Bilder (fast) immer zu viert ausgegeben werden, oder lokale Ergebnisse häufig zu siebt?).

Den Ergebnissen des vorliegenden Retrievaltests zufolge ist zu hoffen, dass die Einbindung von Resultaten spezieller Kollektionen in Zukunft noch stärker und sensibler reguliert wird, denn es konnte gezeigt werden, dass viele dieser zusätzlichen Resultate überflüssig sind und damit den Suchprozess der Nutzer eher behindern, als zu erleichtern.

3.3.7 Einschränkungen und weitere Forschung

In diesem Kapitel wird auf die Einschränkungen dieses Retrievaltests und Anknüpfungspunkte für die weitere Forschung eingegangen.

Der vorliegende Test erfährt Einschränkungen durch die Unbekanntheit der Juroren, die vergleichsweise geringen Grundgesamtheiten der Resultate spezieller Kollektionen, sowie das zugrundeliegende Nutzermodell.

Mit Kenntnis über die Juroren kann die Aussagekraft deren Bewertungen besser eingeschätzt werden. Dies entfällt hier, da die Daten aufgrund eines Fehlers nicht ausgewertet werden konnten. Es wird lediglich angenommen, dass der Großteil der Bewertungen von Studierenden vorgenommen wurde.

Die Grundgesamtheiten der Resultate spezieller Kollektionen sind in diesem Test ziemlich klein; Bilder, die am häufigsten vorkommen, stellen ca. zehn Prozent der Daten, die anderen Kollektionen sind jeweils lediglich mit knapp fünf Prozent vertreten (s. Kapitel *3.3.1 Datenbasis*), was ca. je 40 Ergebnisse bedeutet, die sich auf beide Suchmaschinen verteilen. Von diesen Ergebnissen stammen mehr von Google, da Bing generell weniger von ihnen ausgibt. Dies sollte bedacht und die Aussagekraft der Untersuchungsergebnisse in Hinsicht auf die speziellen Kollektionen daher nicht überschätzt werden.

Der Test basiert auf einem Modell, bei dem der Nutzer eine festgelegte Menge Treffer samt Trefferbeschreibungen berücksichtigt. Im Vergleich mit Retrievaltests, in denen nur die Treffer berücksichtigt werden, stellt das eine Verbesserung dar, doch muss es nicht das echte Nutzerverhalten abbilden. Viele Nutzer klicken Ergebnisse an, bis sie ein passendes oder mehrere passende gefunden haben, und beenden ihre Suche dann (vgl. LEWANDOWSKI 2008C, pdf S.20).

Die Originalität des im Rahmen dieser Arbeit durchgeführten Retrievaltests liegt in der erstmaligen Einbeziehung von Resultaten spezieller Kollektionen. Die Untersuchung ergab, dass relativ viele dieser Ergebnisse weniger oder gar nicht relevant sind. Ziel weiterführender Forschungsarbeiten könnte nun sein, weitere Erkenntnisse über solche Resultate z.B. mithilfe anderer Methoden zu gewinnen. So könnte bspw. ihre Position auf der Ergebnisseite berücksichtigt werden (je weiter oben, desto höher die Klickwahrscheinlichkeit), oder ihr Erscheinungsbild (unsere Aufmerksamkeit wird von Reizen angezogen; Farben, Bilder etc.).

Weiter ist von Interesse, wie Nutzer generell zur Einbindung solcher zusätzlichen Resultate stehen. Dazu könnte bspw. die Methode des Retrievaltests unter Aufsicht mit einer Befragung der Juroren kombiniert werden.

Nicht zuletzt sollten die Kollektionen einzeln genauer untersucht werden, um z.B. Zusammenhänge zwischen Merkmalen und Nutzerwertschätzung aufzudecken.

Literaturverzeichnis

ALPAR 2012
ALPAR, Andre: *Andre's Ansichten: Universal Search – verwirrende Vielfalt?!*. – Stand: 2012-05-30 http://etailment.de/2012/andres-ansichten-universal-search-verwirrende-vielfalt/. – Abruf: 2012-06-15

BACHOR 2011
BACHOR, Matthias: *Fakten rund um Universal Search- und andere Integrationen*. - Stand: 2011-11-29 http://blog.searchmetrics.com/de/2011/11/29/fakten-rund-um-universal-search-und-andere-integrationen/. – Abruf: 2012-06-15

BACHOR 2012
BACHOR, Matthias: *Universal Search : Die aktuellen Daten mit Gewinnern und Verlierern*. - Stand: 2012-03-23 http://blog.searchmetrics.com/de/2012/03/23/universal-search-die-aktuellen-daten-mit-gewinnern-und-verlierern/. – Abruf: 2012-06-16

BERCHER 2006
BERCHER, Andreas: *Google kauft YouTube*. – Stand: 2006-11-13 http://www.zeit.de/online/2006/41/google-tube. - Abruf: 2012-06-16

BERNERS-LEE 1991
BERNERS-LEE, Tim: *World Wide Web : Executive Summary*. – Stand: 1991-08-06 https://groups.google.com/forum/?fromgroups#!msg/alt.hypertext/eCTkkOoWTAY/bJGhZyooXzkJ. - Abruf: 2012-05-11

BRODER 2002
BRODER, Andrei: *A taxonomy of web search*. In: *SIGIR Forum* 36 (2002), S. 3-10 – Online verfügbar unter: http://www.sigir.org/forum/F2002/broder.pdf. - Abruf: 2012-05-24

BRODER ET AL. 2000
BRODER, Andrei; KUMAR, Ravi; MAGHOUI, Farzin; RAGHAVAN, Prabhakar; RAJAGOPALAN, Sridhar; STATA, Raymie; TOMKINS, Andrew; WIENER, Janet: *Graph Structure in the Web*. In: *Computer Networks* 33 (2000) 1-6, S. 309-320 – Online verfügbar unter: http://www9.org/w9cdrom/160/160.html. - Abruf: 2012-05-15

CALDERON-BENAVIDES & GONZALES-CARO & BAEZA-YATES 2010
CALDERON-BENAVIDES, Liliana; GONZALES-CARO, Cristina; BAEZA-YATES, Ricardo: *Towards a Deeper Understanding of the User's Query Intent*. In: CROFT, Bruce; BENDERSKY, Michael; LI, Hang; XU, Gu (Hrsg.): *SIGIR 2010 Workshop on Query Representation and Understanding*, Geneva, Switzerland. S. 21-24 – Online verfügbar unter: http://grupoweb.upf.es/WRG/dctos/Calderon-Gonzalez-Baeza-SIGIR10.pdf. - Abruf: 2012-05-31

CHU 2010
CHU, Heting: *Information Representation and Retrieval in the Digital Age, Second Edition*. 2. Aufl. New Jersey : Information Today, Inc., 2010. – ISBN 978-1-57387-393-2

CLEF 2012A
THE CLEF INITIATIVE (Hrsg.): *Home*. – Stand: 2012-05-12 http://www.clef-initiative.eu/web/clef-initiative/home. - Abruf: 2012-06-10

CLEF 2012B
THE CLEF INITIATIVE (Hrsg.): *Track Series*. – Stand: 2012-05-12 http://www.clef-initiative.eu/track/series. - Abruf: 2012-06-10

CLEVERDON 1962
CLEVERDON, Cyril: *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield, England: College of Aeronautics - Online verfügbar unter: https://dspace.lib.cranfield.ac.uk/bitstream/1826/836/3/1962b.pdf. - Abruf: 2012-06-05

COMSCORE 2010
COMSCORE (Hrsg.): *comScore Reports Global Search Market Growth of 46 Percent in 2009*. – Stand: 2010-01-22 http://www.comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009. - Abruf: 2012-07-11

CRYSTAL & GREENBERG 2006
CRYSTAL, Abe; GREENBERG, Jane: *Relevance criteria identified by health information users during Web searches*. In: *Journal of the American Society for Information Science and Technology* 57 (2006) 10, S. 1368-1382 – Online verfügbar unter: http://www.ils.unc.edu/mrc/pdf/crystal06relevance.pdf [Preprint]. - Abruf: 2012-05-25

CUTRELL & GUAN 2007
CUTRELL, Edward; GUAN, Zhiwei: *Eye tracking in MSN Search: Investigating snippet length, target position and task types*. - Stand: 2007-01 http://research.microsoft.com/pubs/70395/tr-2007-01.pdf. - Abruf: 2012-06-17

DEMARTINI & MIZZARO 2006
DEMARTINI, Gianluca; MIZZARO, Stefano: *A Classification of IR Effectiveness Metrics*. In: LALMAS, Mounia et al. (Hrsg.): *Advances in Information Retrieval : European Conference on IR Research*. London UK : Springer, 2006. – ISBN 978-3-540-33347-0. – S. 488-491 – Online verfügbar unter: http://www.springerlink.com/content/m554q1061pp48256/fulltext.pdf. - Abruf: 2012-05-20

DING & MARCHIONINI 1996
DING, Wei; MARCHIONINI, Gary: *A Comparative Study of Web Search Service Performance*. In: *Proceedings of the ASIS Annual Meeting* 42 (1996) 33, S. 136-142

DOPICHAJ 2009
DOPICHAJ, Philipp: *Ranking-Verfahren für Web-Suchmaschinen*. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen*. Heidelberg : Akademische Verlagsgesellschaft Aka, 2009. – ISBN 978-3898386074. – S. 101-115 – Online verfügbar unter: http://www.theseus.joint-research.org/wp-content/uploads/2011/07/101-115.pdf. - Abruf: 2012-07-18

EMAGNETIX 2012
EMAGNETIX ONLINE MARKETING GMBH (Hrsg.): *Universal Search : Optimierte Suchergebnisse von Google*. http://www.emagnetix.at/de/suchmaschinenoptimierung/universal-search.html. – Abruf: 2012-06-15

ENQUIRO 2005
ENQUIRO (Hrsg.): *Google Eye Tracking Report*. – Stand: 2005-07 http://pages.enquiro.com/whitepaper-enquiro-eye-tracking-report-I-google.html [Whitepaper]. - Abruf: 2012-06-15

FERBER 2003
FERBER, Reginald: *Information Retrieval : Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. 1. Aufl. Heidelberg : Dpunkt Verl., 2003. – ISBN 978-3898642132

FLOTHOW 2009
FLOTHOW, Sebastian: *Eye Tracking : Ein Überblick über Geschichte, Methoden und Anwendungen*. Online verfügbar unter: http://www.cs.hs-rm.de/~linn/fachsem0809/eyetracking/Eye_Tracking.pdf. - Abruf: 2012-06-12

FUGMANN 1993
FUGMANN, Robert: *Subject analysis and indexing: Theoretical foundation and practical advice*. Frankfurt a.M. : Indeks Verl., 1993 – ISBN 3-88672-500-6

GANTERT & HACKER 2008,
GANTERT, Klaus; HACKERT, Rupert: *Bibliothekarisches Grundwissen*. 8., vollst. neu bearb. u. erw. Aufl. München : Saur, 2008. – ISBN 978-3-598-11771-8

GATES 2012
GATES, Fred: *Long Tail Keywords Versus Short Tail Keywords*. – Stand: 2012-05-12 http://www.seojournalist.com/2012/05/12/long-tail-keywords-versus-short-tail-keywords-2/. – Abruf: 2012-06-27

GEY & OARD 2002
GEY, Fredric; OARD, Douglas: *The TREC-2001 Cross-Language Information Retrieval Track : Searching Arabic using English, French or Arabic Queries*. In: NIST (Hrsg.): *Proceedings of the 2001 Text Retrieval Conference* (2002), S. 16-25 - Online verfügbar unter: http://trec.nist.gov/pubs/trec10/papers/clirtrack.pdf. - Abruf: 2012-06-10

GOOGLE 2012A
GOOGLE (Hrsg.): *Der Werdegang von Google*. – Stand: 2012-04-02 http://www.google.com/about/corporate/company/history.html#2001. – Abruf: 2012-06-15

GOOGLE 2012B
GOOGLE (Hrsg.): *Technology overview*. - Stand: 2012-04-02 http://www.google.com/intl/en_uk/about/corporate/company/tech.html. - Abruf: 2012-05-15

GORDON & PATHAK 1999
GORDON, Michael; PATHAK, Praveen: *Finding information on the World Wide Web: the retrieval effectiveness of search engines*. In: *Information Processing & Management* 35 (1999), S. 141-180 – Online verfügbar unter: http://www.jasonmorrison.net/iakm/cited/Gordon_Pathak.pdf. - Abruf: 2012-05-19

GREISDORF & SPINK 2001
GREISDORF, Howard; SPINK, Amanda: *Median measure: an approach to IR systems evaluation*. In: *Information Processing & Management* 37 (2001) 6, S. 843-857 – Online verfügbar unter: http://goanna.cs.rmit.edu.au/~aht/tiger/gs01-median.pdf. - Abruf: 2012-05-31

GRIESBAUM 2000
GRIESBAUM, Joachim: *Evaluierung hybrider Suchsysteme im WWW*. – Online verfügbar unter: http://www.joachim-griesbaum.de/files/evaluierung_hybrider_suchsysteme_im_www.pdf. - Abruf: 2012-06-07

GRIESBAUM 2004
GRIESBAUM, Joachim: *Evaluation of three German search engines : Altavista.de, Google.de and lycos.de*. In: *Information Research* 9 (2004) 4, S. 1-35 – Online verfügbar unter: http://informationr.net/ir/9-4/paper189.html. - Abruf: 2012-06-11

HAWKING ET AL. 2001
HAWKING, David; CRASWELL, Nick; BAILEY, Peter; GRIFFITHS, Kathy: *Measuring Search Engine Quality*. In: *Information Retrieval* 4 (2001), S. 33-59 – Online verfügbar unter: http://es.csiro.au/pubs/hawking_ir01.pdf. - Abruf: 2012-05-19

HAWKING & CRASWELL 2005
HAWKING, David; CRASWELL, Nick: *The Very Large Collection and Web Tracks*. In: VORHEES, Ellen; HARMAN, Donna (Hrsg.): *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, Massachussetts : MIT Press, 2005. - S. 199-231 - Online verfügbar unter: http://es.csiro.au/pubs/trecbook_for_website.pdf [Preprint version]. - Abruf: 2012-06-11

HÖCHSTÖTTER & KOCH 2007
HÖCHSTÖTTER, Nadine; KOCH, Martina: *Standard comparators for Information Searching Behaviour in Search Engines*. In: *Journal of Information Science* XX (2007) X, S. 1-22 – Online verfügbar unter: http://www.topicflux.de/publications_files/Benchmarks%20for%20Measuring%20Information%20Searching%20Behavior%20in%20Search%20Engines.pdf. – Abruf: 2012-05-20

HÖCHSTÖTTER & KOCH 2008
HÖCHSTÖTTER, Nadine; KOCH, Martina: *Standard parameters for searching behaviour in search engines and their empirical evaluation*. In: SCHMIDT-MÄNZ, Nadine (Hrsg.): *Untersuchung des Suchverhaltens im Web : Interaktion von Internetnutzern mit Suchmaschinen*. Hamburg : Dr. Kovac Verl., 2009. - 3-8300-2725-7. – S. 46-65 – Online verfügbar unter: http://www.topicflux.de/publications_files/Benchmarks%20for%20Measuring%20Information%20Searching%20Behavior%20in%20Search%20Engines.pdf. – Abruf: 2012-06-17

HÖCHSTÖTTER & LEWANDOWSKI 2009
HÖCHSTÖTTER, Nadine; LEWANDOWSKI, Dirk: *What users see - Structures in search engine results pages*. In: *Information Sciences* 179 (2009) 12, S. 1796-1812 – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/What_users_see_preprint.pdf. - Abruf: 2012-07-17

JANSEN & ZHANG & ZHANG 2007
JANSEN, Bernard; ZHANG, Mimi; ZHANG, Ying: *The effect of brand awareness on the evaluation of search engine results*. In: *Conference on Human Factors in Computing Systems – Proceedings*. New York, NY, USA : ACM, 2007. – S. 2471-2476

JOACHIMS ET AL. 2005
JOACHIMS, Thorsten; GRANKA, Laura; PAN, Bing; HEMBROOKE, Helene; GAY, Geri: *Accurately Interpreting Clickthrough Data as Implicit Feedback*. Paper präsentiert auf der *Conference on Research and Development in Information Retrieval*, Salvador, Brasilien. S. 154-161 – Online verfügbar unter: http://www.cs.cornell.edu/people/tj/publications/joachims_etal_05a.pdf. - Abruf: 2012-06-12

KEANE & O'BRIEN & SMYTH 2008
KEANE, Mark; O'BRIEN, Maeve; SMYTH, Barry: *Are people biased in their use of search engines?*. In: *Communications of the ACM* 51 (2008), S. 49-52 – Online verfügbar unter: http://irserver.ucd.ie/dspace/bitstream/10197/1643/3/MOB.ACM.v3-1.pdf. - Abruf: 2012-05-24

KLEINBERG 1999
KLEINBERG, Jon: *Authoritative Sources in a Hyperlinked Environment*. In: *Journal of the ACM* 46 5, S. 604-632 – Online verfügbar unter: http://www.cs.cornell.edu/home/kleinber/auth.pdf. - Abruf: 2012-05-15

KORFHAGE 1997
KORFHAGE, Robert: *Information Storage and Retrieval*. 1. Aufl. John Wiley & Sons, 1997. – ISBN 978-0471143383

KRONENBERG 2012
KRONENBERG, Hans: *Hybrid-Local-Results*. – Stand: 2012-05-03 http://www.sistrix.com/blog/1017-hybrid-local-results.html. - Abruf: 2012-07-03

KUNDER 2012
KUNDER, Maurice: *The size of the World Wide Web (The Internet)*. – Stand: 2012-07-10 http://worldwidewebsize.com/. – Abruf: 2012-07-18

LACKES ET AL. 2011
LACKES, Richard; SIEPERMANN, Markus; KOLLMANN, Tobias; SJURTS, Insa: *World Wide Web (WWW)* / GABLER VERLAG (Hrsg.). - http://wirtschaftslexikon.gabler.de/Archiv/74922/world-wide-web-www-v8.html. - Abruf: 2012-05-11

LEWANDOWSKI 2004
LEWANDOWSKI, Dirk: *Bewertung von linktopologischen Verfahren als bestimmender Ranking-Faktor bei WWW-Suchmaschinen*. In: OHLY, Peter; SIEGLERSCHMIDT, Jörn; SWERTZ, Christian (Hrsg.): *Wissensorganisation und gesellschaftliche Verantwortung : Gesellschaftliche, ökonomische und technische Aspekte*. Proceedings der 9. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (Wissensorganisation‘2004) - S. 318-329 – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Linktopologische_Verfahren_bestimmender_Rankingfaktor.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2005A
LEWANDOWSKI, Dirk: *Web Information Retrieval*. In: *IWP – Information: Wissenschaft und Praxis* 56 (2005) 1, S. 5-12 – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Web_Information_Retrieval_IWP2005.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2005B
LEWANDOWSKI, Dirk: *Web Information Retrieval : Technologien zur Informationssuche im Internet*. Frankfurt a.M. : DGI, 2005 (Informationswissenschaft; 7). – ISBN 3-925474-55-2. – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Web_Information_Retrieval_Buch.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2008A
LEWANDOWSKI, Dirk: *Neue Entwicklungen im Bereich der Suchmaschinen(technologie)*. In: *info7* 23 (2008) 2, S. 29-35 – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/info7_Lewandowski.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2008B
LEWANDOWSKI, Dirk: *Spezialsuchmaschinen*. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen*. Heidelberg : Akademische Verlagsgesellschaft Aka, 2009. – ISBN 978-3898386074. – S. 53-69 - Online verfügbar unter: http://eprints.rclis.org/bitstream/10760/12713/1/spezialsuchmaschinen.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2008C
LEWANDOWSKI, Dirk: *The Retrieval Effectiveness of Web Search Engines: Considering Results Descriptions*. In: *Journal of Documentation* 64 (2008) 6, S. 915-937 – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/retrieval_effectiveness_results_descriptions_JDoc2008.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2011A
LEWANDOWSKI, Dirk: *Evaluierung von Suchmaschinen*. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen 2 : Neue Entwicklungen in der Web-Suche*. Heidelberg : Akademische Verlagsgesellschaft Aka, 2011 – ISBN 978-3898386517. – S. 203-228 - Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Evaluierung_von_Suchmaschinen.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2011B
LEWANDOWSKI, Dirk: *Query Understanding*. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen 2 : Neue Entwicklungen in der Web-Suche*. Heidelberg : Akademische Verlagsgesellschaft Aka, 2011 – ISBN 978-3898386517. – S. 55-75 - Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Query_Understanding.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2011C
LEWANDOWSKI, Dirk: *The retrieval effectiveness of search engines on navigational queries*. In: *ASLIB Proceedings* 62 (2011) 4, S. 354-363 – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/ASLIB2009_preprint.pdf. - Abruf: 2012-07-17

LEWANDOWSKI 2012
LEWANDOWSKI, Dirk: *A Framework for Evaluating the Retrieval Effectiveness of Search Engines*. In: JOUIS, Christophe (Hrsg.): *Next Generation Search Engines : Advances Models for Information Retrieval*. Hershey, Pennsylvania : IGI Global, 2012. – ISBN 978-1466603301. – S. 456-479 - Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Evaluating_Next_Gen_Search_Engines_preprint.pdf. - Abruf: 2012-07-17

LEWANDOWSKI & HÖCHSTÖTTER 2007
LEWANDOWSKI, Dirk; HÖCHSTÖTTER Nadine: *Qualitätsmessung bei Suchmaschinen : System- und nutzerbezogene Evaluationsmaße*. In: *Informatik Spektrum* 30 (2007) 3, S. 159-169 – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Qualitaetsmessung_bei_Suchmaschinen_Informatik_Spektrum_2007.pdf. - Abruf: 2012-07-12

LEWANDOWSKI & SÜNKLER 2012
LEWANDOWSKI, Dirk; SÜNKLER, Sebastian: *Relevance Assessment Tool : Ein Werkzeug zum Design von Retrievaltests sowie zur weitgehend automatisierten Erfassung, Aufbereitung und Auswertung der Daten*. In: *Proceedings der 2. DGI-Konferenz: Social Media und Web Science - Das Web als Lebensraum*. Frankfurt a.M. : DGI, 2012, S. 237-249. – Online verfügbar unter: http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/RAT_DGI_Lewandowski_Suenkler_preprint.pdf [Preprint]. - Abruf: 2012-07-01

MACCALL & CLEVELAND 1999
MACCALL, Steven; CLEVELAND, Ana: *A Relevance-based Quantitative Measure for Internet Information Retrieval Evaluation*. In: *Proceedings of the ASIS Annual Meeting*, 36, S. 763-768

MARCHIONINI 2006
MARCHIONINI, Gary: *Exploratory search: From finding to understanding*. In: *Communications of the ACM* 49 (2006) 4, S. 41-46 – Online verfügbar unter: http://www.ischool.utexas.edu/~i385t-sw/readings/Marchionini-2006-Exploratory_Search.pdf. - Abruf: 2012-05-25

MAYER 2007
MAYER, Marissa: *Universal search : The best answer is still the best answer*. – Stand: 2007-05-17 http://googleblog.blogspot.de/2007/05/universal-search-best-answer-is-still.html. - Abruf: 2012-06-15

MICROSOFT 2012
MICROSOFT (Hrsg.): *Microsoft beendet Betaphase von Bing in Deutschland* [Pressemitteilung]. – Stand: 2012-01-27 http://www.microsoft.com/germany/newsroom/pressemitteilung.mspx?id=533470. – Abruf: 2012-06-27

MIßFELDT 2011
MIßFELDT, Martin: *Google Universal Search [1] : Milestones*. Stand: 2011-10-20 http://www.tagseoblog.de/google-universal-search-1-milestones. - Abruf: 2012-06-15

MIZZARO 1997
MIZZARO, Stefano: *Relevance: The whole history*. In: *Journal of the American Society for Information Science* 48 (1997) 9, S. 801-832 – Online verfügbar unter: http://ilps.science.uva.nl/Teaching/0405/AR/part2/rel-hist-jasis.pdf. - Abruf: 2012-05-20

NICHOLSON ET AL. 2006
NICHOLSON, Scott; SIERRA, Tito; ESERYEL, Yeliz; PARK, Ji-Hong; BARKOW, Philip; POZO, Erika; WARD, Jane: *How Much of It is Real? Analysis of Paid Placement in Web Search Engine Results*. In: *Journal of the American Society for Information Science and Technology* 57 (2006) 4, S. 448-461 – Online verfügbar unter: http://bibliomining.com/nicholson/nicholsonads.html. - Abruf: 2012-05-31

NIELSEN 2012
NIELSEN, Jakob: *Usability 101 : Introduction to Usability*. – Stand: 2012-01-30 http://www.useit.com/alertbox/20030825.html. - Abruf: 2012-06-17

NIELSEN & TAHIR 2002
NIELSEN, Jakob; TAHIER, Marie: *Homepage usability : 50 websites deconstructed*. Indianapolis, Ind. : New Riders, 2002. – ISBN 0-7357-1102-X

NIST 2010
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (Hrsg.): *Overview*. – Stand: 2010-08-10 http://trec.nist.gov/overview.html. - Abruf: 2012-06-07

NIST 2012
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (Hrsg.): *TREC Tasks*. – Stand: 2012-01-19 http://trec.nist.gov/tracks.html. - Abruf: 2012-06-07

NOTESS 2007
NOTESS, Greg: *Search Engines Features Chart*. – Stand: 2007-10-01 http://searchengineshowdown.com/features/. – Abruf: 2012-05-28

PSYCHOLOGY48 2010
PSYCHOLOGY48 (Hrsg.): *Interraterreliabilität*. Stand: 2010-06-13 http://www.psychology48.com/deu/d/interraterreliabilitaet/interraterreliabilitaet.htm. - Abruf: 2012-05-25

QUIRMBACH 2009
QUIRMBACH, Sonja: *Universal Search : Kontextuelle Einbindung von Ergebnissen unterschiedlicher Quellen und Auswirkungen auf das User Interface*. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen : Nutzerorientierung in Wissenschaft und Praxis*. Heidelberg : Akademische Verlagsgesellschaft AKA, 2009. – 978-3-89838-607-4. – S. 220-248 – Online verfügbar unter: http://eprints.rclis.org/bitstream/10760/12741/1/universal_search.pdf. - Abruf: 2012-06-17

RISVIK & MICHELSEN 2002
RISVIK, Knut Mage; MICHELSEN, Rolf: *Search engines and Web dynamics*. In: *Computer Networks* 39 (2002) 3, S. 289-302 – Online verfügbar unter: http://www.idi.ntnu.no/~algkon/generelt/se-dynamicweb1.pdf. - Abruf: 2012-07-18

SCHAER & MAYR & MUTSCHKE 2010
SCHAER, Philipp; MAYR, Philipp; MUTSCHKE, Peter: *Implications of Inter-Rater Agreement on a Student Information Retrieval Evaluation*. In: ATZMÜLLER, Martin; BENZ, Dominik; HOTHO, Andreas; STUMME, Gerd (Hrsg.): *LWA2010 : Lernen, Wissen & Adaptivität; Workshop Proceedings; Kassel, 4.-6. Oktober 2010*, *Kassel* - Online verfügbar unter: http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir8.pdf. - Abruf: 2012-05-25

SCHRÄPLER 2009
SCHRÄPLER, Frank: *Google Universal Search : Das Ende der Top-10*. – Stand: 2009-05-05 http://seo-marketing-blog.de/goatix/google-universal-search-das-ende-der-top-10/. – Abruf: 2012-06-14

SCHULZ 2012
SCHULZ, Ursula: *Glossar*. – Stand: 2012-04-30 http://www.bui.haw-hamburg.de/pers/ursula.schulz/worg1/glossar.html#I. – Abruf: 2012-05-23

SPINK 2002
SPINK, Amanda: *A user centered approach to evaluating human interaction with web search engines : an exploratory study*. In: *Information Processing and Management* 38 (2002) 3, S. 401-426 – Online verfügbar unter: http://eprints.qut.edu.au/4993/1/4993.pdf. - Abruf: 2012-05-18

SPINK ET AL. 2006
SPINK, Amanda; JANSEN, Bernard; BLAKELY, Chris; KOSHMAN, Sherry: *A study of results overlap and uniqueness among major web search engines*. In: *Information Processing & Management* 42 (2006), S. 1379-1391 – Online verfügbar unter: http://eprints.qut.edu.au/4755/1/4755.pdf. - Abruf: 2012-05-28

STATISTA 2012
STATISTA (Hrsg.): *Median*. – Stand: 2012-07-05 http://de.statista.com/statistik/lexikon/definition/85/median/. – Abruf: 2012-07-07

STEINER 2010
STEINER, Josef: *Wahrnehmung der SERPs und Suchverhalten bei der Suchmaschine Google*. – Online verfügbar unter: http://evolutionarts.at/tl_files/wahrnehmung_serps_suchverhalten_google.pdf. - Abruf: 2012-07-22

STOCK 2007
STOCK, Wolfgang: *Information Retrieval : Informationen suchen und finden*. München : Oldenbourg Wissenschaftsverl., 2007. – ISBN 978-3486581720

SU 1998
SU, Louise: *Value of Search Results as a Whole as the Best Single Measure of Information Retrieval Performance*. In: *Information Processing & Management* 34 (1998), S. 557-579 – Online verfügbar unter: http://ac.els-cdn.com/S0306457398000235/1-s2.0-S0306457398000235-main.pdf?_tid=a97717b6f513f1dcff6dc8cb865b901b&acdnat=1337506708_550149731c85542c6d1aa404f7207008. - Abruf: 2012-05-19

SULLIVAN 2003
SULLIVAN, Danny: *Searching with Invisible Tabs*. – Stand: 2003-12-01 http://searchenginewatch.com/article/2064036/Searching-With-Invisible-Tabs. - Abruf: 2012-06-14

SULLIVAN 2007
SULLIVAN, Danny: *Search 3.0 : The Blended & Vertical Search Revolution*. – Stand: 2007-11-27 http://searchengineland.com/search-30-the-blended-vertical-search-revolution-12775. - Abruf: 2012-06-14

TAGUE-SUTCLIFFE 1992
TAGUE-SUTCLIFFE, Jean: *The pragmatics of information retrieval experimentation, revisited*. In: *Information Processing & Management* 28 (1992) 4, S. 467-490 – Online verfügbar unter: http://ac.els-cdn.com/030645739290005K/1-s2.0-030645739290005K-main.pdf?_tid=92241e1cad51c3721373661dc43d3a34&acdnat=1337436487_93fb1365ceb7680afb8028e7b94bc1cd. – Abruf: 2012-05-19

TAWILEH & GRIESBAUM & MANDL 2010
TAWILEH, Wissam; MANDL, Thomas; GRIESBAUM, Joachim: *Evaluation of five web search engines in Arabic language*. In: *LWA 2010 – Lernen, Wissen & Adaptivität: Workshop Proceedings. Workshop Information Retrieval*, S. 221-228 – online verfügbar unter: http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir1.pdf. - Abruf: 2012-06-28

TOBII 2010
TOBII TECHNOLOGY (Hrsg.): *Tobii Eye Tracking : An introduction to eye tracking and Tobii Eye Trackers* [Whitepaper]. – Online verfügbar unter: http://www.scribd.com/doc/25907389/Tobii-Eye-Tracking-An-introduction-to-eye-tracking-and-Tobii-Eye-Tracker. - Abruf: 2012-06-12

USTPO 2008
UNITED STATES PATENT AND TRADEMARK OFFICE (Hrsg.): *Interface for a universal search engine*. – Stand: 2008-11-04 http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,447,678.PN.&OS=pn/7,447,678&RS=PN/7,447,678. – Abruf: 2012-06-16

VAN EIMEREN & FREES 2009
VAN EIMEREN, Birgit; FREES, Beate: *Der Internetnutzer 2009 – multimedial und total vernetzt? : Ergebnisse der ARD/ZDF-Onlinestudie 2009*. In: *media perspektiven* 7 (2009), S. 334-348. – Online verfügbar unter: http://www.media-perspektiven.de/uploads/tx_mppublications/Eimeren1_7_09.pdf. - Abruf: 2012-07-11

VAUGHAN 2004
VAUGHAN, Liwen: *New measurements for search engine evaluation proposed and tested*. In: *Information Processing & Management* 40 (2004), S. 677-691 – Online verfügbar unter: http://www.sciencedirect.com/science/article/pii/S0306457303000438. - Abruf: 2012-05-20

VAUGHAN & THELWALL 2004
VAUGHAN, Liwen; THELWALL, Mike: *Search Engine Coverage Bias : Evidence and Possible Causes*. In: *Information Processing & Management* 40 (2004), S. 693-707

W&V 2011
W&V (Hrsg.): *Yahoo-Suche: Ergebnisse über Bing*. – Stand: 2011-08-03 http://www.wuv.de/nachrichten/digital/yahoo_suche_ergebnisse_ueber_bing. - Abruf: 2012-05-25

WEBHITS 2012
WEBHITS (Hrsg.): *Web-Barometer*. – Stand: 2012-06-27 http://www.webhits.de/deutsch/index.shtml?webstats.html. – Abruf: 2012-06-27

WIRTH 2009,
WIRTH, Thomas: *Aufmerksamkeitsgesetze*. – Stand: 2009-07-09 http://www.kommdesign.de/texte/aufmerk4.htm. - Abruf: 2012-06-17

WOLFF 2000
WOLFF, Christian: *Vergleichende Evaluierung von Such- und Metasuchmaschinen*. In: *Proceedings of 7. Internationales Symposium für Informationswissenschaft, Darmstadt, Germany*. Konstanz : Universitätsverl. – S. 31-48 – Online verfügbar unter: http://epub.uni-regensburg.de/11032/1/Informationskompetenz_Basiskompetenz_in_der_Informationswissenschaft.pdf. - Abruf: 2012-05-26

Anhang

A Beigabe: Inhalt der CD

Bachelorarbeit als PDF-Version
Projektbeschreibung
Suchanfragen
Suchanfragen – thematische Einordnung
Bing-Ergebnisseiten\ Bing-Suchergebnisseiten zu den Suchanfragen
Google-Ergebnisseiten\ Google-Suchergebnisseiten zu den Suchanfragen
Rohdaten\ Rohdaten des Retrievaltests (nach *Trefferbeschreibungen* und *Treffern & Resultaten spezieller Kollektionen*)

B Suchanfragen

Suchanfrage	**Informationsbedürfnis**	US Google	BS Bing
medizinstudium	Wo man in Deutschland Medizin studieren kann	X	
eureka serie	Inhalt und Schauspieler der US-Fernseh-Serie	X	
raf	Historisches über die Rote Armee Fraktion (Hintergrund, Mitglieder, Aktionen)	X	X
herr der fliegen	Inhalt und Rezeption des Buchs von William Golding	X	X
gez befreiung	Wie kann man sich von der Gebühreneinzugszentrale befreien lassen?		
tattoopflege	Informationen zum Schutz; Cremes etc.		
ipod ohne itunes nutzen	Lässt sich ein iPod auch ohne iTunes nutzen? (Wie?)		
brienzersee	touristische und geografische Informationen	X	
knotentechnik	Anleitungen zum Knoten (z.B. fürs Segeln oder Bergsteigen)	X	
reiserücktrittsversicherung	Anbieter, Tests und Produktvergleiche	X	
jakobsweg	Wegbeschreibungen und Erfahrungsberichte	X	
ostern	Hintergrund des Osterfests	X	
mode renaissance	Bilder, Texte und Literaturtipps zur Mode der Epoche	X	
friends	Inhalt und Schauspieler der US-amerikanischen Sitcom		
clive barker	Biografisches und Informationen zu Werken des britischen Schriftstellers/Regisseurs/Künstlers	X	X
columbine	Täter, Ablauf, Opfer und Hintergründe des Amoklaufs von 1999 in Colorado, USA	X	X
joachim gauck	Biografisches und Informationen zur Arbeit des Bundespräsidenten	X	X
21 gramm dvd	Bild-/Tonqualität und Extras von DVD-Erscheinungen des Films	X	X
sneakers	Arten des Schuhs mit ihren Merkmalen	X	
lars von trier	Biografisches und Informationen zu Werken des dänischen Regisseurs	X	X
anwalt hamburg	Informationen zu Anwälten in Hamburg (Spezialisierungen, wer, wo etc.)	X	X
zahnarzt hamburg	Informationen zu Zahnärzten in Hamburg (Spezialisierungen, wer, wo etc.)		
star wars	Inhalt des von George Lucas erdachten Heldenepos‘	X	X
farin urlaub akkorde	Gitarrenakkorde zu Liedern von Farin Urlaub		X
esp ltd alexi-200	Techn. Daten, Vergleiche und Bewertungen der E-Gitarre	X	X
freizeithemd	Arten mit ihren Merkmalen	X	
planetarium hamburg	Standort, Veranstaltungen und Preise	X	X
laserdrucker	Techn. Daten, Vergleiche und Bewertungen	X	
angela merkel	Biografisches und Informationen zur Arbeit der Bundeskanzlerin	X	X
philipp rösler	Biografisches und Informationen zur Arbeit des Bundesministers für Wirtschaft und Technologie und Vizekanzlers	X	X
soundsystem	Techn. Daten, Vergleiche und Bewertungen	X	
satelliten receiver	Techn. Daten, Vergleiche und Bewertungen	X	X
katzenurin aus sofa	Wie kann man Katzenurin vom Sofa entfernen?		
samsung r580	Techn. Daten, Vergleiche und Bewertungen des Notebooks	X	X
digitale spiegelreflexkamera	Techn. Daten, Vergleiche und Bewertungen	X	X
spaced	Inhalt und Schauspieler der englischen Serie	X	

molotow hamburg	Standort, Veranstaltungen und Historie des hamburger Musik-Clubs	X	X
fdp	Historie, Ausrichtung und Mitglieder der Partei	X	X
hamburg michel	Standort, Veranstaltungen und Historie der hamburger Kirche	X	X
venetica	Inhalt, techn. Anforderungen und Bewertungen des Computerspiels	X	X
mogwai cd	Tracklisten, Bewertungen und Preise von CD-Erscheinungen der Band Mogwai	X	
king anschlag	Inhalt und Rezeption des Buchs "Der Anschlag" von Stephen King		
berlin mauerpark	Lage und Veranstaltungen des berliner Parks	X	X
dropped c stimmen	Wie bringt man eine Gitarre in die Dropped-C-Stimmung?	X	
waschmaschine a+++	Techn. Daten, Vergleiche, Bewertungen und Preise von Waschmaschinen der Energie-Effizienz-Klasse A+++	X	
schlecker insolvenz	Rekapitulation der Ereignisse und aktueller Stand der Insolvenz des Unternehmens	X	X
elgin nordsee	Rekapitulation der Ereignisse und aktueller Stand des Gaslecks auf der Bohrinsel "Elgin" in der Nordsee	X	X
hamburg panoptikum	Lage, Öffnungszeiten und Bewertungen des hamburger Wachsfigurenkabinetts	X	X
12 monkeys dvd	Bild-/Tonqualität und Extras von DVD-Erscheinungen des Films	X	X
christian wulff	Biografisches und Informationen zur Arbeit des ehem. Bundespräsidenten	X	X

C Suchanfragen – thematische Einordnung

Thema	Anzahl der Suchanfragen
Nicht Jugendfreies & Sex	
Kunst & Kultur	ostern, mode renaissance, clive barker, lars von trier, planetarium hamburg, molotow hamburg, hamburg michel, berlin mauerpark, hamburg panoptikum
Schönheit & Stil	tattoopflege, sneakers, freizeithemd
Autos und Verkehr	
Computer & Internet	
Bildung	medizinstudium
Unterhaltung	eureka serie, herr der fliegen, friends, 21 gramm dvd, star wars, spaced, venetica, king anschlag, 12 monkeys dvd
Musik & Spiele	farin urlaub akkorde, esp ltd alexi-200, mogwai cd, dropped c stimmen
Finanzielles	
Essen & Trinken	
Gesundheit	
Haus & Garten	katzenurin aus sofa
Industrielle Produkte & Services	
Politik & Regierung	joachim gauck, angela merkel, philipp rösler, fdp, christian wulff
Religion	
Wissenschaft & Mathematik	
Sozialwissenschaften	
Sport	
Technologie & Elektronik	ipod ohne itunes nutzen, laserdrucker, soundsystem, satelliten receiver, samsung r580, digitale spiegelreflexkamera, waschmaschine a+++
Reisen	brienzersee, reiserücktrittsversicherung, jakobsweg
Undefiniert	raf, gez befreiung, anwalt hamburg, zahnarzt hamburg, knotentechnik, elgin nordsee, columbine
Arbeit	schlecker insolvenz

Es sei angemerkt, dass diese Einordnung subjektiv erfolgte und keinen Anspruch auf Absolutheit erhebt.

Eidesstattliche Erklärung

Ich versichere, die vorliegende Arbeit selbstständig ohne fremde Hilfe verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt zu haben. Die aus anderen Werken wörtlich entnommenen Stellen oder dem Sinn nach entlehnten Passagen sind durch Quellenangabe kenntlich gemacht.

Hamburg, 30. Juli 2012

_____