Feature Extraction and Representation for Economic Surveys (FERES)

{Mihail.Minev, Christoph.Schommer, Theoharry.Grammatikos}@uni.lu, and Ulrich.Schaefer@dfki.de

UNIVERSITÉ DU LUXEMBOURG

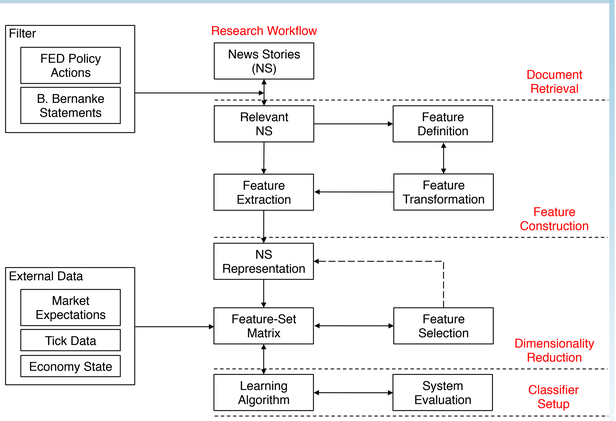
Contribution: Dimensionality reduction using advanced feat.

A novel method for identification and extraction of multi-word features, defined as attribute-value pairs, from official press releases. For the specific purpose, we use linguistic and statistical criteria in conjunction with financial domain knowledge.

Abstract

The study concerns the manifold news articles, which reflect the adjustments in the monetary policy during the financial crisis. In particular we consider official decisions conducted by the Federal Reserve System, but also information leaks in the press. One goal of this work is to retrieve and quantify such information using modern pre-processing and text mining techniques. Further the implications of news on the stock markets are examined by discovering and modelling composite index volatilities as functions of key announcements. A model for the prediction of price trends is targeted, which should reveal the economic value of information. Here, an important aspect is the definition, extraction, and management of topic-related features.

Research Model



Motivation + Research Quest.

M1: Can we quantify the decisions conducted by the Federal Reserve in official announcements without information loss?
M2: Does a measurable link exist between these and a leading stock market index?
M3: In this context, are market movements foreseeable, and if yes, to what extent?
RQ: How to identify and extract relevant terms, in the context of monetary policy decisions, which offer a valuable representation for economic surveys?

References

[Li10] Xiao Li. Understanding the semantic structure of noun phrase queries. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1337–1345, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[MSG12] Mihail Minev, Christoph Schommer, and Theoharry Grammatikos. News and stock markets: A survey on abnormal returns and prediction models. Technical Report UL-ARTICLE-2013-018, University of Luxembourg, August 2012.

[SWZ+07] M. Shafiei, Singer Wang, R. Zhang, E. Milios, Bin Tang, J. Tougas, and R. Spiteri. Document representation and dimension reduction for text clustering. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 770 –779, April 2007.

Dataset

NewsScope refers to an archive containing news texts in English published by Reuters.

- Date range: 2003 - 2012
- Volume: ∼700 MB/month
- Dataset entries: ∼479121 monthly
- Format: CSV, UTF-8
- Nineteen attribute codes, e.g. Date (GMT) (yyyy-mm-dd); Time (hh:mm:ss.000); Headline_alert_text; Take_text.

Tick Data refers to high frequency data of each individual price fluctuation for a security.

- Date range: 2003 - 2012
- Data format: CSV, UTF-8
- Eight attribute codes, e.g. Date [G], Time [G], Volume, Type, Price.

Federal Funds Futures are used to estimate the market expectations of future monetary policy changes conducted by the FED.

Workflow and Document Analysis

Scheme of Tasks

1. Filter
   Retrieval of documents nearly related to monetary policy decisions.
2. Feature construction
   A feature set, incorporating financial domain knowledge, is built to capture the enclosed information.
3. Document representation
   A vector representation where document features are indexed and weighted.
4. Feature-set-matrix
   A set of feature expressions. Include external data sources to further refine the model.
5. Learning algorithm
   Discover feature and stock market dependencies using a classifier.
6. Evaluation
   Model evaluation applying standard quality metrics.

Document → Features

- The initial data set is filtered on announcements conducted by the FED as well as interviews and statements of the central bank chairman Ben Bernanke.
- The representative feature set is built by transformations and combinations of the terms in the relevant news documents.
- Only the obtained set of features is used for the document representation.

Dimensionality Reduction

- For a relevant feature set, where p is number of dimensions and x is a random variable with $x = (x_1, ..., x_p)^T$.
  → Find a representation of lower dimensionality k, $s = (s_1, ..., s_k)^T$ with $k < p$ preserving the information content of the initial data according to a classification metric.

Construction of multi-word features

Noun Phrase identification

(Adjective|Noun)* (Noun Preposition)?
(Adjective|Noun)* Noun

Noun Phrase extraction

- First, *tagging* (POS-tagger & Chunker)
- Next, the features (attribute - value pairs) are determined using heuristic and syntax-based rules

Noun Phrase extension

- Adopts financial domain knowledge

The extended features enable tracking the content of the information using *open class words*, e.g.

$$interest\ \&\ rate\ \rightarrow class\ A$$
$$interest\ rate\ \rightarrow class\ B$$

+ Scalability; Noise reduction, while retaining the meaning; Capture of domain knowledge
− The features are set descriptive, limited; No universally accepted evaluation metrics for contextual relevance

A feature is built upon following schema

1. Intent Head (IH)
2. Intent Modifiers (IM)
3. Class C, which describes the type of IM

Extract from a FED announcement, 20.06.2012:
"However, growth in employment has slowed in recent months, and the unemployment rate remains elevated. Business fixed investment has continued to advance. Household spending appears to be rising at a somewhat slower pace than earlier in the year."

1. [IH :employment]
   [IM :$_{past_and_current_state}$growth_has_slowed]
   [IM :$_{time_period}$recent_months]
2. [IH :unemployment_rate]
   [IM :$_{past_and_current_state}$remains_elevated]
   [IM :$_{time_period}$recent_months]
3. [IH :business_fixed_investment]
   [IM :$_{past_and_current_state}$continued_to_advance]
   [IM :$_{time_period}$unknown]
4. [IH :household_spending]
   [IM :$_{past_and_current_state}$rising_somewhat_slower_pace]
   [IM :$_{time_period}$recent_months]