UNIVERSITÄT
KOBLENZ · LANDAU
Fachbereich 4: Informatik

Studienarbeit

WS 2008/2009

# Beschreibung und Evaluation des Mytag Merging Algorithmus

vorgelegt von

Daniel Grabs

Betreuer: Prof. Dr. Steffen Staab, Dipl.-Inform. Klaas Dellschaft
(Forschungsgruppe ISWeb - Informationssysteme und Semantic Web)

Koblenz, 19.01.2009

# Inhaltsverzeichnis

# 1 Motivation

Im Projektpraktikum „Mytag 2.0“[1] welches von Klaas Dellschaft et al. betreut wurde, kam erstmals die Problemstellung des Merging auf. Merging bedeutet in diesem Fall das Mischen der Ergebnisse verschiedener Web-Plattformen, welche denselben Medientyp liefern. In unserem speziellen Fall handelte es sich um Bookmarks von den Plattformen del.icio.us[2], Connotea[3] und BibSonomy.[4] Produkt dieses Praktikums war unter anderem ein Merge-Algorithmus, der folgende Kriterien erfüllen musste:

- Die Mytag Struktur unterstützen
- Schnell und effizient sein
- Als Ergebnis eine Liste mit sinnvoll sortierten Suchergebnissen hervorbringen

Mytag wurde basierend auf Ruby on Rails 2.0 entworfen. Aufgrunddessen ist im ganzen Projekt ein hoher Grad an Dynamik wiederzufinden, was der Algorithmus auch direkt ausnutzt. Suchanfragen von Mytag gehen entweder über die entsprechende GET-Methode direkt an die API oder erfolgen per RSS Anfrage. In beiden Fällen erhält Mytag ein XML Dokument als Antwort, welches alle Ergebnisse und auch weitere Informationen zu diesen Ergebnissen wie z.B. verwandte Tags enthält.

Merging soll genau dann erfolgen, wenn eine Anfrage an zwei oder mehr Plattformen gesendet wird, welche dieselbe Ressource bereitstellen. Ist dies also der Fall, werden alle Ergebnisse sämtlicher Plattformen zunächst in einem Pool vereint und mittels des Merging-Algorithmus sortiert. Bevor die Problematik des Algorithmus erläutert und das Ziel dieser Studienarbeit definiert wird, soll aber zunächst der Algorithmus selbst vorgestellt werden.

[1] http://mytag.uni-koblenz.de
[2] http://del.icio.us
[3] http://www.connotea.org
[4] http://www.bibsonomy.org

# 2 Vorgehen des Merging Algorithmus

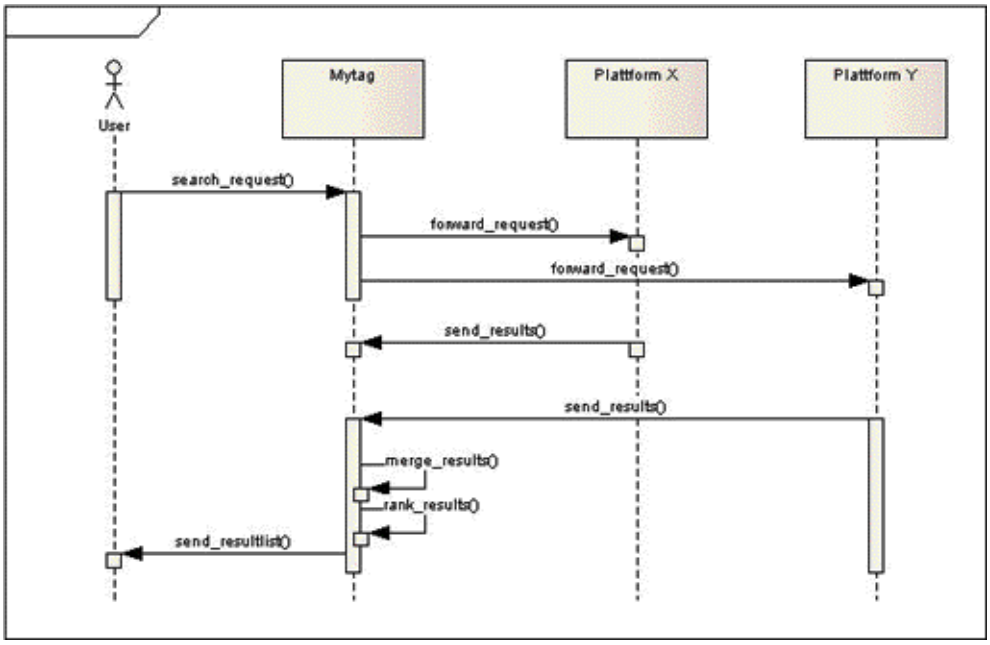


**Abbildung 1:** Ablauf einer Suchanfrage

Der Algorithmus läuft über alle Elemente im Ergebnispool und errechnet für jedes Element einen individuellen Ranking-Wert. Dieser Ranking-Wert setzt sich zum Einen aus ergebnisbezogenen Bewertungskriterien, zum Anderen aus plattformbezogenen Kriterien („Confidence-Faktoren“) zusammen.

## 2.1 Ergebnisbezogene Kriterien

Sie sind für jedes Ergebnis unterschiedlich und werden anhand der zusätzlichen Information, die uns die Plattform liefert, bestimmt. Speziell auf den drei Plattformen, die momentan gemerged werden (Stand Januar 2009), wurden zusätzlich zu den Suchergebnissen folgende brauchbare Informationen zur Ressource geliefert:

- Titel
- URL
- Weitere Tags, mit denen die Ressource versehen wurde

Mit Hilfe dieser Angaben ermittelt der Algorithmus vier unterschiedlich gewichtete Bewertungskriterien (vgl. Abb. 2), aus deren Produkt schließlich das Ranking jedes einzelnen Elementes berechnet wird. Hierbei sind alle vier Faktoren (A-

D) normiert.

Zuerst wird geprüft, ob der Suchbegriff im Titel der Ressource enthalten ist (Wert A). Ist dies der Fall, wird in A der Wert 1,0 gespeichert. Andernfalls erfolgt ein Abzug und A wird nur mit 0,7 belegt.

Wert B wird ähnlich gesetzt. Er beschreibt die Güte der Url der Ressource. Dazu wird überprüft, ob der Suchbegriff Teil der Url ist. Ist dem nicht so, wird B direkt auf 0,09 gesetzt und nicht weiter verfahren. Ansonsten werden mehrere Fälle untersucht und somit B als Summe der Unterfaktoren B1 bis B5 berechnet: Im ersten Schritt wird lediglich geprüft, ob die Url Normgerecht mit dem String *http://* oder *https://* beginnt. Falls ja wird B1 = 0,1. Falls nicht erfolgt ein minimaler Abzug von 0,02. Als nächstes prüft der Algorithmus, ob die Domain exakt dem gesuchten Tag entspricht und wertet B2 mit 0,3 falls dies zutrifft. Falls nicht wird B2 = 0,05. B3 untersucht die Komplexität der Url von der Subdomain bis zur Toplevel Domain: Erfolgt der Link direkt auf die Hauptdomäne, also nicht über eine Subdomain, so nimmt B3 den Wert 0,2 an, falls nicht wird der Wert 0,15 gespeichert. Im nächsten Schritt wird die Toplevel Domain untersucht. Hierbei werden alle unüblichen Domains wie z.B. *.info* schlechter eingestuft. Dies wird realisiert durch prüfen ihrer Länge. Wenn die Länge nun $\geq 4$ ist wird B4 = 0 gesetzt, andernfalls = 0,1. Zu guter Letzt wird geprüft, ob die Url einen Pfad enthält. Ist dies nicht der Fall, wird B5 = 0,3 gesetzt. Wenn nicht, wird wiederum der Pfad genauer untersucht. B5 wird = 0,2 gesetzt. falls der gesuchte Tag im obersten Pfad enthalten ist. Ist er in einem tieferen Pfad enthalten, wird B5 = 0,1. Wenn er gar nicht im Pfad vorkommt setzt der Algorithmus B5 = 0. Die Idee ist, dass eine Url durch einen Pfad unpräziser wird und es demnach Abzüge gibt. B1 bis B5 werden nun aufsummiert. B kann im bestmöglichen Fall 1,0 werden.

Das nächste Faktor, also C, untersucht das Vorkommen des gesuchten Tags in den anderen Tags, mit welchen die Ressource ausserdem versehen wurde. Beim Suchtag *dog* würde es zum Beispiel einen Bonus geben, wenn die Ressource zusätzlich mit dem Tag *dogtrainer* versehen wurde. Hierfür werden jegliche Suchtags mit jeglichen weiteren Tags verglichen und für jede Übereinstimung ein Counter um 1 erhöht. Zuletzt wird der Counter durch die gesamte Anzahl der Tags geteilt. Auch hier kann der Wert bestenfalls 1,0 betragen.

Die Tags werden noch nach einem zweiten Kriterium untersucht: Mytag stellt für jede Suchanfrage eine Tagcloud der 20 häufigsten Tags in der Ergebnisliste auf. Für jede Ressource wird nun überprüft, wie groß der Anteil aller Tags, mit denen sie versehen wurde, an der Tagcloud ist. Auch in diesem Fall ist der beste Wert, den D annehmen kann, 1,0.

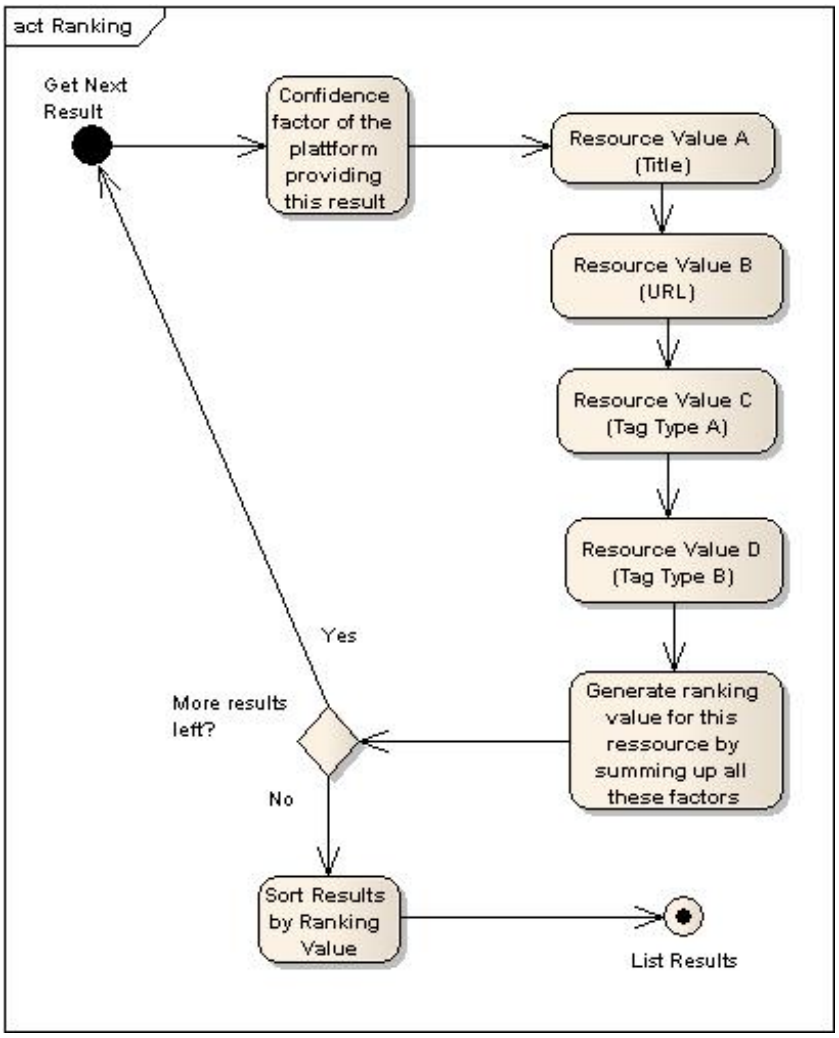


**Abbildung 2:** Vorgehen des Ranking Algorithmus

## 2.2 Plattformbezogene Kriterien

Sind alle ergebnisbezogenen Werte berechnet, multipliziert man sie mit dem Confidence Faktor, welcher der Plattform zugeordnet ist. Dieser Wert ist also für alle Ergebnisse einer Plattform derselbe. Für die Bestimmung dieses Wertes haben wir uns am Kalibrierungsalgorithmus vom Profusion Projekt orientiert.[GWG06] In diesem wird der Confidence Faktor einer Plattform durch Benutzertests ermittelt. Der Benutzer startet eine Suchanfrage und bewertet die ersten 10 Suchergebnisse nach ihrer Relevanz. Diese Auswertung geht dann in folgende Formel ein, mit der der Confidence Wert errechnet wird. Nach genügend Benutzertests wird der Mittelwert gebildet.

$$\left(\frac{\sum_{i=0}^{10} Ni}{10} * \frac{R}{10}\right) \div 0,2929$$

Ni = 0 wenn das Dokument i irrelevant ist, andernfalls 1/i. R ist die Anzahl an relevanten Dokumenten aus einem Set von 10 Ergebnissen. Zuletzt wird der Wert durch Division mit dem Maximalwert normiert.

Im Mytag 2.0 Praktikum wurden pro Plattform 5 Benutzertests durchgeführt, von denen jeder je 2 Suchanfragen beinhaltete. Genauer dokumentiert und mittels Codebeispielen erklärt wird der Algorithmus im offiziellen Mytag Paper. [ASG09]

### 2.3 Verbesserungen des Algorithmus

Im Laufe dieser Studienarbeit soll nun ein alternativer plattformbezogener Faktor für jede Bookmarking Plattform durch eine repräsentative Studie ermittelt werden. Die Zielsetzung besteht darin, einen verbesserten Faktor für die Bookmarking Plattformen zu erhalten und diesen unter verschiedenen Sachverhalten zu untersuchen. Dies würde insgesamt eine höhere Flexibilität und zufriedenstellendere Ergebnisse bedeuten.

## 3 Aufbau der Evaluation

Um die Qualität der Ergebnisse einer retrieval Plattform zu bewerten, hat man im Prinzip nur das Kriterium, eine Ressource als relevant oder irrelevant einzustufen. Dies ist natürlich ein recht subjektives Kriterium. Für eine Person mag ein Ergebnis sehr relevant sein, für die nächste ist sie absolut unbrauchbar. Um trotzdem ein repräsentatives Ergebnis zu erhalten, muss die Studie deshalb sehr umfangreich werden. Die populärsten Methoden, um die Güte von Ergebnislisten zu bestimmen sind Precision und Recall. [KC06]

Recall gibt darüber Aufschluss, wie groß der Anteil der gefundenen, relevanten Dokumente am Anteil aller existierenden, relevanten Dokumente ist. Dieser Wert wird berechnet, in dem man die Schnittmenge aller gefundenen, relevanten Ergebnisse (P), mit der Gesamtmenge aller Dokumente (R) im Pool berechnet und durch die Gesamtmenge aller *relevanten* Dokumente (R') teilt.

$$\frac{R \cap P}{R'}$$

Im Fall von Mytag wäre der Recall Wert aber nahezu unmöglich zu berechnen, da aufgrund der gewaltigen Gesamtmenge von Ergebnissen nicht bestimmt werden kann, wie viele denn relevant wären. So bleibt also nur die Precision, bzw. ein Algorithmus, der auf der Precision basiert, übrig.

Bei der Precision berechnet man anhand von einer festen Anzahl von Ergebnissen, wie zum Beispiel den Top 10 der Ergebnisliste, wie viele Dokumente davon relevant bzw irrelevant sind. Der Unterschied zu Recall besteht darin, dass durch die Menge der *gefundenen* Ressourcen (R") geteilt wird.

$$\frac{R \cap P}{R''}$$

Basierend auf der Precision stehen mehrere Möglichkeiten zu Verfügung, um einen repräsentativen Wert zu ermitteln. Während die meisten Untersuchungen für die Precision binär arbeiten, also dem Nutzer nur die Wahl lassen *ob* das Dokument relevant ist, soll der Nutzer in dieser Evaluation bewerten *wie* relevant das Dokument ist. [MRS08] Damit soll dem Benutzer so viel Freiheit wie möglich gelassen werden, was zugleich ein genaueres Ergebnis zur Folge hat.

Das generelle Ziel dieser Untersuchung ist herauszufinden, ob der Confidence Faktor ein repräsentatives Kriterium ist, um im Algorithmus eingesetzt zu werden. Stellt man beispielsweise bei ein und derselben Plattform bei unterschiedlichen Themengebieten starke Schwankungen fest, sollte der Einsatz desselbigen in Frage gestellt werden. Ebenfalls soll untersucht werden, ob die Ergebnisse nicht doch besser sortiert sind, wenn man statt des Confidence Faktors lediglich die Precision berechnet. Beim direkten Vergleich der Precision und dem in Kapitel 2 vorgestellten Confidence Wert fällt auf, dass man durch den Confidence Wert um einiges flexibler ist, weil man auch Ressourcen verschiedener Plattformen in derselben Ergebnisliste verrechnen kann. Das Risiko liegt aber leider darin, dass eine Plattform schnell durch schlechte Wertungen benachteiligt werden kann, wenn sie bei Suchanfragen überhaupt keine Ergebnisse liefert. Solch schlechte Wertungen sind später nur sehr schwer wieder auszugleichen.

Wie in [Grä99] bemerkt wird, bestehen bei Online-Evaluationen typische Fehlerquellen. Speziell im Fall der hier durchgeführten Evaluation musste darauf geachtet werden, dass alle Befragten ein elementares Verständnis von Suchmaschinen und deren Anwendung haben.

Ein weiterer Kritikpunkt ist sicherlich der große Umfang der Umfrage. Da aber sehr viel Wert auf ein repräsentatives Ergebnis gelegt wird, ist diese Maßnahme leider unumgänglich. Mit Hilfe von interessanten Aufgaben und Entlohnungen soll die Motivation der Probanden erhalten bleiben.

## 3.1 Die erste Evaluation

Bevor man sich die Frage stellt, wie die Evaluation aufgebaut sein soll, muss man sich natürlich über die genauer Ziele, die man erreichen will, im Klaren sein.

In der ersten Evaluation soll zunächst keinerlei Auswertung von Ergebnissen statt finden. Das Primärziel liegt hier bei der Erzeugung von Trefferlisten, die viele repräsentative Suchanfragen dokumentieren. In dem Zusammenhang ist es sehr wichtig, die Umstände, unter denen die Listen erzeugt wurden, genau fest zu halten. Interessant ist zum Beispiel, welche Person sie erzeugt hat, wonach gesucht wurde oder auch ob es die erste Suchanfrage zu dem Thema war oder vielleicht eine Spezialisierung.

Die Studie wird als Laborexperiment durchgeführt werden, was den Vorteil hat, dass zum Verhalten des Probanden auch mal konkret nachgefragt werden kann. Der Nachteil bei Laborexperimenten besteht allerdings generell darin, dass die Ergebnisse leicht verfälscht werden können, da die Benutzer nicht in „natürlicher Umgebung“ arbeiten. [BD06] Speziell im Rahmen dieser Evaluation kann es sein, dass das in den Fragen erschaffene Szenario vom Probanden nicht gut akzeptiert wird und sich so sein Suchverhalten unter Umständen ändert.

Der Proband wird über einen Fragebogen Aufgaben gestellt bekommen, die er durch Nutzung von Mytag lösen soll. Hierbei werden drei Bookmarking Plattformen gleichzeitig angefragt und die Ergebnisse mit dem aktuell in Mytag implementierten Suchalgorithmus sortiert.

Der primäre Zweck des Fragebogens liegt darin, anhand von Aufgaben dem Benutzer möglichst „natürliche“ Suchaufträge zu erteilen. Dabei sollen 4 der wichtigsten Themengebiete im Internet (siehe Kapitel 2.3) abgedeckt werden.

Bei derartigen Aufgaben unterscheidet man zwischen offenen Fragestellungen („Sammle Informationen über die globale Erwärmung“), die keine bestimmte Antwort liefern sollen und geschlossenen Fragestellungen („Finde heraus, in welchem Jahr die Glühbirne erfunden wurde“).[Bie05]

Da es für diese Umfrage von Vorteil ist, dass möglichst unterschiedliche Trefferlisten produziert werden, macht es mehr Sinn, die Aufgabe allgemein zu halten und ein relativ freies Szenario zu erschaffen. Das funktioniert besser mit offenen Fragestellungen. Der Umfang der Studie wird 15 Probanden umfassen. Zeitgleich werden aus ihr Ergebnisse für eine parallel entstehende Studienarbeit zum Suchverhalten mit Navigationshilfen gewonnen.[Sch09]

Die Trefferlisten, welche beim Erfüllen der Suchaufträge generiert werden, sollen primär der Verbesserung des Confidence Faktors helfen und durch die anschließende Auswertung folgende Fragen geklärt werden:

- Bleibt der Confidence Wert auch bei verschiedenen Themengebieten konstant?

- Sind die Ergebnisse durch Einsatz des Confidence Wertes wirklich besser sortiert?
- Inwiefern spiegelt sich die Erfahrung mit Suchmaschinen in der Güte der Ergebnisse wieder?

Die Bewertung der so gewonnenen Trefferlisten wird in einer zweiten, intensiveren Studie durchgeführt. Die Versuchspersonen bekommen die zuvor generierten Trefferlisten aus der ersten Studie vorgesetzt und müssen die dortigen Ressourcen klassifizieren. Wie diese Ergebnisse dann genau verwertet werden, wird in Kapitel 3.3 beschrieben.

## 3.2 Die zweite Evaluation

Die zweite Umfrage stützt sich komplett auf die Resultate aus der ersten. Hier soll eine kleinere Benutzergruppe von drei Personen alle zuvor generierten Ergebnislisten detailliert durchgehen und jede Ressource nach ihrer Relevanz im Hinblick auf den gestellten Suchauftrag mit einer Wertung versehen. Diese Benutzergruppe wird also keinerlei Suchanfragen stellen, sondern bereits durchgeführte Suchanfragen bewerten.

Die Studie wird nicht als Laborversuch stattfinden, da sie für den Einzelnen zeitlich um einiges intensiver verlaufen wird und dem Probanden somit ein freies Zeitlimit gesetzt werden kann. Desweiteren fallen die verfälschenden Laborbedingungen weg. Der Benutzer wird über ein Webinterface die Möglichkeit haben, seine Bewertungen abzugeben. Die Eingaben werden dann in Datenbanktabellen gespeichert und später ausgelesen und weiterverarbeitet.

## 3.3 Realisierung

In folgenden Abschnitten wird erläutert, wie die Suchaufträge der Probanden in der ersten Studie aufgebaut sein werden. Für vier der zehn populärsten Themengebiete im Internet wird eine umfassende Aufgabe gestellt, die jeder Proband lösen muss. Da die Umfrage in dieser Form schon sehr umfangreich ausfällt, können leider nicht alle Gebiete abgedeckt werden.

Die Themengebiete wurden mit Hilfe der Ergebnisse des Open Directory Projects[5] ausgesucht. Dieses Projekt beschäftigt über 81000 Internetnutzer, die Webseiten sammeln und kategorisieren. Auch große Online Services wie Google oder Aol Search beziehen und verwerten Informationen vom ODP.

Die Themengebiete lauten:

- „Wissenschaft“

[5] http://www.dmoz.org

- „Kultur“
- „Spiele“
- „Bildung“
- „Shopping“
- „Berufsleben“
- „Gesundheit“
- „Gesellschaft“
- „Computer“
- „Freizeit“

Der genaue Wortlaut der Aufgaben findet sich im Anhang. Sie decken die Themengebiete Spiele, Shopping, Freizeit und Berufsleben ab. Bei quantitativen Umfragen wie dieser bekommen alle Probanden die selben Aufgaben zu lösen, um ein einheitliches Schema zu schaffen.[BD06]

Viele der Aufgaben werden mehr als eine Ergebnisliste hervorbringen, weil der Benutzer z.B. die Suche verfeinert, auf die nächsten 50 Ergebnisse blättert oder gar die Suchbegriffe komplett austauscht. Diese Informationen werden ebenfalls dokumentiert und in die Auswertung miteinbezogen. Sobald der Benutzer die gewünschte Information gefunden hat, gilt die Aufgabe als erledigt. Insofern wird also bei jeder Aufgabe mindestens eine Trefferliste generiert.
Sobald ein Benutzer eine Suchanfrage losschickt, wird sowohl der Suchtag, die Ergebnisliste, die Plattform auf der gesucht wurde und ein von ihm gewählter Benutzername in einer Tabelle gespeichert. Die Tabelle wird bei der zweiten Evaluation wieder ausgelesen und in einem Webinterface dargestellt. Dem Benutzer wird der Suchauftrag und die entsprechende Trefferliste angezeigt. Nun muss er sich die zugehörigen Ressourcen ansehen und diese nach ihrer Relevanz bewerten. Die Bewertung kann in das Interface eingegeben werden.

Sind alle Ressourcen bewertet, werden die Bewertungen ihrerseits in einer neuen Tabelle gespeichert. Aus dieser Tabelle werden die Bewertungen dann für jede Plattform ausgelesen und zur Berechnung des neuen Confidence Wertes in folgende Formel eingesetzt. Die Formel orientiert sich an der des Profusion Projekts, welche in Kapitel 1 vorgestellt wurde. Wie schon angedeutet, wird sie aber modifiziert, so dass der Benutzer das Suchergebnis nicht nur binär bewerten kann. Ausserdem wird die Gesamtmenge der Ergebnisse von 10 auf maximal 50 erhöht, je nachdem, wie viele Resultate die drei Plattformen liefern.

$$\left(\sum_{i=1}^{k} \frac{j}{i} * \frac{R}{k}\right) \div n$$

In Worten beschrieben verfährt der Algorithmus so: Jedes Ergebnis wird je nach Relevanz entweder mit 0 (irrelevant), 0,5 (mäßig relevant) oder 1 (sehr relevant) versehen. Dieser Relevanzwert wird in j eingesetzt. i steht für den Rang der Ressource in der Ergebnisliste. Somit würde ein sehr relevantes Ergebnis, welches auf Platz 1 steht also einen besseren Wert (=1) erhalten als ein sehr relevantes Ergebnis auf Platz 2 (=0,5). Die Summe dieser errechneten Werte für alle k Ergebnisse der Treffeliste werden mit $\left(\frac{R}{k}\right)$ multipliziert. R steht hier für die Anzahl der relevanten Dokumente, k beschreibt die gesamte Anzahl der Ergebnisse. Zuletzt wird dieses Ergebnis durch Division mit *n* normiert. Der Wert *n* wird für jede Trefferliste neu berechnet, da sich je nach Anzahl der Suchergebnisse auch der bestmögliche Confidence Wert verändert.

Die Formel kombiniert einerseits die Precision, andererseits gewichtet sie aber zusätzlich die Position des Suchergebnisses extra. Da wir in einer Trefferliste Ergebnisse von drei verschiedenen Plattformen haben, werden aus einer Liste auch drei Confidence Faktoren ermittelt. Hierfür wird obige Formel auch einmal für jede Plattform angewandt und jedes mal nur die zugehörigen Ergebnisse aufsummiert. Sollte beispielsweise eine Trefferliste aus zehn Ergebnissen bestehen, von denen aber nur zwei von Plattform X geliefert werden, so fließen für die Berechnung des Confidence Wertes der Plattform X die restlichen 8 Ergebnisse mit dem Wert j = 0 ein. Hier zeigt sich schon ein möglicher Schwachpunkt: Wenn eine Plattform nun häufig bei Anfragen kein einziges Ergebnis liefert, wird sie für jede dieser Anfragen einen Confidence Wert von 0 erhalten, was sich im Endeffekt sehr negativ auswirken könnte.

# 4 Implementierung

## 4.1 Implementierung für die erste Evaluation

Für den Benutzer unterscheidet sich die Nutzung von Mytag im Rahmen der ersten Studie nicht vom herkömmlichen Gebrauch der Suchplattform. Was hier implementiert wurde, ist lediglich eine Protokollierung der Suchdaten. Hierfür wurde zunächst via SQL eine entsprechende Tabelle erstellt, um alles festzuhalten und später wieder darauf zugreifen zu können.

Um die Ergebnisse einem Benutzer zuordnen zu können, wird der Benutzername abgespeichert. Für jeden Probanden wurde ein Mytag Account angelegt und der Benutzername bei jeder Suchanfrage mitübergeben. In der Spalte *Searchtag* werden die Tags eingetragen nach denen gesucht wurde, in *Plattform* wird die Plattform gespeichert, welche die Ressource zur Verfügung stellt. Das ist nötig, um später den Confidence Wert für jede Plattform einzeln zu berechnen. Die Spal-

te *Rank* speichert den Rang der Ressource in der Mytag Ergebnisliste ab. Titel und URL der Ressource werden in den entsprechenden Spalten festgehalten.

Die Speicherung wird unmittelbar nach der Sortierung durch den Merge Algorithmus durchgeführt. Das hat den Vorteil, dass die Ergebnisse schon sortiert sind, man also auf den Rang zugreifen kann, der später in die Formel für den Confidence Faktor einfließt.

Rails bietet für die Interaktion mit Datenbanken sehr gute Funktionen, was die Implementierung sehr erleichterte. Bei der Nutzung von Rails ist jede Tabelle eine Klasse und jeder Eintrag ein Objekt dieser Klasse. Von daher wird für jedes Suchergebnis ein neues Objekt erstellt und seine Attribute, also die Spalten, mit Werten gefüllt.

**Listing 1:** Speicherung der Ergebnisse

```
1  #Protokollierung nur im eingeloggten Fall
2  unless response.user == nil
3            #Schleife über alle Ergebnisse
4            result_list.results.each do |res|
5              #neues Objekt wird erstellt
6              entry = Evalu1.new
7              #Attribute des Objekts werden gesetzt
8              entry.username = response.user
9              entry.searchtag = res.searchtag
10             entry.plattform = res.source
11             entry.rank = rank
12             entry.title = res.title
13             entry.url = res.resource.resource_id
14             #und gespeichert
15             entry.save
16             rank +=1;
17           end
18         end
```

Anhand des Ranges des Ergebnisses, kann man beim Auslesen die Suchanfragen wieder trennen, da er bei jeder Suchanfrage wieder auf 1 zurückgesetzt wird.

## 4.2 Implementierung für die zweite Evaluation

Bei der zweiten Evaluation wird für jeden Evaluator eine neue Tabelle erstellt. Die Tabelle entspricht der gefüllten Tabelle aus Umfrage 1, besitzt aber noch zwei zusätzliche Spalten. *Value* speichert je nach Bewertung den float Wert 0,5 oder 1 (Vgl. Algorithmus in Kapitel 2.3). *Category* speichert die int Werte 1-4, angelehnt

an der entsprechenden Aufgabennummer.

Die Bewertung erfolgt über das in Abb. 3 gezeigte Web-Interface. Der Benutzer bekommt die Tags und den Titel der Ressource angezeigt, die es zu bewerten gilt. Ein Klick auf den Titel öffnet die entsprechende Website in einem neuen Fenster. Nach Sichtung der Seite muss der Proband die Ressource bewerten, indem er eine der drei Checkboxen setzt. Ein Klick auf den *Absenden* Button aktualisiert die entsprechenden Einträge in der Datenbank und lädt 10 neue Ergebnisse.

In das Interface werden in 10er Blöcken alle Einträge der Tabelle geladen, die noch keinen Wert in der Spalte *Value* haben. So kann die Versuchsperson zwischendurch auch eine Pause einlegen und spätere komfortabel an derselben Stelle weiterarbeiten.


**Abbildung 3:** Interface für die zweite Befragung (hier noch ohne Design)

Realisiert wurde das Interface durch in HTML eingebetteten Ruby Code, der dynamisch jeweils die ersten unbewerteten 10 Einträge der Tabelle auflistet und für jeden Eintrag drei Checkboxen generiert.

**Listing 2:** Generierung des Interfaces (Für bessere Übersichtlichkeit wurde der HTML Code für die Formatierung und einige Parameter weggelassen)

```
1  form_tag(:action => "item_check")
2   #Schleife über alle Einträge der Tabelle
3    for entry in @evalu2s
4   #Nur die Einträge, die noch nicht bewertet wurden
5      if entry.value == nil
6        #gesuchte Tags Ausgeben
7          entry.searchtag
```

```
8         #Titel (mit Hyperlink) ausgeben
9           link_to (entry.title, entry.url.to_s)
10        #Checkboxen für alle drei Fälle generieren
11          check_box ("not_rel", [#parameter])
12          check_box ("mid_rel", [#parameter])
13          check_box ("vry_rel", [#parameter])
14  submit_tag "Absenden"
```

Drücken des *Absenden* Buttons ruft eine Funktion auf, die je nach Konfiguration der Checkbox die *Value* Spalte der entsprechenden Tabelle füllt.

**Listing 3:** Füllen der Tabelle

```
1  def item_check
2
3      if things_to_check = params[:not_rel]
4        things_to_check.each do |item_id, act|
5          if act == "yes"
6         Evalu2.update_all value = 0.0, id = #{item_id}
7          end
8        end
9      end
10
11     if things_to_check = params[:mid_rel]
12       things_to_check.each do |item_id, act|
13         if act == "yes"
14    Evalu2.update_all value = 0.5, id = #{item_id}
15         end
16       end
17     end
18
19     if things_to_check = params[:vry_rel]
20       things_to_check.each do |item_id, act|
21         if act == "yes"
22    Evalu2.update_all value = 1.0, id = #{item_id}
23         end
24       end
25     end
26   end
```

## 4.3 Implementierung für die Auswertung

Für die Auswertung wurden mehrere Funktionen geschrieben, da es einige Sonderfälle und Sachverhalte zu untersuchen galt. Diese Funktionen unterschieden

sich teilweise nur in Details, weshalb hier auch nur ein repräsentativer Quelltext vorgestellt wird. Die Funktion beinhaltet die Implementierung der in 3.3 vorgestellten Formel. Die benötigten Werte werden aus der Tabelle aus Studie Zwei ausgelesen und miteinander verrechnet. In diesem Fall wird der allgemeine Confidence Wert von allen Ergebnissen berechnet, ungeachtet von Themengebieten oder Plattformen.

**Listing 4:** Berechnung des Confidence Faktors

```
#deklarieren der benötigten Variabeln
 #Berechneter Confidence Wert für die
 #aktuelle Trefferliste
          conf = 0.0
 #Mittelwert aller Berechneter Confidence Werte
          allconf = 0
 #Berechnete Summe aller (j / i) der Ergebnisse
 #der Trefferliste
          sum = 0.0
 #Anzahl der Ergebnisse der Trefferliste
          k = 1.0
 #Anzahl relevanter Ergebnisse der Trefferliste
          r = 0.0
 #Normierungsfaktor für die aktuelle Trefferliste
          norm = 1.0

    #Alle Einträge der Tabelle werden durchgegangen
      for entry in @entrys

     #Bei neuer Trefferliste confi zu allconfi
     #addieren und dann die Werte zurücksetzen
      if entry.rank == 1

              #Formel für den Confidence Faktor
              conf = (sum * (r / k)) / norm
              allconf += conf
              conf = 0.0
              sum = 0.0
              norm = 0.0
              k = 0.0
              r = 0.0

            end

            #Bewertung der Ressource aus Tabelle auslesen
            j = entry.value
```

```
37              #Rang der Ressource aus Tabelle auslesen
38              i = entry.rank
39              #Sofern das Ergebnis nicht mit "irrelevant"
40              #bewertet wurde, wird r um 1 erhöht
41              if entry.value != 0.0
42                r += 1.0
43              end
44
45        k += 1.0
46              #tatsächliche Summe wird berechnet
47              sum = sum + j / i
48              #bestmögliche Summe wird berechnet
49              norm = norm + 1.0/i
50
51            end
52
53            return allconf / 190.0
```

Nachdem alle benötigten Variablen deklariert und mit Ausgangswerten versehen wurden beginnt die Berechnung des Confidence Faktors. Er wird genau dann berechnet, wenn die aktuelle Trefferliste zu Ende ist und eine neue beginnt. Das ist dann der Fall, wenn der Algorithmus auf ein Ergebnis mit der Platzierung 1 trifft. In diesem Fall werden alle bisher gesammelten Werte zum Confidence Faktor verrechnet, dessen Ergebnis zur Summe aller bisher berechneten Confidence Faktoren addiert wird. Anschließend werden alle anderen Werte wieder auf ihren Ursprungswert zurückgesetzt, da kurz darauf eine neue Trefferliste untersucht wird, die andere Ergebnisse enthält und sehr wahrscheinlich auch eine andere Anzahl an Ergebnissen enthält.

Werte *i* und *j* werden bei jedem Ergebnis neu ausgelesen und ihr Quotient zur Summe aller bisherigen Quotienten der Trefferliste addiert. Beim Beginn einer neuen Trefferliste wird der Wert dieser Summe ebenfalls zurückgesetzt. Gleichzeitig wird für jede Ergebnisliste ein Normierungswert berechnet, durch den der Confidence Faktor für diese Liste geteilt wird. Er kann somit bestenfalls 1 werden. Der Normierungswert wird generiert, in dem jedes Ergebnis als „sehr relevant“ verrechnet wird. In diesem Fall ist r / k auch stets 1 und in der Summe wird stets 1 / i addiert.

# 5 Präsentation und Analyse der Ergebnisse

## 5.1 Confidence Faktor

Das Durchschnittsalter aller teilnehmenden Probanden beträgt 31,07 Jahre. Die folgende Grafik gibt Überblick über die Popularität der in Kapitel 6.1 evaluier-

ten Plattformen. Keiner der Befragten gab an, jemals die Plattformen Altavista, del.icio.us, Connotea oder BibSonomy genutzt zu haben.

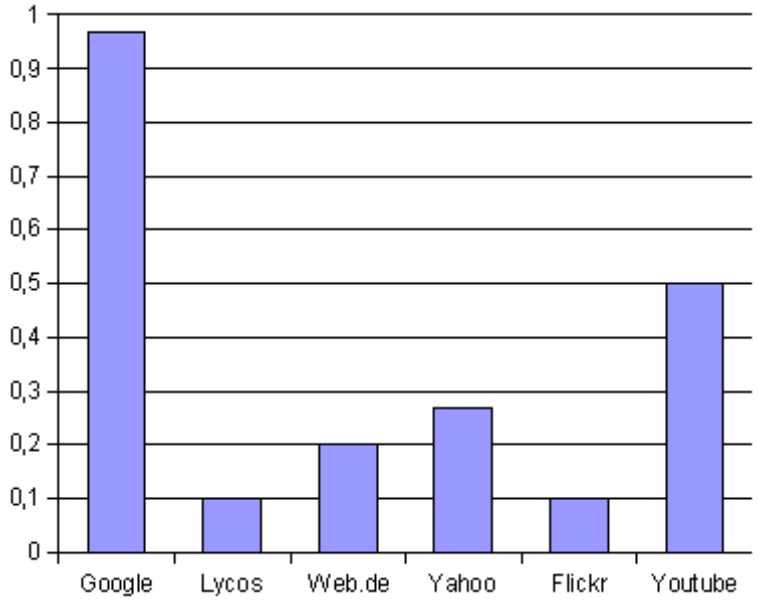


**Abbildung 4:** Prozentuale Nutzung einiger Plattformen. (1,0 entspricht häufiger Nutzung von 100 % der Probanden)

Auf Grund der Angaben bei selbiger Frage wurden die Probanden in zwei Gruppen eingeteilt. Hier wurde angenommen, dass die Probanden, welche viele Services Nutzen, generell mit der Suche im Web erfahrener sind. Bei späteren Untersuchungen, werden diese beiden Benutzergruppen miteinander verglichen. Die weniger erfahrene Gruppe besteht aus 8 Personen, die erfahrenere aus 7.

Insgesamt wurden bei 190 Suchanfragen 3027 Ergebnisse erzielt, davon 1464 individuelle. Pro Suchauftrag stellte ein Nutzer im Schnitt etwa 4 Suchanfragen, von denen jede durchschnittlich etwa 16 Ergebnisse lieferte. Die im Folgenden präsentierten Ergebnisse der Evaluation sind zusätzlich im Anhang in Tabellenform aufgelistet.

Der plattformunabhängige Confidence Faktor von Mytag, welcher durch die in Kapitel 2.3 vorgestellte Formel errechnet wurde, beträgt etwa 0,376 bei einem Maximalwert von 1,0. Er setzt sich aus der Summe der Confidence Werte der einzelnen Plattformen zusammen.

Von den 3027 Ergebnissen wurden 2043 von del.icio.us geliefert, 532 von Connotea und 452 von BibSonomy. Hierbei deutet sich schon an, dass der Confidence Faktor von del.icio.us um einiges höher ausfallen wird als der, der anderen Plattfor-

men, da unter mehr Ergebnissen auch sicherlich mehr relevante sein werden. Die einzelnen Confidence Werte, werden in folgender Grafik visualisiert.

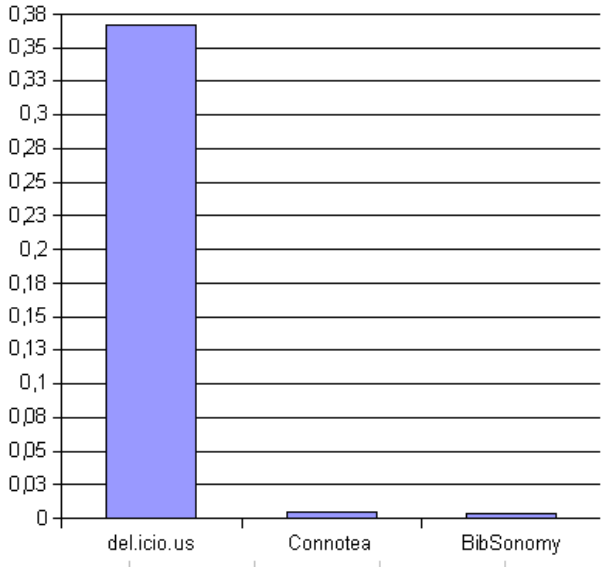


**Abbildung 5:** Confidence Werte der Plattformen über alle Ergebnisse

Es war zwar abzusehen, dass del.icio.us weit vorne liegen würde, was den Confidence Wert engeht; dass die anderen Plattformen aber so wenig relevante Ressourcen zur Ergebnisliste beisteuern ist doch etwas überraschend. Es wird deutlich, dass del.icio.us allein schon 97% Anteil am gesamten Confidence Faktor von Mytag hat. Das ist natürlich auf die Popularität der Plattform zurückzuführen. Grund für den sehr geringen Wert der beiden anderen Plattformen ist, dass bei zahlreichen Trefferlisten 100% der Ergebnisse von del.icio.us geliefert wurden. In dem Fall werden sowohl für Connotea, als auch für BibSonomy für diese Listen Confidence Werte von 0 notiert, was die Gesamtwertung stark nach unten zieht. In den Abbildungen 6-8 wird dargestellt, wie sich die Confidence Faktoren verändern, wenn man die einzelnen Themengebiete voneinander trennt.

Bei del.icio.us fallen sehr starke Schwankungen auf. Das Themengebiet „Shopping“ erzielt einen fast doppelt so hohen Confidence Wert als „Spiele“. Bei dieser Plattform ist eine solche Schwankung in Anbetracht der zahlreichen Ergebnisse sehr aussagekräftig. Auch eine populäre Plattform wie del.icio.us hat also einige Schwachpunkte vorzuweisen und das bei einem doch ziemlich populären Themengebiet. Der Bereich „Freizeit“ weicht nur schwach vom Durchschnittswert ab, „Arbeitswelt“ fällt noch ein wenig geringer aus. Insgesamt liefert del.icio.us aber doch eine sehr zufriedenstellende Anzahl an relevanten Ergebnissen, fällt teilweise positiv auf, ist aber dennoch verbesserungsfähig.

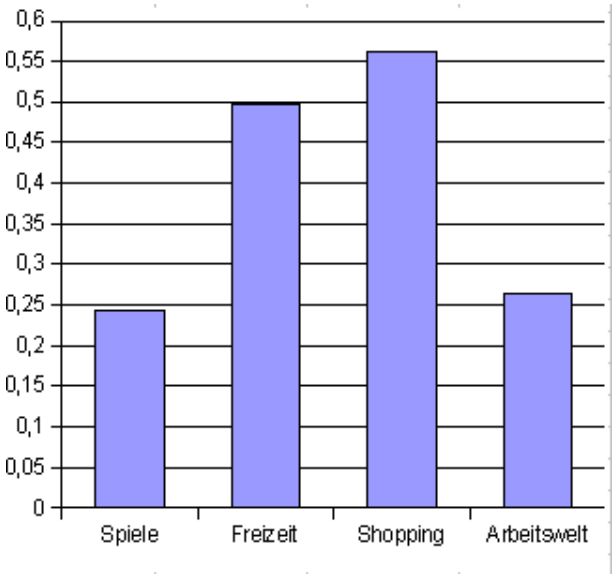


**Abbildung 6:** Confidence Faktor von del.icio.us getrennt nach Themengebieten

| **Themengebiet** | **Anzahl an Ergebnissen** |
|---|---|
| Spiele | 520 |
| Freizeit | 203 |
| Shopping | 434 |
| Arbeitsleben | 886 |

Wie schon erwähnt, hat Connotea bei dieser Umfrage sehr enttäuscht. Neben der Langsamkeit der Plattform ist die doch sehr geringe Anzahl relevanter Ergebnisse stark zu kritisieren. Bei einem solch niedrigen Confidence Wert sind Schwankungen in den Themengebieten leider nicht besonders aussagekräftig, trotzdem fällt auf, dass die eher „seriösen“ Themengebiete um ein vielfaches besser vertreten sind als jene, die mit Zeitvertreib und Spaß zu tun haben. Da Connotea aber primär als wissenschaftliche Plattform bekannt ist, ist dies auch nicht sehr verwunderlich.

Dasselbe kann man im Prinzip auch über BibSonomy sagen. Auffällig ist hier aber, dass neben „Shopping“ auch „Arbeitswelt“ einen Großteil der relevanten Ergebnisse ausmacht, während bei Connotea Abstriche bei der „Arbeitswelt“ zu machen sind.

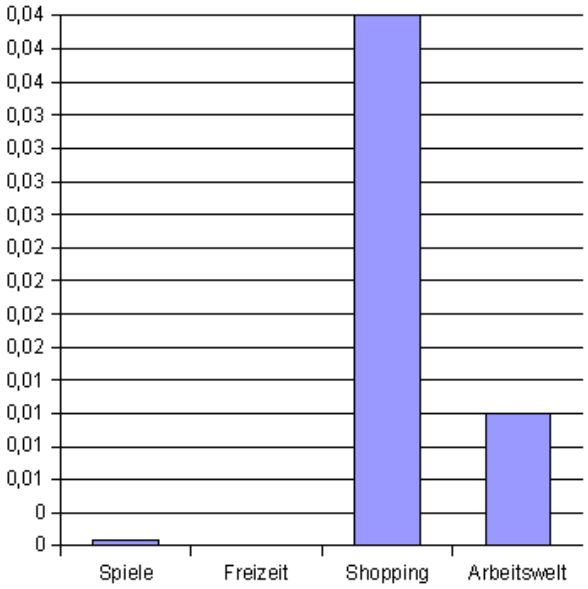


**Abbildung 7:** Confidence Faktor von Connotea getrennt nach Themengebieten

| **Themengebiet** | **Anzahl an Ergebnissen** |
|---|---|
| Spiele | 22 |
| Freizeit | 61 |
| Shopping | 108 |
| Arbeitsleben | 341 |

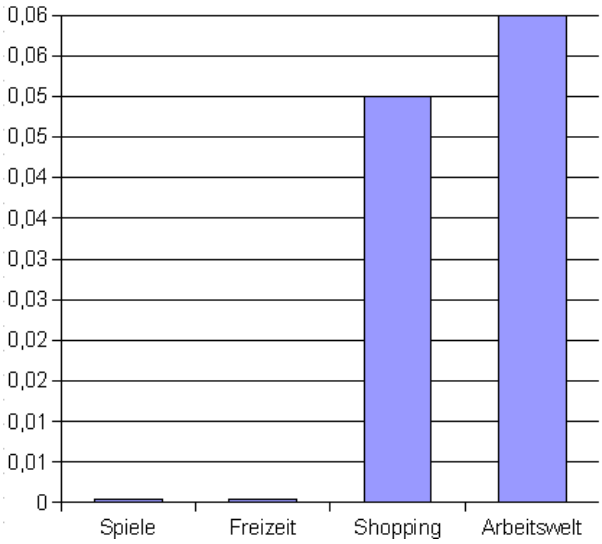


**Abbildung 8:** Confidence Faktor von BibSonomy getrennt nach Themengebieten

| Themengebiet | Anzahl an Ergebnissen |
|---|---|
| Spiele | 112 |
| Freizeit | 18 |
| Shopping | 40 |
| Arbeitsleben | 282 |

Als nächstes werden die beiden Benutzergruppen gegenübergestellt, um beurteilen zu können, ob sich die Erfahrung eines Nutzers mit Suchmaschinen auf die Qualität ihrer Suchergebnisse auswirkt.

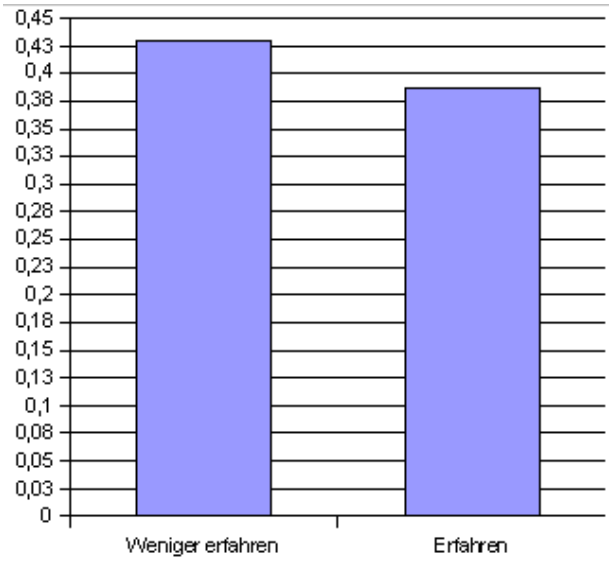


**Abbildung 9:** Vergleich der Trefferlisten von erfahrenen und unerfahrenen Benutzern.

Wie man sieht, gibt es keinen großen Unterschied zwischen den beiden Gruppen. Diejenigen, mit weniger Erfahrung erzielten sogar etwas bessere Ergebnisse. Dass die weniger erfahrenere Benutzergruppe geringfügig bessere Ergebnisse erzielte könnte sich damit begründen lassen, dass sie womöglich unbefangener an Mytag herangegangen sind und einfach intuitiv gehandelt haben, anstatt schon bewährte Suchstrategien anzuwenden. Ergebnis und Schlußfolgerung dieses Sachverhaltes sollen nun noch in einem Signifikanztest untersucht werden.

Die Nullhypothese besteht darin, dass ein Unterschied zwischen der erfahrenen und der weniger ergahrenen Benutzergruppe besteht und die Abweichung vom Durchschnittswert nicht nur zufällig entstanden ist. Dies kann man überprüfen, indem man die Signifikanz der Ergebnisse, also deren Zuverlässigkeit berechnet. Beträgt dieser Wert 5% oder weniger, kann man davon ausgehen, dass die Ergebnisse nicht durch Zufall so entstanden sind. [Wik09]

Die Berechnung wurde einmal für jede Benutzergruppe gemacht. Bei Gruppe

1, den weniger erfahreneren Benutzern liegt die Behauptung vor, dass ihr Confidence Wert stets besser als der durchschnittliche Confidence Wert ist. Das führt zu einem Wahrscheinlichkeitswert von $p$ = 1/2. Die Zahl der Benutzer in Gruppe 1 beträgt $n$ = 8. Variable $a$ wird mit der Anzahl der Benutzer belegt, bei denen die Hypothese zutrifft. Das war bei 4 Personen der Fall. Hier wird bereits deutlich, dass die Signifikanz nahezu unmöglich bestätigt werden kann. Der Vollständigkeit halber soll der Test aber komplett durchgeführt werden. Diese drei Werte setzt man nun in die kumulative Verteilungsfunktion der binomialverteilten Zufallsgröße Z ein.

$$F_p^n(a) = P_p^n(z \leq a) = \sum_{k=0}^{a} B(n; p; k)$$

Das Ergebnis dieser Gleichung lautet 0,6086. Dieses setzt man nun in folgende Formel ein, deren Ergebnis der Signifikanzwert ist und 0,05 nicht überschreiten sollte.

$$1 - F_{p0}^n(c) \leq \alpha = 0,05$$

Wie schon erwartet liegt das Ergebnis mit 0,3914 deutlich über der zulässigen Signifikanz. Man kann also davon ausgehen, dass die Schwankung rein zufällig verlaufen ist und die Erfahrung der Benutzer sich nicht auf den Confidence Wert auswirkt.

Wie man am Anfang des Kapitels sieht, ist die Bewertung von Connotea und BibSonomy sehr schlecht ausgefallen. Zu beurteilen, ob die Platzierung gerechtfertigt ist, oder nicht, bleibt jedem selbst überlassen. Fakt ist aber: Wenn man den Confidence Faktor so, wie er hier berechnet wurde in Mytag einbauen würde, würden alle Ergebnisse von Connotea und BibSonomy ganz ans Ende der Ergebnisliste gesetzt werden. In dem Fall wäre das Merging alles andere als sinnvoll. Im nächsten Unterkapitel soll anhand derselben Ergebnislisten die Precision für alle Plattformen einzeln berechnet werden um zu überprüfen, ob man auf diesem Wege eine sinnvolle Alternative für den Confidence Faktor berechnen kann.

## 5.2 Berechnung von Precision Werten zum Vergleich

Plattformunabhängig beträgt der durchschnittliche Precicion Wert für alle Ergebnislisten 0,53. Der Confidence Wert lag zum Vergleich bei 0,376. Folgende Grafik zeigt die Precicion Werte der einzelnen Plattformen. Zur Berechnung wurden die kombinierten Ergebnislisten wieder aufgespalten.

Hier ist klar zu sehen, dass die Plattformen dichter beeinander liegen. Del.icio.us liegt zwar noch immer klar vorne, aber der Vorsprung ist um einiges geringer. Aus-

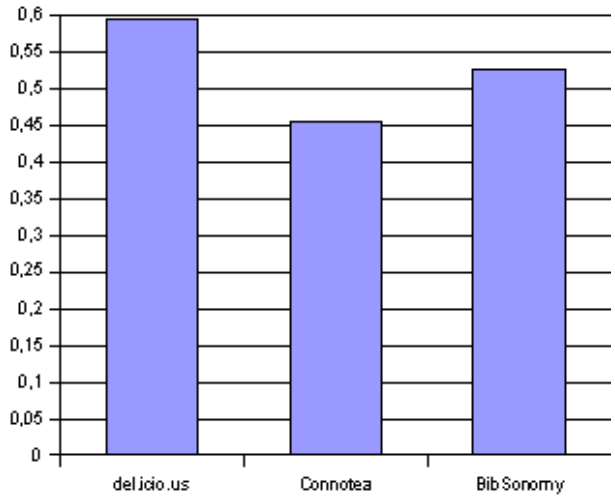


**Abbildung 10:** Precision Werte der Plattformen über alle Ergebnisse

serdem hat BibSonomy Connotea überholt. Bei der Berechnung der Precicion wurden auch wirklich nur die Ergebnisse der entsprechenden Plattform in die Formel aufgenommen. Der Vorteil hiervon ist, dass die Quantität der Ergebnisse kaum noch eine Rolle spielt. Beim Confidence Wert fiel dieser Faktor eindeutig zu stark ins Gewicht.

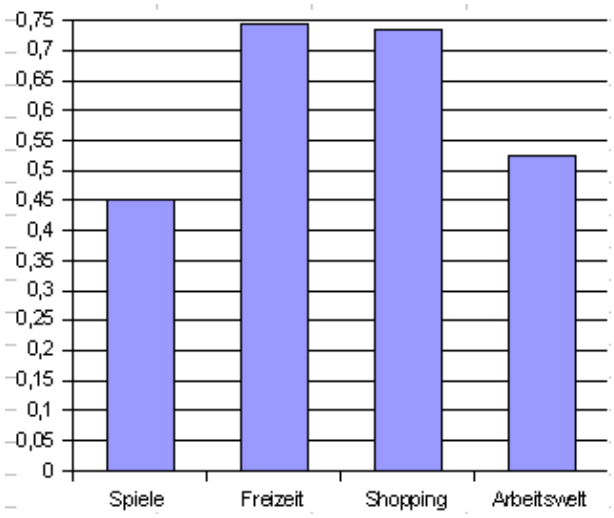


**Abbildung 11:** Precision Wert von del.icio.us getrennt nach Themengebieten

Die Themenverteilung bei del.icio.us ist in etwa gleich geblieben. Lediglich „Freizeit“ hat sich „Shopping“ ein wenig angeglichen.

Bei Connotea fällt ein gewaltiger Sprung der Kategorie „Spiele“ auf, die sich

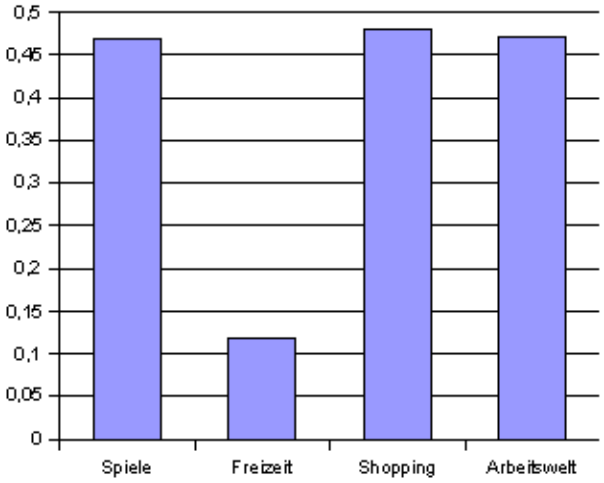


**Abbildung 12:** Precision Wert von Connotea getrennt nach Themengebieten

„Shopping“ und „Arbeitsleben“ angeglichen hat. Das lässt sich damit erklären, dass im Gegensatz zu del.icio.us, Connotea im Bereich „Spiele“ ganz besonders wenig Ergebnisse geliefert hat. Diese waren aber, wie man hier sieht, durchaus brauchbar.

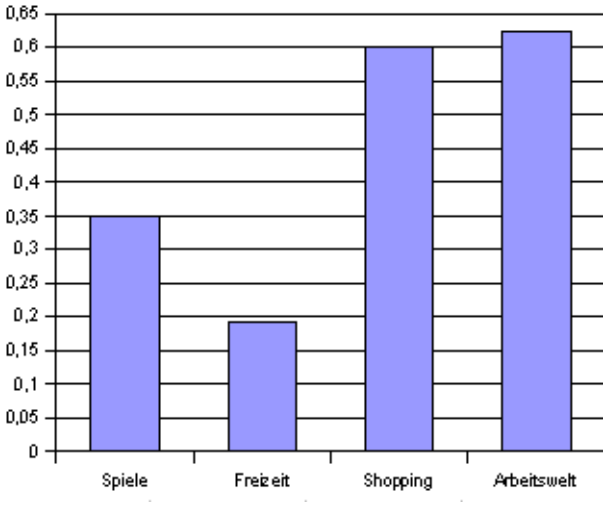


**Abbildung 13:** Precision Wert von BibSonomy getrennt nach Themengebieten

Bei BibSonomy kann man ähnliches feststellen. Während „Shopping“ und „Arbeitsleben“ im Verhältnis gleich blieben, konnte bei „Spiele“ eine Erhöhung des Wertes beobachtet werden.

## 5.3 Confidence Faktor und Precision in der Praxis

Im folgenden Kapitel werden zwei Ergebnislisten derselben Suchanfrage miteinander verglichen. Bei der ersten Liste wurde als plattformbezogener Faktor der in 5.1 vorgestellte Confidence Wert benutzt. In der zweiten Liste wurde für diesen Faktor der Precision Wert aus 5.2 eingesetzt. Angefragt wurde der Begriff *Macbook* und jede Ergebnisliste enthielt 79 Ergebnisse, von denen Mytag die Top 50 darstellte. In folgender Grafik wird die Anzahl der relevanten Ergebnisse in den Top 25 visualisiert.

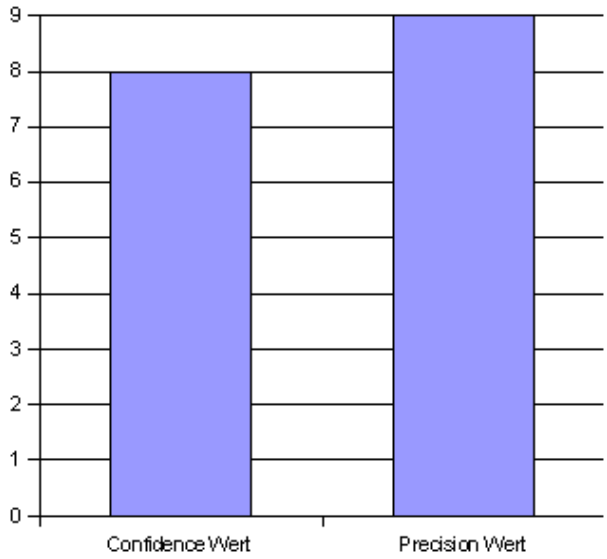


**Abbildung 14:** Relevante Dokumente unter den Top 25 Ressourcen

Abbildung 11 zeigt die Anzahl der relevanten Ressourcen unter den restlichen 25 Ergebnissen.

Beim Precision Wert ist nicht nur die Zahl der relevanten Dokumente in den Top25 höher, sondern auch die gesamte Anzahl der relevanten Dokumente. Das lässt sich dadurch begründen, dass durch den Confidence Wert alle Dokumente von BibSonomy und Connotea ganz ans Ende der Liste gesetzt wurden und so relevante Dokumente aus den Top 50 verdrängt und demnach garnicht aufgelistet wurden. Setzt man den Precision Wert ein, sind dir Ergebnisse also besser sortiert.

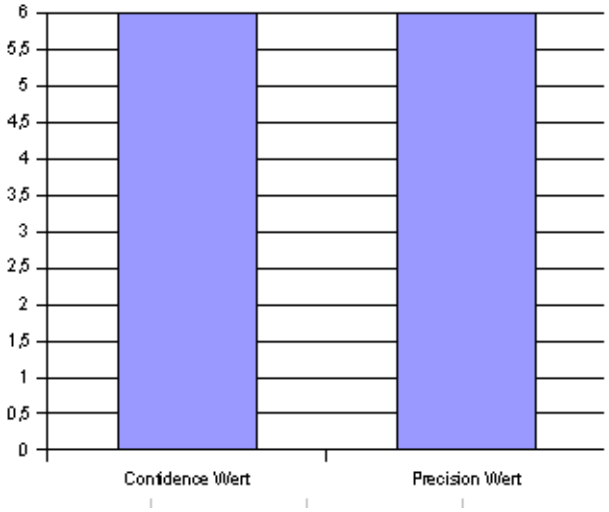


**Abbildung 15:** Relevante Dokumente unter den foldenden 25 Ressourcen

# 6 Erkenntnisse aus der Studie

## 6.1 Fazit

Von den untersuchten Plattformen zeichnet sich del.icio.us deutlich als Sieger heraus. Sehr überraschend ist diese Erkenntnis nicht - bezeichnet sich del.icio.us doch selbst als „the biggest collection of bookmarks in the universe“[6], während Connotea und BibSonomy eher kleine Insiderplattformen darstellen. Hierbei ist dennoch zu bedenken, dass sie eher als Ergänzung von del.icio.us gedacht waren, um auch wissenschaftliche Bookmarks besser anbieten zu können. Durch den Merging Algorithmus werden nun die besten Ergebnisse von allen 3 Plattformen zur Verfügung gestellt. Insgesamt kann man sagen: Wenn del.icio.us Ergebnisse zu einem Thema liefert, kann man davon ausgehen, dass relevante Ressourcen darunter zu finden sind. Dennoch werten Connotea und BibSonomy die Ergebnisliste mit wenigen, aber relevanten Ergebnissen auf. Ihre Einbindung war also doch nicht ungerechtfertigt.

Was während der Studie leider häufig auffiel, war die schlechte Performance von Connotea. Diese Langsamkeit wurde von den meisten Benutzern leider auf Mytag selbst übertragen, von daher sollte entweder Connotea für mehr Schnelligkeit sorgen, oder darüber nachgedacht werden, diese Plattform zu ersetzen.

Wie in Kapitel 4 schon erwähnt, spielen bei allen drei Plattformen die Themengebiete durchaus eine Rolle. Del.icio.us hat im Bereich „Spiele“ etwas enttäuscht - ist es doch nicht zuletzt eine Bookmarking Plattform für Jedermann. Abschließend

[6] http://del.icio.us

kann man aber sagen, dass Mytag sehr gut daran tut, die Ergebnisse von del.icio.us abzufragen. Die del.icio.us API liefert pro Anfrage nur 18 Ergebnisse, was momentan noch ein kleiner Schwachpunkt ist. Trotzdem kamen viele relevante Ressourcen von del.icio.us. Die beiden anderen Plattformen lassen sich, was die Zahl der Ergebnisse angeht, zwar beliebig konfiguieren, wenn aber z.B. nur zwei Ergebnisse vorhanden sind, bringt ein Limit von maximal 50 Ergebnissen, wie es im Moment eingestellt ist, auch keine Vorteile. BibSonomy und Connotea liefern zusammen knapp 25% aller Ergebnisse. In Anbetracht dieser Zahl ist ein Confidence Faktor von knapp 3% sehr mager.

Beim Information Retrieval kann keine Plattform ausschließlich zufriedenstellende Ergebnisse liefern. Bei der Beobachtung von Suchanfragen wurde immer wieder klar, was eigentlich schon offensichtlich ist: Ungenaue Suchanfragen liefern schlechte Ergebnislisten. Wurde für Suchauftrag 2 z.B. nur nach New York gesucht, kamen unzählige, für die Aufgabe irrelevante Ergebnisse in die Trefferliste. Ohne eine gewisse Fachkenntnis des Benutzers, kann auch die beste Suchmaschine keine Wunder bewirken. Wichtig ist es, den Benutzer zu „erziehen“. Er muss intuitiv merken, wie er suchen muss, um relevante Ergebnisse zu erhalten. Durch Mytag haben die Probanden mehr Erfahrung in der Suche auf Tagging Plattformen gewonnen. Das spricht also eindeutig für benutzerfreundlichen Aufbau der Suchmaschine und auch für den Ranking Algorithmus, der durch bessere Anfragen auch bessere Ergebnislisten liefert.

Auch wenn sich in dieser Studie der Confidence Faktor nicht als geeignetes Bewerungskriterium für Plattformen erwiesen hat, kann er in anderen Fällen durchaus sinnvoll sein. Immer dann, wenn es wenige Ergebnisse gab, wurde der Wert sehr schlecht. Wenn man nun also garantiert eine hohe Quantität an Ergebnissen hat und ein gutes Kriterium für die Qualität sucht, ist der Confidence Faktor eine gute Wahl. Hat man es eher mit unpopulären Plattformen zu tun, bietet es sich an, bei der Precision zu bleiben.

## 6.2 Ausblick

Mit den Ergebnisse dieser Studie wird klar, dass es für Mytag mehr Sinn macht, den Precision Wert einer Plattform als plattformbezogenes Kriterium zu nehmen, statt des Confidence Wertes. Die neu ermittelten Werte werden in Kürze in die Mytag Plattform integriert werden und so für bessere Ergebnisse sorgen. Der nächste Schritt könnte nun sein, den verbesserten Ranking Algorithmus im Vergleich zu anderen Algorithmen zu untersuchen.

In weiterer Hinsicht sollte aber bedacht werden, Connotea und BibSonomy durch populärere Bookmarking Plattformen wie z.B. Mr Wong[7] zu ergänzen, da

[7] http://www.mr-wong.de

sie im Vergleich zu del.icio.us quantitativ ein relativ schlechtes Ergebnis in der Studie erzielt haben.

Bisher wurde Mytag auf diversen Konferenzen sehr positiv zur Kenntnis genommen. Durch diese Studie wird es zusätzlich noch eine bessere wissenschaftliche Basis erhalten. Durch Implementierung von neuen Funktionalitäten wie der Suchhilfe und weiterer Plattformen, vielleicht auch im Video- oder Fotobereich hat Mytag Potential in der Web 2.0 Gemeinde gut aufgenommen zu werden.

# 7 Anhang 1: Der Fragebogen

Der erste Teil des Fragebogens sammelt Informationen über den Probanden. Im zweiten Teil werden ihm die Aufgaben mitgeteilt, die er zu erfüllen hat. Hier muss er nichts ausfüllen, da alle Daten elektronisch erfasst werden. Zuletzt werden einige abschließende Fragen über die behandelten Suchmaschinen und über die Studie selbst gestellt. Alle Textstellen in kursiver Schrift kommen auch so im Fragebogen vor.

## 7.1 Allgemeine Fragen

- *Geben Sie bitte Ihr Alter und Geschlecht an.*
- *Wie oft nutzen Sie die folgenden Suchmaschinen/Plattformen?*

| | oft | gelegentlich | nie |
|---|---|---|---|
| *Google* | ☐ | ☐ | ☐ |
| *Altavista* | ☐ | ☐ | ☐ |
| *Lycos* | ☐ | ☐ | ☐ |
| *Web.de* | ☐ | ☐ | ☐ |
| *Yahoo* | ☐ | ☐ | ☐ |
| *del.icio.us* | ☐ | ☐ | ☐ |
| *Flickr* | ☐ | ☐ | ☐ |
| *Youtube* | ☐ | ☐ | ☐ |
| *Connotea* | ☐ | ☐ | ☐ |
| *BibSonomy* | ☐ | ☐ | ☐ |

## 7.2 Die Aufgabenliste der ersten Umfrage

Die Aufgaben wurden in Anlehnung an die in [Bie05] vorgestellten Aufgabentypen erstellt. Jede einzelne wurde im Vorfeld getestet. Die Aufgabe gilt als erfüllt, wenn der Benutzer die gewünschte Information auf der Ressource, die unter der URL zu erreichen ist, finden kann. Das schließt auch alle Seiten ein, die auf derselben Domain der Ressource liegen.

*Bitte führen Sie folgende Suchaufträge mittels Mytag aus. Wie Sie die Aufgabe lösen, also durch welche Suchanfragen, bleibt Ihnen überlassen. Sie können eigene Begriffe verwenden, mit der Suchhilfe oder mit der Tagcloud arbeiten. Die Suchanfragen können auch auf englisch erfolgen, falls die Ergebnisse in deutscher Sprache nicht zufriedenstellend sind. Eine Aufgabe gilt als erfüllt, sobald sie die Url-Adresse einer passenden Seite angeben können.*

**Aufgaben**

- Spiele: *Ihr Neffe (11) hat bald Geburtstag und wünscht sich ein Spiel (egal ob Brettspiel oder Videospiel) von Ihnen. Seine Eltern legen aber sehr viel Wert darauf, dass es angemessen für sein Alter ist. Suchen Sie eine Webseite, die sich mit Jugendschutz/Altersfreigaben von Spielen beschäftigt.*
- Freizeit: *Sie wollen eine Reise nach New York planen. Benutzen Sie MyTag um sich über mögliche Sehenswürdigkeiten zu informieren. Finden Sie Web-Seiten, die Ihnen Informationen über Sehenswürdigkeiten, Restaurantempfehlungen und Hotels liefern. Geben Sie als Lösung zu jeder dieser Kategorien die URL von mindestens einer geeigneten Webseite an.*
- Shopping: *Sie sind in ein abgelegenes Dorf gezogen und dadurch mehr und mehr auf Online Shopping angewiesen. Finden Sie einen Online Shop, der Artikel für ein beliebiges Hobby von ihnen vertreibt. Da Sie sparen müssen suchen Sie noch einen weiteren Shop um die Preise vergleichen zu können.*
- Berufsleben: *Ein guter Bekannter hat kürzlich seine Arbeit verloren. Suchen Sie mittels Mytag je ein deutsches und ein englisches Job-Portal, um ihm zu helfen.*

*Vielen Dank für Ihre Mitarbeit!*

Desweiteren wurden noch Fragen zur Tagcloud und zur Suchhilfe gestellt, die aber im Rahmen der Partnerstudienarbeit verfasst und ausgewertet wurden.

# 8 Anhang 2: Die Ergebnisse auf einen Blick

## 8.1 Nutzung der verschiedenen Plattformen

| Alter | Geschlecht | Google | Altavista | Lycos | Web.de | Yahoo | del.icio.us | Flickr | Youtube | Connotea | Bibsonomy |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | m | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 26 | m | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | w | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0,5 | 0 | 0 |
| 23 | w | 1 | 0 | 0 | 0,5 | 0,5 | 0 | 1 | 0 | 0 | 0 |
| 19 | w | 1 | 0 | 0,5 | 0 | 0,5 | 0 | 0,5 | 1 | 0 | 0 |
| 24 | w | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0,5 | 0 | 0 |
| 48 | w | 1 | 0 | 0,5 | 1 | 0,5 | 0 | 0 | 0,5 | 0 | 0 |
| 48 | m | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 55 | w | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 56 | m | 0,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | w | 1 | 0 | 0,5 | 0 | 0,5 | 0 | 0 | 0,5 | 0 | 0 |
| 30 | m | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0,5 | 0 | 0 |
| 18 | m | 1 | 0 | 0 | 0 | 0,5 | 0 | 0 | 1 | 0 | 0 |
| 25 | m | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0,5 | 0 | 0 |
| 18 | m | 1 | 0 | 0 | 0,5 | 0,5 | 0 | 0 | 0,5 | 0 | 0 |

**Abbildung 16:** Nutzung einiger Plattformen. (1,0 entspricht häufiger Nutzung, 0,5 seltener Nutzung)

## 8.2 Genereller Vergleich der drei Plattformen

| **Plattform** | **Confidence Faktor** | **Precision** |
|---|---|---|
| del.icio.us | 0.367 | 0.595 |
| Connotea | 0.005 | 0.455 |
| BibSonomy | 0.004 | 0.526 |

## 8.3 Entwicklung bei verschiedenen Themengebieten

| Plattform | Themengebiet | Confidence Faktor | Precision |
|---|---|---|---|
| del.icio.us | Spiele | 0.242 | 0.45 |
| del.icio.us | Freizeit | 0.496 | 0.744 |
| del.icio.us | Shopping | 0.563 | 0.736 |
| del.icio.us | Arbeitsleben | 0.263 | 0.526 |
| Connotea | Spiele | 0.0004 | 0.468 |
| Connotea | Freizeit | 0.0001 | 0.119 |
| Connotea | Shopping | 0.04 | 0.481 |
| Connotea | Arbeitsleben | 0.01 | 0.47 |
| BibSonomy | Spiele | 0.0005 | 0.35 |
| BibSonomy | Freizeit | 0.0005 | 0.193 |
| BibSonomy | Shopping | 0.05 | 0.6 |
| BibSonomy | Arbeitsleben | 0.06 | 0.624 |

## 8.4 Einfluss der Erfahrung mit Suchmaschinen

| Benutzergruppe | Confidence Faktor |
|---|---|
| Weniger erfahren | 0.429 |
| Erfahren | 0.387 |

# Literatur


[ASG09] Florian Altherr, Matthias Scharek, and Daniel Grabs. *Personalisierte Suche in MyTag*. Universität Koblenz-Landau, Informationssysteme und Semantic Web, 2009.

[BD06] Jürgen Bortz and Nicola Döring. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler (4. Aufl.)*. Berlin: Springer, 2006.

[Bie05] Andreas Bienz. *Das Benutzerverhalten beim Suchen im WWW*. Universität Basel,Wirtschaftwissenschaftliches Zentrum, 2005.

[Grä99] Lorenz Gräf. *Online Research, Methoden, Anwendungen und Ergebnisse*, chapter Optimierung von WWW-Umfragen, pages 159–177. Hogrefe, 1999.

[GWG06] Susan Gauch, Guijun Wang, and Mario Gomez. Profusion: Intelligent fusion from multiple, distributed search engines. *Journal of Universal Computer Science*, vol. 2, no. 9:637–649, 2006.

[KC06] Anoop Kunchukuttan and Prof. Soumen Chakrabarti. *Evaluation of Information Retrieval Systems*. Department of Computer Science and Engineering,Indian Institute of Technology, Mumbai, 2006.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Sch09] Matthias Scharek. *Optimierung von Suchmaschinen basierend auf dem Suchverhalten von Benutzern im Internet*. Universität Koblenz-Landau, Informationssysteme und Semantic Web, 2009.

[Wik09] Wikipedia. *Signifikanztest - Wikipedia, Die freie Enzyklopädie*. http://de.wikipedia.org/w/index.php?title=Signifikanztest&oldid=55279668, [Online; Stand 11. Januar 2009].