Mögliches Projekt 2 / Bachelorarbeit

"Timeline-Search" - Projektarbeit?/Bachelorarbeit

Was meine ich mit "Timeline-Search":
Es soll eine Web-Applikation geben, wo jeder nach einem bestimmten Datum / Zeitraum und Ort in der Vergangenheit suchen kann. Das Resultat sind jegliche Ereignisse in diesem Zeitraum an diesem Ort oder andere Ereignisse, die die gesuchten Ereignisse beeinflusst haben. Dies dient einem Historiker oder Journalist, einfach nachforschungen zu machen und vor allem, könnte man so neue Erkenntnisse aus der Vergangenheit gewinnen, die man heute noch nicht weiss. Z.B wieso ein Weltkrieg ausgebrochen ist, weil vielleicht 2 Tage vorher noch ein Erdbeben in Österreich stattgefunden hat, welches die Lage angespannt hat, was man heute so noch nicht einbezogen hat.

Wikipedia bietet schon was ähnliches:

- http://de.wikipedia.org/wiki/1986#September
  - Man kann aber nur nach Jahr suchen!!
- Weitere Produkte die es schon gibt?.. Ich habe nichts derartiges gefunden..

Was sind meine Ziele:

- Es soll möglich sein, unabhängige Zusammenhänge früherer Geschehnisse zu erkennen und daraus eventuelle mögliche Zusammenhänge zu erkennen.
- Die Daten sollen für öffentlich Zugänglich sein, nicht nur lokal.
- Es soll eine einfache Web-API geschrieben werden, mit welcher die Daten durch jedermann erfasst werden kann.

Wenn will ich ansprechen?

- Historiker
- Journalisten
- interessierte private Personen

Ausgangslage

- *DWH auf Zeitpunkt fixiert (Vor Jahr 700 gibt es nur noch Jahresereignisse (nicht auf Tag))*
- Faktentabelle
    - Ereignisse
- Dimensionen
    - Zeit,
    - Ort,
    - GPS,
    - …?
- Algorithmus der die Zusammenhänge erkennt (z.B. einfacher Alg. an einem Tag mehr als 10 Ereignisse), damit nicht selbst jedes Datum / Zeitraum durchsucht werden muss

kann beliebig ausgebaut werden

==> Anfrage bei Peter Schwab??

- Generisches Datenmodell (Wie Teradata??) --> Gibt es Konzepte dafür?
    - Laufend neue Attribute - Falls einer Erbbebenstärke eingeben will --> Diese Daten auch erfassen um danach den Algorithmus zu verbessern!
    - Via Applikation lösen, falls übers Web “add Attribut” geklickt wird, automatisch in der DB ein neues Feld hinzufügen

Herausforderungen Daten beschaffen

- (Beschränken auf Zeit, oder Land)
- Erdbeben kann z.B. in einem bestimmten Format heruntergeladen werden
    - http://www.earthquake.ethz.ch/education/150_Jahre_ETH/box_feeder/
    - .hy4 -> dieses Format?? Bringt einem nicht viel..
- Export / Import aus wikipedia??
    - http://en.wikipedia.org/wiki/Wikipedia:Database_download

    - GUI: Wie will man Navigieren:
        - Sicher nach Datum und Ort Suchen
        - Google Maps um nach Ort zu suchen

==> Es muss eine definierte Schnittstelle definiert werden, es kann nicht wie bei gewöhnlichen BI- Projekten für jede Anforderung eine Schnittstelle geschrieben werden, sondern umgekehrt. Die welche Daten laden wollen, müssen in der definierten Schnittstelle imporieren.

Wichtig

- Granularität
    - Zeit:
        - Weltkieg dauerte mehrere Jahre
    - Ort:
        - Weltkrieg ganze Welt, Erdbeben nur lokal in einer Stadt

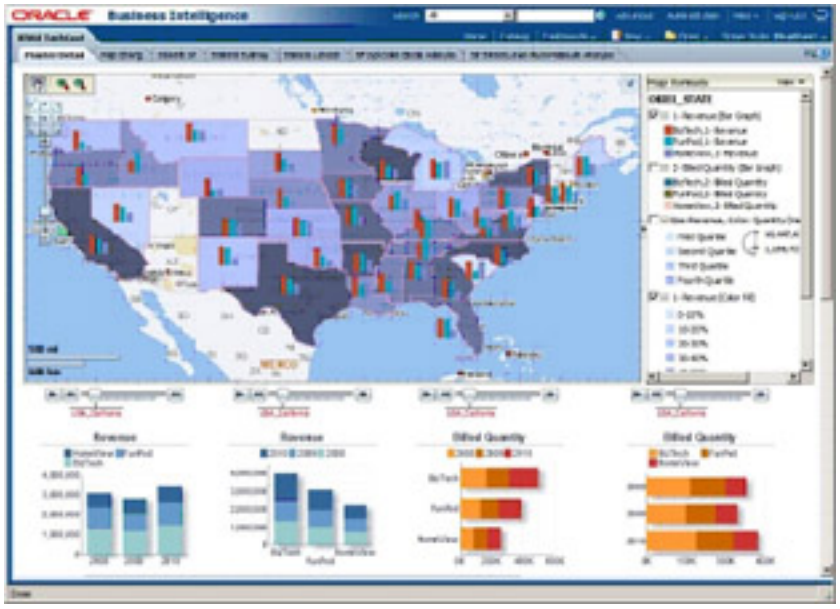
Zusätzliches (Optional)

- Mobile App
    - API um mit dem Handy Daten zu erfassen
    - Reports anzuzeigen
    - Abfragen zu machen
- Google Maps Appliance
    - Nach Ort suchen und sieht alle Ereignisse in diesem Ort
    - Man müsste die Zeit navigieren können (vor und zurück , in einem 1. Schritt vielleicht noch nicht)
    - https://developers.google.com/maps/visualize
    - earthquake magnitudes
    - More Examples: https://developers.google.com/maps/documentation/javascript/demogallery
- Reporting
    - Es sollen Reports mit einem BI-Tool erstellt und automatisch versendet (bursting) werden.

Technisch / Umgebung

- DB
    - MySQL (sspaeti.com / Synology NAS)
    - Oracle XE
        - (im Internet??, nur lokal oder?)
        - auf lokalem PC installieren und port 1251 (?) freigeben
    - Oracle Geo-Variante

    - Oracle Spatial and Graph (Kostet Lizenzgebühren)
    - Billigvariante in Oracle XE enthalten Location Intelligence



- Reporting
    - Falls Oracle -> BI-Publisher (Funktioniert auch auf XE odr MySQL) //Eher nicht
- Web-Sprache
    - Ruby on Rails? (Heroku)
    - Java JSF (Hosting, wo? - Glassfish/Tomcat Server auf Synology NAS)
    - Java Grails / Groovy ?

Microsoft

- SQL Server (?)

Vorgehensweise

1. Sauberes Datenmodell erstellen (Sehr wichtig! Nicht ganz Trivial!)
2. Erste Daten laden um ein Gefühl für das Ergebnis zu bekommen
3. Erste einfache Web-Applikation entwickeln (Ruby on Rails / Java EE mit JSF (Grails?))
4. Algorithmus entwickeln der die zusammenhänge erkennt -> Data Mining?
5. Google Maps Appliance integrieren inkl. einen Schieber für die Zeit, damit in der Zeit vor und zurück gewechselt werden kann.

Offene Fragen:

- Dieses Projekt im Zusammenhang mit BigData.. Eine Anwendungsmöglichkeit oder besser klassisches ERM?

Google Trend

http://www.google.ch/trends

BIG DATA – Projektarbeit?/Bachelorarbeit

Ansprechpartner Trivadis
- Heinz Steiner
- Roland Siposs ??
- PeterWelkenbach
- GuidoSchmutz
- NicholasJagger
- Ilias Ortega

Stichworte:

- MapReduce-Ansatz von Google
  - Open-Source-Software
    - Hadoop
    - MongoDb
  - Kommerziellen Produkten
    - Aster Data
    - Greenplum
    - u. a.

Beispiele

Für Unternehmen bietet die Analyse von Big Data die Möglichkeit zur Erlangung von Wettbewerbsvorteilen, Generierung von Einsparungspotentialen und zur Schaffung von neuen Geschäftsfeldern. Beispiele hierfür sind:

- zeitnahe Auswertung von Webstatistiken und Anpassung von Online-Werbemaßnahmen
- bessere, schnellere Marktforschung
- Entdeckung von Unregelmäßigkeiten bei Finanztransaktionen (Fraud-Detection)
- Einführung und Optimierung einer intelligenten Energieverbrauchssteuerung (Smart Metering)
- Erkennen von Interdependenzen in der medizinischen Behandlung
- Realtime-Cross- und Upselling im E-Commerce und stationären Vertrieb
- Aufbau flexibler Billingsysteme in der Telekommunikation

Verarbeitung von Big Data

Klassische relationale Datenbanksysteme sowie Statistik- und Visualisierungsprogramme sind oft nicht in der Lage, derart große Datenmengen zu verarbeiten. Für Big Data kommt daher eine neue Art von Software zum Einsatz, die parallel auf bis zu Hunderten oder Tausenden von Prozessoren bzw. Servern arbeitet. Hierbei gibt es folgende Herausforderungen:

- Verarbeitung vieler Datensätze
- Verarbeitung vieler Spalten innerhalb eines Datensatzes
- schneller Import großer Datenmengen
- sofortige Abfrage importierter Daten (Realtime-Processing)
- kurze Antwortzeiten auch bei komplexen Abfragen
- Möglichkeit zur Verarbeitung vieler gleichzeitiger Abfragen (Concurrent Queries)

Die Entwicklung von Software für die Verarbeitung von Big Data befindet sich noch in einer frühen Phase. Prominent ist der MapReduce-Ansatz, der in der Open-Source-Software Hadoop, MongoDb, sowie in einigen kommerziellen Produkten (Aster Data, Greenplum u. a.) zum Einsatz kommt.

Kritik

Kritik gibt es an "Big Data" vor allem dahingehend, dass die Datenerhebung und Auswertung oft nach technischen Aspekten erfolgt, also dass beispielsweise der technisch einfachste Weg gewählt wird, die Daten zu erheben und die Auswertung von den Möglichkeiten diese Daten zu verarbeiten begrenzt wird. Statistische Grundprinzipien wie das einer repräsentativen Stichprobe werden oft vernachlässigt. So kritisierte die Sozialforscherin Danah Boyd[1]:

- Größere Datenmengen müssten nicht qualitativ bessere Daten sein
- Nicht alle Daten seien gleich erzeugt
- "Was" und "Warum" seien zwei unterschiedliche Fragen
- Bei Interpretationen sei Vorsicht geboten
- Nur weil es verfügbar ist, sei es nicht ethisch

So ermittelte ein Forscher beispielsweise, dass Nutzer eines sozialen Netzes nicht mehr als 150 Freundschaften pflegen würden - was jedoch lediglich eine technische Begrenzung des Netzwerkes war[1]. Und sicherlich würde nicht jeder alle seine Facebook-Freunde in einem Interview als Freunde benennen - der Begriff eines "Freundes" auf Facebook gibt lediglich eine Kommunikationsbereitschaft an.
Ein anderer kritischer Ansatz setzt sich mit der Frage auseinander, ob Big Data das Ende aller Theorie bedeutet. Chris Anderson, Chefredakteur bei WIRED beschrieb 2008 das Glaubwürdigkeitsproblem jeder wissenschaftlichen Hypothese und jedes Modells bei gleichzeitiger Echtzeitanalyse lebender und nicht lebender Systeme. Korrelationen werden wichtiger als kausale Erklärungsansätze, die sich oft erst später bewahrheiten oder falsifizieren lassen. Chris Anderson in WIRED

Data Mining – Projektarbeit?/Bachelorarbeit

- Daten analysieren
    - irgend ein crawler mit java schreiben, der daten von Webseiten in einer DB speichert --> Ädu abegglen
- was kann man mit diesen anstellen
- vorhersagen (forecasting)