Data Mining: Online

László Pitlik

Institut für Wirtschaftsanalyse
Universität Gödöllő
Páter K. u. 1.
H-2100 Gödöllő,
pitlik@miau.gau.hu

Abstract: Das Ziel des „my-X“ Projektes besteht darin: Analysewerkzeuge online in Form von „pre-paid“ Konstruktion anzubieten. Im folgenden Beitrag werden die wichtigsten Charakteristika über das (seit Januar 2007 erstellte) System dargestellt. Diese Informationen betreffen einerseits kurz die technische Realisierung, andererseits umfangreicher die Erfahrungen der Referenz- und Testprojekte (bis Ende Oktober 2007). Außerdem werden die Eckpunkte eines Vergleiches zu bereits bekannten „Konkurrenten“ zusammengefasst. Die Dienstleistung wird im ersten Quartal 2008 für die ersten Partner freigegeben.

1 Vorgeschichte und technische Realisierung des Projektes

Im GIL - Tagungsbeitrag 2007 wurde bereits das Projektvorhaben dargestellt. Das Projekt wurde also vom ungarischen Innovationsamt gefördert (No.156/2006) mit dem Ziel, data mining Aktivitäten zu demokratisieren, d.h. Analysewerkzeuge in pre-paid Konstruktionen online anzubieten. Die Realisierung läuft grundsätzlich nach Plan. Die online Dienstleistung wurde nach den folgenden strategischen Gedanken programmiert:

- Kernmodule (Ähnlichkeitsanalysen und passende Visualisierungseffekte, bzw. OLAP-Dienste): in PHP, als eigene Entwicklung [PIC06]
- Hilfsmodule (wie Nachrichten- und Dokumentenverwaltung): in PHP, als eigene Entwicklung
- Rahmensystem für Menü und Help: Moodle
- Interaktionsrahmen: MediaWiki
- (Fern-)Administration: in PHP, eigene als Entwicklung
- Mehrsprachigkeit: für die Oberflächen durch automatische Sprachenerkennung (des Browsers), für Inhalte destinationsspezifisch, manuell

Die Entwicklung wird in drei Wellen realisiert. Die letzte (noch nicht abgeschlossene) Welle ist hauptsächlich verantwortlich für die fine tuning (Optimierung von Datenabfragen, Programmverläufen, Benutzerfreundlichkeitseffekten, Schutzmaßnahmen), bzw. für die Realisierung der Zahlungsvarianten.

Das ganze System läuft unter Linux, unterstützt durch PostgreSQL. Es wurde bei der Realisierung bevorzugt, bereits existierende, bekannte Umgebungen (vgl. CMS: Content Management Systems) zu verwenden und nur die Kernmodule neu zu entwickeln. Das System hat ständig eine Notebook-Variante (über Windows) zu Demonstrierungszwecken. Durch Integration von Moodle und MediaWiki erhofft man, für viele Nutzer eine Art sofort übersichtliche Navigationsoberfläche bereit zu stellen.

2 Erfahrungen der Test- und Referenzprojekten

Die online Dienste wurden und werden permanent getestet: einerseits durch Probleme, welche intern (als Tests) generiert werden, andererseits durch die ersten Anwender:

2.1 Referenzprojekte

Paradigmenwechsel bei Prognoseaufgaben?: Das „my-X“ System bringt programmierungstechnisch an sich nichts neues und es muss auch nicht sein. Man braucht ja gewisse Inputformulare, um die zu analysierenden Daten für weitere Manipulationen vorzubereiten, danach können im Hintergrund bereits „beliebige“ Algorithmen laufen. Zum Schluss müssen die Outputs kontextabhängig visualisiert werden. Warum kann man dann bezüglich „my-X“ über Paradigmenwechsel sprechen? Bislang (vgl. künstliche neuronale Netze) wurden Prognosen hauptsächlich in Form von Lern- und Testaufgaben verstanden. Es wurde nach Modellen gesucht, welche „hohe“ Testgenauigkeit haben. Dadurch haben jedoch die Betroffenen meistens auf die letzten (erlernbaren) Inputs verzichten müssen. Versucht man ohne Tests zu arbeiten, können endlich alle Daten erlernt werden. Die Tests werden hierbei durch eine höhere Anzahl von primeren Modellen ersetzt, die fähig sind, die Zukunft aus „Puzzle-Stücken“ möglichst widerspruchsfrei (konsistent – vgl. Grundforschungsprojekt: OTKA 049013) „Pixel“ für „Pixel“ abzuleiten. Eine konkret bestellte Aufgabe wurde im Zusammenhang mit der Börse definiert: Es musste der Verlauf von CHF/HUF für 30 Tage im Voraus charakterisiert werden. Die Lösung bestand aus 8 (beim Erstellen unabhängigen, bei der Interpretation sich gegenseitig verifizierenden) Modellen, die jeweils zwei Antworten liefern konnten (ja/nein). Die Zukunft wurde also durch eine Modellhybrid beschrieben, in welchem jedes Teilmodell ihre konsistenten (1/2^8) Informationen geliefert hat.

Fachbücher aus Daten?: Angenommen, dass Experten (z.B. Erziehungsoffiziere in Strafvollzugsanstalten, oder Sportwettexperten) nie die Möglichkeit haben (oder wahrnehmen), sich über Ihr Wissen in Form von schriftlichen/verbalen Dokumentationen zu äußern, aber viele Entscheidungen treffen, bzw. Schätzungen als Daten (Wer darf z.B. arbeiten? oder Wer gewinnt das nächste Spiel?) hinterlassen und danach plötzlich „verschwinden“, stellt sich die Frage: Lässt sich Ihr Experten-„System“ aus den Rohdaten hervorrufen?

Passende Fachliteratur spricht sicherlich vielseitig über das untersuchte Phänomen, jedoch zu allgemein, ohne an den bekannten Fakten jeweils verifiziert zu sein. Stattdessen lassen sich beinah beliebig genaue angepasste Modelle finden, welche die verborgenen Teilzusammenhänge (vgl. ceteris paribus, expert system) in einer solchen Form schätzen, welche als Grundlage für eine Art Fachliteratur geeignet ist: z.B. Selten können Fußballmannschaften immer gewinnen: welche konkrete Daten zeichnen den Weg zu einer „unerwarteten" Niederlage?

Vereine als „Firewall"?: Eine weniger dienstleistung- und technikbedingte, jedoch sehr wichtige Erfahrung zeigte sich in den juristischen Aspekten der online Datenverwaltung. Es sind kaum kooperationsfähige Partner auf dem Markt (meistens staatlichen Organisationen), welche korrekte, gemeinnützige OLAP-basierte Dienste anbieten, obwohl u.a. im Bereich der GIS bereits Direktiven vorliegen (vgl. INSPIRE: [SA07]), welche vorschreiben, dass kostenlose Meta-Datenbasen vorhanden sein sollten, damit die Kunden mindestens darüber Bescheid wissen, wo/wer/welche Daten verwaltet. Ginge man noch ein Schritt weiter, um nicht nur Meta-Informationen sondern auch die Daten selber im Rahmen eines Mehrwertdienstes (d.h. z.B. beliebig einstellbare Kopfzeilen für die abgefragten Tabellen) anzubieten, kann man „ruhig" davon ausgehen, dass einige Anklagen entstehen werden… Profitorientierte Firmen können solche „lange" Prozesse nicht verkraften. Vereine sind jedoch geeignet, solange als „juristische Firewalls" zu fungieren, bis die Betroffenen endlich Ihre Rechten und Pflichten klar stellen.

Online Datenverwaltung für beliebige Gruppierungen?: Als noch mehr „radikale" Fortsetzung der „Firewoll-Theory" lässt sich folgende Problemstellung formulieren: Viele Daten (vgl. Meteorologiemessungen und -Prognosen) werden für kurze Zeit (z.B. für einen Tag) im Internet freigegeben. Später kann man diese Daten nur kostenpflichtig aus den Archiven abfragen. Durch minimale Management und Technik können solche „Gemeinschaftsdienste" initiiert werden, wo beliebige Leute einer virtuellen Gruppe die täglich freigegeben Daten online so verwalten, dass sofort eine OLAP - Datenbasis entsteht. Dadurch können z. B. Verbraucherschützer die Trefferquoten von Wettervorhersagen permanent ableiten, um die Betroffenen darüber zu informieren.

2.2 Testprojekte

Meteorologie-Vorhersagen: Wiederum als Fortsetzung des vorherigen Punktes können Wettbewerbssituationen sobald entstehen, wenn Gütemerkmale (vgl. Trefferquoten für Wettervorhersagen) bekannt werden. Über einen Duzend Studenten arbeiten im Rahmen eines data mining Kurses z.B. dafür, wie können die täglich publizierten Modellergebnisse des ungarischen und englischen Wetterdienstes durch eigene Modelle „besiegt" werden. Hierfür werden keine Daten offiziell erworben, nur die Daten verwendet, die durch die Gruppe selber zusammengeführt wurden. URL: http://miau.gau.hu/lps/olap4/olap_m.php3

Gemeinsame Schlagkarteien (inkl. Bodenuntersuchungen): Die Situation ist ähnlich im Bereich der Schlagkarteien (und Bodenuntersuchungen), wo Studenten anonymisierte Datensätze sammeln, um zu simulieren, welch riesiges Kraftfeld für weitere Analysen entstehen könnte, falls alle Betroffenen freiwillig mitwirken würden. URL: http://miau.gau.hu/lps/olap2/olap_m.php3

Benchmarking Datenbasen: Wiederum durch studentischen Arbeiten entstehen kleine Produktkatalogen (inkl. Nebenwirkungshäufigkeitskatalog für Medikamente), welche uns ermöglichen, KO-Kriterien (d.h. Filterungen) zu definieren, und danach benchmarking Berechnungen (z.B. Preis-Leistungsvergleiche) abzuleiten. Diplomatisch gesagt: es ist mindestens „komisch", dass bislang wenige gemeinnützigen (Meta)Informationen von Behörden vorliegen, welche alle Informationen über zugelassene Objekte durch OLAP-Dienste (d.h. standardisiert) abfragen lassen und zu welchen Diensten sich die Händler durch ihre Preisangaben anschließen könnten, damit die Kunden nicht nur in Meinungsumfragen sondern in der Realität immer stärker ein Preis-Leistungsoptimum durch Ähnlichkeitsanalysen wahrnehmen könnten.

3 Konkurrenzanalyse

Die folgende kurze Darstellung zeigt uns, welche methodischen Unterschiede im Gegensatz zu den bekanntesten Alternativen (meistens offline Lizenzen) zu beobachten sind [PI04]:

Entscheidungsbäume: Die Entscheidungsbäume (vgl. SPSS) können solche Schätzungen nicht liefern, für welche Beobachtung noch nicht vorliegen. Ähnlichkeitsanalysen anerkennen den Begriff des „genetischen Potentials", d.h. die Leistung unter den besten Umständen (und sie definieren andere, nie gemessene Input-Output-Kombinationen ebenso). Entscheidungsbäume liefern rel. wenig Regeln zum untersuchten Phänomen, während die Ähnlichkeitsanalysen kombinatorisch vollständige Regelsysteme erstellen.

Neuronale Netze: Diese Modelle ergeben meistens sog. „black box" Zusammenhänge, d.h. solche mathematische Beschreibungen, aus denen nur schwerlich Fließtexte zu Interpretation der Modellergebnisse erstellt werden können, bzw. die entdeckten Zusammenhänge sind „oft" Polynome. Ähnlichkeitsanalysen liefern (wie bereits signalisiert) Regelsysteme mit begrenzten Polynomen-Effekten.

Literaturverzeichnis

[SA07] Schopp A.: Európa digitális térképen (Europa in digitalen Karten - INSPIRE) IT_Business, Budapest, 09.10.2007, S. 22.

[PIC06] Pitlik, L., Pető I.: The role of consistency controlled future generating models in the strategic management, SZIE-Bulletin, Gödöllő, 2006. http://miau.gau.hu/miau/91/bulletin_en.doc

[PI04] Pitlik, L: Component-based Object Comparison for Objectivity, 25. GIL - Jahrestagung, Bonn, 08-10.09.2004, http://miau.gau.hu/miau/69/gilfull.doc