# Quantifizierung der Qualität von Suchmaschinen

Marius Riehl

Hochschule Darmstadt - Fachbereich Informatik
marius.riehl@stud.h-da.de

**Zusammenfassung.** Suchmaschinen verwenden Ranking- bzw. Scoring-Algorithmen zur Sortierung der Suchtreffer bezüglich der Relevanz für den Nutzer. Der Betreiber einer Suchmaschine hat aus wirtschaftlichen Gründen das Ziel, die Ergebnisqualität für den Nutzer zu maximieren. Bisherige Arbeiten bemessen die Qualität von Suchmaschinen und die Zufriedenheit derer Nutzer primär an der Relevanz der Suchergebnisse. These dieser Arbeit ist, dass die Güte des Rankings als alleiniger Parameter unzureichend ist um die Qualität von Suchmaschinen zu quantifizieren. Ziel ist es zu belegen, dass andere Qualitätsziele und Komponenten einer Software ebenfalls Einfluss auf die messbare Qualität von Suchmaschinen haben.Dazu wird analysiert, wie die Qualität des Rankings bisher experimentell bestimmt wurde und welche Probleme die aktuelle Verfahrensweise kennzeichnen. Durch die Kombination verwandter Arbeiten konnte nachgewiesen werden, dass das Ranking nur einer von vielen Parametern für die umfassende Quantifizierung der Qualität von Suchmaschinen darstellt.

## 1 Einleitung

Die alltägliche Verwendung von Suchmaschinen im Internet ist für Millionen von Menschen zum Alltag geworden [PHJ+07, S. 801]. Historisch war die Beschaffung von Informationen noch nie so leicht wie heute, weshalb Suchmaschinen einen bedeutenden, stets wachsenden Einfluss auf die Kultur haben. Dieser ist bereits in der Sprache, zum Beispiel an Wortneuschöpfungen wie "googeln" zu erkennen, was dabei primär als Synonym für das Finden von Information im Internet, allerdings auch für das Suchen und Finden im Allgemeinen steht. Internetsuchmaschinen ersetzen die klassische Bibliothek, die vor 25 Jahren noch die erste Anlaufstelle zur Informationsbeschaffung war. Informationen und Daten jeglicher Form sind nun ständig und in sekundenschnelle zugänglich. Google alleine, mit etwa 71 Prozent Marktanteil weltweit und 95 Prozent in Deutschland, verarbeitet täglich mehr als 6 Milliarden Suchanfragen.

Die Sortierung der Suchtreffer nach ihrer Relevanz, im Folgenden auch als "Ranking" bezeichnet, ist dabei hauptsächliches Kriterium für die Bewertung der Qualität von Suchmaschinen in der Literatur und Wirtschaft [VC07], [BJPBY10], [HCBG01], [KWM+08]. Der Anbieter Yahoo!, weltweiter Marktanteil von 7 Prozent und Marktführer bis 2002 [ZT11], definiert das ultimative Ziel einer Suchmaschine als das Zurückgeben von für den Nutzer relevanten Treffern als Antwort auf eine Suchanfrage. [BJPBY10, S. 1].

Diese Seminararbeit beschäftigt sich mit der Quantifizierung der Qualität von Suchmaschinen. Zentrale These ist, dass die in der Literatur und Wirtschaft verwendeten Kriterien zur Bemessung der Qualität von Suchmaschinen und der Zufriedenheit der Nutzer mit diesen Suchmaschinen unzureichend sind und keinen quantitativen Vergleich verschiedener Suchsysteme oder die Feststellung der möglichen Verbesserung eines Systems nach einer entsprechenden Maßnahme zulassen. Ziel ist es aufzuzeigen, dass die Nutzerzufriedenheit nicht in ausschließlichem Zusammenhang mit der Relevanz der Treffer einer Suchanfrage steht. Dazu muss bewiesen werden, dass andere Faktoren ebenso Einfluss auf die Zufriedenheit der Nutzer haben. Die Ergebnisse dieser Arbeit könnten Auswirkungen auf die zukünftige Entwicklung und Qualitätsoptimierung von Suchmaschinen haben und zu einer Verschiebung der Prioritäten in der Entwicklung führen.

Zunächst werden in Kapitel 2 verwandte Arbeiten präsentiert, durch deren Verknüpfung für die These dieser Arbeit argumentiert werden kann. Wie Informationsbeschaffungssysteme entstanden und Ranking zu ihrem wichtigsten Qualitätsparameter wurde, wird in Kapitel 3 betrachtet. Die Motivation für den Betreiber einer Suchmaschine, seine Qualität kontinuierlich zu steigern, wird zu Beginn des 4. Kapitels betrachtet, welches zudem beschreibt, wie die Qualität von Ranking anhand zweier Parameter experimentell quantifiziert werden kann. Wie Ranking zum Vergleichen und Verbessern der Qualität von Suchmaschinen verwendet werden kann, ist ebenfalls Teil dieses Kapitels. Was Qualität im Sinne der Betriebswirtschaftslehre, der Software im Allgemeinen und bei Suchmaschinen im Speziellen bedeutet, ist Thema des 5. Kapitels. Die Beweisführung zur These dieser Arbeit ist Inhalt des 6. Kapitels, wozu die in Kapitel 2 genannten Arbeiten referenziert werden. Kapitel 7 fasst die Arbeit zusammen, beantwortet die Forschungsfrage und gibt einen Ausblick über nächste Schritte und mögliche Ziele weiterführender Arbeiten.

## 2 Verwandte Arbeiten

Dieses Kapitel verweist auf verwandte Literatur, die im Laufe dieser Seminararbeit referenziert wird, um die Eingangsthese zu beweisen. Die hier dargestellten Arbeiten behandeln hauptsächlich die Qualität von Suchmaschinen und die Parameter, an denen diese quantifiziert wird. Sergey Brin und Lawrence Page, die Gründer von Google im Jahr 1998, beschrieben in ihrer Arbeit "*The Anatomy of a Large-Scale Hypertextual Web Search Engine*" neben den konzeptionellen Grundzügen ihres Suchmaschinenprototypen auch die Qualitätsprobleme der damals verfügbaren Suchmaschinen [BP98]. Ihr Ziel war, die Nutzerzufriedenheit durch qualitative Suchtreffer zu steigern. Dazu präsentierten sie den "PageRank" [BP98] Algorithmus, der die Hyptertextstruktur des Internets nutzen soll, um das Ranking zu optimieren. Die Steigerung der Qualität von Suchmaschinen stellte dabei ihre Motivation dar. PageRank approximiert die relative Relevanz eines Dokuments an der Anzahl der Hyperlinks, die auf es verweisen [PBMW98, S. 1-6]. Dokumente können als Knoten und Links als Kanten interpretiert werden, so wird die Ähnlichkeit zu rekursiven Routing-Algorithmen deutlich [YKS+01,

S. 96]. Die Studie *"Measuring Search Engine Quality"*, erschienen 2001 im *Information Retrieval Journal* (siehe Kapitel 3) beschäftigte sich ausführlich mit der Qualität von Suchmaschinen und quantifizierte diese bei den damals 20 populärsten Anbietern anhand der Relevanz der Suchtreffer [HCBG01].

Das 2009 in den USA an Microsoft ausgestellte Patent *"SYSTEM AND METHOD FOR MEASURING AND IMPROVING SEARCH RESULT RELEVANCE BASED ON USER SATISFACTION"* bemaß die Qualität von Suchmaschinen ebenfalls am Ranking [KWM+08]. Microsoft schlug vor, die Zufriedenheit der Nutzer mehr in die Bemessung der Relevanz von Suchtreffern einfließen zu lassen. Dazu modellierten die Autoren die Interaktion von Nutzern mit einer Suchmaschine und versuchten aus dem Verhalten dieser Nutzer Rückschlüsse über ihre Zufriedenheit mit den präsentierten Suchergebnissen abzuleiten. Yahoo definierte in dem 2010 an sie ausgestellten Patent *"METHOD AND SYSTEM FOR QUANTIFYING THE QUALITY OF SEARCH RESULTS BASED ON COHESION"* ebenfalls das Zurückgeben von relevanten Treffern als ultimatives Ziel ihrer Suchmaschine [BJPBY10].

In ihrer Studie *"Is Relevance Relevant? Market, Science and War: Discourses of Search Engine Quality"* aus 2007 führte Elizabeth Van Couvering Interviews mit den Betreibern von Suchmaschinen durch, welche das Ranking als Schlüsselparameter der Suchmaschinenqualität identifizierten [VC07]. Van Couvering betrachtete das Thema von einer wirtschaftlich-sozialen Seite und sah die ausschließliche Quantifizierung von Qualität anhand des Rankings als kritisch. Ebenso kritisch gegenüber der Relevanz von Ranking waren die Autoren der Studie *"In Google We Trust: Users' Decisions on Rank, Position, and Relevance"*, deren These es ist, dass die Nutzer ein blindes Vertrauen in Googles Ranking-Algorithmus haben und diesen nicht hinterfragen [PHJ+07, S. 803]. Dies impliziert, dass auch weniger relevante Treffer an den obersten Positionen der Suchtreffer den Nutzer zufriedenstellen, was in Kapitel 6.2 ausführlich diskutiert wird.

Mit der Qualität von Software im Allgemeinen beschäftigte sich die ISO-Norm 9126 [ISO01] (Grundlage von- und ersetzt durch ISO/IEC 25000) und die IEEE-Studie *"Measuring Software Product Quality: A Survey of ISO/IEC 9126"* [JKC04]. Das nachfolgende Kapitel erläutert die Entstehung von Suchmaschinen als Informationsbeschaffungssystem und zeigt auf, welche historischen Einflüsse Ranking zur meist verbreitetsten Metrik für Qualität gemacht haben.

## 3 Historische Entwicklung von Informationsbeschaffungssystemen

***Information Retrieval*** (deutsch: "Informationsbeschaffung") ist ein wissenschaftliches Teilgebiet der Informatik und beschreibt den Prozess des Durchsuchens einer Kollektion von Dokumenten, um einen bestimmten Informationsbedarf abzudecken [LM06, S. 1]. Die ersten "Information Retrieval Systems" [SC12, S. 1444] werden auf das 3. Jahrhundert vor Christus datiert, als der griechische Philosoph Callimachus erstmals Schriften katalogisierte [SC12, S. 1445]. Ähnliche

hierarchische Katalog- oder Karteikartensysteme waren bis zum Aufkommen von computergestützten System auch in neuzeitlichen Bibliotheken zu finden [LM06, S. 2]. Ab 1891 wurden erste Patente zu elektromechanischen Information Retrieval Systems ausgestellt. Diese verwendeten unter anderem Mikrofilm als Speichermedium für den Katalog [SC12, S. 1445]. In seinem 1945 erschienenen Artikel "As We May Think" beschreibt Vannevar Bush den "Memex", eine fiktive, tischgroße Maschine, die in der Lage ist, tausende Dokumente auf "Supermikrofilm" zu Durchsuchen und über eine integrierte Kamera automatisch neue Dokumente zur Kollektion hinzuzufügen [Bus45, S. 101-108]. Dieser Apparat weißt große Ähnlichkeit mit heutigen Internetsuchmaschinen, beziehungsweise der Struktur des Internets im Allgemeinen, auf [LM06, S. 3]. Mit dem Aufkommen von Computern und der exponentiell wachsenden Speicher- und Rechenleistung (vergleiche hierzu "mooresches Gesetz" [Sch97, S. 53]) waren herkömmliche Methoden zur Information Retrieval erschöpft und erste computergestützte Systeme wurden entwickelt [SC12, S. 1446]. Das Internet ist unstrukturiert, dynamisch und die größte jemals existente Dokumentenkollektion, was bisher die größte Herausforderung für Informationsbeschaffungssysteme darstellt. Das Dokumentenreichtum des Internets stellt für die Betreiber von Suchmaschinen ein Problem dar, denn eine Suchanfrage generiert teilweise mehrere Millionen Treffer (Beispiel Suchanfrage "Hochschule Darmstadt" 3,17 Mio. Ergebnisse, google.de am 03.03.2018), von denen nur ein Bruchteil für den Nutzer relevant sind. Dieser sucht sprichwörtlich "die Nadel im Heuhaufen" [LM06, S. 3]. Suchmaschinen verwenden deshalb Ranking- bzw. Scoring-Algorithmen, welche die Reihenfolge der Suchtreffer bezüglich der Relevanz für den Nutzer optimieren. Sergey Brin und Lawrence Page haben bereits 1998 erkannt: "The biggest problem facing users of web search engines today is the quality of the results they get back." [BP98, S. 15]. Deshalb hat der Betreiber einer Suchmaschine zum Ziel, die Ergebnisqualität für den Nutzer zu maximieren. Ranking, welches als hauptsächlicher Faktor für die Qualität von Suchmaschinen angenommen wird (siehe Kapitel 2 und [VC07], [BJPBY10], [HCBG01], [KWM+08]) erlangte seinen Stellenwert also historisch aus den technischen Limitationen, der Größe des Internets und den damaligen Anforderungen der Nutzer. Das Ranking nicht ausschließlicher Faktor für die Quantifizierung der Qualität von Suchmaschinen heutzutage sein kann und sollte, ist These dieser Seminararbeit (siehe Kapitel 1). Im folgenden Kapitel wird zunächst betrachtet, was die Betreiber zur ständigen Erhöhung der Qualität ihrer Suchlösung motiviert, und wie das Ranking (und damit die Qualität) verschiedene Suchmaschinen miteinander verglichen wird.

## 4 Vergleich von Suchmaschinen

Um die These, dass die aktuellen Kriterien zur Feststellung der Qualität von Suchmaschinen unzureichend sind um a) verschiedene Suchsysteme miteinander zu vergleichen und b) zu analysieren, ob ein Versionsupdate tatsächlich eine Verbesserung der Qualität erzielt hat, muss zunächst betrachtet werden wie eine Suchmaschine mit sich selbst, und Anderen verglichen wird.

### 4.1 Motivation der Suchmaschinenbetreiber

Der Markt für Internet-Suchmaschinen ist kompetitiv, da fast jede Person mit Internetzugang bereits Nutzer einer Suchmaschine ist. Die Möglichkeit, den Nutzer durch niedrige Qualität an eine Suchlösung der Konkurrenz zu verlieren, macht Microsoft in ihrer Studie zur Suchtrefferqualität deutlich: "[...] poor result ranking may cause the user [...] to switch to another search system before encountering highly relevant results." [KWM+08, S. 2]. Zu einer deckungsgleichen Konklusion kam auch Van Couvering: "[...] satisfied customers will recommend other satisfied customers [...], whereas dissatisfied customers will leave [...]" [VC07, S. 872]. Der entscheidende Faktor für den Nutzer ist seine Zufriedenheit (User Satisfaction), welche wiederum von der Qualität der entsprechenden Suchmaschine abhängig ist. Die Betreiber haben also ein wirtschaftliches Interesse an der Steigerung der Nutzerzufriedenheit, beziehungsweise der Qualität ihrer Suchlösung: "[...] quality is linked to [the] satisfaction [...] of customers, which in turn is linked to revenue [...]" [VC07, S. 873]. Um die Qualität der eigenen Suchmaschine und der, der Konkurrenz zu quantifizieren, wird die Relevanz der Suchtreffer als ausschlaggebender Faktor verwendet: "Relevance [is] being used as a competitive measure to judge the quality of other search engines." [VC07, S. 881]. Die Motivation der Hersteller von Suchmaschinen ist also wirtschaftlicher Natur. Damit besteht für diese auch ein Interesse, sich mit anderen Anbietern quantitativ zu vergleichen und die Effizienz ihres eigenen Ranking-Algorithmus zu messen. Welche Parameter wiederum die Relevanz ausmachen, wird im Folgenden analysiert.

### 4.2 Quantifizierung von Relevanz

Die Relevanz von Suchtreffern ist definiert als "[...] the extent to which the result [of a search] correlates to the intent of the user performing the search." [KWM+08, S. 1] und wird an den Faktoren ***Precision*** und ***Recall*** gemessen. Precision ist dabei die "Reinheit" oder "Genauigkeit" einer Trefferliste und beschreibt den Anteil der relevanten Dokumente in der Ausgabe einer Suchmaschine [KWM+08, S. 1]. Der selten verwendete redundante Wert "Fallout" (deutsch: "Ausfall") gibt an, wie gut eine Suchmaschine die Ausgabe von irrelevanten Dokumenten vermeidet. Mathematisch betrachtet ist die Precision der Quotient aus relevanten Dokumenten und der Gesamtzahl der angezeigten Treffer [LM06, S. 204]. Recall hingegen steht für die "Vollständigkeit" [KWM+08, S. 1] der Trefferliste und kann als Verhältnis von der Anzahl der relevanten zurückgegeben Dokumente zu der Anzahl der relevanten Dokumente im gesamten Index beschrieben werden [LM06, S. 204]. Würde beispielsweise die Rückgabe der Suchanfrage "Tesla Sportwagen" auch Suchtreffer (englisch "Hits") über Nikola Tesla enthalten, so wäre die Präzision gering. Würde die Trefferliste allerdings auch für den Nutzer relevante Ergebnisse zu Sportwagen anderer Marken einschließen, so wäre eine hohe Vollständigkeit gegeben. Einerseits soll also vermieden werden, irrelevante Dokumente in der Ergebnismenge zu inkludieren um eine hohe Präzision zu gewährleisten, andererseits sollten viele Dokumente enthalten

sein, um umfassendere Suchergebnisse zu präsentieren. Wird also versucht, den Recall durch mehr disjunktive Treffer in der Ergebnismenge zu steigern, so ist dies nur im Austausch mit verminderter Präzision möglich. Generell besteht also ein Zielkonflikt zwischen den beiden Parametern der Relevanz: "[...] there is a trade-off between precision and recall. [...] the higher the the precision, the lower the recall [and vice versa.]" [KWM+08, S. 1]. Wie die beiden Parameter bei Suchmaschinen gemessen werden, wird im nächsten Kapitel behandelt.

### 4.3 Text REtrieval Conference

Die Text REtrieval Conference (TREC) ist eine seit 1992 jährlich stattfindende wissenschaftliche Konferenz auf dem Gebiet der Informationsbeschaffung. Zu ihren Errungenschaften zählt u.a. die Verdopplung der Effektivität (Güte des Rankings) von Suchmaschinen in den Jahren bis 1998 [Web10, S. 3]. Um diese Verdopplung feststellen zu können, verwendet die TREC eine experimentelle Methode zur Quantifizierung der Parameter Precision und Recall, also der Relevanz von Suchergebnissen. Dazu werden zunächst eine Kollektion von Dokumenten und Suchanfragen mit wohlbekannten Eigenschaften benötigt, gegen die getestet werden kann [SNC01, S. 66] [LM06, S. 8]. Dies entspricht dem Cranfield Paradigma, dessen Kern es ist die Evaluierung von *Retrieval Systems* durch Auslagerung und Minimalisierung der menschlichen Komponente reproduzierbar und einfacher zu machen [Voo02, S. 355ff]. So besteht die Testkollektion der TREC beispielsweise aus 18.5 Millionen Dokumenten und 10.000 Beispiel - Queries [HCBG01, S. 34]. Bei kleineren Dokumentensammlungen war es üblich, für jedes Dokument manuell die Relevanz als binären Wert zu bestimmen. Dazu wurden möglichst viele Testperson befragt, ob ein Dokument für eine bestimmte Suchanfrage relevant oder irrelevant ist. Bei beispielsweise 5000 Testdokumenten und 100 Suchanfragen müssten die Testsubjekte (auch "Richter", englisch "Judges") im Vorfeld für jedes Dokument bestimmen, ob es irrelevant (0) oder relevant (1) für die einzelnen Suchanfragen ist. Dies Beurteilungen werden als ***Relevance Judgements*** bezeichnet. So wird exakt definiert, welche Dokumente in der Ergebnisliste einer bestimmten Suchanfrage enthalten sein sollten. Die Suchmaschinenhersteller optimieren ihre Algorithmen dann so, das die Werte für Precision und Recall möglichst groß sind, wobei beide Parameter negativ voneinander abhängig sind. Es werden nicht erneut menschliche Testpersonen benötigt, was den Test zunächst deterministisch, und damit reproduzierbar macht. Würde der zu testende Ranking-Algorithmus keine irrelevanten Dokumente zurückliefern, hätte er eine Precision von 1.0 (100 Prozent). Wären alle Dokumente der Testkollektion enthalten, so beträgt der Recall 1.0 (100 Prozent), was allerdings dem Zweck eines Information Retrieval Systems widersprechen würde [BG94, S. 12]. Im Idealfall hätten also beide Parameter einen Wert von 1, dieser Zustand kann allerdings aufgrund des im vorherigen Kapitels beschriebenen Zielkonfliktes in der Realität unmöglich erreicht werden [LCS97, S. 68]. In folgendem Sonderfall jedoch würde der Idealzustand erreicht: a) alle Dokumente im Index sind relevant für die Suchanfrage und b) alle Dokumente im Index werden dem Nutzer angezeigt. Im Durchschnitt sind 5,35 Prozent der Dokumente im gesamten Index

relevant für eine Suchanfrage [SNC01, S. 67, TREC-8] innerhalb der TREC-Testkollektion.

Für größere Kollektionen wie die der TREC in späteren Instanzen, werden manuelle Relevance Judgements lediglich für einen Pool aus den 100 ersten Dokumenten durchgeführt, die die populärsten Suchmaschinen zurückgeben [SNC01, S. 66]. David Hawking et. al. haben in *"Measuring Search Engine Quality"* 1999 die jeweilige Performanz der 20 führenden Suchmaschinen gemessen. So lag Google auf Platz 3 und Yahoo!, damaliger Marktführer, auf Platz 16 [HCBG01, S. 45]. Da Ranking als primäres Qualitätsmerkmal von Suchmaschinen betrachtet wird, wurde damit die Qualität von Google als besser im Vergleich zu Yahoo! angenommen. Suchmaschinenhersteller vergleichen also ihre Qualität, beziehungsweise die angenommene Zufriedenheit ihrer Nutzer bis heute quantitativ über die Messmethodik der Text REtrieval Conference. Die große Problematik bei dieser Methodik liegt bei der menschlichen Komponente: "[...] people usually disagree about relevance. [...] Even a single individual may be inconsistent in judging relevance." [SNC01, S.66]. Außerdem spiegeln die statischen Testkollektionen die dynamische Struktur des Internets nur unzureichend wieder. Suchmaschinen und deren Ranking-Algorithmen nähern sich asymptotisch der maximal möglichen Optimierung von Precision und Recall an. In Zukunft wird der Aufwand der nötig ist, um eine für den Nutzer wahrnehmbare Verbesserung des Rankings zu generieren, überproportional ansteigen. Außerdem werden die *Relevance Judgements* statt wie vorgeschlagen von möglichst Vielen, nur von einer Person durchgeführt [SNC01, S. 67]. Zur Berechnung von Precision und Recall muss die Anzahl (Betrag) der relevanten Dokumente bekannt sein. Im Internet kann dieser Wert nicht eindeutig bestimmt werden, weshalb der er approximiert werden muss. Aufgrund all dieser Schwierigkeiten, lässt sich das TREC-Verfahren zur Quantifizierung der Qualität von Relevanz nur bedingt auf das Internet übertragen.

### 4.4 Verbesserung von Suchmaschinen

Wie in Kapitel 4.1 festgestellt, besteht für den Betreiber einer Suchmaschine ständig die Motivation, die Qualität seiner Suchmaschine zu erhöhen. Um also feststellen zu können, ob eine entsprechende Maßnahme, zum Beispiel ein Versionsupdate, tatsächlich die gewünschte Wirkung auf die Nutzerzufriedenheit erzielt hat, muss ebenfalls die Qualität quantifiziert werden. Dazu kann die Methodik der Text REtrieval Conference kongruent verwendet werden, da Precision und Recall nicht nur den Vergleich werden verschiedener Systeme ermöglichen, sondern auch als Maßstab dienen um die Effizienz der eigenen Suchmaschine zu steigern [LM06, S. 8].

## 5 Qualität und Nutzerzufriedenheit

In diesem Kapitel wird betrachtet, was Qualität im Sinne der Betriebswirtschaftslehre, der Software im Allgemeinen und bei Suchmaschinen im Speziellen, bedeutet.

### 5.1 Qualität im Allgemeinen

Das Gabler Wirtschaftslexikon definiert Qualität als die "[...] Übereinstimmung von Leistungen mit Ansprüchen" [Win01]. Die Ansprüche werden dabei nicht nur von der Zielgruppe eines Produktes, sondern auch von dessen Entwicklern, gestellt. Die Leistungen unterteilen sich dann in naturwissenschaftlich messbare Werte (z.B die Ladezeit einer Webapplikation oder eben das Ranking bei Suchmaschinen) und die subjektiv wahrgenommene Qualität, welche sich in der Nutzerzufriedenheit widerspiegelt. Die Wichtigkeit und Problematik der Messung von Qualität beschreibt Van Couvering: "Quality is an important normative issue, yet despite its importance, it is problematic to study emperically". Die Zufriedenheit der Endnutzer verhält sich oft gegensätzlich, teilweise antiproportional, zu den Annahmen der Entwickler. So ist der Nutzer oft mit der aktuellen Lösung zufrieden und wünscht explizit gar keine Veränderung des Produkts.

### 5.2 Software-Qualität nach ISO-9126 im Allgemeinen

Die Qualität von Software im Allgemeinen wird in der ISO-Norm 9126 *"Software Product Quality"* anhand der 6 Kategorien "Funktionalität", "Zuverlässigkeit", "Benutzbarkeit", "Effizienz", "Wartbarkeit" und "Portabilität" gemessen, welche sich wiederum in insgesamt 27 Unterkategorien aufteilen [JKC04, S. 89] [ISO01]. Großer Wert wird dabei, durch die Kategorien "Wartbarkeit" und "Portabilität", auf die Qualität des Source-Codes gelegt. Diese Qualitätseigenschaften haben keinen direkten Einfluss auf die Zufriedenheit des Nutzers, da diese für ihn nicht wahrnehmbar sind. Lediglich indirekte Effekte, wie gesteigerte Performance durch nachgebesserten Source-Code, haben Einfluss auf die Nutzerzufriedenheit.

### 5.3 Qualität von Suchmaschinen im Speziellen

Wie in Kapitel 2 bereits deutliche wurde, wird die Qualität von Suchmaschinen in der Literatur und Wirtschaft daran bemessen, wie relevant die zurückgegebenen Treffer bei einer Suche sind. Je relevanter diese sind, desto größter ist die Qualität der Suchmaschine. Die Betreiber von Suchlösungen verwenden die Güte des Rankings als quantitativen Maßstab, um ihre Systeme mit denen der Konkurrenz zu vergleichen (siehe Kapitel 4). Die Nutzerzufriedenheit wird mit der Qualität und der Relevanz der Suchtreffer gleichgesetzt: "[...][the] customer is satisfied through greater relevance" [VC07, S. 879] und "[...] quality as relevance [...]" [VC07, S. 880]. Gutes Ranking bedeutet also hohe Qualität und damit, so die Theorie, auch hohe Nutzerzufriedenheit.

### 5.4 Nutzerzufriedenheit

Qualität bei Suchmaschinen wird im Zuge dieser Seminararbeit, wie auch in der verwandten Literatur (siehe Kapitel 5.3) mit der Nutzerzufriedenheit gleichgesetzt. Kapitel 5.2 macht zwar deutlich, das bestimmte Qualitätseigenschaften von Software im Allgemeinen keinen direkten Einfluss auf die Zufriedenheit der

Nutzer haben, allerdings wird die Qualität von Suchmaschinen in diesem Kontext an der Nutzerzufriedenheit gemessen, weshalb beide Begriffe im Folgenden äquivalent verwendet werden. Die Nutzerzufriedenheit wird im Wesentlichen durch Umfragen ermittelt, (vergleiche hierzu *"Measurement of Computer User Satisfaction"* [PB80]) allerdings sind nach Huang et. al. auch indirekte Methoden über Analyse des Nutzerverhaltens möglich, vergleiche [HCHL08].

# 6 Beweisführung

In diesem Abschnitt wird auf die in Kapitel 2 verwiesenen Arbeiten referenziert, um die Eingangsthese dieser Seminararbeit zu belegen. Dazu wird die These in die Teilaussagen "Ranking ist unzureichend um festzustellen, ob eine Suchmaschine verbessert wurde" und "Ranking ist unzureichend, um verschiedene Suchsysteme miteinander zu vergleichen" dividiert.

## 6.1 Quantifizierung der Verbesserung eines Suchsystems

Nachfolgend wird eine Studie zur Nutzerzufriedenheit zurate gezogen, welche die Software AG an ihrer unternehmensinternen Suchmaschine *"iSearch"* durchgeführt hat. Diese lässt die Argumentation zu, dass Ranking nicht allein ausschlaggebend für die Bemessung der Qualität von Suchmaschinen sein kann. Im November 2016 wurde eine Befragung mit einem Stichprobenumfang von 48 Personen durchgeführt, um die Zufriedenheit der Nutzer mit der Version 1.0 der Suchmaschine zu quantifizieren. Diese ergab eine durchschnittliche Zufriedenheit von 70 Prozent. Der gleiche Personenkreis wurde im Januar 2018 zum Release der Version 1.1 erneut befragt und beurteilte die Qualität der Suchmaschine nun mit durchschnittlich 90 Prozent. Version 1.1 enthielt umfangreiche Änderungen am User Interface, der Usability und der User Experience. Der Ranking-Algorithmus hingegen blieb unverändert. Wäre die Nutzerzufriedenheit alleine abhängig von der Relevanz der Suchtreffer, so hätten die Änderungen der Version 1.1 nicht zu einer Steigerung der Nutzerzufriedenheit in diesem Ausmaß führen dürfen. Daraus lässt sich ableiten, dass die Nutzerzufriedenheit auch von anderen Faktoren abhängig ist und nicht allein der Qualität vom Ranking unterliegt. Wird das Ranking also alleine als Parameter für Qualität gewählt, so ist dies nachweislich unzureichend um die Nutzerzufriedenheit zu quantifizieren.

## 6.2 Quantitativer Vergleich verschiedener Suchsysteme

Die in Kapitel 2 vorgestellten Studien können miteinander verknüpft werden, um die zweite Teilaussage der These ebenfalls zu belegen. Die Autoren des Patents [KWM+08] haben die Interaktion von Nutzern mit einer Suchmaschine modelliert und versuchen aus dem Verhalten dieser Nutzer Rückschlüsse über ihre Zufriedenheit mit den präsentierten Suchergebnissen abzuleiten. Die Interaktion *"Accept"* beschreibt das implizite Ende einer Interaktion mit der Suchmaschine,

da diese als abgeschlossen gilt, sobald der Nutzer eines der angezeigten Dokumente als Antwort für seinen Informationsbedarf *akzeptiert*. Ein akzeptierter Hit ist relevant ("The user [...] decided [the result] was [...] related to his intent" [KWM+08, S. 7]) und damit zufriedenstellend ("[...] satisfaction may be indicated by accepting the [...] result" [KWM+08, S. 8]). Würde ein Nutzer allerdings auch wenig relevante Treffer akzeptieren, impliziert dies, dass die Nutzerzufriedenheit weniger stark mit dem Ranking korreliert, als bisher angenommen.

Zur "Relevanz der Relevanz" haben Pan et. al. in der Studie [PHJ+07] geforscht. Die Autoren verwendeten ein Gerät zur Aufzeichnung der Augenbewegungen um analysieren zu können, welche Positionen vom Nutzer angesehen, beziehungsweise angeklickt wurden. "Position" beschreibt dabei die Stelle eines Treffers innerhalb der Ergebnisliste, Position 1 ist daher, so die Theorie, der am meisten relevante Treffer für eine Suchanfrage. Den Nutzern wurde dabei jeweils ein Set aus 10 Suchtreffern in drei verschiedenen Konfigurationen präsentiert. Der Zustand "Normal Condition" gab die Treffer in der gewohnten Sortierung nach Relevanz aus, bei der "Swapped Condition" waren jeweils der erste und zweite Suchtreffer in ihrer Reihenfolge vertauscht und in der "Reversed Condition" wurde das Ranking invertiert [PHJ+07, S. 809]. Bei der Anzahl der neuformulierten Suchen konnten zwischen den drei Zuständen kein Unterschied festgestellt werden [PHJ+07, S. 812]. Zwischen den beiden Extremfällen "Normal" und "Reversed" konnte ein Unterschied in der Anzahl der angesehenen Suchtreffer und der Erfolgsrate gemessen werden, die Versuchssubjekte vermuteten den Fehler jedoch bei sich selbst, und nicht bei der Reihenfolge der Suchtreffer [PHJ+07, S. 812f]. Bei vollständig invertiertem Ranking wählten etwa 50 Prozent der Nutzer die wenig relevanten Treffer an den Positionen 1 und 2 (9 und 10 nach Ranking) aus. Die beiden relevantesten Treffer 1 und 2 an den Positionen 9 und 10 wurden nur von 20 Prozent der Nutzer angesehen und von 10 Prozent angeklickt [PHJ+07, S. 814f].

Die Signifikanz dieses Experimentes wird in der Verknüpfung mit der erstgenannten Studie deutlich. Es wird angenommen, dass ein Nutzer zufrieden ist, sobald er einen Suchtreffer *akzeptiert*. Laut [PHJ+07] werden auch Ergebnisse akzeptiert die, nach der in Kapitel 4 beschriebenen Messmethode über Recall und Precision, wenig relevant sind. Nach den Erkenntnissen von [KWM+08] können damit auch vermindert relevante Treffer den Nutzer zufriedenstellen. Das bedeutet, dass Ranking weniger Aussagekraft über die Nutzerzufriedenheit hat, als in der Literatur (siehe Kapitel 5.3) definiert. Suchmaschinenhersteller vergleichen ihre Systeme über die Zufriedenheit ihrer Nutzer, und bemessen diese wiederum an der Relevanz, welche durch die Argumentation in diesem Kapitel an Signifikanz verliert. Diese Vergleichsmethode ist wenig präzise und Ranking als Parameter für diesen Vergleich nachweislich unzureichend.

### 6.3 Wichtigkeit von Relevanz in verwandter Literatur

Auch in verwandter Literatur wird die Validität von Ranking infrage gestellt. So wird kritisiert, dass Ranking die am weitesten verbreitete Eigenschaft für Qualität ist, und damit andere Qualitätsziele außer Acht gelassen werden [VC07, S.

882f]. Langville und Meyer merken an, dass eine Suchmaschine mit hochgradig optimierten Recall- und Precision-Werten auch dann unqualitativ ist, wenn diese 30 Minuten pro Request benötigt [LM06, S. 8]. Sie artikulieren damit andere Qualitätsziele, wie die Ladezeit, welche auch in der ISO-Norm 9126 als Eigenschaft von qualitativer Software im Allgemeinen deklariert ist, vergleiche hierzu [JKC04] [ISO01]. Andere Herangehensweisen setzen die Nutzerzufriedenheit in den Mittelpunkt, diese sind jedoch weniger verbreitet als Precision und Recall als Metrik zur Quantifizierung der Qualität von Suchmaschinen, vergleiche hierzu *"Information Storage and Retrieval"* [Kor97].

## 7 Fazit und Ausblick

Dieses Kapitel dient der Zusammenfassung der bisherigen Erkenntnisse und der abschließenden Beantwortung der Forschungsfrage dieser Seminararbeit. Kapitel 3 stellte fest, dass das Ranking historisch durch die Anforderungen der Nutzer an rudimentäre Suchmaschinen zu Beginn des Internets und deren technischen Limitationen zum primären, quantifizierbaren Parameter für Qualität wurde. Der kompetitive Markt für Suchlösungen ist Motivation für die Anbieter, ihre Qualität sukzessiv zu steigern. Dazu bemessen sie die Relevanz der Hits über die Parameter Precision und Recall nach der Methode der TREC, siehe dazu Kapitel 4. Das Ergebnis dieser Messung wird zum quantitativen Vergleich mit anderen Suchmaschinen und als Referenzwert für die Verbesserung der eigenen Suchmaschine durch Updates, genutzt. In Kapitel 5 wurde definiert, was Qualität bei Software und bei Suchmaschinen im Speziellen bedeutet, und dass diese mit der Zufriedenheit der Nutzer und der Güte des Rankings gleichgesetzt wird. Über die Evaluation der Software AG internen Suchmaschine iSearch konnte in Kapitel 6 belegt werden, dass Ranking allein unzureichend ist, um die Verbesserung eines Systems durch ein Update zu quantifizieren. Durch Verknüpfung des Suchmaschinen-Interaktionsmodells und seiner Annahmen über die User Satisfaction mit der Studie zum Nutzerverhalten bei invertiertem Ranking konnte belegt werden, dass das Ranking allein unzureichend für die Vergleichbarkeit von verschiedenen Suchsystemen ist. Die Forschungsfrage dieser Seminararbeit lässt sich also abschließend wie folgt beantworten: Die Qualität einer Suchmaschine und die Zufriedenheit derer Nutzer kann nicht alleine über die Güte des Rankings quantifiziert werden. Ranking ist also unzureichend als isolierter Qualitätsparameter für Suchmaschinen. Die in Kapitel 4.3 aufgezeigten Probleme der Messung von Precision und Recall (TREC) zeigen auf, dass die Güte des Rankings bisher nicht präzise, und für die dynamische Struktur des Internets repräsentativ, quantifiziert werden kann. In Kapitel 6.1 konnte belegt werden, dass weitere Faktoren, wie zum Beispiel die Qualität des User Interfaces auch Auswirkungen auf die Zufriedenheit der Nutzer von Suchsystemen haben. In Zukunft müssen also neue Parameter identifiziert werden, über die Suchmaschinenqualität präziser bemessen werden kann. Dazu könnten die in Kapitel 5.2 vorgestellten allgemeinen Kriterien für Qualität von Software verwendet werden, solange sich dabei an der betriebswirtschaftlichen Definition von Qualität aus

Kapitel 5.1 orientiert wird. Der Autor schlägt also vor, in weiteren Studien zu bestimmen, welche Eigenschaften zur Zufriedenheit der Nutzer von Suchmaschinen beitragen. Diese neuen Qualitätsparameter können dann in Zukunft das Ranking als Maßstab für Suchmaschinen ablösen. Der Autor betrachtet die Verwendung von Ranking zur Bestimmung der Qualität als nicht mehr zeitgemäß. Zudem wird das Optimierungspotential der Ranking-Algorithmen in Zukunft immer geringer, bis es gegen null geht, weshalb Kosten und Aufwand zur Erhöhung der Qualität langfristig ihren Nutzen übersteigen werden. Aufgrund der Wichtigkeit des Internets für die Kultur und den Fortschritt der Menschheit, sieht der Autor den Bedarf qualitative, zufriedenstellende Suchmaschinen zu entwickeln und ist der Meinung, dass dafür Suchsysteme umfassender betrachtet werden müssen.

# Literaturverzeichnis


[BG94] Michael Buckland and Fredric Gey. The Relationship between Recall and Precision. *Journal of the American Society for Information Science*, pages 12–19, 1994. John Wiley and Sons Inc.

[BJPBY10] Luciano Barbosa, Flavio Junqueira, Vassilis Plachouras, and Ricardo Baeza-Ytes. METHOD AND SYSTEM FOR QUANTIFYING THE QUALITY OF SEARCH RESULTS BASES ON COHESION. *United States Patent / Yahoo! Inc.*, 2010.

[BP98] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. 1998. Stanford University.

[Bus45] Vannevar Bush. As We May Think. *The Atlantic Monthly*, pages 101–108, 1945. The Atlantic.

[HCBG01] David Hawking, Nick Craswell, Peter Bailey, and Kathleen Griffiths. Measuring Search Engine Quality. *Information Retrieval*, 4:33–59, 2001. Kluwer Academic Publishers.

[HCHL08] Te-Yuan Huang, Kuan-Ta Chen, Polly Huang, and Chin-Laung Lei. A Generalizable Methodology for Quantifying User Satisfaction. *IEICE TRANS. COMMUN.*, 91:1260–1268, 2008. The Institute of Electronics, Information and Communication Engineers.

[ISO01] ISO/IEC 9126-1:2001 Software Engineering – Product Quality – Part 1: Quality model . Technical report, International Organization for Standardization, Geneva, CH, 2001.

[JKC04] Ho-Won Jung, Seung-Gweon Kim, and Chang-Shin Chung. Measuring Software Product Quality: A Survey of ISO/IEC 9126. *IEEE Software*, pages 88–92, 2004.

[Kor97] Robert Korfhage. *Information Storage and Retrieval.* Wiley, 1997.

[KWM+08] Kuldeep Karnawat, Thomas D. White, Mark B. Mydland, Steven C. Fox, and Thomas A. Taylor. SYSTEM AND METHOD FOR MEASURING AND IMPROVING SEARCH RESULT RELEVANCE

BASED ON USER SATISFACTION. *United States Patent / Microsoft Corporation*, 2008.

[LCS97] Dik L. Lee, Huei Chuang, and Kent Seamons. Document Ranking and the Vector-Space Model. *IEEE Software*, 14:67–75, 1997. IEEE.

[LM06] Amy N. Langville and Carl D. Meyer. *PageRank and Beyond: The Science of Search Engine Ranking*, volume 18. Princeton University Press, 2006.

[PB80] Sammy W. Pearson and James E. Bailey. Measurement of computer user satisfaction. 1980. Ariziona State University.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. 1998. Stanford University.

[PHJ+07] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In Google We Trust: Users' Decisions on Rank, Position and Relevance. *Journal of Computer-Mediated Communication*, 12:801–823, 2007. International Communication Association.

[SC12] Mark Sanderson and W. Bruce Croft. The History of Information Retrieval Research. *Proceedings of the IEEE*, 100:1444–1451, 2012. IEEE.

[Sch97] Robert R. Schaller. Moore's law: past, present and future. *IEEE Spectrum*, 34:52–59, 1997. IEEE.

[SNC01] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking Retrieval Systems without Relevance Judgments. *SIGIR Special Interest Group on Information Retrieval*, pages 66–73, 2001.

[VC07] Elizabeth Van Couvering. Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality. *Journal of Computer-Mediated Communication*, 12:866–887, 2007. International Communication Association.

[Voo02] Ellen M. Voorhees. The Philosophy of Information Retrieval Evaluation. *CLEF*, pages 355–370, 2002. Springer-Verlag Berlin Hamburg.

[Web10] William Webber. Evaluating the Effectiveness of Keyword Search. 2010. IEEE.

[Win01] Eggert Winter. *Gabler Wirtschaftslexikon*. 2001. Springer.

[YKS+01] Sejong Yoon, Doohyun Ko, Koh Sanghoon, Heungwoo Nam, and Sunshin An. PR-RAM: The Page Rank Routing Algorithm Method in Ad-hoc Wireless Networks. *Consumer Communications and Networking Conferenc*, 2001. IEEE.

[ZT11] Wugang Zhao and Edison Tse. Competition in Search Engine Market. *Journal of Business Strategies*, 2011. Center for Business and Economic Research.