

Chapter 8: Analysis of Multiple Process Perspectives

Contents

1 Introduction

2 Social network analysis

3 Organizational mining

4 Decision mining

5 Text mining

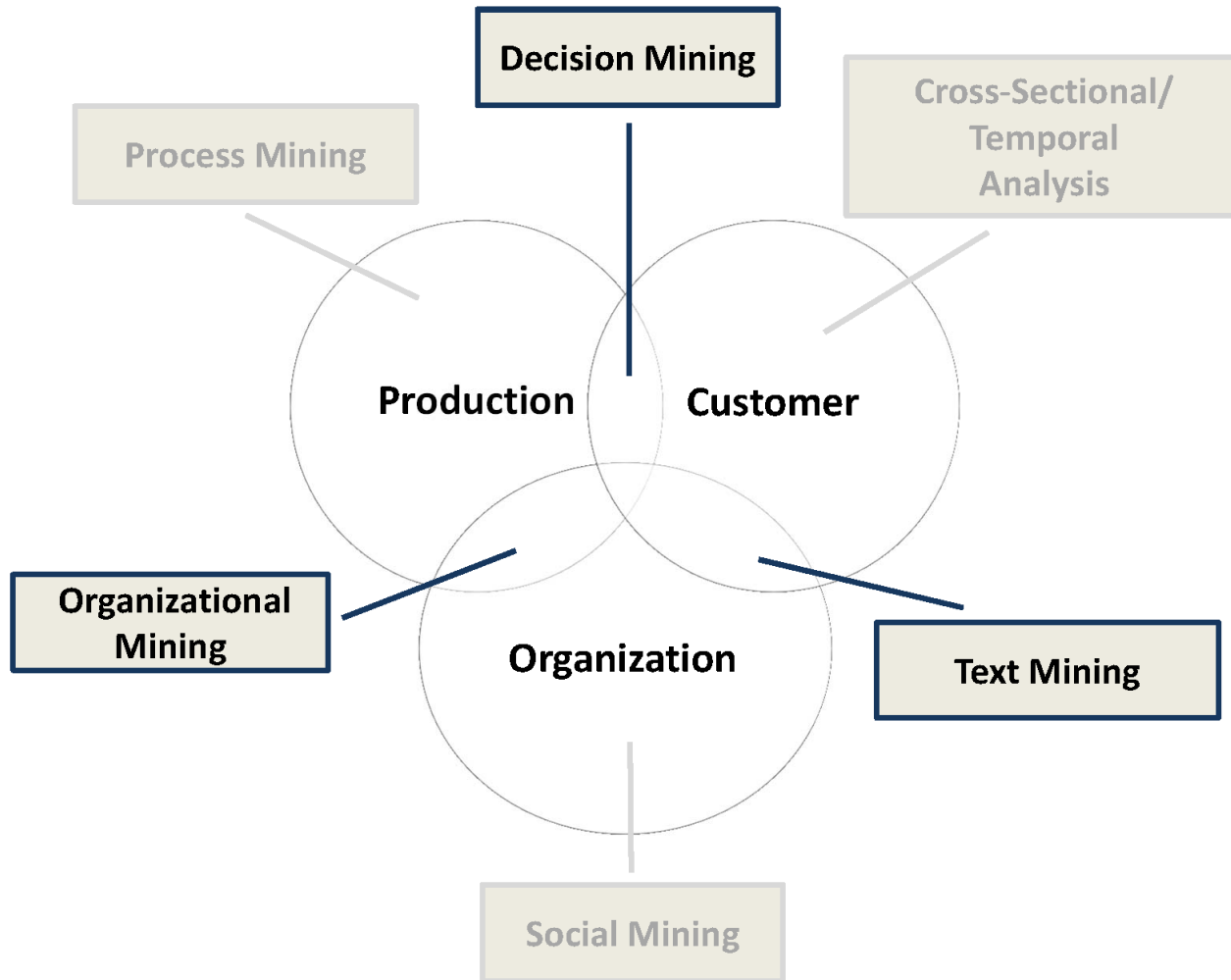
6 Summary & outlook

References

1 Introduction

- Many analysis questions can be answered by using one type of technique, e.g., cross-sectional analysis
- Other questions may require the combined application of analysis techniques
- Why? Because they touch more than one analysis perspective (see next slide)
- Examples:
 - Which roles do participate in the higher education process?
 - → Touches process and organizational perspectives
 - For which containers did the vehicle return to the origin?
 - → Touches production and customer perspectives

1 Introduction



© 2015 Springer-Verlag Berlin Heidelberg

Contents

1 Introduction

2 Social network analysis

3 Organizational mining

4 Decision mining

5 Text mining

6 Summary & outlook

References

2 Social network analysis

- Enormous amounts of „social data“ available through, e.g., social networks
- Possibility for asking new questions:
 - Who is interacting with whom?
 - Whom am I interacting with?
- Where „interacting“ can be any kind of „social relation“, e.g., owe money, hands over work, etc.
- Recall the three BI perspectives
 - Customer
 - Organization
 - Production
- → Social network analysis focuses on organizational perspective

2 Social network analysis

- *Data*: database containing social entities and their relations
- *Analytical goals*:
 - Visualization of relations between entities
 - Describing relationships by summary measures
 - Finding patterns in relationships
- *Modeling task*: first generate data matrix with social entities as nodes and relations as edges; then generate graphical representation as sociogram¹
- *Analysis task*: analyze sociogram using different metrics
- *Evaluation and reporting task*: visualize sociogram and descriptive measures

¹John Scott: Social Network Analysis. SAGE (2012)

2 Social network analysis

Data + Model

Example 1: Building model from relational data

Students	<u>SID</u>	Name	enrolled	<u>SID</u>	<u>UID</u>	University	<u>UID</u>	Name
	S1	Simon		S1	U1		U1	Univie
	S2	Maria		S2	U1		U2	TUWien
	S3	Frank		S1	U2		U3	WUWien
	S4	Sally		S3	U3			
	S5	Bert		S3	U2			
				S2	U2			

		Cases				
		S1	S2	S3	S4	S5
Cases	S1	-	2	1	-	-
	S2	2	-	1	-	-
	S3	1		-	-	-
	S4	-	1	-	-	-
	S5	-	-	-	-	-

2 Social network analysis

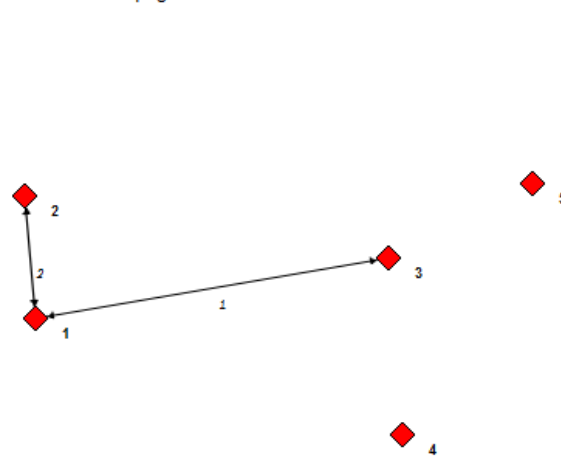
Data + Model

Example 1: Building model from relational data

		Cases				
		S1	S2	S3	S4	S5
Cases	S1	-	2	1	0	0
	S2	2	-	1	0	0
	S3	1		-	0	0
	S4	0	1	0	-	0
	S5	0	0	0	0	-

```
*Network
*Vertices 5
1 "Simon"
2 "Maria"
3 "Frank"
4 "Sally"
5 "Bert"
*Edges
1 2 2
1 3 1
```

SocNetV: RelUni.png



2 Social network analysis

Data + Model

Example 2: Building model from log data

```
<AuditTrailEntry>
  <WorkflowModelElement>Evaluate presentation 1</WorkflowModelElement>...
  <Originator>person001-lecturer</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>Evaluate presentation 1</WorkflowModelElement>...
  <Originator>person003-lecturer</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>plus</WorkflowModelElement>...
  <Originator>person003-lecturer</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>plus</WorkflowModelElement>...
  <Originator>person004-lecturer</Originator>
</AuditTrailEntry>.000+01:00</Timestamp>
```

**Event Type and Time Stamp
omitted**

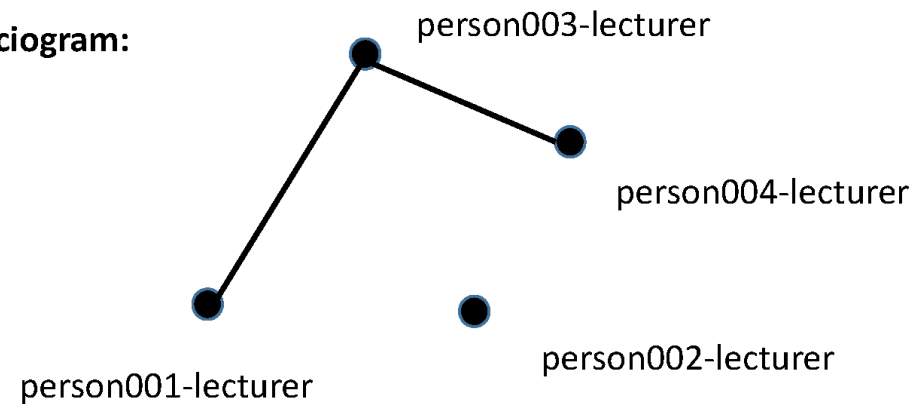
2 Social network analysis

Data + Model

Data Matrix:

	Evaluate presentation 1	plus
person001-lecturer	1	0
person002-lecturer	0	0
person003-lecturer	1	1
person004-lecturer	0	1

Sociogram:



© 2015 Springer-Verlag Berlin Heidelberg

2 Social network analysis

Model

- As mentioned before, the basic model is the sociogram
- Model structures for SNA (based on [GrRi])
 - *Undirected graphs*: an undirected graph G is defined as $G = (V;E)$ with set of nodes V and set of undirected edges E .
 - *Directed graphs*: Opposed to undirected edges, directed edges establish a relation that reflects a causal relation or a relation that is directed from one to another entity.
 - *Weighted Graphs*: It can be also useful to assign weights to the edges in the graph, i.e., a weight $w(e)$ expressing some kind of quantitative measure for the relation.
 - *Connected Subgraphs*: Special connected subgraphs might be of interest. A subgraph consisting of two nodes (with or without relations between them) describes a *dyad*, a sub-graph consisting of three nodes of interest a *triad* respectively.
 - *Dyad / triad*: Two / three actors who are connected by a relation in the social network

2 Social network analysis

Measures (selection)¹

- **Local:** related to single nodes in the sociogram
 - *Degree centrality:* degree, indegree, outdegree
 - Example: $\text{degree}(\text{person003_lecturer})=2 \rightarrow$ connected to 2 other nodes
 - Also relative to overall number of nodes
 - Example: $\text{reldegree}(\text{person003_lecturer})=2/4=0.5$
 - *k-path centrality:* number of paths of length k originating from the node
 - *Closeness:* how many shortest paths between to other nodes
- **Global:** refers to the entire sociogram $G=(V, E)$
 - Density: $\text{dens}(G) = \frac{2*|E|}{|V|*(|V|-1)}$
 - Example: $\text{dens}(\text{logsociogram}) = 1/3$

¹John Scott: Social Network Analysis. SAGE (2012)

Contents

1 Introduction

2 Social network analysis

3 Organizational mining

4 Decision mining

5 Text mining

6 Summary & outlook

References

3 Organizational mining

Example 2: Building model from log data

```
<AuditTrailEntry>
  <WorkflowModelElement>Evaluate presentation 1</WorkflowModelElement>...
  <Originator>person001-lecturer</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>Evaluate presentation 1</WorkflowModelElement>...
  <Originator>person003-lecturer</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>plus</WorkflowModelElement>...
  <Originator>person003-lecturer</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>plus</WorkflowModelElement>...
  <Originator>person004-lecture</Originator>
</AuditTrailEntry>.000+01:00</Timestamp>
```

Goal: evaluate both perspectives, i.e., production and organization together!
→ **Combine process mining with social network mining**

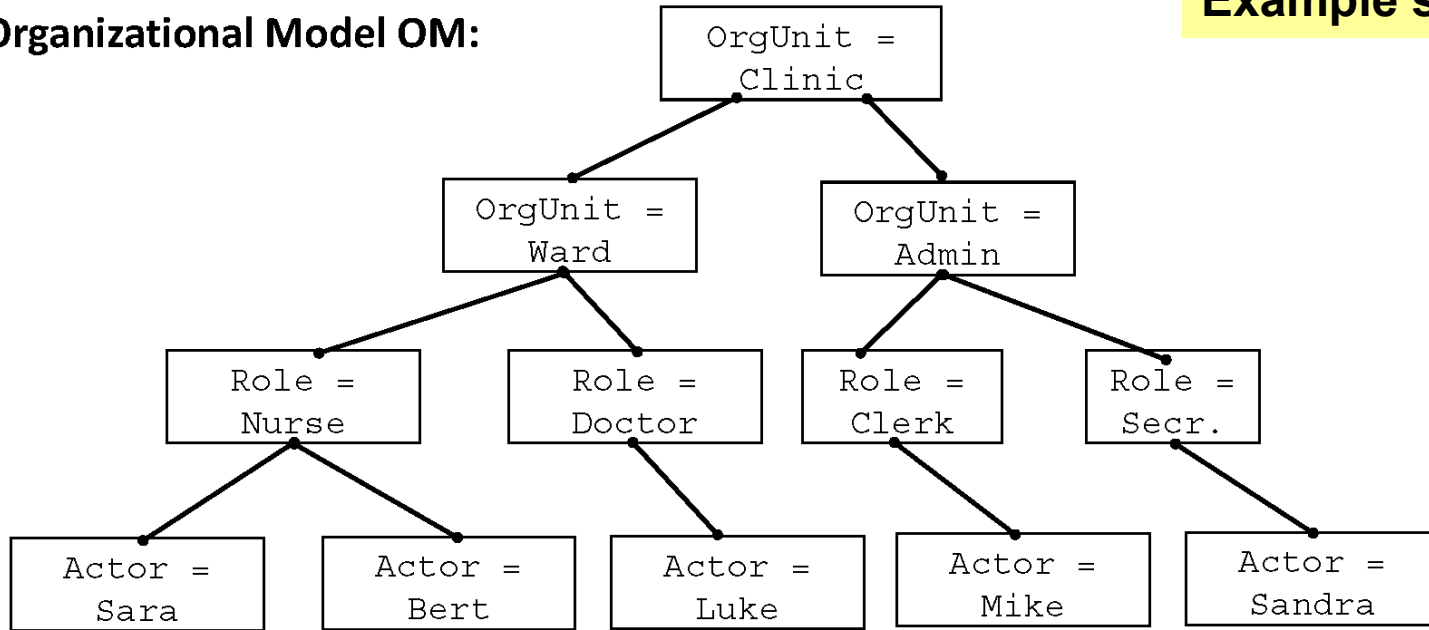
3 Organizational mining

- *Data*: process log containing information on originator, possibly additionally organizational model, organigram
- *Analytical goals*:
 - Describing and visualizing the relation between originators
 - Deriving organizational structures behind the process (organigram)
- *Modeling task*: generate actor profiles and sociogram from the log
- *Analysis task*: cluster analysis on the actor profiles, applying measures to sociogram
- *Evaluation and reporting task*: visualize sociogram and descriptive measures, visualize organigrams

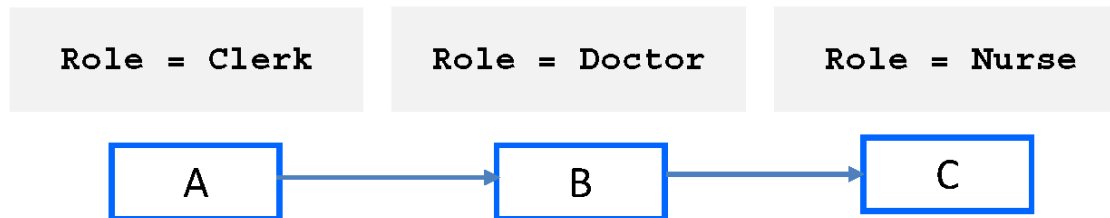
3 Organizational mining

Example setting

Organizational Model OM:



Process Model PM:



© 2015 Springer-Verlag Berlin Heidelberg

3 Organizational mining

Example setting

10 instances, producing the following log snippet

```
<AuditTrailEntry>
  <WorkflowModelElement>A</WorkflowModelElement>
  <EventType>start</EventType>
  <Timestamp>...</Timestamp>
  <Originator>Mike</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>B</WorkflowModelElement>
  <EventType>start</EventType>
  <Timestamp>...</Timestamp>
  <Originator>Luke</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <WorkflowModelElement>C</WorkflowModelElement>
  <EventType>start</EventType>
  <Timestamp>...</Timestamp>
  <Originator>Sara</Originator>
</AuditTrailEntry>
```

3 Organizational mining

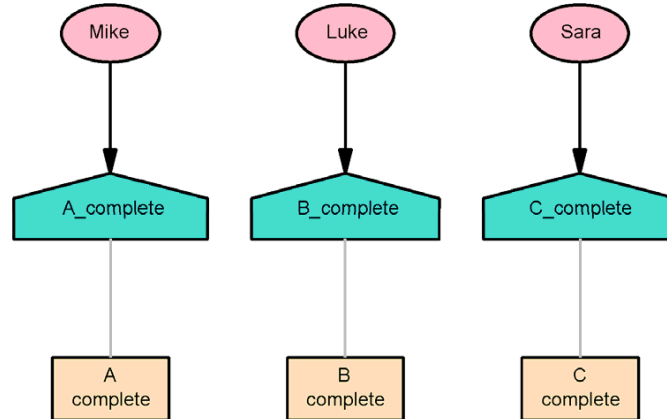
Example setting

- Let the actor profiles look as follows:
 - Mike (10,0,0) or $10*(1,0,0)$
 - Luke (0,10,0) or $10*(0,1,0)$
 - Sara (0,0,10) or $10*(0,0,1)$
- Meaning: Mike performed task A for all 10 instances, but did not work on tasks B and C
- Observation: Bert has never performed C, even though he qualifies for it
- Cluster analysis is applied to profiles, e.g., k-means²
- Results for example on next slide

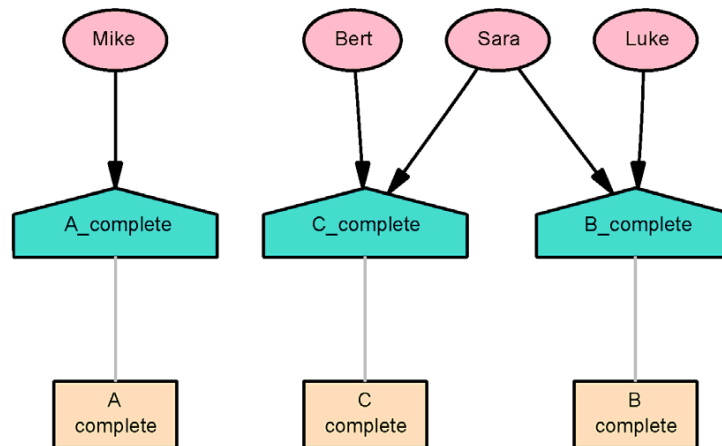
²Minseok Song, Wil M. P. van der Aalst: Towards comprehensive support for organizational mining. Decision Support Systems 46(1): 300-317 (2008)

3 Organizational mining

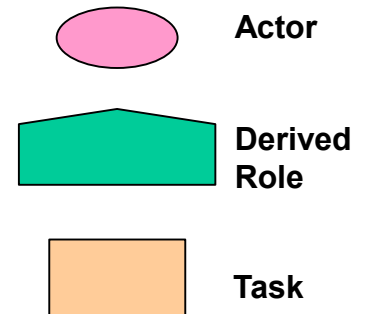
a) Organizational Miner, Profiles Mike(10,0,0), Luke(0,10,0), Sara(0,0,10)



b) Organizational Miner, Profiles Mike(10,0,0), Bert(0,0,5), Luke(0,5,0), Sara(0,5,5)



Using ProM 5.2



Example Setting with Changed User Profiles

© 2015 Springer-Verlag Berlin Heidelberg

3 Organizational mining

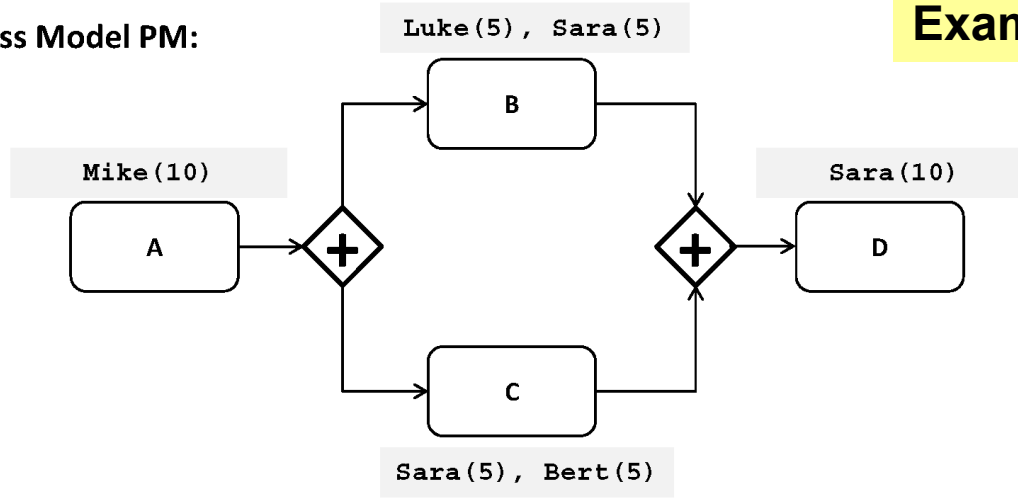
Example setting

- For both settings role labels have to be adapted, e.g.,
A_complete to Clerk
- a) confirms task / role assignment
- For b) three interpretations are conceivable, i.e.,
 - Sara has two roles Doctor and Nurse OR
 - Task B has an actor assignment $B \leftarrow \text{Role} = \text{Doctor AND Role} = \text{Nurse}$ OR
 - Runtime deviations

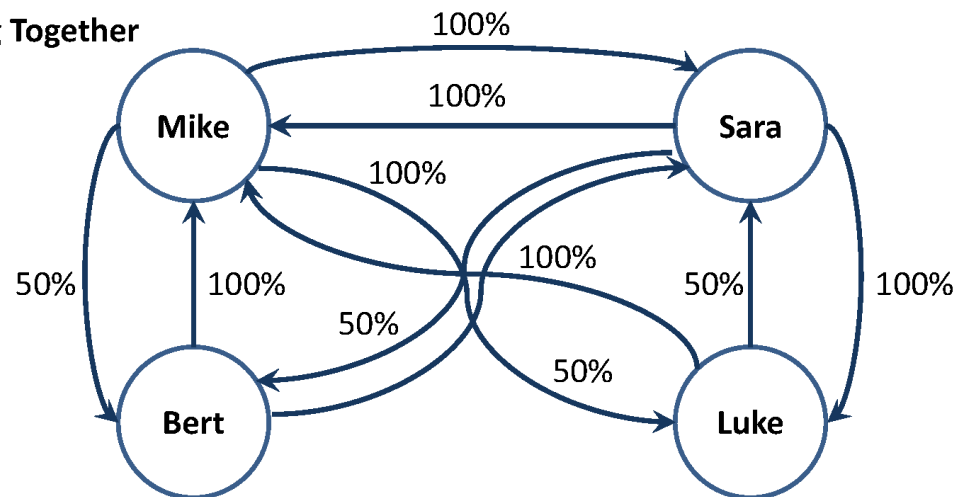
3 Organizational mining

Example setting

Process Model PM:



Working Together



© 2015 Springer-Verlag Berlin Heidelberg

Contents

1 Introduction

2 Social network analysis

3 Organizational mining

4 Decision mining

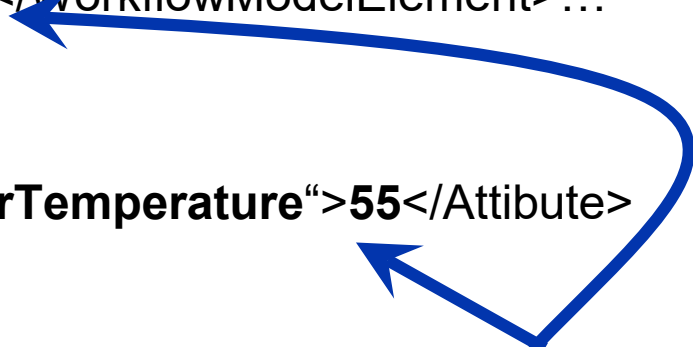
5 Text mining

6 Summary & outlook

References

4 Decision mining

```
<AuditTrailEntry>  
  <WorkflowModelElement>Move to</WorkflowModelElement>...  
  <Originator>unknown</Originator>  
  <Data>  
    <Attribute name=“ContainerTemperature”>55</Attribute>  
  </Data>  
</AuditTrailEntry>
```

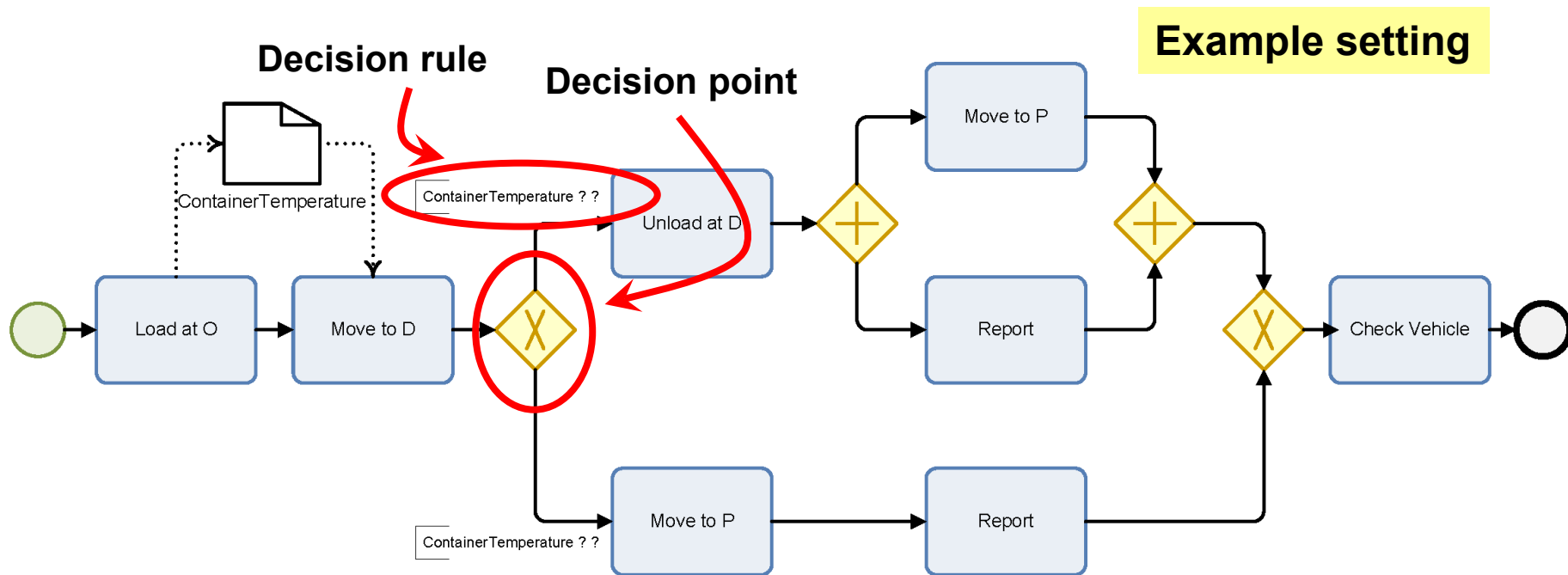


Goal: evaluate both perspectives, i.e., production and customer, together!
→ **Combine process mining with cross-sectional analysis**

4 Decision mining

- *Data*: process log containing information on data attributes and values written during execution
- *Analytical goals*:
 - Deriving decision points by classification (process mining)
 - Deriving decision rules at these points by cross-sectional analysis methods
- *Modeling task*: generate process models with decision points and cross-sectional models
- *Analysis task*: process discovery and decision trees
- *Evaluation and reporting task*: visualize process with decision points, visualize results of cross-sectional analysis, display decision rules

4 Decision mining



© 2015 Springer-Verlag Berlin Heidelberg

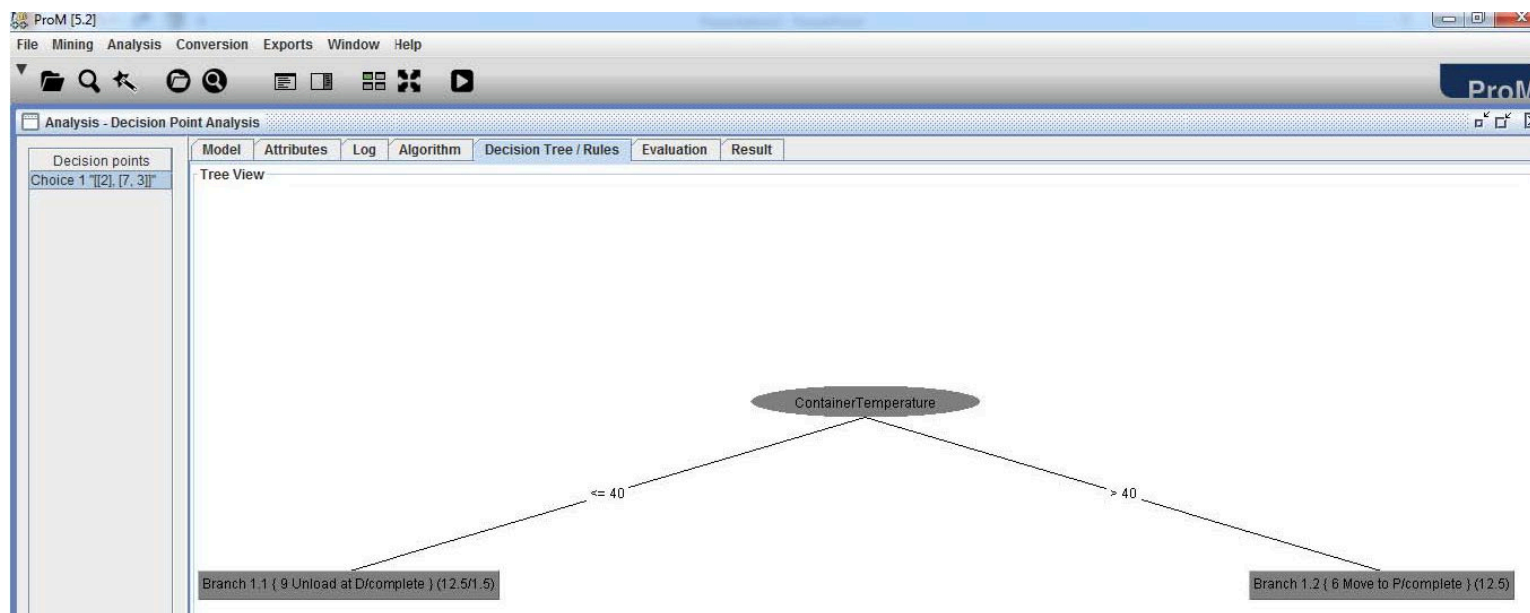
```

<AuditTrailEntry>
  <WorkflowModelElement>Move to</WorkflowModelElement>...
  <Originator>unknown</Originator>
  <Data>
    <Attribute name="ContainerTemperature">55</Attribute>
  </Data>
</AuditTrailEntry>
  
```

4 Decision mining

Example setting

- Build decision tree at decision point based on the available data attributes and values³
- In the example: container temperature

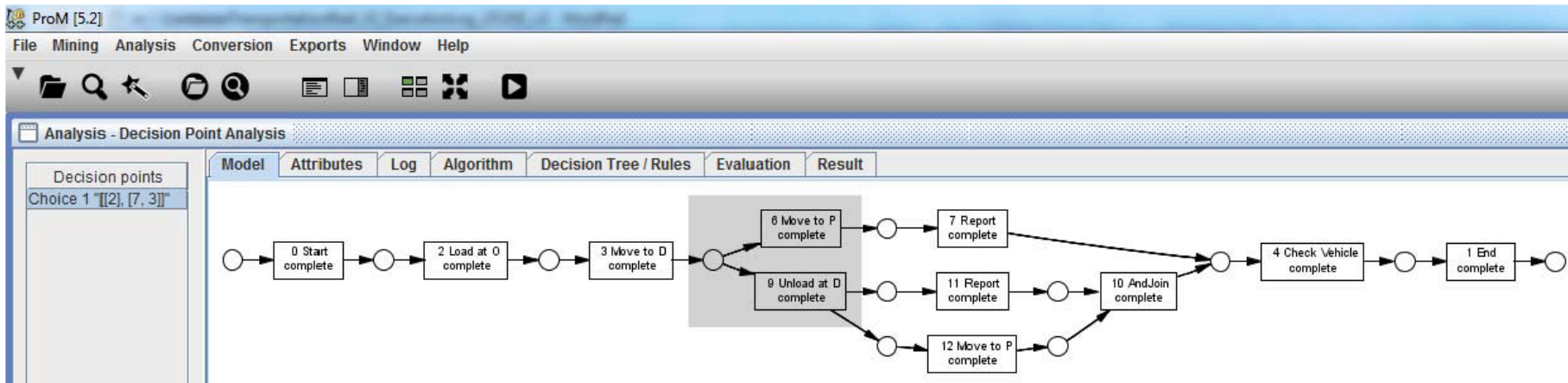


© 2015 Springer-Verlag Berlin Heidelberg

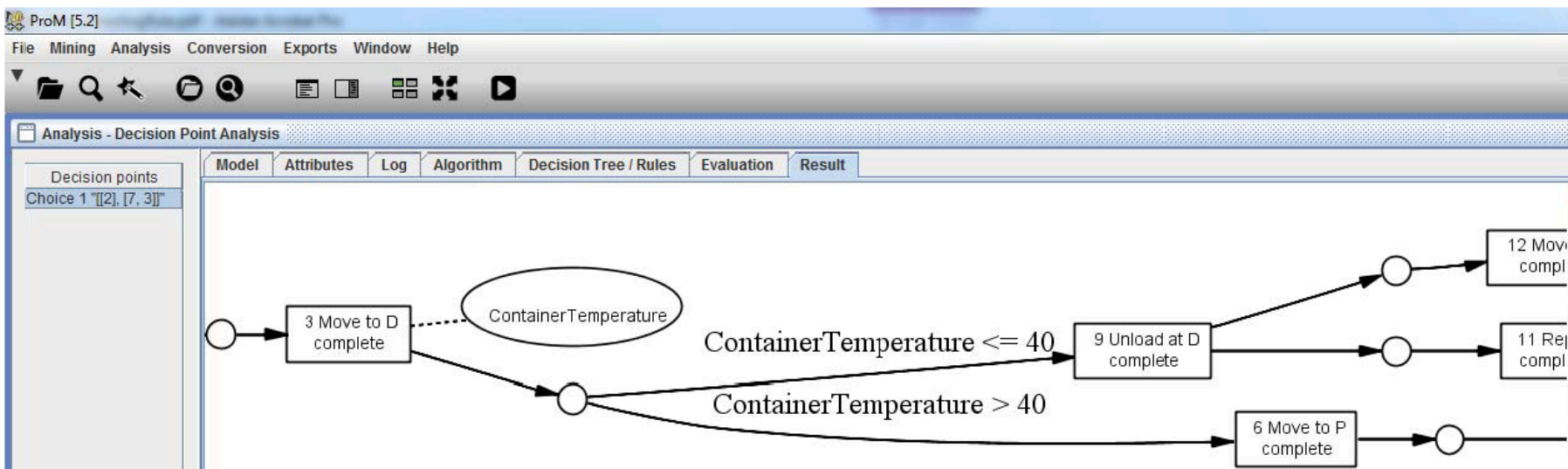
³Anne Rozinat, Wil M. P. van der Aalst: Decision Mining in ProM. Business Process Management 2006: 420-425

4 Decision mining

Example setting



© 2015 Springer-Verlag Berlin Heidelberg



© 2015 Springer-Verlag Berlin Heidelberg

Contents

1 Introduction

2 Social network analysis

3 Organizational mining

4 Decision mining

5 Text mining

6 Summary & outlook

References

5 Text mining

Introduction and Terminology, Data

- Text documents may be of different origin
 - Reports, abstracts, journal articles, blogs, tweets, email,...
- There are many different formats
 - .txt, .pdf, .doc, html, xml,...

5 Text mining

Introduction and Terminology, Approaches

- Metadata view: a description of the document
- There exist a number of standards for describing resources
- One popular standard is the Dublin Core Metadata Initiative (DCMI): <http://dublincore.org/>

- Example for Metadata using R:

author: Rinderle-Ma, Grossmann

datetimestamp: 2014-09-28 08:09:19

description: Brief Task description

heading: 1.3.5 Evaluation and Reporting Task

id:11

language: en

origin : Fundamentals of Business Intelligence

V1.0Text Mining (Text Analytics view)

5 Text mining

Introduction and Terminology, Approaches

- Text analytics combines knowledge and techniques from different areas⁴, including
 - Data Mining
 - AI and Machine Learning
 - Computational Linguistics
 - Databases
 - Information Retrieval
 - Web Mining

⁴Miner et al. Practical Text Mining, Elsevier (2012)

5 Text mining

Introduction and Terminology, Levels of Text Mining

- Text Mining can be done at different levels
 - Word level
 - Sentence level
 - Document level
 - Corpus level = Collection of documents
- A Document can be defined in different ways:
 - Sections of a document, paragraphs in text, and so on
 - A tweet, an email, and so on

5 Text mining

- *Data*: text corpus, defined by a collection of text documents
- *Analytical goals*:
 - Description of the documents in the corpus
 - Clustering the documents in the corpus
 - Finding topics of the corpus
 - Classification of documents based on rules derived from a training corpus
- *Modeling task*: definition of a document term matrix
- *Analysis and evaluation*: see next slide

5 Text mining

- *Analysis task:*
 - *Description of corpus:* determination of type-token relation and association measures, visualization of the content in the corpus using word clouds and correlation plots
 - *Clustering documents:* using cluster analysis methods
 - *Topic models:* define a number of topics and find the probability of assignment of the documents to the topics
 - *Classification:* learn classification rules for assignment of new documents
- *Evaluation and reporting task:* represent analysis results by word clouds, correlation plots, and by characterization of topics with terms

5 Text mining

Data Preparation and Modeling, Transformations

- Usually not the original text is used for text mining, but a transformed (purged) text
- Basic standard transformations
 - Removal operations (punctuation, numbers, special characters (@, /,...), email address
 - White space operations
 - Lower case letters
 - Stop words (articles, prepositions,...)
 - Stemming (words without endings)
 - Example sentence: *Its main goals are the interpretation of the results in reference to domain knowledge and coming to a decision of how to proceed further.*
 - Transformed sentence: *main goals interpretation results reference domain knowledge coming decision proceed*

5 Text mining

Data Preparation and Modeling, Document Term Matrix

- After the transformations the corpus consists of a number of documents with preprocessed terms
- These terms are organized in a list of tokens and the frequency of the tokens obtained by tokenization
- A token is defined by a n-gram = n contiguous words in the document, usually 1-grams (one term) or bigrams (2 words)

5 Text mining

Data Preparation and Modeling, Document Term Matrix

- The basic unit for analysis is the document term matrix (DTM)

$$DTM = (t_{ij}), i = 1, \dots, d, j = 1, \dots, n$$

where t_{ij} = frequency of term j in document i

- Sometimes also the transposed matrix is used and called TDM (term document matrix)
- Other name for the DTM: Bag of words
- An alternative to the DTM is often to replace the frequency simply by an indicator

$$DTMI = (d_{ij}), i = 1, \dots, d, j = 1, \dots, n$$

$$d_{ij} = \begin{cases} 1 & \text{if term } j \text{ occurs in document } i \\ 0 & \text{otherwise} \end{cases}$$

5 Text mining

Data Preparation and Modeling, Document Term Matrix

- Usually the DTM has many columns and contains many terms with low frequency
- General assumption:
 - Frequency of a term informs about the importance of the term for the contents
 - There are terms occurring frequently due to linguistic reasons, for example verbs like “have”, “is”, etc.

5 Text mining

Data Preparation and Modeling, Document Term Matrix

- Solution of the problem:
 - Define upper and lower thresholds for the terms
 - Use instead of the DTM the TF-IDF = Term frequency– inverse document frequency matrix
- Inverse document frequency (IDF) = Number of documents divided by the frequency of the documents which contain the term
 - Reduces the importance of terms which occur in many documents

5 Text mining

Formulas:

- $D = \{d_1, d_2, \dots\}$ Documents, $W = \{w_1, w_2, \dots\}$ Words

$$IDF_{ij} = \frac{|D|}{1 + DF_{ij}}, DF_{ij} = \text{card}\{d_i : w_j \in d_i\}$$

$$TF - IDF_{ij} = ti_j * \log(IDF_{ij})$$

- TF-IDF is of special interest for keywords differentiating between documents

5 Text mining

Descriptive Analysis of the DTM, Word Clouds

- A useful representation of a DTM is using a word cloud
 - Representation of the terms in the DTM with size according to the frequency of the terms
 - Usually the most frequent terms are in the center
 - Terms can be also rotated and colored

5 Text mining

Descriptive Analysis of the DTM, Word Clouds

- For comparison of documents a *comparison cloud* is a useful tool
- The documents are organized in an outer circle in the graphic
- Terms are shown with size according to their frequency and are positioned according to their occurrence in the documents
- Let the DTM contain M documents, calculate deviation of the relative frequencies d_{ij} of a term w_i in a document d_j from the main relative frequency of the document as

$$d_{ij} = (p_{ij} - p_j), p_{ij} = \frac{t_{ij}}{\sum_j t_{ij}}, p_j = \sum_i \frac{p_{ij}}{M}$$

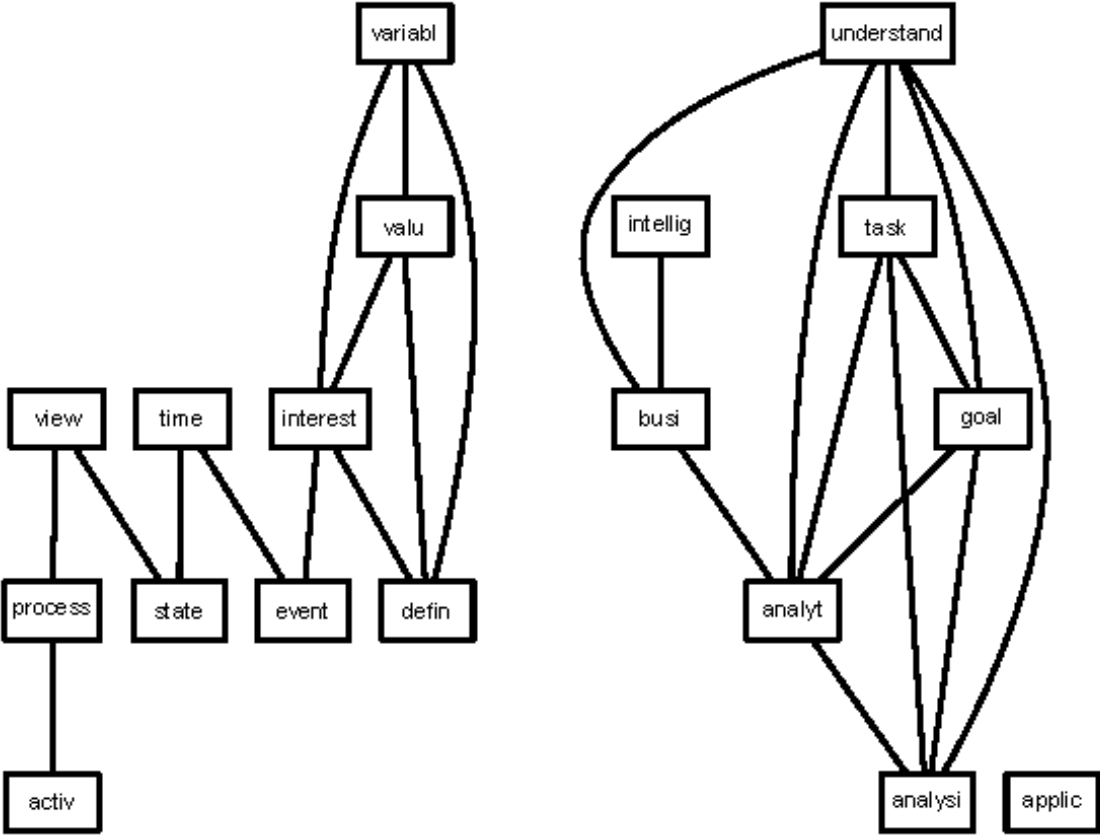
5 Text mining

Descriptive Analysis of the DTM, Associations between terms

- Another way to describe the contents of a document is to use correlation between the term frequencies in the different documents
- One can use also the indicator matrix for such associations
- Example DTM (Chapters 1 and 2 of the book):
 - Look for terms with a correlation to the term “business” > 0.6
 - Results (selection): “intelligence” with 0.76 and “understanding” with 0.72

5 Text mining

Associations between frequent terms in Chapters 1 and 2 of the book, produced using R package tm



© 2015 Springer-Verlag Berlin Heidelberg

5 Text mining

Analysis of a Text Corpus, Clustering

- Cluster analysis of text data is based on the definition of similarities between documents
- For definition of the similarity the most popular measure is the cosine measure of the term frequencies in the documents

$$\text{sim}(d_i, d_j) = \frac{t_{i\bullet} * t_{j\bullet}}{\|t_{i\bullet}\| * \|t_{j\bullet}\|}$$

where $t_{i\bullet}$ = frequency vector of terms in document i

5 Text mining

Analysis of a Text Corpus, Clustering

- Based on this distance one can apply any cluster analysis algorithm (hierarchical or k-means)
- Many other methods have been proposed
- Co-Clustering:
 - Interpret the DTM as a bipartite graph: Terms and documents
 - Partition the two sets in such a way that the edges between different clusters are minimized

5 Text mining

Analysis of a Text Corpus, Classification

- Classification of documents can be done by interpretation of the terms as variables (features) describing the documents
 - Hence the DTM is a classical feature matrix for the documents and we can apply any classification algorithm
 - Frequently the indicator version DTMI is used instead of the DTM
 - Classical application: Spam detection in emails

5 Text mining

Analysis of a Text Corpus, Topic Models

- Topic models is an advanced method for grouping documents and terms into topics
- Model:
 - Define a number of topics
 - For each document a distribution of the topics is assumed
 - For each topic the terms have a topic specific characteristic distribution
- A topic model estimates the parameters of the distributions of the topics within the different documents and identifies the most frequent terms in each topic
- Usually the algorithm is applied for a different number of topics and the results are compared

5 Text mining

Further Aspects of Text Mining, Analysis at the Word Level

- Words allow the representation of concepts with different words (synonyms)
- Concepts have many times and ordering
 - Hypernyms: terms representing a narrower concept
 - Hyponyms: Terms representing a broader concept
 - Part of relation between concepts
- Representation of such relations in a database for words
- For English terms WordNet (<http://wordnet.princeton.edu>) is an
- important resource which is free available
 - Examples: business, model, busy

5 Text mining

Further Aspects of Text Mining, Analysis at the Sentence Level

- Analysis at the sentence level allows the syntactic analysis of a sentence
- POS = Part of Speech Tagging
- Taggers identify the role of the words in a sentence
- Apache Open NLP is a frequently used tool (available in R) (<https://opennlp.apache.org/>)
- For tagging a standard are the Penn Treebank Tags (<http://web.mit.edu/6.863/www/PennTreebankTags.html>)

5 Text mining

Further Aspects of Text Mining, Analysis at the Sentence Level

- Example sentence: *The evaluation and reporting task looks at the analysis results from a global business perspective.*

{(TOP

(S

(NP (DT The) (NN evaluation) (CC and)
(NN reporting) (NN task))

(VP (VBZ looks)

(PP (IN at) (NP (DT the) (NN analysis)
(NNS results))))

(PP (IN from) (NP (DT a) (JJ global)
(NN business) (NN perspective))))})

5 Text mining

Further Aspects of Text Mining, Keyword Extraction

- Keyword extraction is usually done in a number of steps for creating features
 - TF-IDF for keyword search in a corpus
 - POS
 - Identifying words at the beginning of a text
 - Relation of the words to words in a thesaurus
 - Using such features each word gets a score and high scores define keywords
 - Learning the scores is based on supervised learning

5 Text mining

Further Aspects of Text Mining, Opinion Mining

- Opinion Mining and Sentiment Analysis
- General setup
 - Basic unit is a document (cf. questionnaire)
 - Opinion holder : author of the document (cf. surveyed person)
 - Objects and features about which an opinion is stated (cf. questions in the questionnaire)
 - Polarity of the opinion (cf. answers in the questionnaire)
- Main tasks in opinion mining
 - Finding in a document all opinionated sentences
 - Identify the objects and features about which an opinion is stated
 - Classify the opinion (polarity)

5 Text mining

Further Aspects of Text Mining, Opinion Mining

- Finding opinionated sentences
- Different cases
 - Direct opinion (frequently based on adjectives like good, nice, bad, ...)
 - Negation (non, not, negative prefixes)
 - Comparative opinion (better worse, comparative form of adjectives, ...)
- POS is of utmost importance
- Identification of objects and features
 - Many times only one object, in simple cases in the header of the document (metadata)
 - One of the first nouns in a document represent the objects
 - Features can be identified using data bases which characterize objects, for example products or movies
 - Knowledge about synonyms, hypernyms, and hyponyms is necessary (WordNet)

5 Text mining

Further Aspects of Text Mining, Opinion Mining

- Opinion classification
 - Polarity is usually based on wordlists or dictionaries for adjectives stating the polarity (twitter) <http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon>
 - A more elaborated dictionary is SentiWordNet (<http://sentiwordnet.isti.cnr.it/>)
 - Other methods are based on statistical classification
- Problems with opinion mining
 - Opinion is not so well defined as objective features of products
 - Opinion is based on common sense (cf. SenticNet for such an approach <http://sentic.net/sentics/>)
 - Terms like precision and recall are difficult to apply
 - What is the group of opinion holders?
 - Can we identify spam opinion?

Contents

1 Introduction

2 Social network analysis

3 Organizational mining

4 Decision mining

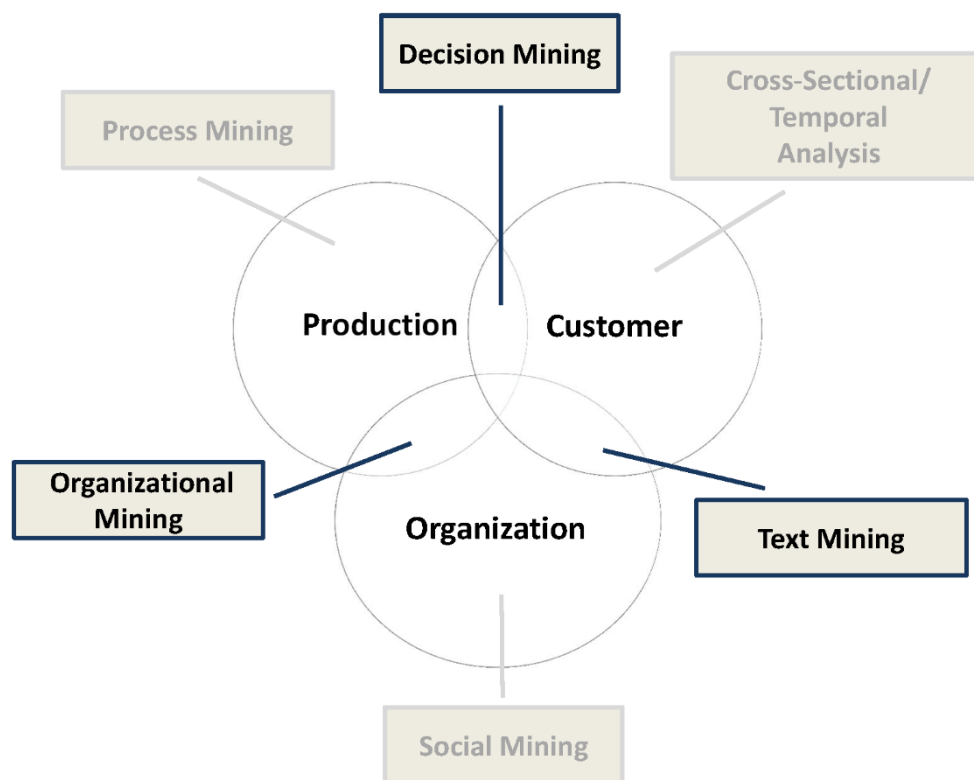
5 Text mining

6 Summary & outlook

References

6 Summary & outlook

- Combination of different techniques allows to answer „new“ questions at the interfaces of analysis perspectives



© 2015 Springer-Verlag Berlin Heidelberg

6 Summary & outlook

