

Ariadne: ein fokussierter Web-Crawler mit adaptiver Klassifikation der Hyperlinks

Martin Ester, Matthias Groß
Institut für Informatik, Ludwig-Maximilians-Universität München,
Oettingenstr. 67, D-80538 München
{ester|gross}@dbs.informatik.uni-muenchen.de

Extended Abstract

1. Einleitung

Das World-Wide Web wächst so rasant, daß die Suchmaschinen nur einen relativ kleinen Teil aller Webseiten indizieren können. Aus diesem Grund ist der *Recall* (der Anteil der für eine Anfrage gelieferten relevanten Webseiten an der Menge aller relevanten Webseiten) der Suchmaschinen relativ gering. Zusätzlich sind im allgemeinen viele Antworten veraltet oder URLs nicht mehr gültig. Oft ist außerdem die *Precision* (Anteil der relevanten Webseiten an der Menge aller gelieferten Antworten) der Suchmaschinen schlecht, d.h. neben den relevanten Antworten wird ein Vielfaches an irrelevanten Webseiten geliefert. Fokussierte Crawler sind eine Alternative zu den konventionellen Suchmaschinen, die diese Schwächen vermeiden.

Ein *fokussierter Crawler* erhält eine Menge von Webseiten (*Trainingsseiten*), die ein für den Benutzer interessantes Thema spezifizieren. Auf der Suche nach weiteren relevanten Seiten geht der Crawler von den Trainingsseiten aus und besucht eine geeignet eingeschränkte Teilmenge der durch Hyperlinks direkt oder indirekt verbundenen Webseiten. Dieser Ansatz basiert auf der Beobachtung, daß durch einen Hyperlink miteinander verbundene Seiten sehr viel wahrscheinlicher das gleiche Thema behandeln als zwei zufällig gewählte Seiten. Zentrale Aufgabe beim fokussierten Crawling ist eine Anordnung der Hyperlinks nach absteigender "Qualität", so daß die besten Links jeweils zuerst verfolgt werden können.

In [2] wird gezeigt, daß bereits mit einer einfachen Bewertung des Texts der Quellseite und des Ankertexts des Hyperlinks wesentlich bessere Ergebnisse erzielt werden als bei zufälliger Wahl der Priorität der Links. In [1] wird zur Ordnung der Links sowohl die Relevanz der Quellseite als auch ihre "Zentralität" bestimmt. Die *Zentralität* einer Webseite wird rekur-

siv durch die Zentralität der benachbarten Seiten definiert. Die Berechnung erfolgt mit Hilfe der Adjazenzmatrix eines Teiles des Webgraphen, wobei die Links initial mit der Relevanz ihrer Zielseite gewichtet werden. Diese Methode hat sich experimentell bewährt, erfordert aber zur Bestimmung der Zentralität das Laden zahlreicher Webseiten, die für das benutzerdefinierte Thema irrelevant sind.

2. Der fokussierte Crawler Ariadne

Wir betrachten als maßgeblichen Kostenfaktor das Laden der Webseiten und verfolgen daher ein alternatives Konzept: die einzelnen Hyperlinks werden unter Nutzung von Ankertext und Ziel-URL themenspezifisch klassifiziert, ohne die jeweilige Zielseite schon zu laden. Der Link mit der höchsten Bewertung wird als nächster verfolgt. Dieser Link-Klassifikator ist adaptiv, d.h. er wird im Lauf des Crawls aufgrund der Erfahrung mit den tatsächlich geladenen Webseiten verbessert. Das neue Konzept wird momentan im System *Ariadne* (Algorithm Regarding Interesting Anchortext for Directed Neighborhood Expansion) prototypisch implementiert und evaluiert.

Ariadne läßt sich in folgende Phasen gliedern:

- *Vorverarbeitung*
Aus den Webseiten im HTML-Format werden die reinen Texte extrahiert. Dann erfolgt ein Stemming der Terme sowie eine Elimination von Stoppwörtern.
- *Feature-Extraktion*
Nach der Zählung der Häufigkeiten aller enthaltenen Terme werden seltene Terme eliminiert. Die Auswahl der für die Klassifikation der Texte und der Hyperlinks wichtigsten Terme (Features) geschieht in Kooperation mit dem Benutzer, der einen Parameter eingibt und die vorgeschlagenen Features modifizieren kann.
- *Crawl*
Der Crawler startet mit den Trainingsseiten und

durchsucht einen Teil des Webgraphen. Die dabei gefundenen für den Benutzer "relevanten" Webseiten werden ausgegeben.

Die zentrale Phase des eigentlichen Crawls wiederholt für die jeweils aktuelle Webseite folgende Schritte:

- *Klassifikation des Textes der Webseite*
Aus der geladenen Webseite wird wie in der Vorverarbeitung der reine Text extrahiert und die Häufigkeiten der Features gezählt. Mit Hilfe eines einfachen Textklassifikators wird nun die Wahrscheinlichkeit bestimmt, daß die aktuelle Seite relevant ist.
- *Klassifikation der enthaltenen Hyperlinks*
Ein zweiter Klassifikator nutzt den Text der aktuellen Webseite sowie die Ankertexte und die URLs, um die in der aktuellen Seite enthaltenen Hyperlinks für die Zwecke des Crawl zu bewerten. Alle diese Hyperlinks werden mit ihrer Bewertung in eine sortierte Liste der *offenen Hyperlinks* eingefügt.
- *Verfolgen des Links mit der besten Bewertung*
Nach Abarbeiten der aktuellen Webseite wird als nächstes der beste offene Hyperlink, d.h. der Hyperlink mit der global höchsten Bewertung verfolgt. Die Webseiten werden also nur nach Prioritäten geordnet, es werden keine Seiten explizit von der Suche ausgeschlossen.

3. Klassifikation der Hyperlinks

Der zentrale Schritt des fokussierten Crawlers ist die Klassifikation der Hyperlinks, die im folgenden genauer erläutert wird.

Jedem Link wird eine Zahl zwischen 0 und 1 zugeordnet, die als Wahrscheinlichkeit dafür interpretiert wird, daß die Zielseite für das Anfragethema relevant ist. Diese Bewertung ergibt sich als gewichtete Summe von Einzelkriterien wie Themenbezug der Quellseite sowie Bewertung von Ankertext und Ziel-URL.

Beim Start des Crawlers sind nur die Trainingsseiten bekannt. Es wurden noch keine Hyperlinks verfolgt, so daß der Link-Klassifikator noch nichts lernen konnte. In der initialen Phase des Crawls wird daher zur Bewertung der Ankertexte und URLs ein statischer Klassifikator verwendet.

Im weiteren Verlauf des Crawls soll die Klassifikation der Hyperlinks adaptiv aufgrund der Erfahrungen mit den geladenen Webseiten verbessert werden. Der Link-Klassifikator soll insbesondere von Fällen lernen, in denen seine Vorhersage stark von der späteren Klassifikation des Textes der Zielseite abwich. Während der initialen Phase sollen deshalb für die Zwecke des Lernens auch genügend viele als "schlecht" bewertete Links verfolgt werden.

Am Ende der initialen Phase werden aus der Menge aller bisher geladenen Webseiten zwei Typen von Trainingsseiten ausgewählt:

1. *einfache Seiten*, d.h. Seiten, bei denen die Bewertung durch den Link-Klassifikator und durch den Text-Klassifikator gut übereinstimmte.
2. *schwierige Seiten*, d.h. Seiten, bei denen die Klassifikation des Links stark von der späteren Klassifikation des Textes der Webseite abwich.

Beide Typen von Webseiten müssen angemessen berücksichtigt werden. Mit Hilfe dieser Trainingsseiten wird nun ein Naiver Bayes-Klassifikator für Ankertexte und URLs trainiert, der im Vergleich zum einfachen statischen Klassifikator im allgemeinen eine wesentlich größere Anzahl von Features berücksichtigt. Nach dem Trainieren des Link-Klassifikators müssen alle offenen Links mit Hilfe dieses Klassifikators neu bewertet werden, womit im allgemeinen ihre Ordnung geändert wird.

Die Adaption des Link-Klassifikators wird während des Crawls zu geeigneten Zeitpunkten wiederholt. Der Text-Klassifikator ist dagegen statisch, d.h. das für den Benutzer relevante Thema wird während des Crawls als invariant angenommen.

4. Ausblick

Während der Implementierung wurden erste Experimente durchgeführt. Die Precision wird automatisch mit Hilfe des Text-Klassifikators berechnet und ist erfolgversprechend. Der Recall soll durch Vergleich mit den Ergebnissen einer Suchmaschine ermittelt werden, was sich erst nach Integration von Ariadne mit einem Datenbanksystem effizient durchführen lassen wird. Die Integration mit einem relationalen Datenbanksystem, das sowohl die offenen Links als auch die geladenen Webseiten effizient verwaltet, ist weitgehend abgeschlossen.

Ein experimenteller Vergleich von Ariadne mit Verfahren aus der Literatur ist geplant. Ein wichtiges Thema unserer weiteren Forschung soll die Integration des Benutzers in den Lernprozeß des Crawlers sein, damit sein Feedback möglichst früh zur Fokussierung der Suche genutzt werden kann.

Referenzen

- [1] Chakrabarti S., van den Berg M., Dom B.: "Focused Crawling: a new Approach to Topic-Specific Web Resource Discovery", Proc. WWW 1999.
- [2] Cho J., Garcia-Molina H., Page L.: "Efficient Crawling Through URL Ordering", Proc. WWW 1998.