Konzeption und Entwicklung eines intelligenten Agenten zum Internet Content Mining

Diplomarbeit

Zur Erlangung des Grades eines Diplom-Ökonomen des
Fachbereichs Wirtschaftswissenschaften der
Universität Hannover

Vorgelegt von

Patrick Bartels

[redacted]

Aufgabenstellung/Betreuung: Prof. Dr. Michael H. Breitner

Hannover, den 7. Oktober 2002

Inhaltsverzeichnis

1. Einleitung

1.1 Motivation

Der Verlauf der Geschichte der Arbeit lässt erkennen, dass die Bedeutung des Wissens für Unternehmen zukünftig immer wichtiger wird. Zuerst war die Muskelkraft kritischer Erfolgsfaktor für die ersten Unternehmer wie bspw. Handwerker, die, je stärker sie waren, ihre Arbeit umso leichter verrichten konnten und so einen Vorteil gegenüber den Konkurrenten genossen. Später wurden, neben der körperlichen Kraft, auch die geistigen Fähigkeiten gefordert. Es ging nicht nur darum, Tätigkeiten schnell zu erledigen, sondern die diversen Inputfaktoren mussten möglichst intelligent miteinander kombiniert werden. Heute stehen die Gesellschaften der höher entwickelten Länder in der Situation, dass für einen erfolgreichen Wettbewerb als kritischer Erfolgsfaktor Informationen und Wissen über Zusammenhänge immer wichtiger werden. Spätestens zu Beginn des dritten Jahrtausends ist festzustellen, dass das Vorhalten von Informationen nicht mehr ausreichend ist. Vielmehr ist das schnelle und effiziente Auffinden von Informationen und Wissen der Schlüsselfaktor, der einen Wettbewerbsvorteil gegenüber Konkurrenten bedeutet. Dieses so genannte Metawissen, das Wissen über Wissen, respektive die Metainformationen, Informationen über Informationen, sind ein breites Anwendungsfeld, in dem auf Software basierende Agenten eine zunehmend bedeutendere Rolle spielen.

Viele der Informationen, die Unternehmen benötigen, werden im Internet angeboten, wobei ein Großteil sogar kostenlos zur Verfügung gestellt wird. Gerade diese Gratis-Verfügbarkeit von Informationen macht den Bereich des automatisierten Web-Content-Mining durch Agenten interessant, da so eine große und wichtige Informationsquelle kostengünstig erschlossen werden kann.

Die stetig wachsende Problematik auf Informationen im Internet schnell und komfortabel zuzugreifen, in Kombination mit den Nutzenpotenzialen intelligenter Agenten, lässt die Forschungsrichtung Information Brokering in ihrer Bedeutung stetig steigen.[1]

[1] Vgl. KOENEMANN/THOMAS 1998. S. 1ff.

1.2 Zielsetzung und Vorgehensweise der Arbeit

Zielsetzung dieser Arbeit ist es, einen Agenten zu konzipieren und zu entwickeln, der in der Lage ist, Inhalte des Internet zu sammeln und lokal abzuspeichern. Der Begriff Konzeption wird dabei in der Weise interpretiert, dass der Grobaufbau eines Agenten zum Web-Content-Mining erstellt wird, wobei hier die Beziehungen zwischen den einzelnen Komponenten und deren jeweilige Aufgabe im Vordergrund stehen. Entwicklung bedeutet ferner, dass der konzipierte Agent in einer ausgewählten Programmiersprache realisiert und getestet wird und dabei eine vorgegeben Aufgabe erfüllt. Dazu wird zuerst in Kapitel 2 eine Abgrenzung vorgenommen, welche Art von Programmen unter der Bezeichnung Agent subsumiert werden können. Es folgt eine Erläuterung der für diese Arbeit relevanten Typen und, da die Entwicklung eines intelligenten Agenten gefordert ist, zusätzlich eine Betrachtung der Frage, wann ein Agent als intelligent gilt und welche Anforderungen daraus an die Konzeption abgeleitet werden können. Hauptaufgabe des Agenten ist es, Web-Content zu „minen“, weshalb eine Abgrenzung erfolgt, welche Teile des Internet als Web-Content betrachtet werden und wie Mining abgegrenzt wird, nämlich als das Auslesen von Internetinhalten und das Abspeichern auf einem lokalen Rechner.
Die eigentliche Systementwicklung des Agenten, der PISA genannt wird (Patricks intelligenter Software Agent), erfolgt in Kapitel 3 und basiert, da es sich bei Agenten letztlich auch nur um Programme handelt, auf einem Standardmodell der Systementwicklung mit vier Phasen.
1. Phase: Analyse – Jede Entwicklung eines Programms beginnt mit einer detaillierten Analyse des Umfeldes und der Aufgaben der geplanten Software, was in Abschnitt 3.1 erfolgt. Im Falle eines Web-Content-Mining-Agenten ist die Umwelt das Internet, welches daher bezüglich Aufbau und Funktionsweise untersucht wird, wobei eine erste Eingrenzung erfolgt, wie Daten ausgelesen werden können. Betrachtet wird unter anderem der Aufbau des Internets, das heißt, wie kommt ein Browser an die Daten der Seiten und welche Fähigkeiten muss die Programmiersprache, in welcher der Agent erstellt werden soll, folglich besitzen, um diese Funktion wahrzunehmen.
Kernstück des Internets ist die Möglichkeit, nicht nur per bekannter Adresse eine Seite aufzurufen, sondern auch per Mausklick so genannten Links zu folgen, die eine Verknüpfung zwischen Seiten herstellen. Daraus resultiert für einen Agenten

die Chance nicht nur Daten einer Seite zu extrahieren, sondern auch von den verbundenen Seiten, indem auch die im Quelltext als Verknüpfungen eingetragenen Seiten aufgerufen und bearbeitet werden. Dieser Vorgang ist als Crawling bekannt. Da die Adressen im Quelltext oft in einer verkürzten Schreibweise dargestellt und vom Browser in eine vollständige Adresse umgewandelt werden, muss die zugrunde liegende Systematik dem Agenten bekannt sein. Deswegen erfolgt eine Betrachtung der verschiedenen Notationen der Adressierung von Internetseiten.

Sollen, wie in dieser Arbeit, Daten aus Internetseiten ausgelesen werden, muss bekannt sein, wie die Seiten beschrieben sind, die die Inhalte liefern. Die am häufigsten eingesetzte Sprache zur Auszeichnung von Internetseiten ist derzeit HTML (Hypertext Markup Language), wobei die Zukunft sicher XML (e Xtensible Markup Language) gehören wird. Beide Sprachen werden in ihrer Funktionsweise dargestellt sowie einige ausgewählte alternative Darstellungsarten, wobei eine weitere Abgrenzung der Möglichkeiten der Datenextraktion anhand der technischen Realisierbarkeit erfolgt.

Zur Entwicklung ist ein Beispiel-Einsatzgebiet sinnvoll, damit die Funktionsweise vor einem realistischen Hintergrund getestet werden kann. Zwei der wichtigsten Vorteile von Agenten im Gegensatz zum Einsatz von Menschen sind die Fähigkeiten große Datenmengen pro Zeiteinheit und über einen sehr langen Zeitraum zu verarbeiten. Die Ermittlung von Aktienkursen und Optionsscheinen in festgelegten Zeitintervallen zum Aufbau einer Kursdatenbank erfordert genau diese Fähigkeiten. Ferner kann so die Transparenz der Informationen aus dem Internet erhöht werden, da nicht nur einzelne Daten ausgelesen werden können, sondern beliebig viele, die in anderen Anwendungen wie Microsoft Excel miteinander verbunden und ausgewertet werden können. So können bspw. die Preise für Währungsoptionsscheine und die Währungskurse von verschiedenen Seiten ausgelesen und einander gegenübergestellt werden. Daher werden die beiden genannten Einsatzgebiete in Abschnitt 3.1.2 näher erläutert und die auszulesenden Informationen festgelegt.

2. Phase: Konzeption – Die Konzeption in Kapitel 3.2 basiert auf den Ergebnissen der Analysephase und hat zum Ziel, einen Grobaufbau für den Agenten zu erstellen, weshalb zunächst die beiden wichtigsten Techniken zur Extraktion von Information aus Text dargestellt werden: Extraktion anhand natürlicher Sprachmuster (NLP) und anhand einer gegebenen Textstruktur. Aufgrund der für

NLP-Techniken notwendigen Texte in natürlicher Sprache, eignet sich der Ansatz der Informations-Extraktion anhand der bestehenden Struktur der HTML-Seiten besser, was in Abschnitt 3.2.1.2 näher begründet wird. Daraus wird der Aufbau des Agenten abgeleitet und die einzelnen Komponenten werden in ihrer Funktion in Abschnitt 3.2.2 erläutert.

Die Funktionsweise des Agenten basiert auf der Idee, dass die Tags, die die Inhalte einer Internetseite auszeichnen, nach bestimmten Regeln gefiltert werden können. Dazu muss der Quelltext der Seite geladen und von dem Agenten bearbeitet werden. Im einfachsten Fall könnte die Vorgehensweise eine einfache Untersuchung auf Vorkommen bestimmter Muster sein, die bspw. alle zweistelligen Dezimalzahlen repräsentieren, welche ggf. abgespeichert werden könnten. Dies ist für eine präzise Beschreibung der zu extrahierenden Muster nicht ausreichend, da auf einer Seite mehrere Fundstellen des Musters existieren können. Deswegen wird im Abschnitt 3.2.2.4 die Komponente zur Auswertung des Quelltextes so entworfen, dass die Position des zu findenden Musters relativ zu einem weiteren Muster angegeben werden kann.

Aus dem beschriebenen Vorgehen leiten sich spezifische Anforderungen an eine Programmiersprache ab, die im Abschnitt 3.2.3 zusammengefasst werden.

3. Phase: Entwicklung – In der Entwicklungsphase muss zuerst eine Programmiersprache gefunden werden, welche die Anforderungen, die in den vorherigen Phasen ermittelt wurden, erfüllt, weshalb ausgewählte Sprachen diesbezüglich in Abschnitt 3.3.1 miteinander verglichen werden. Die Module des Agenten, die in der Konzeptionsphase ermittelt wurde, werden dann in Java, der als dafür am geeignetsten angesehenen Sprache, realisiert und in Abschnitt 3.3.3 beschrieben. Vorher erfolgt in Abschnitt 3.3.2 eine kurze Einführung in die verwendeten Programmpakete und Datentypen, die Java beiliegen.

4. Phase: Programmtest – Der Programmtest soll abschließend zeigen, ob der Agent die gestellten Aufgaben zufrieden stellend ausführt und wie Leistungsfähigkeit und Robustheit einzustufen sind, wobei die Bereiche Aktien und Optionsscheine getrennt betrachtet werden.

Aktien – Als Testaufgabe werden bei den Aktienkursen die 30 größten deutschen Unternehmen gewählt, die von der Deutschen Börse AG im DAX 30 zusammengefasst werden. Dabei werden zum einen die jeweiligen Kurse des XETRA-Handels ausgelesen sowie zusätzlich die Kurse des außerbörslichen Handelssystems der Lang & Schwarz AG, deren Daten im Gegensatz zu den XETRA-Kursen

in Echtzeit dargestellt werden. Bei den XETRA -Daten handelt es sich bei den kostenlosen Angeboten in der Regel um 15 Minuten zeitversetzte Kurse. Um die Zuverlässigkeit der angebotenen Kurse beurteilen zu können, werden alle Werte jeweils außer von den Seiten der Deutschen Börse AG und der Lang & Schwarz AG von zwei weiteren abgefragt, die dieselben Werte anbiete n. So veröffentlicht die Comdirect Bank die Daten der Deutschen Börse und der Finan zdienstleister Onvista die der Lang und Schwa rz AG. Daraus folgt, dass für jedes Unternehmen vier Werte abgespeichert we rden (Je zweimal der Kurs der Deutschen Börse und von Lang & Schwarz).

Daraus lässt sich zum einen entnehmen, ob die Daten der Zweitanbieter (Comd i-rect und Onvista) korrekt und zuv erlässig sind und inwiefern die Kurse des Ha n-delssystems den Werte n des XETRA -Handels entsprechen. Darüber hinaus kann anhand der Dauer der Abfragen die Leistungsfähigkeit ermittelt werden. Der Test-zeitraum erstreckt sich dabei über einen Tag und wird in verschiedenen Konfigu-rationen durchgeführt, damit die Grenzen des Agenten ermi ttelt werden können. Aus den ermittelten Daten werden abschließend mögl iche Weiterentwicklungen und Verbesserungen abgeleitet.

Optionsscheine – Neben Aktien werden Optionsscheinpr eise ermittelt, womit be-sonders gut die mit Hilfe des Agenten gesteigerte Transparenz der Daten verdeu t-licht werden kann. Ein Benutzer des Internet, der auf der Suche nach entspreche n-den Kursen ist, kann die von vergleichbaren Produkten zweier oder mehr Em itten-ten nicht komfortabel vergleichen, da die Preise immer nur für einen Zeitpunkt gestellt werden. Selbst wenn ein Preis niedriger ist, muss dies nicht bedeuten, dass dies die Regel ist. Mit Hilfe des Agenten können die Preise für mehrere Produkte (nicht nur Optionsscheine) ausgelesen und abgespeichert werden, um diese mit Hilfe eines Werkzeuges wie Microsoft Excel grafisch gegenüber zu stellen.

Wie bei den Aktienkursen erfolgt der Test über den Zeitraum von einem Ha n-delstag. Es werden drei Emittenten (Ci tibank, Deutsche Bank und UBS Warburg) betrachtet, deren Optionsscheinpreise zu fünf verschiedenen Basispreisen (0,90 Cent, 0,95 Cent, 1,00 Cent, 1,05 Cent, 1,10 Cent) und jeweils vier verschiedenen Fälligkeitsterminen ermittelt werden. Dabei werden zur Üb erprüfung der Daten, wie bei den Aktien, neben den Seiten der Emittenten die der Comdirect Bank und Onvista ausgewertet.

Abschließende Beurteilung – Am Ende der Arbeit erfolgt eine Zusammenfa s-sung der Ergebnisse des Test und der Erfahrungen bei Konzeption und Entwic k-

lung des Agenten. Ferner wird ein Ausblick auf die weitere Entwicklung des Web-Content-Mining unter Berücksichtigung der Ergebnisse vorgenommen.

2. Agenten

In diesem Kapitel wird die Bedeutung von intelligenten Agenten zur Informat i-onsgewinnung dargestellt. Des Weiteren erfolgen eine Abgrenzung von verschi e-denen Agenten und die genaue Zielfestlegung des Agenten, der im Rahmen dieser Arbeit entwickelt werden soll.

2.1 Bedeutung von Agenten

In den letzten Jahren ist eine ständig zunehmende Vernetzung von verschiedenen Informationssystemen mit dem Internet zu beobachten. Einer der wesentlichen Gründe liegt im kommerziellen Potenzial der Entwicklung von „elektronischen Märkten“. Zwei bedeutende Bereiche, in denen Unternehmen sich durch Nutzung des Interne ts große Gewinne versprechen , sind der internationale Aktienhandel und Informationsdienste.

- Der internationale Aktienhandel wird in Zukunft vermehrt über das Internet erfolgen. Geschäfte können online innerhalb von wenigen Minuten erledigt werden und erlauben Reaktionen auf kleinste Marktschwankungen.[2] Als Folge des Wettbewerbes zwischen den Finanzdienstleistern, die solche Dienste an-bieten, stellen fast alle Wettbewerber Wertpapierkurse im Internet bereit. Zum einen Echtzeitkurse gegen Entgelt, zum andere n kostenlose zeitversetzte Ku r-se.
- Diverse Informationsdienste wie Zeitungen oder Fernsehsender stellen ihr Angebot ins Internet , um damit der wachsenden Akzeptanz dieses Mediums gerecht zu werden. Vielfach stehen die Inhalte (noch) kostenlos zur Verf ü-gung.[3]

In der Folge ist das Internet die größte existierende Bibliothek. Mit vielen Millia r-den Seiten, die (fast) überall auf der Erde zugän glich sind, bietet das Internet auf nahezu alle Fragen eine Antwort. Allerdings gibt es keine zentrale Instanz, die Webseiten katalogisiert und somit systematisch durchsuchbar macht. [4] Mit ste i-

[2] Vgl. KLUSCH/BENN 1998, S. 8.
[3] Vgl. KLUSCH/BENN 1998, S. 8.
[4] Vgl. MURCH/JOHNSON, 2000, S. 46f.

4. Fazit und Ausblick

Ziel dieses Abschnittes ist es, den aktuellen Stand des Agenten bewertend zusammenzufassen sowie die Verbesserungsmöglichkeiten aufzuzeigen, die über die im Abschnitt 3.4.3 genannten hinausgehen.

Der in dieser Arbeit erstellte Agent PISA ermöglicht es, Daten aus Webseiten auszulesen und auf dem lokalen System abzuspeichern, wobei während der Aufgabenausführung keine Interaktion mit dem Benutzer nötig ist. Einzig die Festlegung der auszulesenden Daten in der Konfigurationsdatei muss durch den Anwender vorgenommen werden. Insgesamt kann das Programm diese Aufträge autonom ausführen und erfüllt damit die Mindestanforderung an einen Agenten. Die sozialen Fähigkeiten beschränken sich dabei auf die Entgegennahme von Benutzerbefehlen, mit deren der Zustand des Agenten angezeigt wird.

Die Intelligenz eines Agenten, und wann diese angenommen werden kann, sind in der Agenten-Wissenschaft die am meisten diskutierten Themen. Im Rahmen dieser Arbeit wurde ein Agent erstellt, der in der Lage ist, mit einer gegebenen Wissensbasis, Aufträge auszuführen und die Ergebnisse in vorgegebener Art und Weise abzuspeichern. Schlussfolgerungen werden dabei vom Agenten insofern getroffen, dass durch im Programmcode vorhandene Fall-Unterscheidungen die Vorgehensweise beeinflusst wird. Daraus können in keiner Weise Verbesserungen auf die zukünftige Effizienz des Agenten abgeleitet werden, weshalb dieser nicht als lernfähig bezeichnet werden kann. Folglich handelt es sich im Sinne der in Abschnitt 2.3 dargestellten Definition um eine wenig ausgeprägte Intelligenz, die kein proaktives Verhalten oder eigene echte Schlussfolgerungen auf Basis von Denkvorgängen ermöglicht.

Das vorangegangene Kapitel hat deutlich gemacht, dass mit der Benutzung von Agenten das Informationspotenzial des Internet effizient ausgenutzt werden kann. Dabei kann ein Software-Agent die Aufgaben der Informationsextraktion besser und effizienter wahrnehmen als eine menschliche Arbeitskraft, was ebenfalls deutlich geworden ist. Durch die Leistungsfähigkeit moderner Rechner ist es zudem möglich, die Daten aus vielen Quellen zusammenzutragen und einander vergleichend (Wertpapiere) oder ergänzend (Nachrichtenartikel) gegenüber zu stellen, wodurch die Transparenz im Vergleich zu r einzelnen Information gesteigert werden kann.

Das Anfragen der Seiten erfolgt bisher ausschließlich auf Basis des HTT-Protokolls, was für die meisten Anwendungszwecke ausreichend ist. Innerhalb der empfangenen Seiten werden die enthaltenen relativen und absoluten Links korrekt verarbeitet, so dass auch Unterseiten aufgerufen werden können. Dabei ist sichergestellt, dass keine Seite mit identischer Adresse doppelt aufgerufen wird. Alle rdings erfolgt derzeit keine Überprüfung, ob zwei unterschiedlich geschriebene Adressen dennoch auf dasselbe Ziel verweisen. Diese ließe sich in Form eines Abgleichs mit regulären Ausdrücken realisieren, wobei auch damit nicht alle möglichen Fälle erfasst werden können.

Die empfangenen Seiten können genau nach den HTML-Tags durchsucht werden, die als Klasse dem Programm bekannt sind. Die derzeit fehlenden können ggf. leicht nachgerüstet werden, indem die entsprechenden Klassen hinzugefügt werden. Es können momentan die Attributwerte der Tags nicht in die Auswahl zur Feldliste einbezogen werden. Eine Möglichkeit dieses nach zurüsten ist die, jeder Tag-Klasse eine Methode hinzuzufügen, die das öffnende Muster auf Übereinstimmung mit einem übergebenen Attribut und dessen Ausprägung überprüft. Die Funktionsweise einer solchen Funktion ist anhand der Methode *isAttribute()* der Klasse FONT zu erkennen, die allerdings vom Agenten derzeit nicht verwendet wird. Auch ohne die Auswertung der Attributwerte konnten die hier gestellten Aufgaben gut bewältigt werden.

Die XML-Unterstützung für die Datenausgabe erzeugt syntaktisch korrekte XML-Dateien, die zur Weiterverarbeitung problemlos in andere Anwendungen wie Datenbanken importiert werden können. Allerdings bezieht sich dies nur auf die Erstellung der Inhaltsdateien, denn die Document-Type-Definitions müssen derzeit vom Benutzer erstellt werden, wozu Grundkenntnisse der Funktionsweise von XML notwendig sind. In einer späteren Version des Agenten sollte dies automatisch anhand der verwendeten *patternName*-Angaben geschehen.

Das Abspeichern in Textdateien ist zwar sehr plattform- und anwendungsunabhängig, jedoch sind allein bei der Auswertung der Optionsscheine über 20 Graf iken zu erstellen gewesen, wobei über 120 Kurse berücksichtigt werden mussten. Für größere Aufträge sollte daher ein Weg gefunden werden, die Daten automatisch aufzubereiten. Die leistungsfähigste Lösung sind Datenbanken, deren Unterstützung zur Ansteuerung über eine Standard-Schnittstelle wie ODBC dem Agenten in einer nächsten Version eingebaut werden sollte. In Java ist eine solche Unterstützung über die JDBC-to-ODBC-Bridge möglich und einfach nachzurüsten.

Die derzeitige Benutzerschnittstelle ist sehr zweckorientiert und könnte durch eine einfacher zu bedienende Oberfläche ersetzt werden. Am geeignetsten erscheint dazu als Lösung ein Java-Programm, das in Form eines Applets oder einer einfachen Applikation realisiert werden kann. Beide können über ein Netzwerk auf Basis verschiedener unterstützter Protokolle auf den auf einem Server liegenden Agenten zugreifen und dessen Funktionen steuern.

Darüber hinaus können in Java mit dem *java.net*-Paket Verbindungen zu Servern hergestellt werden, die eine Anmeldung über einen Benutzernamen und dem entsprechenden Kennwort erfordern. In Zukunft sollten auch solche Seiten abgefragt werden können, um ein weiteres Hindernis der Informationsextraktion aus dem „versteckten Internet“ zu beheben, denn durch die Anmeldung, die auf vielen Seiten kostenlos erfolgt, können auch die geschützten Bereiche ausgewertet werden.

Da die Suche nach Informationen und deren Filterung so weit wie möglich automatisiert werden sollte, sind Neuronale Netze eine mögliche Erweiterung. Die Muster für die Suche nach Wertpapierkursen und Nachrichten könnten, nachdem ein Trainingsdatenbestand für ein neuronales Netz angesammelt wurde, automatisch erstellt werden. Damit könnte auch die Leistungsfähigkeit des Agenten gesteigert werden, denn es müssten nicht mehr alle Muster vom Benutzer erstellt und getestet werden.

Zusammenfassend lässt sich feststellen, dass der Agent, obwohl an einigen Stellen noch Verbesserungspotenzial zu erkennen ist, allgemeine Aufgaben sehr gut lösen kann.