Data Integration: Financial Domain-Driven Approach

Caslav Bozic[1], Detlef Seese[2], and Christof Weinhardt[3]

[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

[1] LENZERINI, M. (2002): Data integration: a theoretical perspective. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, New York, 233–246.

Keywords
DATA INTEGRATION, DATA PREPROCESSING, FINANCE

Data integration: Financial Domain-Driven Approach

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

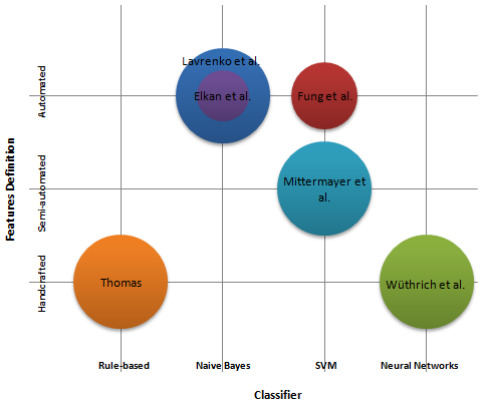
Applied Informatics and Formal Description Methods (AIFB)
Information Management and Market Engineering (IME)
Karlsruhe Institute of Technology (KIT), Germany

Agenda

FINDS Project

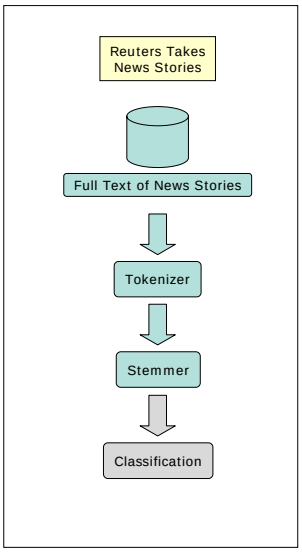
- Financial News and Data Service
  - Conducting innovative research on the analysis of quantitative and qualitative information from financial markets
  - Amount of financial data available (previous trades, news stories) makes it impossible for a human trader to process it in whole
  - Services to 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



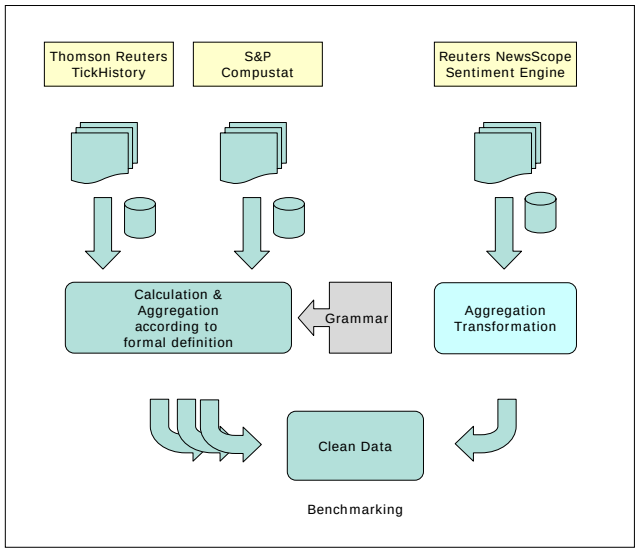
Agenda

(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



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

Data Processing

- Simple grammar for calculated fields definition

```
<definition> ::= <source_table_name>
                 <partitioning_columns>
                 <date_columns>
                 <field_definitions>
                 { <additional_predicate> }

<source_table_name> ::= source <table>

<partitioning_columns> ::= partition <column> {, <column>}

<date_columns> ::= date <column> {, <column>}

<filed_definitions> ::= define <variable> = <expression>
                            {; <variable> = <expression> }

<additional_predicate> ::= where <sql_where>
```

```
<expression> ::= <add_exp>

<add_exp> ::= <add_exp> <add_op> <mult_exp> | <mult_exp>

<add_op> ::= + | -

<mult_exp> ::= <mult_exp> <mult_op> <unary_exp> | <unary_exp>

<mult_op> ::= * | / | div | mod

<unary_exp> ::= <unary_op> <unary_exp> | <base_exp>

<unary_op> ::= + | -

<base_exp> ::= <variable> | <function> ( <expression> ) |
               <number> | <string> | null
```

```
<variable> ::= <identifier>

<function> ::= <identifier>

<table> ::= <identifier>

<column> ::= <identifier>
```


Data Processing

- Definition file example

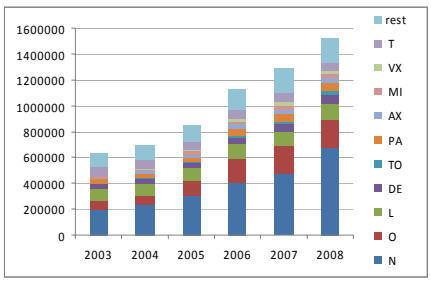
```
source instage.ins_quanti_mcc
partition "RIC"
date "Date[G]"
define
  "spread" = ("Ask"-"Bid")/("Ask"+"Bid");
  "cc" := ln("Last"/lag1("Last"));
  "oc" := ln("Open"/lag1("Last"));
  "oo" := ln("Open"/lag1("Open");
  "co" := ln("Last"/"Open");
  "Vol" := "Volume"
where
  not ("Open" is null or "Last" is null or "Bid" is null
  or "Ask" 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

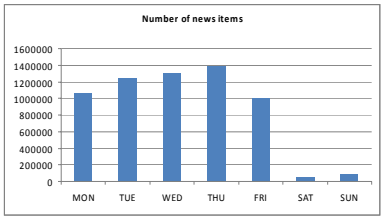
- 6 Mio records about 10,000 different companies
- 2.5 times increase in yearly volume in period 2003 – 2008
- 2 biggest US markets (NYSE & NASDAQ)
  - 40% in 2003
  - 60% in 2008

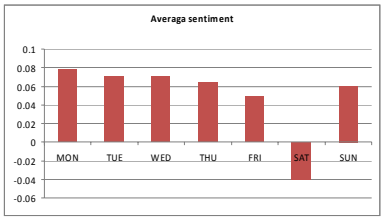


Number of records per year

Sentiment Data

- Negative sentiment on Saturdays

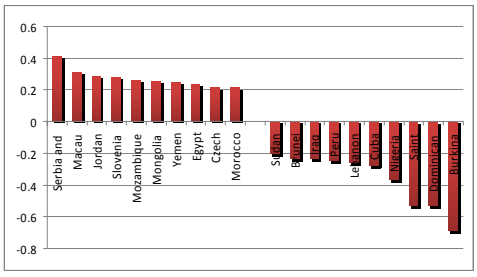




Number of news items and average sentiment per days of week

Sentiment Data

- Example of aggregation: sentiment per country



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

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

- [1] Hevner, A.R., March, S.T., Park, J. & Ram, S., Design Science in Information Systems Research, MIS Quarterly, Management Information Systems Research Center, University of Minnesota, 2004, Vol. 28(1), pp. 75-105
- [2] FINDS - Integrative services, Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on, 2009, pp. 61-62
- [3] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. & Allan, J., Mining of Concurrent Text and Time-Series, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000
- [4] Gidófalvi, G. & Elkan, C., Using news articles to predict stock price movements, Department of Computer Science and Engineering, University of California, San Diego, 2003
- [5] Pui Cheong Fung, G., Xu Yu, J. & Lam, W., Stock prediction: Integrating text mining approach using real-time news, Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on, 2003, pp. 395 - 402
- [6] Mittermayer, M.-A. & Knolmayer, G.F., NewsCATS: A News Categorization and Trading System, Data Mining, IEEE International Conference on, IEEE Computer Society, 2006, Vol. 0, pp. 1002-1007
- [7] Thomas, J., News and trading rules, 2003
- [8] Schulz, A., Spiliopoulou, M. & Winkler, K., Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informations\überlastung im Mobile Banking, Wirtschaftsinformatik, 2003, Vol. 2, pp. 181-200
- [9] Antweiler, W. & Frank, M.Z., Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, The Journal of Finance, Blackwell Publishing for the American Finance Association, 2004, Vol. 59(3), pp. 1259-1294

References

- [10] Das, S. & Chen, M., Yahoo! for Amazon: Sentiment extraction from small talk on the web, Management Science, INFORMS, 2007, Vol. 53(9), pp. 1375-1388
- [11] Tetlock, P., Giving Content to Investor Sentiment: The Role of Media in the Stock Market, THE JOURNAL OF FINANCE, 2007, Vol. 62(3)
- [12] Tetlock, P., Saar-Tsechansky, M. & Macskassy, S., More Than Words: Quantifying Language to Measure Firms' Fundamentals, Journal of Finance, American Finance Association, 2008, Vol. 63(3), pp. 1437-1467
- [13] Pfrommer, J., Hubschneider, C. & Wenzel, S., Sentiment Analysis on Stock News using Historical Data and Machine Learning Algorithms, Term Paper, 2010
- [14] Mittermayer, M. & Knolmayer, G., Text mining systems for market response to news: A survey
- [15] Wüthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J. & Lam, W., Daily prediction of major stock indices from textual www data, 1998
- [16] LENZERINI, M., Data integration: a theoretical perspective, Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.,ACM, New York, 233–246. 2002