Data Integration:
Financial Domain-Driven Approach

Caslav Bozic¹, Detlef Seese², and Christof Weinhardt³

¹ IME Graduate School, Karlsruhe Institute of Technology (KIT) bozic@kit.edu
² Institute AIFB, Karlsruhe Institute of Technology (KIT) detlef.seese@kit.edu
³ Institute IISM, Karlsruhe Institute of Technology (KIT) weinhardt@kit.edu

Abstract. Finance practitioners and researchers rely heavily on accurate and accessible historical data. Practitioners require the data to evaluate trading and investing decisions. Researchers may use data to test market quality and efficiency. Unfortunately data is not error-free and is difficult to access and integrate with other sources. The ongoing project FINDS (Financial News and Data Service) is designed to fill this gap and provide clean, integrated and accessible data to both practitioners and researchers in finance. We achieve these goals via flexible data preprocessing and novel data preparation methods presented below.

Data integration includes the task of combining data residing at different sources and providing the user with the unified view of this data [1]. Formally, we can provide an integrated view - schema G - over several data sources - schemata S - using mappings M. Source S consists of data provided by Thomson Reuters TickHistory and Reuters NewsScope Sentiment Engine, as well as Compustat Data. What hampers analysis of underlying financial phenomena is the fact that the data is related but not linked in electronic form. We propose a generic interface that generates and stores mappings M and a simple grammar to define fields, which allows operations on numerous fields across S and the definition of new fields by means of arithmetic and lagging operators. Metadata repository stores types and formats of the source fields, and it is used for automated interface generation. Parser follows the definition of the calculated fields created by user, and generates code that provides complex calculations. This enables high performance preprocessing and provides financial data linked in the way that was not possible to achieve by available systems.

References

Keywords
DATA INTEGRATION, DATA PREPROCESSING, FINANCE
Data integration: Financial Domain-Driven Approach

Caslav Bozic (bozic@kit.edu), Detlef Seese, Christof Weinhardt

Applied Informatics and Formal Description Methods (AIFB)
Information Management and Market Engineering (IME)
Karlsruhe Institute of Technology (KIT), Germany

Agenda

(i) Motivation
(ii) Data Integration
(iii) Data Processing
(iv) Examples
(v) Summary
(vi) References
FINDS Project

Financial News and Data Service

- Conducting innovative research on the analysis of quantitative and qualitative information from financial markets
- Amount of financial data available (previous trades, news stories) makes it impossible for a human trader to process it in whole
- Services to help traders by
  - filtering important news releases
  - suggesting buy-sell decisions
  - allowing making subjective connections within the data

Text Mining Approaches
FINDS Text Classification Systems

- 3 classifiers
  - Bayes – Fisher
  - SVM
  - Neural Network

---

**Agenda**

(i) Motivation  
(ii) Data Integration  
(iii) Data Processing  
(iv) Examples  
(v) Summary  
(vi) References
Data Sources

- Thomson Reuters TickHistory
  - order book (some)
  - best bid and ask (most)
  - trades (all major exchanges)
  - indices values
- Standard&Poor's Compustat Database
  - fundamental data
- Reuters NewsScope Sentiment Engine
  - sentiment measure for all English-language news published through Reuters NewsScope in period 2003-2008
- Reuters Takes
  - full text of news stories for 2003

Data Flow

Thomson Reuters TickHistory → S&P Compustat → Reuters NewsScope Sentiment Engine

- Calculation & Aggregation according to formal definition
- Aggregation Transformation
- Clean Data
- Benchmarking
Data Integration

- Data integration includes the task of combining data residing at different sources and providing the user with the unified view of this data (Lenzerini 2002)
- data integration system \( I : \) triple \((G, S, M)\)
  - \(G\): global schema
  - \(S\): source schema
  - \(M\): mapping
- \(G\): final benchmarking dataset
- \(S = S_1 \cup S_2\): source databases
  - \(S_1\): Thomson Reuters TickHistory
  - \(S_2\): S&P Compustat
- \(M\): target mappings (global-as-view)

Agenda

(i) Motivation
(ii) Data Integration
(iii) Data Processing
(iv) Examples
(v) Summary
(vi) References
**Data Processing**

**Simple grammar for calculated fields definition**

\[
\text{<definition> ::= } \text{<source_table_name> } \\
\text{<partitioning_columns> } \\
\text{<date_columns> } \\
\{ \text{<field_definitions> } \} \\
\text{<additional_predicates> ::= } \text{<source_table_name> } \\
\text{<partitioning_columns> ::= } \text{<partitioning_columns> } \{ \text{<column> } \} \\
\text{<date_columns> ::= } \text{<date_columns> } \{ \text{<column> } \} \\
\text{<field_definitions> ::= } \text{<define> } \text{<variable> } = \text{<expression> } \{ ; \text{<define> } \text{<variable> } = \text{<expression> } \} \\
\text{<additional_predicates> ::= } \text{<where> } \text{<sql_where> } \\
\text{<expression> ::= } \text{<add_exp> } \\
\text{<add_exp> ::= } \text{<add_exp> } \text{<add_op> } \text{<mult_exp> } | \text{<mult_exp> } \\
\text{<add_op> ::= } + | - \\
\text{<mult_exp> ::= } \text{<mult_exp> } \text{<mult_op> } \text{<base_exp> } | \text{<base_exp> } \\
\text{<mult_op> ::= } * | / | \text{div} | \text{mod} \\
\text{<base_exp> ::= } \text{<base_exp> } \text{<unary_op> } \text{<base_exp> } | \text{<unary_exp> } \\
\text{<unary_op> ::= } + | - \\
\text{<unary_exp> ::= } \text{<unary_op> } \text{<base_exp> } | \text{<function> } \{ \text{<expression> } \} \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<variable> } \\
\text{<variable> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<function> } \\
\text{<function> ::= } \text{<identifier> } \\
\text{<identifier> ::= } \text{<table> } \\
\text{<table> ::= } \text{<identifier> } \\
\text{<column> ::= } \text{<identifier> } 
\]

**Definition file example**

```bash
source  instage.ins_quanti_mcc
partition "MIC"
date "Date[0]

define 
  "spread" = ("Ask"-"Bid")/("Ask"+"Bid");
  "oc" = ln("Last")/lag1("Last");
  "co" = ln("Open")/lag1("Last");
  "oc" = ("Last"/"Open");
  "Vol" = "Volume"

where not ("Open" is null or "Last" is null or "Bid" is null or "Ask" is null or "Last" is null) and not ("Open"=0 or "Last"=0)
```
Agenda

(i) Motivation
(ii) Data Integration
(iii) Data Processing
(iv) Examples
(v) Summary
(vi) References

Sentiment Data

- 6 Mio records about 10,000 different companies
- 2.5 times increase in yearly volume in period 2003 – 2008
- 2 biggest US markets (NYSE & NASDAQ)
  - 40% in 2003
  - 60% in 2008

Number of records per year
Sentiment Data

- Negative sentiment on Saturdays

Number of news items and average sentiment per days of week

Best and worst average sentiment for countries with over 1000 mentions
Regression Results

- Comparison
  - Classifiers trained on first 9 months 2003
  - 5 big technology companies: IBM, Oracle, Microsoft, Apple, SAP
  - Tested on 3 last months 2003 – 729 news stories
  - 9 variables – 10 lagged values each
  - Statistically relevant relation could be proven for
    - Bayes-Fisher: 2 values
    - SVM: 1 value
    - Neural network: 1 value
    - RNSE: 7 values
  - Not enough data to draw certain conclusions
  - Longer period needed, more companies
  - RNSE data for NYSE and NASDAQ in period 2003 - 2008
    - Statistical relevance improved

Agenda

(i) Motivation
(ii) Data Integration
(iii) Data Processing
(iv) Examples
(v) Summary
(vi) References
Summary

- FINDS Project
- Variety of financial text mining approaches creates the need for benchmarking method
- Proposed framework and implemented system for
  - Flexible integration of new data sources
  - Formal definition of calculated fields and aggregations

Data integration: Financial Domain-Driven Approach

Caslav Bozic (bozic@kit.edu), Detlef Seese, Christof Weinhardt

Applied Informatics and Formal Description Methods (AIFB)
Information Management and Market Engineering (IME)
Karlsruhe Institute of Technology (KIT), Germany

Thank you for your attention.

Discussion
Agenda

(i) Motivation
(ii) Data Integration
(iii) Data Processing
(iv) Examples
(v) Summary
(vi) References

References

References


