Was bringt Tagging? Eine methodologische Herangehensweise an die Evaluation von Social-Tagging-Systemen

Diana Jurjević

Zentrum für Bildungsinformatik
Pädagogische Hochschule Bern
Muesmattstraße 29
CH-3012 Bern
diana.jurjevic@phbern.ch

Abstract: Das Angebot an Online-Informationen nimmt weiter rasant zu. Die neuen Informationsmedien, speziell die Dienste im Web 2.0, stellen die Nutzenden vor neue Möglichkeiten und Herausforderungen. Zur Erforschung der neuen Informationsdienste sind geeignete Zugänge und Methoden gefragt. Der Artikel beschreibt und analysiert eine nutzerorientierte Methode um Social-Tagging-Systeme auf ihre Suchfunktion hin zu evaluieren. Dazu werden klassische Retrievalmaße aus dem Information Retrieval um die sozialen Einflussgrössen in Social-Tagging-Systemen erweitert.

1 Einleitung

Bislang wurden in der Informatikausbildung die Technologien mehr oder weniger getrennt von ihren Nutzerinnen und Nutzern behandelt. Der Bereich Human Computer Interaction fristet gerade an sehr technisch orientierten Hochschulen oft ein Schattendasein. Mit dem Aufkommen des Web 2.0 gerät diese Haltung vermehrt in die Kritik. Das Verhalten der Online-Nutzenden bei der Entwicklung von Informatiksystemen kann nicht mehr ausgeblendet werden. Ben Shneiderman fordert deshalb, das Forschungsfeld der Informatik um soziale Einflussfaktoren zu erweitern. Entwicklungen von Web-Technologien müssen Nutzerbedürfnisse berücksichtigen und deren Bedarf analysieren [Sh07]. Tim Berners Lee, der Erfinder des WWW, geht noch einen Schritt weiter. Er und weitere Wissenschaftler fordern, der Forschung über das WWW mehr Gewicht zu verleihen, indem neben der Informatik (Computer Science) die neue Disziplin „Web Science“ geschaffen wird [HS08]. Der tabellarische Vergleich der Disziplin „Web Science“ mit der bedeutend älteren Disziplin „Computer Science“ manifestiert den Trend im Internet, nicht mehr die Daten und Programme würden im Vordergrund stehen, sondern die Menschen, die sich für diese Daten und Programme interessieren.

Computer Science	Web Science
Topics	
Computer Networks	Social networks
Packet Switching	Voice over IP, music sharing
Information	Relationships
Programming languages	Wikis, blogs, tagging
Databases, operating systems, compilers	E-commerce, e-learning, e-government, medical informatics, financial analysis
3D graphics, rendering algorithms, computational geometry, object modelling	Creating and sharing video, animation, music, photos, maps

Abbildung 1: Computer Science vs. Web Science [Sh07]

Im Vordergrund unseres Forschungsinteresses stehen die Menschen, die selbst Meta-Daten, in Form von Tags, generieren und sich auf diese Weise am Aufbau der Web 2.0-Dienste beteiligen. Forschungsschwerpunkt sind die Social-Tagging-Syteme, insbesondere die Möglichkeit mit diesen nach Informationen im Web zu suchen.

Dabei interessieren uns die Fragen: Haben tagbasierte Suchsysteme Einfluss auf die Recherche von Informationen im Internet? Wenn ja, wie wirken sich diese auf die Internetrecherche aus und welche Suchstrategien sind erfolgreich in Social-Tagging-Systemen? Welche sind die Vor- und Nachteile von tagbasierten im Vergleich zu volltextbasierten Suchsystemen, wie wir sie von den großen, kommerziellen Suchmaschinen kennen? Um diese Fragen zu beantworten wurde eine Verfahren entwickelt, welche es ermöglicht, Social-Tagging-Systeme, unter Einbeziehung ihrer nutzergenierten Inhalte, zu untersuchen.

2 Evaluation von Suchmaschinen

Die Forschungsdisziplin Information Retrieval untersucht und evaluiert Online-Suchdienste nach definierten Kriterien.

Ziel der Evaluierungen von Suchmaschinen ist es herauszufinden wie gut ein System relevante Dokumente in einer Kollektion von Dokumenten findet. Grundlegende Aufgabenstellungen an die Systeme können z.B. sein:

1. Finde das beste Dokument
2. Finde alle relevanten Dokumente

Theoretisch wird im Information Retrieval die ganze Kollektion rangiert. In der Praxis ist ein Cutoff Level, λ, definiert. Der Cutoff Level ist der Anhaltepunkt in der Suche und markiert das Suchende. Alle Dokumente, die in der Rangierung größer als λ sind werden erscheinen nicht in der Trefferliste und fließen nicht in die Bewertung ein [BV05].

Die Relevanzbewertungen der Suchergebnisse werden „qrels“ genannt. Diese bestimmen für jedes Thema welche Dokumente gefunden werden sollen. Der einfachste Typ der

Bewertung ist die binäre Unterscheidung: relevant und nicht-relevant. Gefundene Dokumente werden nach diesen beiden Merkmalen bewertet [BV05].

Zentraler Maßstab für die Qualität eines Suchsystems ist das optimale Verhältnis von Ausbeute und Präzision. Auf eine Suchanfrage soll ein Suchdienst möglichst alle zur Verfügung stehenden relevanten Informationen anzeigen (Ausbeute), aber auch genau nur diese und keine irrelevanten (Präzision).

3 Evaluation von Suchmaschinen im Web

Die klassischen Retrievalmaße stammen aus der Zeit vor dem großen Internetdurchbruch in den 1990er Jahren und berücksichtigen nicht die Veränderungen, die mit den Entwicklungen im Internet einhergegangen sind. Im Vordergrund stehen statistische Maße, das Nutzungsverhalten der User wird praktisch vollständig ausgeblendet. Das optimale Verhältnis von Ausbeute und Präzision ist der entscheidende Faktor bei der Evaluation. In der Praxis sagen diese Maße nur wenig über die Qualität einer Suchmaschine im Web aus. Die klassischen Retrievalmaße lassen sich nicht einfach auf das Web übertragen. 83% aller User, die eine Suchmaschine in Europa nutzen, schauen sich nur die erste Seite der Trefferlisten an. Diese Zahlen ermittelten Jansen und Spink [JS05] anhand von Webprotokollen. Sie konnten auch einen Abwärtstrend festmachen, es werden immer weniger Treffer angesehen. Durch die Beschränkung auf die erste Trefferseite fallen alle relevanten Dokumente weg, die erst weiter hinten in den Trefferlisten erscheinen. Für die Evaluation von Suchsystemen bedeutet das, dass es nicht ausreicht, wenn ein Suchsystem relevante Dokumente findet, diese müssen auch auf der ersten Seite angezeigt werden. Alle relevanten Dokumente, die nicht auf der ersten Ergebnisseite erscheinen, nimmt die große Mehrheit der Nutzenden gar nicht wahr.

Röhle [Rö07] hält fest, dass Web-Suchmaschinen zwar eindeutig zu den Nachfolgern früherer Information-Retrieval-Systeme gehören, es aber gravierende Unterschiede gibt, welche die Forschung vor neue Fragen und Herausforderungen stellt. Frühere Information-Retrieval-Systeme wurden hauptsächlich von kompetenten Nutzern für spezifische Recherchen in homogenen und vollständigen Datenbanken eingesetzt.

„Völlig anders stellt sich die Situation bei den Suchmaschinen dar: Als integraler Bestandteil der Internetnutzung […] betreffen ihre Relevanzkriterien einen wesentlich grösseren Kreis von Nutzern, die mit sehr unterschiedlichen Motivationen […] eine Auswahl des Datenbestandes durchsuchen, ohne zu wissen worauf sich diese Auswahl gründet. Verschärft wird diese Situation durch die starke Konzentration auf dem Suchmaschinenmarkt, die oftmals geringe Suchkompetenz der Nutzer und die Popularität der Suchmaschinen bei Multiplikatoren wie Journalisten und Wissenschaftlern […]."

Lewandowski [Le07] weist anhand von Beispielen aus Retrievaltests Defizite am Präzisions-Maßstab auf. Er benutzt die Beispiele, um die Notwendigkeit hervorzuheben, weitere Retrievalmaße für die Bewertung von Suchdiensten im Internet einzusetzen. Auch andere Forscher folgen diesem Ansatz. Viele von ihnen kommen aus den Informations-

wissenschaften. Sie haben neue, webspezifische Maßstäbe zur Evaluation von Suchmaschinen im Internet aufgestellt. Für eine Übersicht siehe [Le07].

Um die Systeme nachhaltig zu verbessern, müssen sich die Evaluationen vermehrt an der Realität der Suchenden und ihren Interaktionen orientieren. Auf diese Weise könnten neue, wirksame und nachhaltige Forschungsergebnisse gefunden werden. Bereits 1995 hielt Saracevic fest:

"The issue and challenge for any and all IR evaluations are the broadening of approaches and getting out of the isolation and blind spots of single level, narrow evaluation. How can interaction be ignored in IR evaluation at any level?" [Sa95]

Spink entwickelte einen Nutzer-zentrierten Ansatz, um den Suchinteraktionen der Nutzer mehr Gewicht zu verleihen. Das Nutzer-zentrierte Verfahren wurde eingesetzt, um eine Suchmaschine auf Usability und Effektivität zu testen. Zur Datenauswertung wurden zusätzlich zu den üblichen Logfiles Fragebögen und Relevanzbewertungen der Nutzer herangezogen. Auf diese Weise konnten die klassischen Retrievalmaße ermittelt werden und zusätzlich Nutzerdaten erhoben werden.
Ein zentrales Ergebnis der Studie war, dass der Maßstab Präzision nicht mit den Nutzer-zentrierten Messdaten korreliert. Einige User haben große Fortschritte im Suchprozess beschrieben, obwohl die Suchmaschine auf ihre Suchanfragen eine niedrige Präzision erzielte und umgekehrt stellten einige User nur geringe Fortschritte fest bei einer hohen Präzision der Ergebnisse [Sp02]. Präzisionswerte alleine sagen noch nichts über den Fortschritt des Suchvorgangs aus.

4 Evaluation von Suchmaschinen im Web 2.0

Mit dem Aufkommen des Taggings im Web 2.0 kommt, neben der Kollektion und den Nutzenden, eine neue Einflussgrösse in die Evaluation moderner Suchsysteme hinzu. Suchende nehmen nicht nur durch ihr Nutzungsverhalten Einfluss auf die Retrievalmaße, sondern sie bauen auch an neuen Arten der Informationserschließung im Internet mit. Web-2.0-Dienste stützen sich stark auf das Tagging ab.

Austauschplattformen wie Flickr, YouTube oder der Social-Bookmarking-Dienst Delicious sind eine Mischform zwischen von Menschen erstellten Themenkatalogen und algorithmischen Suchmaschinen. Betrachtet man diese unter den klassischen Retrievalmaßen von Ausbeute und Präzision, hat dies Vor- als auch Nachteile: Auf der einen Seite erschließen diese semantischen oder sozialen Suchdienste weniger Informationen als die großen algorithmischen Suchmaschinen, die mit ihren Crawlern das Internet möglichst in der ganzen Breite und Tiefe absuchen. Rein statistisch gesehen nimmt damit die Ausbeute ab. Dieser Nachteil fällt aber kaum ins Gewicht, da die meisten User nur die ersten Treffer eines Suchdienstes inspizieren. Auf der anderen Seite bieten Suchdienste wie Delicious gewissermaßen von anderen Usern handverlesene Informationen an. Die Community dient als Filter für die Inhalte. Damit erhöht sich die Präzision bei der Suche.

Eine Evaluation von Social Tagging Systemen, welche nur die Kollektion und die Suchinteraktionen von Seiten der Nutzer berücksichtigt, greift zu kurz, weil der Nutzer in Social-Tagging-Systemen zwei Funktionen erfüllt. Der Nutzer kann Konsument und Produzent sein. Als Konsument kann er nach einem Dokument suchen und als Produzent beteiligt er sich durch das Hinzufügen von Meta-Daten am Aufbau des Systems. Damit ist der Nutzer maßgeblich für Qualität des Systems verantwortlich. Die Qualität hängt Tagging ab. Je besser die Tags zum Dokument passen, desto besser ist das Suchsystem. Damit kommt dem User, als Produzenten, eine wesentliche Rolle bei der Evaluation zu. Der Beitrag der Users am System manifestiert sich in seinen Tags. Diese sagen nicht nur etwas über das Dokument aus, sondern auch über das Vorwissen des Users, z.b. über seine Sprache, Fachkenntnisse oder Vorlieben.
Zur Evaluation von Social-Tagging-Systemen gehören also nicht nur die beiden Dimensionen Information Retrieval und Suchinteraktionen der Nutzer, sondern neu auch die Ordnungsfunktion der User durch ihre Tags.

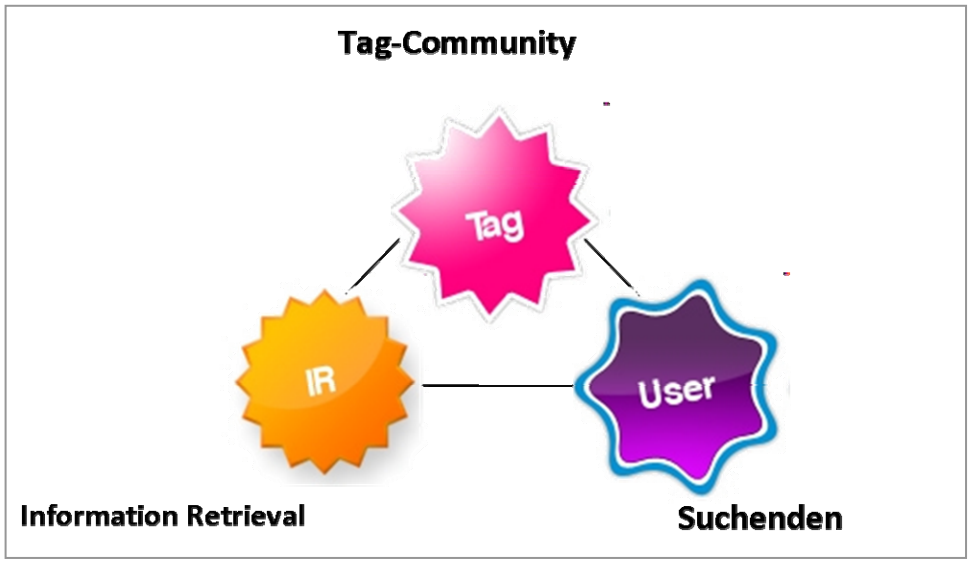

Abbildung 2: Drei Dimension der Suche in Social-Tagging-Systemen

5 Forschungsdesign

Angelehnt an die drei Dimensionen IR, User und Tags wurde ein Forschungsdesign entwickelt. Ziel der Untersuchung ist es charakteristische Merkmale von tagbasierten Suchsystemen im Vergleich zu algorithmische Suchsysteme herauszuarbeiten, um auf diese Weise Handlungsempfehlungen für bessere Suchstrategien in Social-Tagging-Systemen abzuleiten. Dafür musste zunächst ein Verfahren entwickelt werden, welches

es ermöglicht beide Systeme miteinander zu vergleichen. Dabei war die zentrale Frage, wie Daten aus allen drei Dimensionen generiert werden können.

Zum momentanen Zeitpunkt kommt ein direkter Vergleich von konventionellen Suchmaschinen und Social-Bookmarking-Diensten im Internet nicht in Frage. Konventionelle Suchmaschinen sind den Social-Bookmarking-Diensten weit überlegen. Sie sind älter und in ihrer Entwicklung viel fortgeschrittener als die neuen Social-Bookmarking-Dienste. Die Menge an Informationen, systeminterne Faktoren und die Geheimhaltung der Algorithmen machen es unmöglich, Volltext- vs. Tagging-systeme miteinander zu vergleichen und dabei andere Einflussgrößen auszuschließen.

Zum Vergleich der beiden Suchverfahren wird zunächst eine Testkollektion mit definierten Rangierungsprinzipien verwendet, wie sie typisch für Evaluationen im Information Retrieval sind. Die Testkollektion kann mit beiden Suchsystemen durchsucht werden. Auf diese Weise gibt es zwei Suchsysteme für dieselbe Dokumentenkollektion. Von Seiten der Nutzer braucht es Probanden, die mit dem System interagieren und Bewertungen abgeben. Dabei werden Daten vor, während und nach der Suche erhoben. Für die Erhebung der Tags braucht es eine Community, welche die Dokumente in der Testkollektion vertagt.

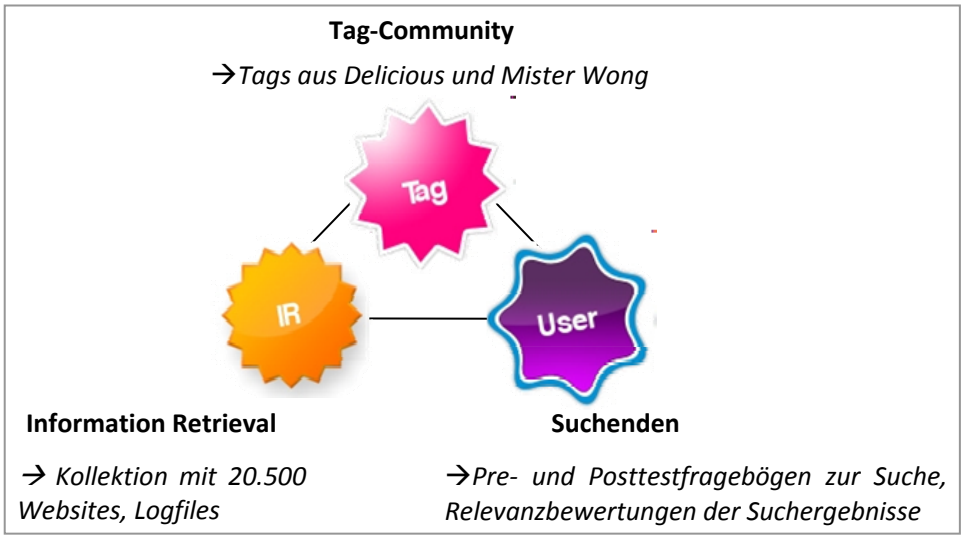

Abbildung 3: Zuordnung der Datenerhebungen

6 Tagidex

Für die Vergleichsstudie wurde die Webapplikation Tagidex entwickelt. Tagidex erlaubt es, auf einer vergleichbar überschaubaren Kollektion von echten Websites zu einem Themenbereich mit Tagging- und Indexsuchverfahren zu suchen. Da, wie bereits angesprochen, Social-Tagging-Dienste einen deutlich kleineren Datenbestand aufweisen, umfasst eine Kollektion in Tagidex nur Webseiten, die bei einem Anbieter wie Delicious oder Mister Wong bereits von Nutzern eingetragen wurden. Für die Indexsuche werden die dazugehörigen HTML-Rohdaten der vertaggten Webseite bezogen und zusammen mit den Tags in einer seperaten Tagidex-Datenbank abgelegt. Wir stützen uns somit auf reale Daten der Tagging-Community. Ein Prototyp von Tagidex findet sich auf [HJ09]. Man kann wählen, mit welchem Retrievalsystem die Kollektion durchsucht werden soll. Bei der tagbasierten Suche werden die dazugehörigen Tags von Social-Tagging-Diensten wie Delicious und Mister Wong bezogen. Beispielsweise kann derselbe Suchbegriff in die Volltextsuche und in die tagbasierte Suche eingegeben werden, die Ergebnislisten können miteinander verglichen und charakteristische Unterschiede herausgearbeitet werden.

Abbildung 4: Screenshot Startseite Tagidex

Bei der Tag-Suche mit Tagidex wird für jeden Datensatz eine Tag-Liste gespeichert. Diese wird vom jeweiligen Anbieter (Delicious und Mister Wong) bezogen. Bei der Tag-Suche werden die eingegebenen Such-Tags mit dieser Tag-Liste verglichen. Zurückgeliefert werden alle Datensätze, die mindestens eines der Such-Tags enthalten. Anschließend werden diese nach Anzahl gefundener Such-Tags geordnet. Da in der Regel sehr viele Seiten die gleiche Anzahl getroffener Tags aufweisen, wird sekundär nach der Benutzerzahl geordnet (der Anzahl Benutzer, die diese Seite beim Anbieter vertagt haben). Als Treffer gelten auch Tags, die ein Such-Tag enthalten – etwa den Such-Tag „Schule“ findet auch „Schule2.0“.

Bei der Volltextsuche mit Tagidex wird auf die etablierte MySQL-Volltextsuche zurückgegriffen. Diese führt eine natursprachliche Suche nach einer Zeichenkette in einer Textsammlung durch. Für jeden Datensatz wird ein Relevanzwert zurückgeliefert, d. h. eine Maßangabe für die Ähnlichkeit zwischen der Such-Zeichenkette und dem Text in diesem Datensatz. Die Suche wird ohne Unterscheidung der Groß-/Kleinschreibung durchgeführt. Die gefundenen Datensätze werden automatisch nach absteigender Relevanz sortiert. Relevanzwerte sind nichtnegative Fließkommazahlen. Nullrelevanz bezeichnet keinerlei Ähnlichkeit. Die Relevanz wird auf der Basis der Anzahl Wörter im Datensatz, Anzahl eindeutiger Wörter im Datensatz, der Gesamtanzahl Wörter in der Sammlung und der Anzahl der Dokumente (Datensätze) berechnet, die ein bestimmtes Wort enthalten.

Einige Wörter werden bei der Volltextsuche ignoriert. Dies sind alle Wörter, die zu kurz sind (weniger als 3 Zeichen) und solche, die auf der Liste der Stoppwörter stehen. Ein Stoppwort ist ein Wort wie „dass“ oder „der“, das so verbreitet ist, dass sein semantischer Wert als vernachlässigbar betrachtet wird.

Genauere Details zum Verfahren finden sich unter: http://dev.mysql.com/doc/refman/5.1/de/fulltext-search.html

Die Testkollektion enthält 20’500 Dokumente. Eine größere Kollektion wäre aus Performancegründen nicht sinnvoll und letztlich für die Suchenden nachteilig. Alle Dokumente in der Kollektion sind mit Tags versehen. Auf diese Weise konnte eine Kollektion zusammengestellt werden, die komplett von Usern der Delicious und Mister Wong Community getaggt ist und den essentiellen Aspekt der Online-Communities bei Social-Tagging-Systemen aufnimmt. Der User als Produzent fließt in Form der Tags mit in die Testkollektion ein.

7 Aufbau der Untersuchung

7.1 Einführung

Im Vorfeld der Befragung und der Suchinteraktion mit Tagidex gibt es eine Schulung für alle Probanden. Danach werden die Probanden in zwei Gruppen geteilt. Die Gruppen versuchen persönliche Fragestellungen mit Tagidex zu beantworten. Eine Gruppe verwendet dafür die tagbasierte, die andere Gruppe die indexbasiert Suche. Vorerst werden alle in die Grundlagen der Internetrecherche, indexbasierten Suche und tagbasierten Suche eingeführt. Es kann davon ausgegangen werden, dass die allermeisten User bereits mit den konventionellen, indexbasierten Suchmaschinen gearbeitet und Erfahrungen gesammelt haben. Bei den Social-Bookmarking-Diensten gilt eher der umgekehrte Fall. Deshalb gibt es eine Einführung in die Grundlagen der Internetrecherche, Suchstrategien und Unterschiede tag- und indexbasierter Suchsysteme. Die Probanden sammeln in einer Übungsphase (20 min) Recherche-Erfahrungen mit einem Social-Bookmarking-Dienst. Die Schulung im Vorfeld der Suchinteraktion gleicht die unterschiedlichen Vorkenntnisse der Probanden zu einem gewissen Anteil aus.

7.2 Fragestellung

Die Probanden wählen selbst eine Frage für die Suchinteraktion aus. Lediglich der Themenbereich der Frage wird eingegrenzt. Die Fragestellung muss dem Themenbereich der Testkollektion entsprechen. An dieser Stelle ist es sinnvoll ein Themenbereich zu wählen, welcher die Probanden interessiert. In unserem Fallbeispiel waren die Probanden Lehramtsstudierende und der Themenbereich „Schule und IKT“. Damit soll die persönliche Relevanz für die Probanden und ihre Motivation sichergestellt werden. Das Lösen einer selbst gewählten Frage kommt dem Recherchieren im Internet, wie die Probanden es aus ihrem Alltag kennen, am Nächsten.

7.3 Pretest

Nach der Einführung und noch vor der Suchinteraktion wird ein Pretest mit den Probanden durchgeführt. In diesem beantworten sie Fragen nach ihrem Informationsbedürfnis, Motivation, Vorkenntnisse, Fachkompetenz im Bezug auf das Fachgebiet ihrer Frage, etc. Sie vermerken ihre Fragestellung und halten ihren momentanen Stand im Suchprozess fest.

27. Wo im Suchprozess befinden Sie sich zum gegenwärtigen Zeitpunkt mit Ihrer Frage? *(Bitte wählen Sie eine der folgenden Kategorien, die am Besten zu Ihrer Frage passt.)*

☐ Anstoß – Ich habe erkannt, dass ich an dieser Stelle meiner Arbeit Informationen brauche.
☐ Auswahl – Ich habe erkannt auf welchem allgemeinen Gebiet ich suchen muss.
☐ Erforschung – Ich bin dabei spezifische Informationsquellen ausfindig zu machen, von denen ich denke, dass sie mir nutzen werden.
☐ Formulierung – die Informationen, die ich bisher gefunden habe, haben mir geholfen das Informationsbedürfnis einzugrenzen und genau zu bestimmen.
☐ Sammlung – Nachdem ich mein Informationsbedürfnis genau bestimmt habe, sammle ich jetzt die dazugehörigen Informationen.
☐ Präsentation – An dieser Stelle meiner Arbeit bin ich dabei das Sammeln von Informationen zu beenden.

Abbildung 5: Fragebogenausschnitt

7.4 Suchinteraktion

Die Suchinteraktion der Probanden mit Tagidex erfolgt wie bei konventionellen Suchmaschinen und orientiert sich an Anbietern wie Google, mit denen sie bereits vertraut sind, um eine möglichst kurze Eingewöhnungszeit sicherzustellen.

Die Probanden geben Suchbegriffe in ein Eingabefeld und erhalten eine Trefferliste. Jede Sucheingabe ergibt zehn Treffer. Die Zahl wurde bewusst gewählt, um damit dem typischen Nutzerverhalten, das Anschauen der ersten Ergebnisseite möglichst nahe zu kommen. λ ist definiert nach typischem Suchverhalten von Suchmaschinen-Nutzern:

$$\lambda=1$$

Tagidex zeigt an mit welchem Suchverfahren die Probanden arbeiten. Der Unterschied in der Suche ist nur bei der Trefferliste ersichtlich: die tagbasierte Trefferliste zeigt für jeden Treffer zusätzlich die dazugehörigen Tags.

Abbildung 6: Screenshot Trefferliste Tagidex

7.5 Relevanzbewertung

Während der Suchinteraktion mit Tagidex bewerten die Probanden im Browser die Relevanz der Treffer für Ihre Frage. Für den Relevanzfragebogen werden vier Messwerte verwendet: relevant, teilweise relevant, teilweise irrelevant und irrelevant, wie sie auch von [SG01], [Su03] und [Su03] in ihren Studien zur methodologischen Herangehensweise bei Relevanzbefragungen dargelegt werden.
Nach jedem angeklickten Treffer erscheint unter dem Link ein kleines Feld mit vier Radiobuttons, jeweils mit den vier Optionen. Nach dem Anklicken einer Option wird das Feld geschlossen.

Abbildung 7: Screenshot Relevanzbewertung in Tagidex

Während des gesamten Suchprozesses werden Logfiles generiert.

7.6 Posttest

Nach der Suchinteraktion mit Tagidex wird ein Posttest durchgeführt. In diesem beantworten die Probanden unter anderem Fragen nach den Veränderungen im Suchprozess aufgrund der Suchinteraktion mit Tagidex, der Klärung ihrer Frage, den Schwierigkeiten bei der Relevanzbewertung, etc. Wie bereits im Pretest werden sie wieder nach ihrem Stand im Suchprozess nach der Suchinteraktion gefragt. Damit wird die Differenz aufgrund der Suche ersichtlich und der Interaktion der Probanden Rechnung getragen. Das Fragebogendesign lehnt sich an dem Nutzerzentrierten Ansatz von Spink [Sp02] an.

In der Datenauswertung können die Selbsteinschätzungen der Probanden aus den Pre- und Posttest mit den Daten aus den Logfiles abgeglichen werden. Dies hat einen entscheidenden Vorteil für die Interpretation der Daten im Vergleich zu einseitigen Erhebungsverfahren. Die Selbsteinschätzungen der Probanden können mit den Auswertungen der Logfiles verglichen werden. Umgekehrt gilt für das Arbeiten mit Logfiles, dass die Datenauswertung schwierig ist, weil die Deutungen der Nutzer daraus nicht direkt ableitbar sind. In unserem Fall können die Pre- und Posttests zu den Logfiles herangezogen werden.

8 Erste Erprobung und Ausblick

Das Forschungsdesign wurde im März 2009 mit einer Gruppe von 17 Studierenden getestet. Die Gruppe war fächerübergreifend organisiert, alle Probanden hatten bereits ein Studium abgeschlossen und studierten zu diesem Zeitpunkt auf Lehramt. Während der Untersuchung wurden 455 Logfiles, 34 Fragebögen und 358 Relevanzbewertungen erhoben. Gegenwärtig liegen noch keine Studienergebnisse vor, die Datenanalyse läuft derzeit. Ziel ist es, die Ergebnisse für eine Verbesserung der Forschungsmethode nutzen zu können. Für Herbst 2009 und Frühling 2010 sind zwei weitere, größere Untersuchungen geplant.

Literaturverzeichnis

[HS08] Hendler, J.; Shadbolt, N.; Hall, W.; Berners-Lee, T.; Weitzner, D.: Web Science: An Interdisciplinary Approach to Understanding the Web. Communications of the ACM, 51 (7), 2008; S. 60-69

[HJ09] Hielscher, M.; Jurjević, D.: www.r2d2.ch/tagidex/ (Stand 10.05.2009)

[JS05] Jansen, B. J.; Spink, A.: How are we searching the World Wide Web? A comparison of nine large search engine transaction logs. Information Processing and Management, 42(1), 2005; S. 248-263

[Sa95] Saracevic, T.: Evaluation of evaluation in information retrieval. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Special issue of SIGIR Forum, 1995; S. 138-146

[Sh07] Shneiderman, B.: Web science: A Provocative Invitation to Computer Science. Communications of the ACM, 50(6), 2007; S. 25-27

[Sp02] Spink, A.: A user-centered approach to evaluating human interaction with web search engines: an exploratory study. Information Processing and Management 38(3), 2002, S. 401-426

[Le07] Lewandowski, D.: Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen? In: Marcel Machill / Markus Beiler (Hrsg.): Die Macht der Suchmaschinen / The Power of Search Engines . Köln, 2007

[SG01] Spink, A.;Greisdorf, H.: Regions and levels: Mapping and measuring users' relevance judgments. Journal of the American Society for Information Science, 52(2), S. 161-173

[Su03] Su, L. T.: A comprehensive and systematic model of user evaluation of web search engines: I. theory and background. In: Journal of the American Society for Information-Science and Technology, 13, 2003, S. 1175-1192

[Su03] Su, L. T.: A comprehensive and systematic model of user evaluation of web search engines: II: an evaluation by undergraduates. In: Journal of the American Society forInformation Science and Technology, 13, 2003, S. 1193-1223

[Rö07] Röhle, Theo: Machtkonzepte in der Suchmaschinenforschung. In: Machill Marcel, Beiler Markus (Hrsg.): Die Macht der Suchmaschinen. Köln, 2007, S. 127-128

[BV05] Buckley, C.; Voorhees, E., M.: Retrieval System Evaluation. In: Voorhees, E., M.; Harman, D., K.(Hrsg.): TREC. Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, Massachusetts. S. 53-58