Language Orientated Olap - Lolap

Matthias Faix, W. Voßhall, W. Stolz
IPM Köln GmbH
Köln

Schlüsselworte:
Olap, UMLS, Thesaurus, Hyperion, Data Mining

Einleitung
Effiziente Erweiterung der Suchfunktion in Oracle in Bezug auf die deutsche Sprache

„Eine Maschine kann Keine Symphonien schreiben“
„Können Sie Es“

Isaac Asimov – I Robot

Relationale Datenbanken sind im Grenzwertfall eindeutig. Im Paradies der normalisierten Datenbank gibt es weder doppelte Datenhaltung noch unsaubere Daten.

Die Abfragessprache SQL basiert auf der Annahme der „Ein Eindeutigkeit“. In exakten Wissenschaften wie der Mathematik mag man das Paradies erreichen. Je weiter man in den Bereich der Lebenswissenschaften eindringt desto mehr stößt man in das Fegefeuer der unsauberen und doppelten Daten vor. Ob das im Bereich der Sicherheitsdatenblätter in der Chemie Rezepturen sind oder auf der anderen Seite die Auswertung von Arztbriefen. Ein menschliches Gehirn kann Mehrdeutigkeiten, Unsauberkeiten in der Sprache und Wortverdrehungen erkennen. Eine Maschine nicht. Das Sinnhafte Erfassen von Texten ist dementsprechend Thema der KI.

Wir haben zwar keine Mind Machine produziert, mit unserem Oracle Package LoLap können wir aber mindestens die Trefferquote von Anfragen sinnvoll erhöhen. Zudem können OLAP Mechanismen davon profitieren. Erfolgreich haben wir Lolap im Cleansing Prozess eingesetzt. Wir haben die Funktion in den Transform Prozess (T) des ETL Verfahrens eingebettet.

Textorientierte Materialien können so der Auswertung mit OLAP Programmen wie Hyperion geöffnet werden. Es bieten sich so der weiter Aufschluss von Materialien in Orbis an.

Methoden
Unser Package erleichtert die Suche in Volltextsegmenten. Records werden nach bestimmten Kriterien mit in die Ergebnismenge übernommen. Wir haben zunächst die Soundex Funktion ersetzt. Unsere Funktion „Koelsch“ entspricht dem Aufruf von Soundex, das Ergebnis auf der Funktion entspricht aber der Definition des Koelner Algorithmus.

klang:= Koelsch(„String“)

Der Koelner Algorithmus kommt dem Klang des geschriebenen Wortes näher als die Funktion Soundex, die auf das Lautbild der amerikanischen Sprache optimiert ist. Das Ergebnis des Funktion kann dann auch zur Transformation im ETL Prozess genutzt werden.

Unser Package ist in PL/SQL geschrieben, damit die allgeneine Lesbarkeit des Codes gewährleistet ist. Zugleich kann man sich des Java Pool in Oracle sparen.
Eine ungetestete Java Variante liegt vor.
Unser Package enthält zudem einen graphischen Theasaurus auf Basis des transformierten Klangwerts.
In den Live Sciences sind Synonyme geläufig. Das gilt speziell auch für die Medizin – und dort auch für die Dimensionen.

Die Begriffe Flusssäure, Fluss-Säure und HF sind Synonym zu betrachten.

Die Funktion

lolap_syn(„string§)

gibt eine Liste von Synonymen aus. Mit der Funktion IN kann man dann auf Existenz prüfen.

Wir testen die hierachische Organisation des Thesauruses unter Verwendung von „Connect By“ Operator. Viele Dimensionen in der Biolgie sind in Hierarchien geordnet.

Das Lolap Paket enthält auch einen kleinen Stringverbinder, man kann damit alle Felder eines Record analysieren.

Diese beiden Ebenen des Lolap Paketes funktionieren und sind im Einsatz. Lolap kann so eine Bereicherung einer relationalen Datenbankstruktur in eine Textdatenstruktur hinein sein. Im Gegensatz zu Textmining Tools wie die Tools von Google sind so auch Datenanalysen möglich.

Wir arbeiten nun an einer fehlertoleranten Erweiterung mit Hilfe des Needleman - Wunsch Algorithmus – Wortdreher würden dann auch toleriert werden.

Gegenüber dem bewährten Package von Oracle Context gibt es Vorteile in der Transparenz und in der Geschwindigkeit. Die Synomye sind ladbar. spezielle Arbeitsgruppen können sich so austauschen.

Die ladbaren Synonyme sind das eigentlich spezielle an dem System. Man kann ein wartbares Thesaurensystem erstellen dass auch Mandantenfähig ist.
Das Thesaurensystem ist in gewisser Hinsicht wiederum hierarchisch geordnet. Wir haben ein allgemeines Thesaurensystem, das für alle Anwendungen gültig ist. Zugleich kann per Mandantensteuerung die jeweilige spezielle Thesaurenumgebung aktiviert werden. Die Thesaurenmenge Basis wird so durch die spezielle Thesaurenmenge erweitert. Speziell von Interesse sind die Suchanfragen ohne Ergebnis. Diese Suchanfragen werden speziell gesammelt und erweitern den Thesaurus dann „On Request“

Die Idee ist, durch die Menge an Daten intelligente Wissensmaschinen entstehen. Eine spezielle Anwendungsform für das Suchen ist das Suchen in Arztbriefen. Arztbriefe enthalten auch interessante wissenschaftliche Hinweise. Man denke an das Entdecken der HIV Erkrankung oder an die Entdeckung des Zusammenhang zwischen Sprue und Lymphomen. Aus gehäuften Symptomen wurden an dieser Stelle dann ein Krankheitsbild und zumindest ein Verdachtsmoment, dem man dann mit anderen statistischen Mitteln nachgehen kann.

Idee des Text Data Minings
Ein Anwendungsmoment für das PL/SQL Package Lolap ist sicherlich die komfortable Suche. Suchen und das Richtige Finden, hierbei unterstützt das Modul den Anwender. Es gibt dafür sicherlich

eine große Anzahl von Anwendungsmöglichkeiten, sicherlich auch im kommerziellen Bereich. Man denke hierbei an die guten Vorschläge vom Amazon. Die Vorschläge von Amazon heizen sicherlich den Verkauf an. Ein anderer und für uns wesentlicher Aspekt ist aber das Reporting unter besonderer Berücksichtigung der Erweiterung durch das Lolap Paket. Man kann so nach Nierenerkrankungen suchen und bekommt zusätzlich noch alle Erkrankungen die mit Nephro gekennzeichnet sind mit in die Zielmenge. Das kann im gesteigerten Bereich auch für Medikamente gelten. Es wird nach Aspirin gesucht, man bekommt dann auch die Medikamente mit in der Zielmenge angezeigt, die ASS oder Acetylsalicylsäure enthalten. Das gilt aber auch in der Hierarchie. Im gegebenen Beispiel ist ASS und Acetylsalicylsäure genauso wie Aspirin dem Generikum Acetylsalicylsüre zugeordnet. Das ist ein Schmerzmittel und so der Medikamentengruppe 45 in der Roten Liste zugeordnet.
Unsere These ist, mit Lolap kann man aus unstrukturierten Informationen, wie man sie in Arztbriefen aber auch in Online Befragungen zum Teil bekommt, gute Auswertungen erzeugen. Durch Reporting Werkzeuge kann man dann die Reizwörter bekommen, die man schließlich auch für die Finale Analyse braucht. Für eine Statistik sind Freitexte Felder für die Analyse im Normalfall schlecht zu gebrauchen, man benötigt reale Menschen, die die Vorstudien auswerten. Oft ist man auf die Wahrnehmung der Auswerter angewiesen. Lolap und die hierarchischen Thesauren erweitert so die Datenbasis auch von sehr vagen Papieren.

Ankopplung an Auswertetools
PL/SQL kann innerhalb von SQL als Abfrageerweiterung genutzt werden und steht im Schema dem User zur Verfügung. Wenn man also ein

Select lolap(‚String’) from t_text

Absendet, dann wird der String nach den Methoden von lolap bearbeitet. Das git auch für Verdichtungs- und Gruppierungswerkzeuge. Im Moment nutzen wir das Paket direkt über das PL/SQL Interface. Wir haben Views erstellt, die Lolap basiert sind und diese wiederum über Tools wie Materialized Views beschleunigt und optimiert. Grundsätzlich sehen wird aber auch die Möglichkeit, über Reporting Tools wie Oracle Reports oder Crystal Reports an die Daten zu gelangen und diese zu verdichten. Das nächste Projekt ist der Einsatz von Hyperion. Hyperion ist ein Data Ware House Tool, das speziell für AD Hoc Abfragen geeignet ist. Im Datenmodell von Hyperion wird mit Hierarchien gearbeitet. Der spezifische, hierarchische Thesaurus ist hier genau das Modell, dass in in das Denkmodell von Hyperion passt. Wir arbeiten an der automatisierten Überführung des hierarchischen Lolap Modells in das Hyperion Modell.

Realisierung
Wie haben Lolap als Paket realisiert. Das Package besteht aus einer Reihe von Funktionen, der Parameterraum und die Variabeln sind im Package gebunden. Wir arbeiten ausschließlich mit Funktionen, diese sind besser in SQL Code zu Integrieren als Prozeduren. Der Aufruf erfolgt mit der Vorsilbe lolap und das der Funktion. Lolap_Syn bildet Synonmye, Lolap_Koeln stellt den Aufruf des Kölner Algorithmus dar. Die Mandantenfähigkeit wird derzeit durch die Verwendung unterschiedlicher Schemen dargestellt und nicht durch Parameter.

Oberfläche
Wie verwenden nativ SQL als Oberfläche zu Lolap. Alternativ haben wir aber auch APEX als Fenster zur Anwendung im Einsatz.

Ergebnisse

Lolap ist ein Erweiterungs- Package dass sowohl die Suche als auch die Datenanalyse nach Scheer eröffnet. Zwar wird kein Textverständnis erzeugt aber doch die Erschließung von Texten ermöglicht – auf einfache und smarte Art und Weise. Im medizinischen Bereich böte sich Einbettung des UMS Methathesaurus
speziell für Orbis an. Orbis ist ein klinisches Arbeitsplatzsystem, das unter Oracle läuft. Eine Reihe von Auswertungen wäre über die erfassten Arzbriefe möglich. Man könnte zusätzlich noch ein Lolap Modul zur Anonymisierung erstellen. Das Problem bei der Auswertung von Krankheitsbildern ist, speziell im Orbis / KAS Umfeld, dass man ethische Richtlinien und Datenschutzaspekte zu beachten hat. Der große Vorteil einer solchen Analyse wäre aber, dass Erkenntnisse ohne spezielle Forschung zu gewinnen wären.

Kontaktadresse:

Matthias Faix
IPM Köln GmbH
Altenberger Str 19 - 21
D-50668 Köln

Telefon: +49 (0) 221 650 3657 10
Fax: +49 (0) 221 650 3657 20
E-Mail Matthias.Faix@ipm-koeln.de
Internet: www.ipm-koeln.de