Fachhochschule Potsdam
Institut für Information und Dokumentation
Kurs A 1999

Statistische Auswertung von Suchfragen als Mittel zur Pflege einer Schlagwortliste

Abschlußarbeit
im Lehrgebiet Inhaltliche Erschließung
Betreuerin: Dr. Christa Ladewig

vorgelegt von

Klaus Ebner

Hamburg, September 1999

Kurzreferat

Thesauri und Schlagwortlisten müssen in bestimmten Zeitabständen überarbeitet werden. Gründe dafür sind die Weiterentwicklung des Fachgebiets, Änderung der Fachsprache, neue Benutzerkreise, Änderung der Benutzerinteressen und der Wunsch der Benutzer nach besserer Indexierung. Am Deutschen Elektronen-Synchrotron DESY wird die Literatur zur Hochenergiephysik für die Literaturdatenbank SPIRES-HEP indexiert. Für die Indexierung wird eine Schlagwortliste verwendet. Bei der Überarbeitung der Schlagwortliste sollen die Benutzerinteressen berücksichtigt werden. Zu diesem Zweck wurden 28 000 Suchfragen der Benutzer an SPIRES-HEP statistisch ausgewertet. Die Suchfragen werden laufend in WWW-Gateway-Logfiles protokolliert. Mit Hilfe eines Computerprogramms wurden die Suchhäufigkeiten von Begriffen im Titel- und im Deskriptorenfeld bestimmt. Es wurden die besonders häufig im Titelfeld recherchierten Begriffe ermittelt und daraus Kandidaten für neue Deskriptoren, Nichtdeskriptoren und Vorzugsbenennungen gewonnen. Es wurden besonders häufig recherchierte Deskriptoren mit sehr allgemeiner Bedeutung ermittelt, die für eine nähere Spezifizierung in Frage kommen. Die sehr häufig recherchierten Deskriptoren wurden mit den sehr häufig bei der Indexierung verwendeten Deskriptoren verglichen. Dabei wurde ein Unterschied zwischen den Interessenschwerpunkten der Benutzer und dem Angebot an neuen Dokumenten in der Datenbank festgestellt. Außerdem wurden Erkenntnisse über das Verhalten der Benutzer bei der Recherche im Titel- und im Deskriptorenfeld gewonnen. Damit die Auswertung von Suchfragen für die Pflege der SPIRES-HEP-Schlagwortliste effektiv eingesetzt werden kann, müßte eine größere Zahl von Suchfragen untersucht werden.

Inhaltsverzeichnis

1 Einleitung **7**

2 Literaturdokumentation zur Hochenergiephysik bei DESY **9**

2.1 Entwicklung 9

2.2 Indexierung 11

2.3 Recherche in der Datenbank SPIRES-HEP 13

3 Bestimmung der Suchhäufigkeiten **16**

4 Statistische Auswertung der Suchhäufigkeiten **18**

5 Ergebnisse **20**

5.1 Suchfragen im Titelfeld 20

5.2 Suchfragen im Deskriptorenfeld 21

5.3 Vergleich der Recherche nach Deskriptoren mit der Indexierung . 22

6 Schlußfolgerungen **25**

7 Zusammenfassung **27**

Literatur **29**

1 Einleitung

In der Dokumentationsabteilung des Deutschen Elektronen-Synchrotrons DESY wird die weltweit erscheinende Literatur zur Hochenergiephysik indexiert und klassiert. Die Dokumentationseinheiten werden in der Literaturdatenbank SPIRES-HEP gespeichert. Für die Indexierung der Dokumente wird ein kontrolliertes Vokabular in Form einer Schlagwortliste verwendet. Diese Schlagwortliste steht auch den Benutzern für die Recherche zur Verfügung.

Thesauri und Schlagwortlisten müssen in bestimmten Zeitabständen überarbeitet werden. Eine Einführung zu diesem Thema geben Gaus [1], Ladewig [2] und Buder u.a. [3]. Die Gründe für eine Überarbeitung sind z. B.

- Weiterentwicklung des Fachgebiets
- Änderung der Fachsprache
- neue Benutzerkreise kommen hinzu
- Interessen der Benutzer ändern sich
- Wunsch der Benutzer nach besserer Indexierung.

Folgende Änderungen können bei der Überarbeitung einer Schlagwortliste vorgenommen werden:

- Einführung neuer Deskriptoren bzw. Äquivalenzklassen
- Einführung neuer Nichtdeskriptoren als Zugangsvokabular
- neue, aussagefähigere Vorzugsbenennungen innerhalb von bestehenden Äquivalenzklassen
- Entlastung von sehr häufig benutzten Deskriptoren (durch zusätzliche Unterbegriffe, verwandte Begriffe, Einschränkung des Bedeutungsumfangs)
- Streichen von sehr selten oder gar nicht benutzten Deskriptoren.

Bei der Überarbeitung der Schlagwortliste für SPIRES-HEP wurden Änderungsvorschläge bisher im Kreis der für die Dokumentation tätigen Physikerinnen und Physiker gesammelt oder durch statistische Auswertung der Indexate gewonnen. Es besteht der Wunsch, die Interessen der Benutzer bei der Überarbeitung der Schlagwortliste zu berücksichtigen. Zu diesem Zweck sollen in dieser Arbeit Suchfragen der Benutzer an die Datenbank SPIRES-HEP statistisch ausgewertet werden.

Ziel der statistischen Auswertung ist es, die folgenden Listen zu erzeugen:

- Begriffe, nach denen besonders häufig im Titelfeld recherchiert wird:
  Aus der Liste dieser Begriffe sollen Kandidaten für neue Deskriptoren, Nichtdeskriptoren und Vorzugsbenennungen gewonnen werden.

- Deskriptoren, nach denen besonders häufig recherchiert wird:
  Hier sollen Deskriptoren mit sehr allgemeiner Bedeutung ermittelt werden, die näher spezifiziert werden können, z. B. durch Einführen von Unterbegriffen.

  Darüberhinaus geben die sehr häufig recherchierten Deskriptoren Auskunft über die Interessenschwerpunkte der Benutzer. Sie sollen mit den sehr häufig bei der Indexierung verwendeten Deskriptoren verglichen werden. Auf diese Weise soll untersucht werden, inwieweit das Angebot an neuen Dokumenten in der Datenbank der Nachfrage durch die Benutzer entspricht.

- Deskriptoren, nach denen sehr selten oder gar nicht recherchiert wird:
  Bei diesen Deskriptoren ist zu überlegen, ob sie durch allgemeinere Begriffe abgedeckt sind und gestrichen werden können.

Die Suchfragen der Benutzer werden laufend in WWW-Gateway-Logfiles protokolliert. Im Rahmen dieser Arbeit wird ein Computerprogramm entwickelt, das die Suchhäufigkeiten von Begriffen im Titel- und im Schlagwortfeld bestimmt. Die Häufigkeiten werden dann statistisch ausgewertet.

2 Literaturdokumentation zur Hochenergiephysik bei DESY

2.1 Entwicklung

Das Deutsche Elektronen-Synchrotron DESY ist ein mit öffentlichen Mitteln finanziertes Forschungszentrum mit Sitz in Hamburg und einem Teilinstitut in Zeuthen (Brandenburg). DESYs Auftrag ist die naturwissenschaftliche Grundlagenforschung mit den Schwerpunkten

- Hochenergiephysik: Untersuchung der fundamentalen Bausteine der Materie
- Anwendung von Synchrotronstrahlung in den Wissenschaften Oberflächenphysik, Materialwissenschaften, Chemie, Molekularbiologie, Medizin und Geophysik.

DESY wurde 1959 in Hamburg als selbständige Stiftung bürgerlichen Rechts gegründet. Seit dem 1. Januar 1992 ist das Institut für Hochenergiephysik in Zeuthen ein Teilinstitut von DESY. DESY ist Mitglied der Hermann von Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF). Weitere Informationen zu DESY gibt es im WWW unter der URL http://www.desy.de.

Der Etat von DESY wird zu 90 % vom Bund und zu 10 % von den Ländern Hamburg bzw. Brandenburg getragen. Bei DESY sind 1390 Mitarbeiter in Hamburg und 160 Mitarbeiter in Zeuthen angestellt.

Im Jahr 1963 wurde bei DESY mit der Literaturdokumentation begonnen. Das Ziel damals wie heute ist es, die Literatur zur Hochenergiephysik umfassend nachzuweisen. Im einzelnen wird die Literatur der folgenden Fachgebiete dokumentiert:

- Elementarteilchenphysik (Experiment, Phänomenologie)
- Quantenfeldtheorie
- Physik und Technik der Beschleuniger und Detektoren
- Hochenergie-Kernphysik
- Hochenergie-Astrophysik.

Folgende Literaturarten werden dokumentiert:

- Artikel aus Zeitschriften und Konferenzabhandlungen
- Lehrbücher und Monographien
- graue Literatur: Preprints[1], Forschungsberichte, Diplomarbeiten, Dissertationen und Habilitationsschriften.

Die Erfassung der Literatur erfolgt in englischer Sprache.

Die bibliographischen Nachweise wurden von Beginn an in der Zeitschrift „High Energy Physics Index“ veröffentlicht. Die Zeitschrift erschien 14-tägig und wurde durch jährliche Register von Autoren, Sachbegriffen und Berichtsnummern ergänzt. Das Erscheinen des gedruckten „High Energy Physics Index“ wurde Ende 1996 eingestellt.

Ende der 60er Jahre wurde bei DESY begonnen, die Dokumentationseinheiten auf Magnetbändern zu speichern. Dadurch wurde Offline-Retrieval im Datenbestand möglich. 1974 begann die Zusammenarbeit von DESY mit dem Forschungszentrum SLAC[2] auf dem Gebiet der Dokumentation mit dem Austausch von Daten.

1988 wurden die Datenbestände von DESY und SLAC zusammengeführt; es entstand die heutige Datenbank SPIRES-HEP. Bei der Produktion von SPIRES-HEP übernimmt SLAC die formale Erfassung der Literatur, während DESYs Aufgabe die Indexierung und Klassierung ist. Die Datenbank ist vom SLAC-Rechner auf einen Rechner bei DESY gespiegelt. Die Daten werden täglich aktualisiert. Seit 1996 bietet DESY den Zugriff auf SPIRES-HEP über das WWW an (http://www.desy.de → DESY Library → SPIRES database system). Bei den meisten seit 1992 bearbeiteten Dokumentationseinheiten besteht Zugriff auf die Abstracts und die Volltexte.

Die Literaturdatenbank SPIRES-HEP enthält zur Zeit über 400 000 Nachweise. Jedes Jahr kommen ca. 20 000 Nachweise hinzu. SPIRES-HEP ist damit die größte Literaturdatenbank zur Hochenergiephysik. Die Benutzung von SPIRES-HEP ist kostenlos.

[1] Preprints sind wissenschaftliche Aufsätze, die schon vor dem Erscheinen in einer Fachzeitschrift durch den Autor verbreitet werden.

[2] Stanford Linear Accelerator Center, Stanford, USA

In der Organisationsstruktur von DESY bildet die Dokumentation zusammen mit der Bibliothek eine Gruppe. In der Dokumentation arbeiten vier Physikerinnen und Physiker und zwei Dokumentationsassistentinnen. Sechs weitere Physikerinnen und Physiker und fünf Korrekturleserinnen und Korrekturleser sind als freie Mitarbeiter für die Dokumentation tätig.

2.2 Indexierung

Zweck der Indexierung von Dokumenten ist, Informationen einfacher, schneller und treffsicherer auffindbar zu machen. Außerdem soll die Gesamtheit der Deskriptoren den Inhalt eines Dokuments in Kurzform wiedergeben.

In der Dokumentation bei DESY wird die Literatur zur Hochenergiephysik von Physikerinnen und Physikern indexiert. Die Indexierung erfolgt ausschließlich intellektuell. Es wird ein kontrolliertes Vokabular verwendet, das in Form einer Schlagwortliste vorliegt. Die Indexierer vergeben zwischen 10 und 15 Deskriptoren pro Dokument. Es gibt die Möglichkeit, je zwei Deskriptoren zu präkoordinieren. Unter Präkoordination versteht man die Kombination von Deskriptoren beim Indexieren mit dem Zweck, einen komplexen Sachverhalt auszudrücken. Bei der Indexierung für die Datenbank SPIRES-HEP werden zwei Deskriptoren präkoordiniert, indem sie durch ein Komma und ein Leerzeichen miteinander verknüpft werden.

Zusätzlich zu den Deskriptoren wird den Dokumenten ein Klassifikationscode zugeordnet. Dieser Code gibt an, in welches Sachgebiet der Hochenergiephysik ein Dokument fällt. Es gibt 14 Sachgebiete, die zu den drei Hauptgebieten „Experimentelle Physik“, „Instrumente und Methoden in der Hochenergiephysik“ und „Theoretische Physik“ zusammengefaßt sind. Pro Dokument können ein Hauptcode und mehrere Nebencodes vergeben werden.

Die Schlagwortliste zur Literaturdatenbank SPIRES-HEP enthält 2140 Deskriptoren. Nach ihrem Inhalt kann man die Deskriptoren unterscheiden in

- Sachbegriffe
- Namen von Teilchen und chemischen Elementen
- Namen von Instrumenten (Beschleuniger, Detektoren)
- Begriffe, die die Form, Charakter oder Teile eines Dokuments beschreiben (z.B. „book“, „experimental results“, „bibliography“).

Nach der Art ihrer Verwendung beim Indexieren kann man die Deskriptoren unterscheiden in

- Hauptdeskriptoren (im Manual der Schlagwortliste „regular keywords“ genannt):
  Hauptdeskriptoren können alleine stehen; sie können aber auch mit einem anderen Hauptdeskriptor oder mit einem Nebendeskriptor präkoordiniert werden.

- Nebendeskriptoren („standardized non-keywords“):
  Nebendeskriptoren sind unselbständige Deskriptoren. Sie können nur bei der Präkoordination an einen Hauptdeskriptor angehängt werden.

- Nichtdeskriptoren:
  Sie dienen als Zugangsvokabular bei der Recherche. Jedem Nichtdeskriptor ist ein Verweis auf einen Deskriptor angefügt.

Die Unterscheidung zwischen Haupt- und Nebendeskriptoren hat ihren Ursprung in der Indexierung für die gedruckte Bibliographie „High Energy Physics Index“. Im „High Energy Physics Index“ erfolgte die Recherche nach Sachbegriffen in einem gedruckten Sachregister. In diesem Register sind die Deskriptoren alphabetisch sortiert; zu jedem Deskriptor sind die laufenden Nummern der Dokumentationseinheiten sowie die Titel der Dokumente aufgelistet, an die der Deskriptor vergeben wurde.

Den Benutzern sollte die Möglichkeit gegeben werden, in der Bibliographie auch nach speziellen Sachverhalten recherchieren zu können. Gleichzeitig sollte aber die Menge der zu vergebenden Deskriptoren überschaubar bleiben. Aus diesem Grund entschied man sich für die Indexierung mit Präkoordination. Dabei wird einem Hauptdeskriptor mit weitem Begriffsumfang ein ergänzender Nebendeskriptor angefügt. Neben dem detaillierten Zugriff auf Sachverhalte liegt ein weiterer Vorteil der Präkoordination darin, daß die Deskriptoren eines größeren Themenbereichs im Register untereinander stehen.

Um das gedruckte Sachregister möglichst kurz zu halten, verzichtete man auf die Permutation der Deskriptoren, falls ein Hauptdeskriptor mit einem Nebendeskriptor präkoordiniert worden war. Somit ist der Sucheinstieg nur über die Hauptdeskriptoren möglich.

Die Speicherung der Dokumentationseinheiten in der Datenbank SPIRES-HEP erweiterte die Recherchemöglichkeiten der Benutzer. Hier erfolgt die Recherche

im Index des Deskriptorenfelds, in dem die Hauptdeskriptoren, die Nebendeskriptoren und die präkoordinierten Deskriptoren stehen. Zum einen ist damit die direkte Suche nach Nebendeskriptoren möglich. Zum anderen gibt es die Möglichkeit, in einer Suchfrage mehrere Deskriptoren mit Boole'schen Operatoren zu verknüpfen.

Seitdem das Erscheinen des „High Energy Physics Index“ Ende 1996 eingestellt wurde, fällt eine Reihe von Beschränkungen für die Indexierung weg. Dies hat zur Folge, daß die Unterschiede zwischen Haupt- und Nebendeskriptoren verwischt werden: manche Nebendeskriptoren werden jetzt wie Hauptdeskriptoren verwendet.

In der Schlagwortliste sind die Hauptdeskriptoren durch Fettdruck gekennzeichnet. Den Nebendeskriptoren ist ein Stern und den Nichtdeskriptoren ein Bindestrich vorangestellt. Manchen Deskriptoren sind Anmerkungen in Klammern hinzugefügt. Mögliche Anmerkungen sind z. B.

- Verweis von einem Nichtdeskriptor auf einen Deskriptor
- Scope Note
- Erläuterung eines Begriffs
- Beispiel für die Verwendung eines Nebendeskriptors.

Die Schlagwortliste zur Datenbank SPIRES-HEP findet man im WWW unter der URL http://www-spires.desy.de/library/schlagw2.html .

2.3 Recherche in der Datenbank SPIRES-HEP

Die aktuelle WWW-Suchoberfläche der Literarurdatenbank SPIRES-HEP hat die URL http://www-spires.desy.de/library/cgi-bin/spires.pl . Von der Startseite der DESY-Bibliothek gibt einen Link auf die Suchoberfläche.

Im oberen Teil der Suchoberfläche bietet ein Pull-down-Menü die Möglichkeit, außer HEP noch andere Datenbanken auszuwählen, die im Datenbanksystem SPIRES verfügbar sind. So befindet sich z. B. der OPAC der DESY-Bibliothek in der Datenbank BOOKS. Außerdem gibt es Datenbanken mit Konferenzen, mit Forschungsinsituten und mit Abstracts von Preprints.

Für die Eingabe der Suchfragen steht ein großes Eingabefenster zur Verfügung. Unter dem Eingabefenster gibt es einen Link zu einem Hilfetext über die Suche in SPIRES, von dort wiederum führt ein Link zur SPIRES-HEP-Schlagwortliste.

Um SPIRES-HEP zu benutzen, müssen Befehle in das Eingabefenster eingegeben werden. Die wichtigsten Befehle und ihre Syntax werden im folgenden erläutert:

- `show index`
  Mit diesem Befehl lassen sich die suchbaren Felder der Datenbank anzeigen. Es werden auch alle Abkürzungen eines Feldnamens angezeigt. Beispiele für suchbare Felder sind
  - Autor
  - Institut
  - Kollaboration
  - Titel
  - Deskriptor
  - Veröffentlichung in Zeitschrift
  - Berichtsnummer
  - Archivnummer des Los Alamos-Preprint-Archivs[3]
  - Konferenznummer
  - Literaturhinweis im Literaturverzeichnis des Dokuments.

- `browse` *Feldname Wert*
  Mit dem `browse`-Befehl kann der Index eines suchbaren Felds an der Stelle *Wert* aufgeblättert werden.

- `find` *Feldname Wert*
  Sucht alle Dokumente, die den Ausdruck *Wert* in einem bestimmten Feld enthalten. Mit den Boole'schen Opreatoren AND, OR, (AND) NOT können mehrere Suchen im gleichen oder in verschiedenen Datenbankfeldern miteinander verknüpft werden. Außerdem gibt es die Möglichkeit, Ausdrücke mit dem Operator # rechts zu trunkieren.

Von den genannten Befehlen können jeweils die ersten drei Buchstaben als Kurzform eingegeben werden.

Das Ergebnis einer Recherche in SPIRES-HEP liefert eine Liste der gefundenen Dokumente. Zu jedem Dokument werden die formalen Daten angezeigt.

[3]Das Los Alamos-Preprint-Archiv ist ein vollautomatisches Archiv für elektronische Preprints aus den Fächern Physik und Mathematik (URL: http://xxx.lanl.gov).

Außerdem gibt es Links zu

- Deskriptoren
- Literaturhinweisen
- zitierender Literatur
- Abstract
- Volltext.

Im Zuge der Migration des Datenbanksystems SPIRES auf einen anderen Rechner bei DESY wird zur Zeit eine neue WWW-Suchoberfläche eingeführt. Diese Suchoberfläche bietet unter anderem einen einfachen Suchmodus, bei dem die gesuchten Ausdrücke ohne SPIRES-Befehle eingegeben werden können. Eine Reihe von Pull-down-Menüs erleichtert die Suche zusätzlich. Die Testversion dieser Suchoberfläche findet man unter der URL http://www-library.desy.de/spires/hep .

3 Bestimmung der Suchhäufigkeiten

Voraussetzung für die statistische Auswertung von Suchfragen im Titel- und im Deskriptorenfeld ist die Bestimmung der Häufigkeiten, mit denen Begriffe in diesen Feldern gesucht werden. Ausgangspunkt für die Bestimmung der Suchhäufigkeiten sind Logfiles, in denen die Suchfragen der Benutzer an SPIRES-HEP protokolliert werden.

Die Logfiles werden von dem CGI-Programm[4] erzeugt, das die Schnittstelle zwischen dem Webserver und dem Datenbanksystem SPIRES bildet. Jeder Zugriff auf SPIRES wird in einer Zeile des Files protokolliert. Festgehaltene Daten sind:

- Datenbank in SPIRES, auf die zugegriffen wird
- Name des Hosts, der auf SPIRES zugreift
- Datum und Uhrzeit
- Suchfrage an die Datenbank.

Am Monatsende wird automatisch das Logfile geschlossen und ein neues angelegt.

Zur Bestimmung der Suchhäufigkeiten wurde im Rahmen dieser Arbeit ein Programm entwickelt. Die folgenden Anforderungen werden an das Programm gestellt:

1. Extrahieren aller Begriffe, die im Titel- bzw. im Deskriptorenfeld gesucht werden, aus den Suchfragen
2. Bestimmung der Häufigkeiten der Begriffe
3. Sortieren der Begriffe nach Häufigkeit
4. Bestimmung der Häufigkeitsverteilung, d. h. der Anzahl der Begriffe mit einer bestimmten Häufigkeit
5. Bestimmung des empirischen Mittelwerts der Suchhäufigkeit eines Begriffs.

[4] CGI (Common Gateway Interface) ist ein Standard, der bestimmte Eigenschaften einer Schnittstelle zwischen einem Webserver und einem externen Programm (z. B. Datenbanksystem) festlegt.

Bei Suchfragen im Deskriptorenfeld wird zusätzlich überprüft, ob es sich bei dem gesuchten Begriff um einen Haupt- oder Nebendeskriptor oder um die Präkoordination von zwei Deskriptoren handelt. Präkoordinierte Deskriptoren werden dann getrennt und einzeln gezählt.

Außerdem sollen für jedes Logfile die folgenden Daten ermittelt werden:

- Zahl der Zugriffe auf SPIRES-HEP
- Zahl der Suchfragen, d.h. Zahl der `find`-Befehle
- Zahl der Suchfragen, die eine Suche im Titel- bzw. Deskriptorenfeld enthalten.

Wegen der unterschiedlichen Anforderungen wurde für die Behandlung von Suchfragen im Titel- und im Schlagwortfeld je ein eigenes Programm erstellt. Die Programme wurden in der Interpretersprache awk geschrieben, die auf UNIX-Rechnern verfügbar ist. awk bietet eine Reihe von speziellen Funktionen für die Bearbeitung von Listen und ist damit für die Lösung des hier untersuchten Problems gut geeignet.

4 Statistische Auswertung der Suchhäufigkeiten

Das Ziel der statistischen Auswertung ist es, aus einer Liste von Begriffen diejenigen herauszufinden, nach denen sehr häufig oder sehr selten recherchiert wurde. Dazu benötigt man ein Verfahren, das es ermöglicht, die Suchhäufigkeit eines Begriffs mit dem empirischen Mittelwert der Suchhäufigkeit zu vergleichen. Das statistische Testverfahren, das hier zur Anwendung kommt, ist der Vergleich eines Erwartungswerts mit einem vorgegebenen Wert (siehe DIN 53 804 Teil 2 [4]).

Zunächst sollen die Voraussetzungen für den Test definiert werden (zu den benutzten statistischen Begriffen siehe DIN 55 350 Teil 12, Teil 14, Teil 21–24 [5–10]). Die Einheit, d. h. der Gegenstand der Betrachtung, ist der Begriff. Im Fall der Recherche im Titelfeld ist die Grundgesamtheit die Menge aller Begriffe, die im Untersuchungszeitraum im Titelfeld gesucht wurden. Im Fall der Recherche im Deskriptorenfeld ist die Grundgesamtheit die Menge aller Deskriptoren in der SPIRES-HEP-Schlagwortliste.

Das diskrete Merkmal, durch das sich die Begriffe unterscheiden, ist die Suchhäufigkeit: die Zahl der Recherchen nach einem Begriff. Es wird vorausgesetzt, daß das Verhalten des Merkmals „Suchhäufigkeit“ in der Grundgesamtheit durch die Poisson-Verteilung beschrieben wird. Die Poisson-Verteilung ist eine Wahrscheinlichkeitsverteilung. Sie gibt an, mit welcher Wahrscheinlichkeit eine diskrete Zufallsgröße die Werte 0, 1, 2, ... annimmt. Einziger Parameter der Poisson-Verteilung ist der Erwartungswert $\mu > 0$. Der Erwartungswert gibt an, wo das Zentrum der Wahrscheinlichkeitsverteilung liegt. Er ist unbekannt und kann nur anhand von Stichproben geschätzt werden.

Entnimmt man einen Begriff als Stichprobe aus der Grundgesamtheit, so ist der Zählwert x der Suchhäufigkeit des Begriffs zum einen ein Schätzwert für den Erwartungswert μ. Zum anderen liefert der Zählwert der Suchhäufigkeit einen Schätzbereich für den Erwartungswert: den Vertrauensbereich. Der Vertrauensbereich enthält den Erwartungswert μ der Poisson-Verteilung mit der vorgegebenen Mindestwahrscheinlichkeit (Vertrauensniveau) $1 - \alpha$. Der Vertrauensbereich wird begrenzt durch die untere Vertrauensgrenze μ_{un} und die obere Vertrauensgrenze μ_{ob}. Eine Tabelle der Vertrauensgrenzen μ_{un} und μ_{ob} für verschiedene Zählwerte x auf dem Vertrauensniveau $1 - \alpha = 95\,\%$ findet man in DIN 53 804 Teil 2 [4]. Außerdem sind dort Formeln zur Berechnung der Vertrauensgrenzen angegeben.

Bei einem statistischen Test wird zunächst die zu prüfende Vermutung als Alternativhypothese H_1 formuliert. Die Verneinung der zu prüfenden Vermutung wird als Nullhypothese H_0 formuliert. Der statistische Test besteht darin, mit Hilfe von Stichprobenergebnissen zu entscheiden, ob die Nullhypothese zugunsten der Alternativhypothese verworfen wird oder nicht.

Mit dem hier verwendeten Test, dem Vergleich eines Erwartungswerts mit einem vorgegebenen Wert, wird geprüft, ob der Erwartungswert μ der Suchhäufigkeit vom empirischen Mittelwert K der Suchhäufigkeit abweicht. Die Nullhypothese lautet

$$H_0 : \mu = K \tag{1}$$

Die Alternativhypothese lautet

$$H_1 : \mu \neq K \tag{2}$$

H_0 wird nicht verworfen, wenn K innerhalb des Vertrauensbereichs von μ liegt, d.h. wenn

$$\mu_{\text{un}} \leq K \leq \mu_{\text{ob}} \tag{3}$$

H_0 wird zugunsten von H_1 verworfen, wenn K außerhalb des Vertrauensbereichs von μ liegt, d.h. wenn

$$K < \mu_{\text{un}} \qquad \text{oder} \qquad \mu_{\text{ob}} < K \tag{4}$$

Ein Begriff, zu dessen Suchhäufigkeit x der Vertrauensbereich mit den Grenzen μ_{un} und μ_{ob} gehört, wird also sehr häufig oder sehr selten gesucht, wenn der empirische Mitttelwert K der Suchhäufigkeit außerhalb dieses Bereichs liegt.

5 Ergebnisse

Es wurden Suchfragen untersucht, die die Benutzer während eines Zeitraums von 15 Monaten von April 1998 bis Juni 1999 an SPIRES-HEP gestellt haben. In diesem Zeitraum gab es 30 683 Zugriffe auf SPIRES-HEP auf dem Webserver bei DESY. Die Zahl der Suchfragen war 28 355.

Bei der Untersuchung der Logfiles fielen zwei Dinge auf:

1. Manche Benutzer gaben in einem kurzen Zeitraum dieselbe Suchfrage mehrmals ein.

2. Manche Benutzer verknüpften bei der Suche einen bestimmten Ausdruck nacheinander mit verschiedenen anderen Ausdrücken. Die Suche nach diesem Ausdruck wurde dadurch nur modifiziert.

In beiden Fällen kam es dazu, daß ein Ausdruck bei der Bestimmung der Suchhäufigkeit mehrmals gezählt wurde, obwohl sich nur ein Benutzer zu diesem Zeitpunkt für diesen Ausdruck interessiert hatte. Die so bestimmten Suchhäufigkeiten sind also mit einem Fehler behaftet. Dieser Fehler kann bei Häufigkeiten in der Größenordnung 10, wie sie in dieser Untersuchung vorkommen, nicht vernachlässigt werden.

Der erste Teil dieses Problems wurde mit Hilfe eines Programms gelöst, das identische Wiederholungen einer Suchfrage löscht, die innerhalb eines bestimmten Zeilenbereichs des Logfiles vorkommen. Die zweite Teil des Problems entzog sich einer Lösung durch Programmierung. Es wurden die Stellen in den Logfiles ausfindig gemacht, an denen das Problem auftrat und die Suche im Titel- oder Deskriptorenfeld betraf. Dann wurden die überzähligen Suchfragen von Hand gelöscht. Die so entstandenen korrigierten Logfiles liegen allen Aussagen zugrunde, die über die Suchhäufigkeiten von Begriffen im Titel- und im Deskriptorenfeld gemacht werden.

5.1 Suchfragen im Titelfeld

Während des Untersuchungszeitraums haben die Benutzer 5 868 Suchfragen eingegeben, bei denen nach einem oder mehreren Begriffen im Titelfeld gesucht wurde. Dies entspricht 21 % aller Suchfragen.

Aus den korrigierten Logfiles wurden 5 597 Suchen nach einzelnen Begriffen ermittelt. Dabei wurden 3 384 verschiedene Begriffe gesucht. Man erhält daraus

einen empirischen Mittelwert der Suchhäufigkeit von 1,65. Wendet man das in Abschnitt 4 beschriebene Verfahren an, erhält man als Ergebnis, daß Begriffe mit einer Suchhäufigkeit ≥ 6 besonders häufig gesucht wurden. Es wurde eine Liste dieser Begriffe erstellt, bei der auch die Singular- und Pluralform sowie die Trunkierung eines Begriffs berücksichtigt wurde. Diese Liste enthält 120 Begriffe. Der am häufigsten gesuchte Begriff hat eine Suchhäufigkeit von 60.

Aus der Liste der sehr häufig gesuchten Begriffe wurden die folgenden Kandidaten ermittelt:

- Kandidaten für neue Deskriptoren:

  DESY TESLA Linac

  charge coupled device

  microvertex

- Kandidaten für Nichtdeskriptoren als Zugangsvokabular:

  -emittance USE beam emittance

  -MSSM USE supersymmetry, minimal

  -electron gun USE electron, particle source

  -dijet USE final state, (2jet)

  -NLO USE perturbation theory, higher-order

  -resummation USE perturbation theory, higher-order

- Kandidat für neue Vorzugsbenennung:

  top USED FOR -truth.

Es wurden nur solche Begriffe als Kandidaten genommen, die auch bei den in SPIRES-HEP nachgewiesenen Dokumenten häufig im Titel vorkommen.

5.2 Suchfragen im Deskriptorenfeld

Die Benutzer haben während des Untersuchungszeitraums 4 623 Suchfragen an SPIRES-HEP gestellt, bei denen nach einem oder mehreren Begriffen im Deskriptorenfeld gesucht wurde. Dies entspricht 16 % aller Suchfragen. 45 % der Begriffe, nach denen im Deskriptorenfeld gesucht wurde, waren Deskriptoren aus der SPIRES-HEP-Schlagwortliste. Dies bedeutet, daß nur 45 % der Suchen nach einzelnen Begriffen erfolgreich sein konnten. Betrachtet man die Suchen nach Deskriptoren, so findet man, daß in 17 % dieser Suchen präkoordinierte Deskriptoren verwendet wurden.

Aus den korrigierten Logfiles wurden 1931 Suchen nach einzelnen Deskriptoren ermittelt. Präkoordinierte Deskriptoren wurden dabei getrennt. Die Suchen verteilen sich auf die 2140 Deskriptoren der Schlagwortliste. Man erhält daraus einen empirischen Mittelwert der Suchhäufigkeit von 0,902.

Die Anwendung des statistischen Verfahrens aus Abschnitt 4 ergibt, daß Deskriptoren mit einer Suchhäufigkeit ≥ 4 besonders häufig gesucht wurden. Diese Menge umfaßt 155 Deskriptoren. Der am häufigsten gesuchte Deskriptor hat eine Suchhäufigkeit von 34.

Unter den sehr häufig gesuchten Deskriptoren sind die folgenden 6 Deskriptoren mit allgemeiner Bedeutung:

RF system, programming, statistics, optics, vacuum system, electronics.

Bei diesen Deskriptoren ist zu überlegen, ob sie näher spezifiziert werden sollten.

Deskriptoren, die besonders selten gesucht wurden, konnten bei der Anwendung des statistischen Tests nicht gefunden werden. Der Grund: zur Suchhäufigkeit $x = 0$ gehört der Vertrauensbereich für den Erwartungswert μ mit den Grenzen $\mu_{\text{un}} = 0$ und $\mu_{\text{ob}} = 3,69$ (Vertrauensniveau 95 %). Der empirische Mittelwert der Suchhäufigkeit $K = 0,902$ liegt innerhalb dieses Vertrauensbereichs. Die Nullhypothese, daß der Erwartungwert mit dem empirischen Mittelwert der Suchhäufigkeit übereinstimmt, kann nicht verworfen werden. Mit anderen Worten: es kann nicht entschieden werden, ob ein Deskriptor mit der Suchhäufigkeit Null besonders selten gesucht wurde.

5.3 Vergleich der Recherche nach Deskriptoren mit der Indexierung

Es wurden die Indexate der Preprints ausgewertet, die in der Zeit von April 1998 bis Juni 1999 in die Datenbank SPIRES-HEP eingegeben wurden. Das Ziel war, die Indexierhäufigkeit der einzelnen Deskriptoren zu bestimmen und die besonders häufig bei der Indexierung verwendeten Deskriptoren mit den besonders häufig recherchierten Deskriptoren zu vergleichen.

Der Anteil der Preprints an allen Dokumenten, die in diesem Zeitraum für die Datenbank indexiert wurden, beträgt 71 %. Die Zuordnung der Klassifikationscodes zu den Preprints und zur Gesamtheit der im Untersuchungszeitraum indexierten Dokumente wurde verglichen. Es zeigt sich, daß die relativen

Häufigkeiten der einzelnen Codes bei den Preprints und bei der Gesamtheit der Dokumente gut übereinstimmen. Aus diesem Grund kann man davon ausgehen, daß die Menge der Preprints eine repräsentative Stichprobe aus der Menge aller indexierten Dokumente ist.

Im Untersuchungszeitraum wurden 14 038 Preprints indexiert. Dabei wurden 135 609 Begriffe vergeben, präkoordinierte Begriffe mitgerechnet. Dies entspricht einer durchschnittlichen Zahl von 10 Begriffen pro Dokument. Es wurden 3 774 einzelne Begriffe vergeben, darunter 1 774 Deskriptoren aus der SPIRES-HEP-Schlagwortliste. Die restlichen Begriffe sind ergänzende Angaben wie z. B.

- Orts- und Zeitangaben von Konferenzen
- Kombinationen mehrerer Teilchen
- Produkte von Symmetriegruppen
- Zahlenwerte.

Die Summe der Indexierhäufigkeiten der einzelnen Deskriptoren ist 214 119. Die Zahl der Deskriptoren in der Schlagwortliste ist 2 140. Daraus ergibt sich ein empirischer Mittelwert der Indexierhäufigkeit von 100.

Auch hier wird das statistische Verfahren aus Abschnitt 4 angewendet, um die sehr häufig und die sehr selten bei der Indexierung verwendeten Deskriptoren zu ermitteln. Besonders häufig benutzt wurden Deskriptoren mit einer Indexierhäufigkeit ≥ 121 (391 Deskriptoren). Besonders selten benutzt wurden Deskriptoren mit einer Indexierhäufigkeit ≤ 80 (1 605 Deskriptoren). An der Spitze der Häufigkeitsliste stehen zwei Deskriptoren, die ein Dokument formal beschreiben: „numerical calculations“ (5 647) und „talk“ (3 621). An dritter Stelle folgt der erste Deskriptor, der ein Dokument inhaltlich beschreibt: „quark“ mit einer Indexierhäufigkeit von 3 234.

Für den Vergleich von Recherche und Indexierung wurden jeweils die 100 am häufigsten recherchierten bzw. bei der Indexierung vergebenen Deskriptoren ausgewählt. Um einen Überblick darüber zu bekommen, in welchen Sachgebieten die Schwerpunkte bei der Recherche und bei der Indexierung liegen, wurden Deskriptoren, sofern dies möglich war, eindeutig einem von vier Sachgebieten zugeordnet.

Diese Sachgebiete sind:

- Feldtheorie
- Teilchen und ihre Eigenschaften
- Wechselwirkungen zwischen Teilchen
- Instrumente und Methoden.

Die Deskriptoren jedes Sachgebiets wurden dann für Recherche und Indexierung gezählt.

Der Vergleich ergibt, daß die beiden Listen der 100 am häufigsten recherchierten bzw. bei der Indexierung verwendeten Deskriptoren 25 Deskriptoren gemeinsam haben. Ein Vergleich der bei Recherche und Indexierung abgedeckten Sachgebiete ergibt das folgende Bild:

	Recherche	Indexierung
Feldtheorie	12	24
Teilchen und ihre Eigenschaften	22	26
Wechselwirkungen zwischen Teilchen	11	18
Instrumente und Methoden	28	4

Es fällt auf, daß im Sachgebiet „Instrumente und Methoden“ der Unterschied zwischen Recherche und Indexierung groß ist: vielen recherchierten Deskriptoren stehen wenige bei der Indexierung benutzte Deskriptoren gegenüber. Umgekehrt verhält es sich bei den Sachgebieten „Feldtheorie“ und „Wechselwirkungen“: hier wurden deutlich mehr Deskriptoren bei der Indexierung benutzt, als von den Benutzern gesucht wurden.

6 Schlußfolgerungen

In diesem Abschnitt sollen die gewonnenen Ergebnisse bewertet und Folgerungen für die Praxis gezogen werden.

Bei der Auswertung von Suchfragen im Titelfeld konnten aus einer Liste von 120 sehr häufig gesuchten Begriffen 10 ernstzunehmende Kandidaten für neue Deskriptoren, Nichtdeskriptoren oder Vorzugsbenennungen gewonnen werden. Angesichts der geringen Suchhäufigkeiten bei dieser Untersuchung ist dieses Ergebnis als positiv zu bewerten. Die Auswertung einer größeren Zahl von Suchfragen könnte weitere Kandidaten liefern.

Das Ziel, besonders selten gesuchte Deskriptoren zu ermitteln, konnte nicht erreicht werden. Der Grund dafür ist die geringe Zahl von Suchfragen im Deskriptorenfeld während des Untersuchungszeitraums und der daraus resultierende niedrige empirische Mittelwert der Suchhäufigkeit. Um mit dem hier verwendeten statistischen Verfahren besonders selten gesuchte Deskriptoren zu finden, muß der Mittelwert der Suchhäufigkeit größer als 3,69 sein (siehe [4]). Es müßte eine dementsprechend größere Zahl von Suchfragen ausgewertet werden.

Unter den 155 besonders häufig recherchierten Deskriptoren sind 6 Deskriptoren, die eine sehr allgemeine Bedeutung haben und die näher spezifiziert werden können. Um einen Deskriptor zu spezifizieren, bieten sich zwei Möglichkeiten an. Zum einen kann man dem Deskriptor Unterbegriffe zuordnen. Zum anderen bietet die SPIRES-HEP-Schlagwortliste die Möglichkeit, ihn mit einem ergänzenden Deskriptor zu präkoordinieren. Die Präkoordination wird zwar bei der Indexierung sehr oft angewendet, bei der Suche jedoch werden präkoordinierte Deskriptoren kaum benutzt. Es gibt hier also ein Verbesserungspotential, das zuerst ausgenutzt werden sollte, bevor man sich an Änderungen in der Schlagwortliste macht. Zu diesem Zweck müßte man die Benutzer besser über die Suchmöglichkeiten mit präkoordinierten Deskriptoren informieren.

Die sehr häufig gesuchten Deskriptoren wurden mit den sehr häufig bei der Indexierung verwendeten Deskriptoren verglichen. Dabei wurde festgestellt, daß bei drei von vier Sachgebieten die Zahl der gesuchten Deskriptoren von der Zahl der bei der Indexierung verwendeten Deskriptoren deutlich abweicht. Dieses Ergebnis weist auf einen Unterschied zwischen den Interessenschwerpunkten der Benutzer und dem Angebot an neuen Dokumenten in der Datenbank hin. Um genauere Ergebnisse zu erhalten, müßte noch ein Problem bei der Bestim-

mung der Indexierhäufigkeiten gelöst werden: es kommt oft vor, daß innerhalb eines Indexats ein bestimmter Deskriptor mit mehreren anderen Deskriptoren präkoordiniert ist. Mit dem hier verwendeten Verfahren zur Bestimmung der Indexierhäufigkeiten wird dieser Deskriptor mehrfach gezählt. Die so bestimmte Indexierhäufigkeit ist also mit einem Fehler behaftet. Dieses Problem ist durch Programmierung lösbar.

Über die Pflege der Schlagwortliste hinaus wurden in dieser Arbeit auch Erkenntnisse über das Rechercheverhalten der Benutzer gewonnen. Eine wichtige Erkenntnis besteht darin, daß nur knapp die Hälfte der Suchen im Deskriptorenfeld potentiell erfolgreich war. Eine bessere Unterstützung der Benutzer bei der Recherche ist daher notwendig. Erster Schritte in diese Richtung könnten ein deutlicher Hinweis auf die SPIRES-HEP-Schlagwortliste und eine benutzerfreundlichere Hilfe für die Deskriptorensuche sein.

Die in dieser Arbeit gewonnenen Erkenntnisse beziehen sich nur auf die Recherche in SPIRES-HEP auf dem Webserver bei DESY. Der Großteil der Benutzer aus aller Welt verwendet zur Recherche in SPIRES-HEP den Webserver bei SLAC. Ein Zahlenvergleich verdeutlicht dies: 15 000 Zugriffen auf den SLAC-Server pro Tag stehen 100 Zugriffe auf den DESY-Server gegenüber. Um sich ein umfassendes Bild von den Suchfragen an SPIRES-HEP zu machen, müßten die Logfiles des SLAC-Servers in die Auswertung einbezogen werden.

7 Zusammenfassung

In dieser Arbeit wurde untersucht, wie die statistische Auswertung von Suchfragen zur Pflege einer Schlagwortliste eingesetzt werden kann. Es wurden Suchfragen der Benutzer an die Literaturdatenbank SPIRES-HEP auf dem Webserver bei DESY ausgewertet. Die Suchhäufigkeit von Begriffen im Titel- und im Deskriptorenfeld wurde bestimmt. Aus den sehr häufig im Titelfeld recherchierten Begriffen konnten Kandidaten für neue Deskriptoren, Nichtdeskriptoren und Vorzugsbenennungen ermittelt werden. Das Ziel, besonders selten recherchierte Deskriptoren zu ermitteln, konnte nicht erreicht werden, weil die Zahl der Suchfragen während des Untersuchungszeitraums zu gering war. Es wurden besonders häufig gesuchte Deskriptoren mit sehr allgemeiner Bedeutung ermittelt, die näher spezifiziert können. Dabei ist zu prüfen, inwieweit diese Spezifizierung durch Präkoordination mit ergänzenden Deskriptoren erreicht werden kann. Die sehr häufig gesuchten Deskriptoren wurden mit den sehr häufig bei der Indexierung verwendeten Deskriptoren verglichen. Es ergab sich ein Unterschied zwischen den Interessenschwerpunkten der Benutzer und dem Angebot an neuen Dokumenten in der Datenbank. Die Auswertung der Suchfragen zeigte, daß nur knapp die Hälfte der Suchen nach Deskriptoren potentiell erfolgreich war. Um den Erfolg bei der Suche nach Deskriptoren zu erhöhen, sollte den Benutzern mehr Hilfe angeboten werden. Um die statistische Auswertung von Suchfragen effektiv für die Pflege der SPIRES-HEP-Schlagwortliste einzusetzen, müßte eine größere und repräsentative Menge von Suchfragen untersucht werden.

Literatur

[1] Gaus, Wilhelm: *Dokumentations- und Ordnungslehre*. Berlin: Springer, 1995, 2. Auflage.

[2] Ladewig, Christa: *Grundlagen der inhaltlichen Erschließung*. Potsdam: Institut für Information und Dokumentation (Hrsg.), 1997.

[3] Buder, Marianne (Hrsg.); Rehfeld, Werner, (Hrsg.); Seeger, Thomas (Hrsg.); Strauch, Dietmar (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. München, Saur, 1997, 4. Auflage.

[4] DIN 53804-2: Statistische Auswertungen: Zählbare (diskrete) Merkmale. März 1985.

[5] DIN 55350-12: Begriffe der Qualitätssicherung und Statistik: Merkmalsbezogene Begriffe. März 1989.

[6] DIN 55350-14: Begriffe der Qualitätssicherung und Statistik: Begriffe der Probennahme. Dezember 1985.

[7] DIN 55350-21: Begriffe der Qualitätssicherung und Statistik: Begriffe der Statistik, Zufallsgrößen und Wahrscheinlichkeitsverteilungen. Mai 1982.

[8] DIN 55350-22: Begriffe der Qualitätssicherung und Statistik: Begriffe der Statistik, Spezielle Wahrscheinlichkeitsverteilungen. Februar 1987.

[9] DIN 55350-23: Begriffe der Qualitätssicherung und Statistik: Begriffe der Statistik, Beschreibende Statistik. April 1983.

[10] DIN 55350-24: Begriffe der Qualitätssicherung und Statistik: Begriffe der Statistik, Schließende Statistik. April 1983.

Danksagung

Diese Arbeit wurde in der Gruppe „Bibliothek und Dokumentation“ am Deutschen Elektronen-Synchrotron DESY in Hamburg angefertigt.

Ich möchte den Mitarbeiterinnen und Mitarbeitern dieser Gruppe für ihre wertvollen Anregungen und ihre Unterstützung bei meiner Arbeit danken.

Ich versichere, daß ich diese Arbeit selbständig angefertigt habe und dabei ausschließlich die angegebenen Quellen verwendet habe.

Hamburg, 1. September 1999

Klaus Ebner