HIP: Intelligente Suche nach Fachinformationen für das Handwerk

Sergej Sizov*, Klaus Meier**, Gerhard Weikum*
*Universität des Saarlandes , **Handwerkskammer des Saarlandes
{*sizov,weikum*}*@cs.uni-sb.de, k.meier@hwk-saarland.de*

Abstract: Angesichts des exponentiell wachsenden Informationsangebots im World Wide Web hat sich die Suche nach relevanten Ressourcen und Datenquellen mit der Zeit zu einem eigenständigen Problem entwickelt. Allgemeine Web-Suchmaschinen verwenden für die Erstellung der Rangliste der Treffer Autoritätswerte, die durch Linkanalyseverfahren auf repräsentativen Web-Ausschnitten bestimmt werden (ggf. kombiniert mit textbasierter Dokument-Query Ähnlichkeit). Diese Vorgehensweise scheitert jedoch oft bei sehr spezifischen fachlichen Anfragen mit insgesamt kleinem Recall. Darüber hinaus bleiben zahlreiche 'Hidden Web' - Informationsquellen (z.B. die Datenbanken der Informationsportale) für konventionelle Crawler nicht zugänglich.

Die Web-Suchmaschine des Projektes HIP (**H**andwerks-**I**nformations-**P**ortal), eines Kooperationsprojektes der Universität des Saarlandes, der saarländischen Handwerkskammer und der saarbrücker Hochschule für Technik und Wirtschaft) kombiniert die Vorteile eines fokussierten Crawlers mit automatischer Erweiterung der Trainingsbasis, eines Frameworks für automatisch erkannte, klassifizierte und als Web Services gekapselte 'Hidden Web'-Informationsquellen sowie einer Suchmaschine mit erweiterten Ranking-Möglichkeiten für Web-Expertensuche. Dieses Papier beschreibt die Architektur des HIP-Frameworks, einzelne Komponenten des Suchsystems sowie die ersten Ergebnisse der Evaluation des Prototyps.

1 Motivation

Die effektive und effiziente Suche nach Informationen im Web gehört heute zu den wichtigsten Problemen der Informatik. Konventionelle Suchmaschinen liefern typischerweise viele Treffer mit einer gewissen Relevanz für das gesuchte Thema, allerdings muss der Benutzer oft durch die Nachbarschaft dieser Ergebnisse manuell navigieren, um die gewünschte Seite zu entdecken. Oft können die besten Ergebnisse vielmehr über thematische Portale erreicht werden, die den Zugriff auf ihre Verzeichnisse in Form einer intellektuell aufgebauten Ontologie bieten. Allerdings ist die Erstellung und die Wartung einer solchen Ontologie sehr aufwendig und kostspielig, da enorme Datenmengen manuell gefiltert und klassifiziert werden müssen.

In dieser Arbeit wird die automatische Generierung eines Portals mit Informationen für Handwerksbetriebe beschrieben, das dem Benutzer gegenüber konventionellen Suchmaschinen deutlich bessere Suchmöglichkeiten bietet. Die Erstellung eines solchen Informationsportals stößt auf folgende problemspezifische Schwierigkeiten:

- Ein Handwerks-Portal vereinigt Web-Inhalte sehr unterschiedlicher Natur (Gesetzessammlungen, Förderprogramme, regionale Handwerksorganisationen, Homepages ein-

zelner Betriebe), wobei die gegenseitige Verlinkung einzelner Themenbereiche (die Grundlage für ein Autoritäts-Ranking der Suchergebnisse) deutlich schwächer ist als z.B. in den Branchen PC-Hardware, Städtereisen oder Informationswissenschaften;
- Die große Vielfalt der Themen macht es für einzelne Portaladministratoren schwierig, die gesamte Ontologie manuell zu pflegen. Darüber hinaus ist es oft nicht einfach, zu jedem Thema hinreichend viele Trainingsdokumente für präzise automatische Klassifikation zu bestimmen;
- Viele inhaltlich wertvolle Informationsquellen (Online-Gesetzessammlungen, Kataloge der Fördermaßnahmen, etc.) sind in den Datenbanken entsprechender Informationssysteme gespeichert, haben keine expliziten Link-Referenzen und sind somit für klassische Web-Crawler unzugänglich;
- Die Zielgruppen des Handwerks-Portals haben potentiell sehr unterschiedliche Interessen (z.B. Lehrlinge vs. Meister). Die Anpasung des Portals an die Informationsbedürfnisse der Zielgruppen erfordert parallele Führung von unterschiedlichen Themenhierarchien;
- Die Suchmaschine soll hohe thematische Präzision bei vergleichsweise einfacher Bedienung ermöglichen.

Das **H**andwerks-**I**nformations-**P**ortal, kurz: HIP, entsteht im Rahmen eines Pilotprojektes in Kooperation zwischen der Universität des Saarlandes, der saarländischen Handwerkskammer und der saarbrücker Hochschule für Technik und Wirtschaft. Hauptziel des HIP-Projektes ist es, intelligentes Suchen nach Fachinformationen für das deutsche Handwerk im Web zu ermöglichen.

Der Rest der Arbeit ist wie folgt organisiert. In Abschnitt 2 werden Aspekte und Probleme der Gestaltung eines Handwerks-Informationsportals besprochen. Abschnitt 3 beschäftigt sich mit der Architektur des HIP-Prototyps und beschreibt die wichtigsten HIP-Komponenten. In Abschnitt 4 werden einige praktische Aspekte der Implementierung erläutert.

2 Gestaltung des Portals

Maßgeschneidert für die Bedürfnisse des Handwerks bietet die HIP-Suchmaschine 3 unabhängige Ansichten der Portalinformationen. Jede Ansicht enthält z.Zt. eine Themenhierarchie aus 3 Ebenen mit 5 bis 12 Hauptkategorien.

- *Berufe des Handwerks*. Diese Ansicht enthält allgemeine Informationen zu den Berufsgruppen des Handwerks sowie technische Fachinformationen (technologische und technische Informationen, Normen und Gesetze, regionale Innungen, Homepages einzelner Betriebe) zu 150 Handwerksberufen, aufgeteilt in 9 thematische Kategorien;
- *Typische Abläufe und Probleme eines Handwerk-Betriebs*. Diese Ansicht beschäftigt sich mit den typischen Routineabläufen innerhalb eines Handwerksbetriebs (Arbeitsschutz, Buchführung, Personalwesen, Existenzgründung, Fördermaßnahmen..) sowie allgemeinen handwerksrelevanten Themen wie Denkmalpflege, RAL-Gütezeichen, Umweltschutz und thematischen Gesetzessammlungen;
- *Ausbildung und Karriere im Handwerk*. Die Zielgruppe dieser Ansicht sind primär Schüler und Auszubildende. Die entsprechende Themenhierarchie enthält allgemeine Informationen zu den Karrieremöglichkeiten im deutschen Handwerk, Praktikanten-

und Lehrstellenbörsen, Musterverträgen, Bewerbungen, etc.

Die Suche nach relevanten Informationen erfolgt entweder durch Navigation im gewählten Themenbaum oder durch keyword-basierte Abfragen. Die Suchmaschine unterstützt die Sortierung der Ergebnis-Rangliste nach unterschiedlichen Kriterien: Klassifikationsgüte, Autorität der Webdokumente (HITS-Linkanalyse [Kle99]), Cosinus- Ähnlichkeit, Bewertung durch andere Besucher, Häufigkeit der bisherigen Zugriffe, letzte Aktualisierung der Datei. Die letzte Option ist insbesondere nützlich bei der Suche nach neuen Informationen und aktualisierten Inhalten. Die Suche kann nach einer Ergänzung oder Änderung der Query auf bestehender Ergebnismenge (statt Neuausführung) fortgesetzt werden. Zu jedem angezeigten Treffer besteht zusätzlich die Möglichkeit zur Fortsetzung der Suche auf:

- seiner *Nachbarschaft* (alle bekannten Vorgänger und Nachfolger, die vom gegebenen Dokument referenziert werden bzw. Referenzen auf ihn enthalten);
- seinem *Ursprung* (alle erfassten Dokumente vom gleichen Webserver);
- ähnlichen Dokumenten.

Für die Verbesserung der Qualität des Datenbestandes sieht das System Schnittstellen für Benutzer-Feedback vor: Vorschlagen der neuen Webseiten und thematischen Kategorien, Melden der Klassifikationsfehler, Bewertung der angezeigten (und besuchten) Ergebnisse.

3 Systemarchitektur

Die wichtigsten Komponenten der Suchmaschine sind:

- ein fokussierter Crawler für die gezielte thematische Suche nach thematisch relevanten Informationen im Web;
- ein Dokument-Analysator für das Parsing der gefundenen Dokumente und die Generierung von Feature-Vektoren;
- ein hierarchischer Klassifikator mit entsprechenden Trainingsdaten;
- ein Linkanalyse-Modul für das Autoritäts-Ranking der Suchergebnisse;
- ein Clustering-Modul für die automatische Generierung und Annotation von neuen Kategorievorschlägen;
- ein Web-Interface für die Steuerung und Administration des Systems;
- eine Suchmaschine für den gesammelten Datenbestand.

Vorbereitung der Daten. Das System wird initialisiert durch Einlesen einer hierarchisch organisierten Ausgangsmenge von inhaltlich hochwertigen Beispieldokumenten (z.B. in Form einer Bookmark-Datei). Der Dokument-Analysator, ein erweiterter Parser für HTML, PDF und andere gängigen Dateitypen, verarbeitet die Trainingsdokumente. Stoppwörter werden anhand vordefinierter Listen entfernt und die verbleibenden Terme werden durch Stemming [Sno] auf die morphologischen Stammformen reduziert. Termbasierte Gewichtsvektoren mit RTF oder TF*IDF Gewichtung [BYRN99] repräsentieren die Dokumente und dienen als direkte Eingabe für den Klassifikator. Der Analysator generiert Vektoren aus allen Eingabedokumenten unter Berücksichtigung des Feature-Raumes des jeweiligen Hierarchieknotens.

Feature Selektion. Als Feature-Selektionskriterium hat sich die Mutual Information (MI, bekannt auch ale relative Entropie) [BYRN99, YP97] bewährt. MI stellt ein statistisches Maß für die Diskriminationsgüte im Hinblick auf die verfügbaren "konkurrierenden"Klassen einer Baumstufe der Ontologie dar. Als Entropiemaß lässt sich MI zur Generierung der Feature-Räume sowohl von Blattklassen als auch von inneren Knoten des Hierarchiebaums verwenden.

Der Crawler. Das HIP-Framework verwendet den fokussierten Crawler BINGO! [SBG+03] [SSTW02] zum Sammeln der themenspezifischen Informationen im Web. Die geladenen Dokumente werden an den Klassifikator für die Zuordnung in die spezifizierte Themenhierarchie weitergegeben. Die Links aus positiv klassifizierten Dokumenten werden in die URL-Queue des Crawlers übernommen und für die Fortsetzung der Suche verwendet. Bei positiv klassifizierten Dokumenten dient die Klassifikationsgüte als Maß für die Priorisierung der URLs in der Queue; die Links der zurückgewiesenen Dokumente werden gar nicht bzw. nur bis zur limitierten Tiefe mit minimaler Priorität weiterverfolgt (Tunneling).

Der Klassifikator. Die hierarchische Klassifikation der neuen Dokumente erfolgt mit Hilfe der linearen Support Vector Machines (SVM) [Vap98, Bur98, Joa02]. Zur Klassifikation von Dokumenten in eine mehrstufige Ontologie wird der Hierarchiebaum rekursiv traversiert, wobei in jedem inneren Knoten eine binäre SVM-Klassifikation über die Wahl des Nachfolgers entscheidet. Der Abstand von der trennenden Hyperebene dient als Maß für die Klassifikationskonfidenz. Die Menge SVM_A der Dokumente jeder Klasse mit der höchsten SVM-Konfidenz bildet die erste Gruppe potentieller themenspezifischer *Archetypen* für eine Erweiterung der initialen Trainingsbasis.

Autoritätsranking der Suchergebnisse. Die Analyse der Linkstruktur zwischen den Dokumenten einer Klasse und der gesamten Themenhierarchie liefert einen weiteren Anhaltspunkt, wie gut einzelne Suchergebnisse das Thema reflektieren. HIP verwendet das HITS-Verfahren [Kle99] auf dem Hyperlink-Graphen der Dokumente aller Ontologieklassen an. Die resultierende Gewichtung (Hub- und Authority-Scores der Dokumente) wird beim Ranking der Suchergebnisse verwendet.

Ausserdem dient diese Methode der Identifikation einer Menge $HITS_A$ bester Authorities, die besonders wertvolle Dokumente mit den besten Hyperlink-Referenzen des vom Crawler erzeugten Web-Ausschnitts enthält - eine zweite potentielle Quelle für neue *Archetypen* als Ergänzung der Trainingsbasis.

Retraining des Klassifikators. Die Themenpalette des HIP-Portals ist sehr breit: allein in Deutschland gibt es weit über 100 handwerkliche Berufe mit den dazugehörigen Technologien, rechtlicher Basis, Ausbildungsmöglichkeiten, etc. Auch für Experten ist es schwierig, den vollen Überblick über alle Themenbereiche zu behalten und die gesamte Ontologie manuell zu pflegen. Darüber hinaus ist es oft schwer, zu jedem Topic hinreichend viele Trainingsdokumente für präzise automatische Klassifikation zu finden.

Die HIP-Suchmaschine versucht die engen Grenzen einer Trainingsbasis mit wenigen intellektuell kategorisierten Dokumenten zu überwinden und in einer automatisierten Lern- und Wachstumsphase selbständig eine breitere Trainingsbasis durch die Identifikation the-

	Laufzeit 90 Min	Laufzeit 12 Std
Besuchte URLs	130,726	8,221,399
Besuchte Hosts	4,254	18,106
Links extrahiert	1,434,228	82,224,710
Positiv klassifiziert	12,711	216,091
Max Crawlingtiefe	9	98

Abbildung 1: Websuche des HIP-Crawlers

menspezifischer *Archetypen* zu generieren [SBG$^+$03]. Wenn eine Klasse der Ontologie einen gewissen Füllstatus von N_{max} Dokumenten erreicht hat, kann ein Re-Training des Klassifikators automatisch durchgeführt werden. Um das ungewünschte Phänomen des Themen-Drifts zu vermeiden, werden die besten Dokumente aus $HITS_A \cap SVM_A$ aufgenommen. Die anschließende Wachstumsphase vervollständigt die Ontologie mit verbesserter Präzision.

Automatische Erweiterung der Ontologie. Ein wichtiges Ziel des HIP-Frameworks besteht darin, die Pflege der Ontologie mit minimalem menschlichen Aufwand zu ermöglichen. Das System bietet die Möglichkeit zur automatischen Erweiterung durch Clustering der 'OTHERS'-Klassen der Themenhierarchie. Die aktuelle Implementierung benutzt die Projektion der Matrix der Dokumentvektoren in den Raum der potentiellen Themen durch Singular Value Decomposition (SVD) [MS99, DHS01, Cha02] und anschliessender Anwendung des klassischen *K-Means* Verfahrens [MS99, DHS01] auf resultierende neue Feature-Vektoren der Dokumente. Die Annotation der Klassenvorschläge erfolgt durch die charakteristischen Terme jedes Clusters (nach Mutual Information, see Abschnitt 3) und ein repräsentatives Dokument mit der größten Autorität aus dem HITS-Verfahren [Kle99].

4 Implementierung

Die aktuelle Implementierung des HIP-Frameworks läuft auf einem Intel 3.0 GHz Server mit 4 GB Hauptspeicher unter Windows 2000 Server OS, in Verbinding mit einer MySQL-Datenbank auf dem gleichen Rechner. Jede Kategorie der Themenhierarchie wird initialisiert durch 10-15 intellektuell klassifizierte Trainingsdokumente: Anleitungen und Dokumentationen für Handwerker, Texte von Gesetzen und Verordnungen, sowie themenspezifische Homepages einzelner Handwerksbetriebe. In der Lernphase wird die Suche auf die Hosts der Startdokumente und maximale Tiefe 3 beschränkt.

Nach dem Re-Training mit automatischer Selektion der zusätzlichen Archetypes wird der fokussierte Crawl ohne Restriktionen fortgesetzt. Dabei werden pro Lauf innerhalb von 12 Stunden ca. 10 million Webseiten von mehreren Tausend Hosts ausgewertet, bis zu 10.000 positiv klassifizierte Dokumente den Themen der Hierarchien zugeordnet (Abbildung 1) und anschließend in der MySQL-Datenbank gespeichert.

Die Suchmaschine des HIP-Portals besteht aus einer Sammlung der Java Servlets und PHP-Seiten. Abbildung 2 zeigt am praktischen (modellierten) Beispiel das Verhalten der Suchmaschine bei 1, 5 oder 10 parallel auszuführenden Anfragen mit unterschiedlicher Selektivität und variabler Anzahl der Suchbegriffe.

Anfragen parallel	Anz. Such-begriffe	Anz. Treffer	Antwortzeit Max (ms)	Antwortzeit Min (ms)	Antwortzeit Avg (ms)
1	1	15508	15250	47	170
1	3	22492	16984	78	330
5	1	16185	110547	78	738
5	3	10798	248984	78	1527
5	6	19876	303891	219	4575
10	1	14053	398781	93	1319
10	3	11924	290718	422	3886

Abbildung 2: Lastprofile der HIP-Suchmaschine

5 Zusammenfassung und Ausblick

Die Technologien des HIP-Frameworks ermöglichen die automatische Generierung eines thematischen Informationsportals mit problemspezifischen und präzisen Suchmöglichkeiten. Durch automatisches Re-Training ist die Einrichtung und Pflege des Datenbestandes mit reduziertem menschlichen Aufwand möglich. In Zukunft planen wir den Ausbau des HIP-Frameworks (größere Crawling-Sessions, Skalierbarkeit) sowie weitere Experimente zur Präzision der Expertensuche (insbesondere adaptive, personalisierte Auto-Ranking Schemas) sowie zu alternativen Verfahren für Clustering und Generierung von Kategorienvorschlägen. Die erste betriebsreife Version des im Rahmen des HIP-Frameworks generierten Handwerks-Portals soll Ende 2003 für öffentliche Benutzung freigegeben werden.

Literatur

[Bur98] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), 1998.

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[Cha02] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, 2002.

[DHS01] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, New York, 2001.

[Joa02] T. Joachims. *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.

[Kle99] J.M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5), 1999.

[MS99] C.D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[SBG+03] S. Sizov, M. Biwer, J. Graupmann, S. Siersdorfer, M. Theobald, G. Weikum, and P. Zimmer. The BINGO! System for Information Portal Generation and Expert Web Search. *The 1st Semiannual Conference on Innovative Data Systems Research (CIDR), Asilomar(CA)*, 2003.

[Sno] Snowball. The string processing language for stemming algorithms. *http://snowball.tartarus.org/*.

[SSTW02] S. Sizov, S. Siersdorfer, M. Theobald, and G. Weikum. The BINGO! focused crawler: From Bookmarks to Archetypes. *IEEE Computer Society International Conference on Data Engineering (ICDE), San Jose, California*, 2002.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[YP97] Y. Yang and O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. *International Conference on Machine Learning (ICML)*, 1997.