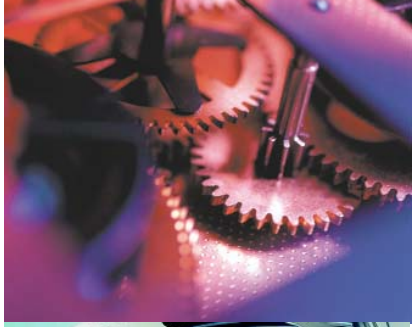


FHM
Fachhochschule des Mittelstands

Unsupervised Neural Net based automatic trend analysis for Weblogs

Johannes Fordemann
Manfred Leisenberg
Timo Timm
Julia Wolff

Motto

“Irrationality of human communication
Might convert into a swarming attitude
after a while”

[Lehmann2005]

…. Let us discover and understand *swarming attitudes*

Agenda

- Introduction
- Initial situation
- Trend analysis based on Neural Networks
- Experimental results
- Summary

Trends are reflected by Weblogs

- Weblog based communication produces and reflects socio- cultural streams
- Streams might produce „Weak Signals“
- „Weak Signals“ might indicate trends
- Trends are based on pattern
  - From such pattern we might learn the future direction of individual action

- **Task**
  - Pattern Identification and Classification
    (which is an native AI (Artificial Intelligence) task)

Identification of trends

- Trend occurs, if
  - Social coincidences **turn** into collective phenomena
  - Weak Signals **turn** into Tipping Points

→ Effective computer based detection methods for future trends have **not** been found yet [Lehmann 2005]

→ This research project aims at solving above detection problems

Market for Automatic Trend analyses

Portal	Advantages	Disadvantages
technorati	• statistics provided • blog search	• no trend analysis
blogpulse	• multilingual • statistical monitoring based analysis (featured and individual)	• no prediction capabilities
ArgYou	• Offer-Demand Comparison • Measurement of potential demand • Supervised statistical monitoring on existing trends	•Not specialized on trend prediction •Part-manual process
blogscout	•Statistical analysis of impressions and visits	•no trend analysis and prediction
gridpatrol	• Automatic monitoring and result analyses • Based on sophisticated grid technology	• Trend analyses for financial applications

Initial Situation

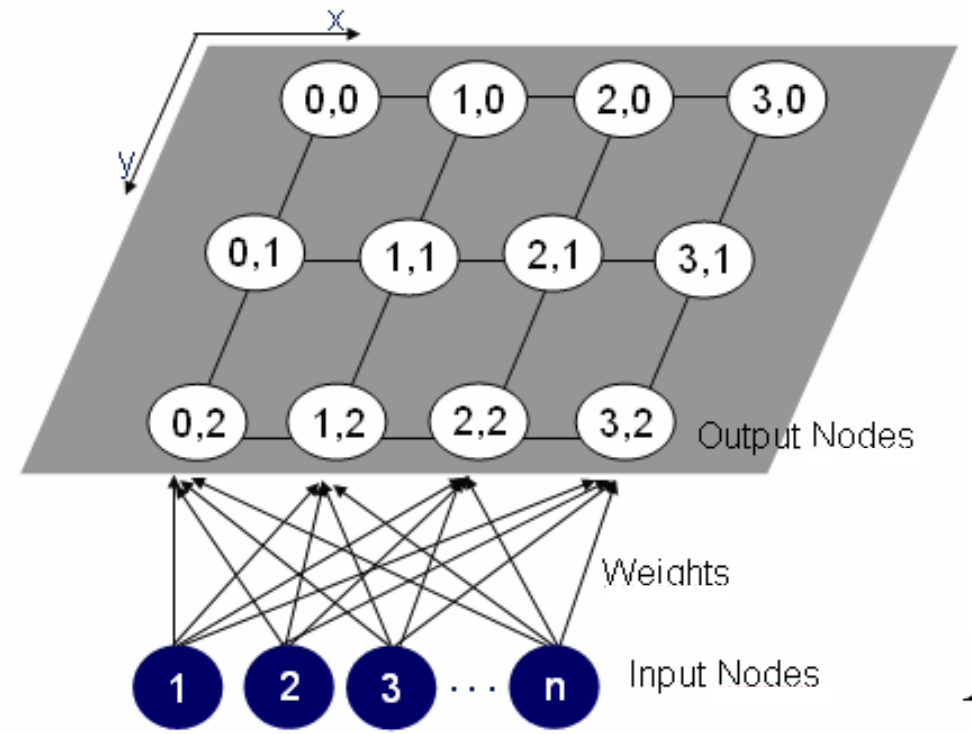
- Disadvantages of „classical“ statistical methods
  - Supervised processing
  - Not sensitive to weak signals
  - Redundancies , imprecision

- Therefore:
  - Unsupervised Analysis of Weblogs based on Self Organizing Maps (SOM)
  - Investigation of capabilities of SOM concerning trend analyses

Trend analyses based on SOM

- Self Organizing Maps (SOM)
  - Unsupervised learning based classifier
  - Classes are not known prior to the training process
  - Capable of identifying clusters/ classes in a multidimensional feature vector space automatically
  - Data for training and test are described by multidimensional feature vectors
  - Classification tendencies might be interpreted as s „Weak Signals“ or „Tipping Points“ in terms of trend analyses

SOM architecture



$$X_i = \{x_{1i}, x_{2i}, \ldots, x_{ki}, \ldots x_{ni}\}$$

- Similiar feature combinations are represented by close regions on the map

Trend analyses based on SOM

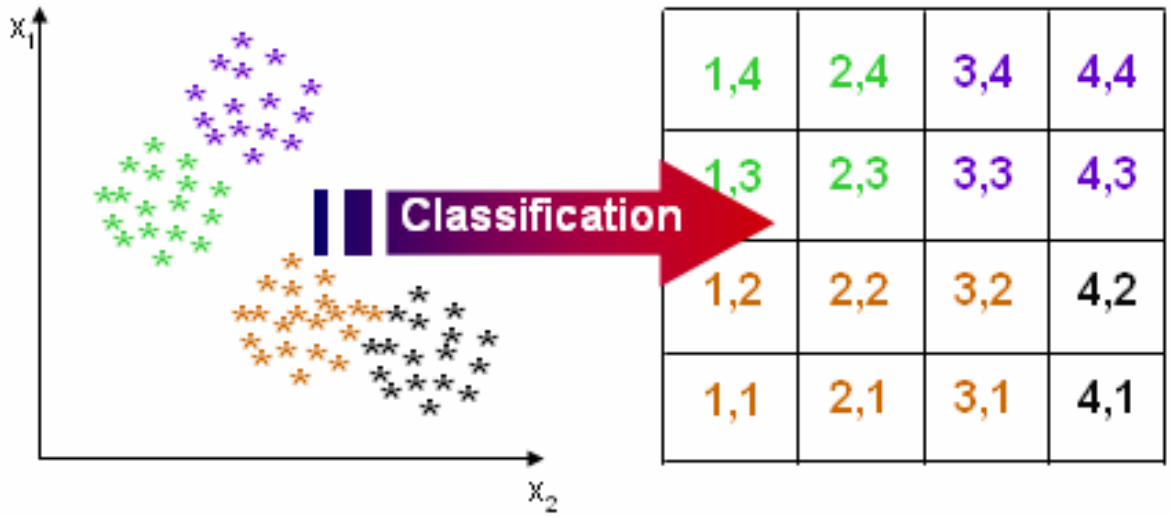
- Features are
  - Multidimensional quantative descriptions of sequences of words
  - Represented by feature vectors

$$X_i = \{x_{1i}, x_{2i}, \ldots, x_{ki}, \ldots x_{ni}\}$$

- Example
  - Context based description of „Items“
    - X_1=Item code, x_2=Owner, x_3=Price, X_4=Colour, …
  - Sequence oriented description
    - X_1=Wordcode1, x_2=Wordcode2, x_3=Wordcode3, …

SOM training process

- SOM capable of finding clusters in the input feature vector space
- Clusters might be labelled
- SOM represents distribution of feature vectors X_i



for presentation purposes:
2- dimensional input feature vector space

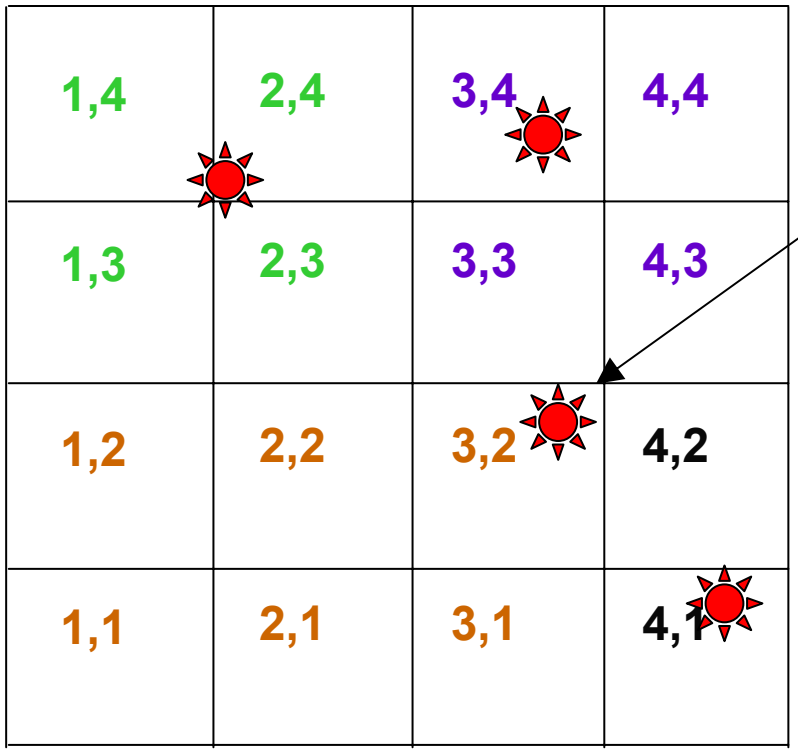
How to find trends by SOM?

- Output nodes represent Weblog clusters
- Weak Signal
  - Indicated by deviation from the centre of class during test mode
- Trend, „Tipping Point“:
  - statistically **significant** deviation from the centre of class during test mode
  - Threshold to be defined

SOM based trend analyses

ector for training

$$X_i = \{x_{1i}, x_{2i}, \ldots, x_{ki}, \ldots x_{ni}\}$$



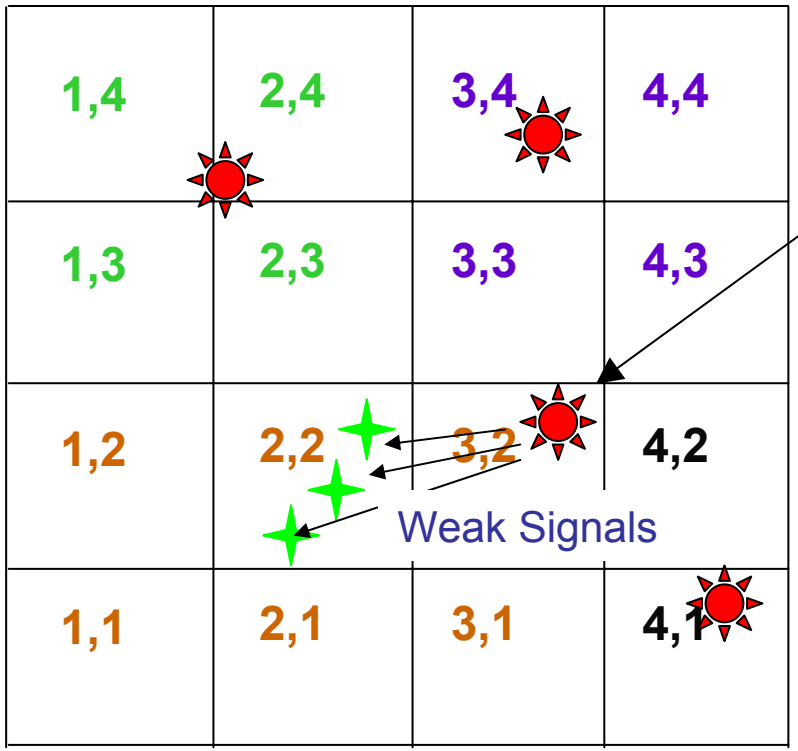
Trained SOM
$t=t_0$

Centre of class
(Strong after- training representatio
of a particular feature)

SOM based trend analyses

ector for training

$$X_i = \{x_{1i}, x_{2i}, \ldots, x_{ki}, \ldots x_{ni}\}$$



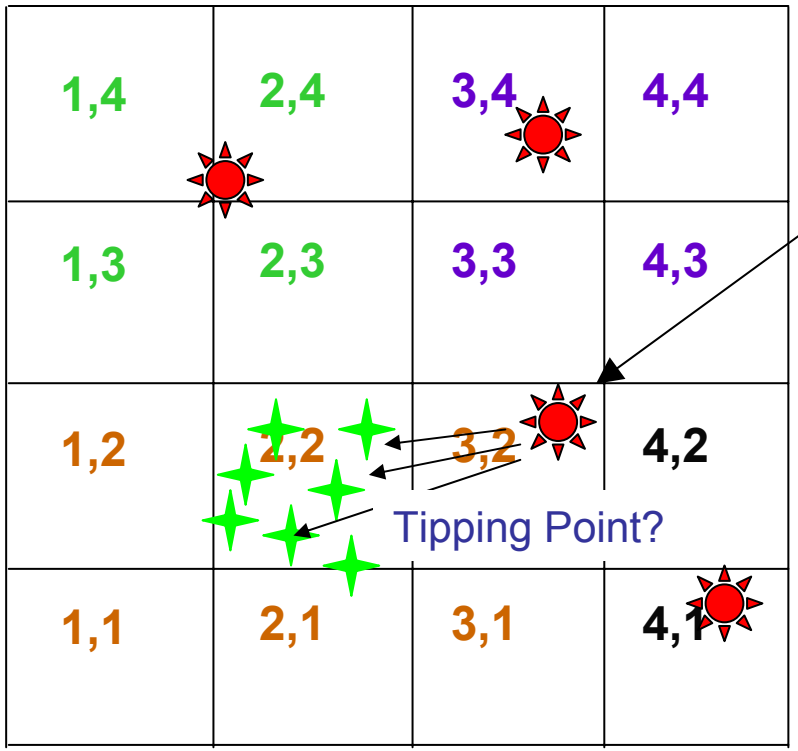
SOM in **Testmode**
$t=t_0+t_{t1}$

Centre of class
(Strong after- training representatio
of a particular feature)

SOM based trend analyses

ector for training

$$X_i = \{x_{1i}, x_{2i}, \ldots, x_{ki}, \ldots x_{ni}\}$$



SOM in **Testmode**
$t=t_0+t_{t1}+t_{t2}$

Centre of class
(Strong after- training representatio
of a particular feature)

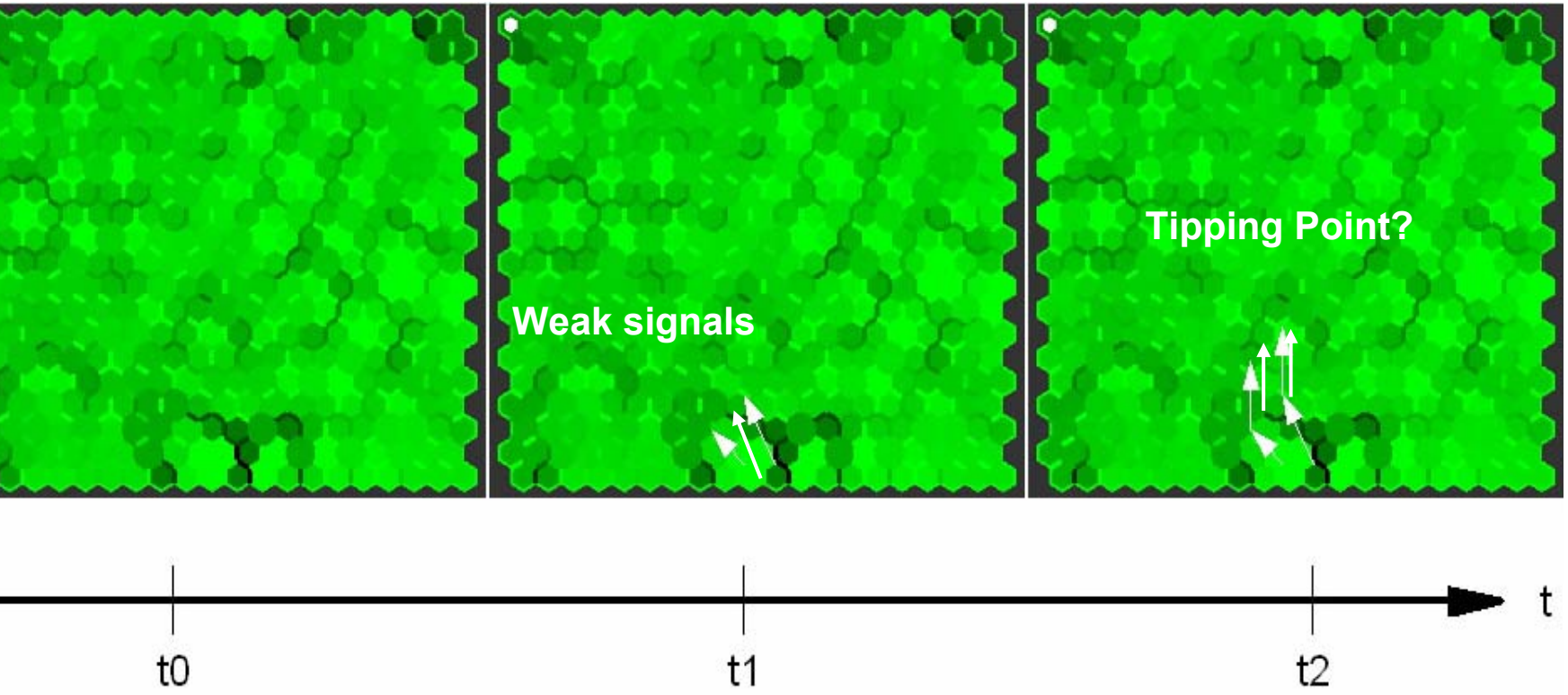
Trend?!

Experimental Result

- Procedure

1. Setup corpus for training and test
2. Identify and extract features
3. Define SOM parameters
4. Training of SOM
5. Testing of SOM
6. Identify Weak Signals, Tipping Points

Experimental Result



10 different souce texts
4*5=20 features (Sequence oriented description)
SOM 20*20

Summary

- Weblogs reflect socio- cultural streams -> trends
- SOM based identification of trends is a promising method
- Most important feature of the method
  - Unsupervised processing
  - Classes do not have to be known prior to training
- First experimental results support technical idea
- Problem: effective feature extraction