Context-driven Data Mining and Knowledge Extraction

V.Gorodetsky
(Co-authors: V.Samoylov, S.Serebryakov)
Practical Reasoning, Inc. under the auspices of
St. Petersburg Institute for Informatics and Automation
gor@iias.spb.su
http://space.iias.spb.su/gorodetsky

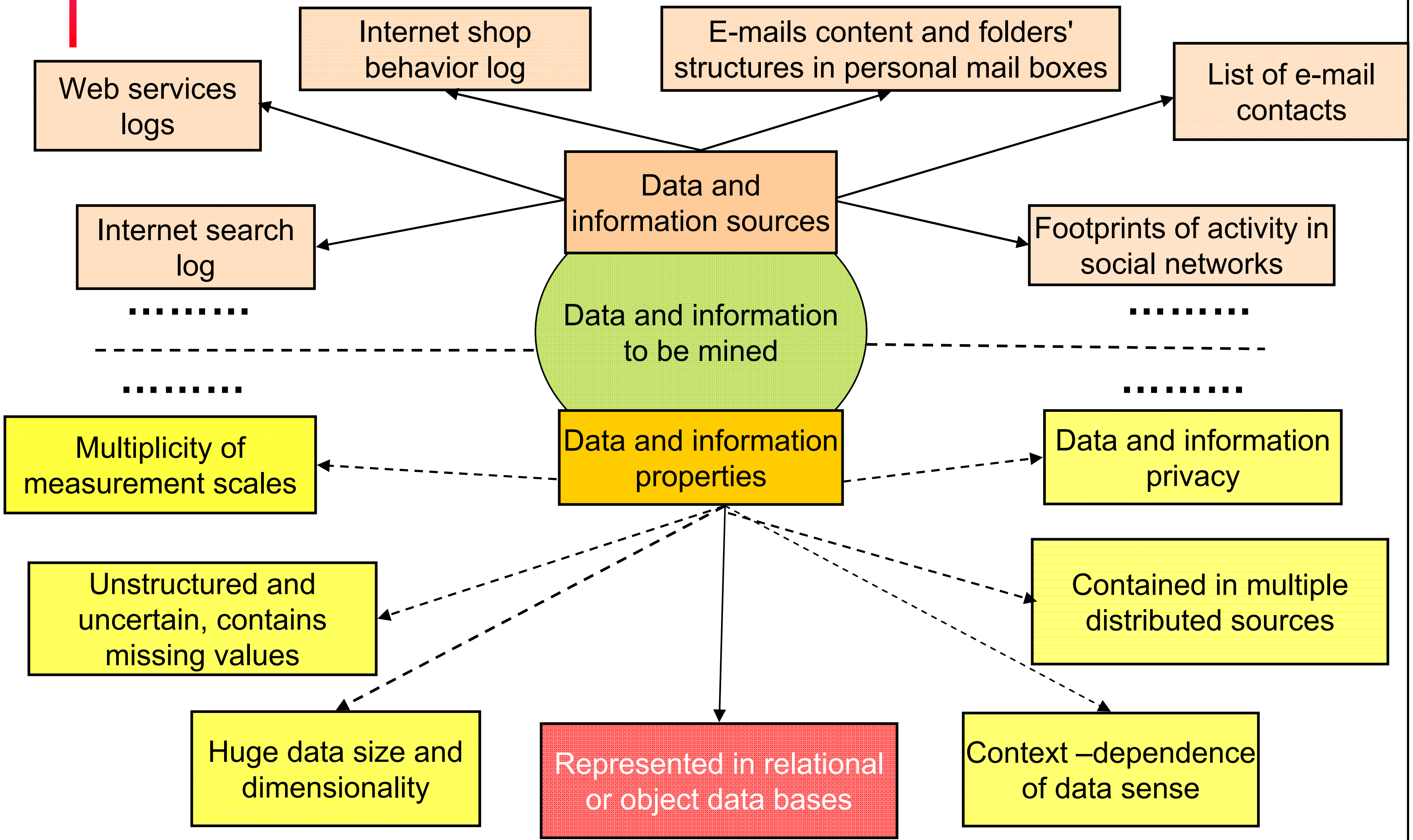
Key Words: What is presentation about?

- Context–driven data mining,
- Context representation
- Feature extraction and selection
- Causality
- Heterogeneous information fusion
- Personalized recommendations
- …………………………………

Content (maximal)

1. Introduction: Problem motivation

Where the Challenges Come from?: From Data and Information Key Features



Peculiarities of Object Data

- It is bad since every object instance can be specified by different *structure*, and *attributes*, and, therefore, by *different features* from data mining viewpoint;
- This is good since these differences *explicitly reflect different contexts* of various object instances;
- This is good since object is specified in in terms of *ontology concepts* and relations between them, therefore, each object *implicitly contains domain knowledge introduced* by knowledge engineer (ontology developer);
- It is very good, since *ontology enriches learning data* sample with expert knowledge and therefore can significantly *enrich knowledge* that can potentially be *extracted* from object data base;
- But this is bad, since it make the data mining problem *much more complex*.

Example of oblect instance: E-mail Instance

EMailItem

EntryID *000000003465C1F6148B1C40 932E6C94E9F0490224D52100*
Size *477097*
Importance 1
BodyFormat 2
Conversation Topic KelwinMag Reseller Agreemen
Read Receipt Request false
ReceivedTime *38567.903020833335*
CreationTime *38567.905411597225*
FolderID *240*
Body Content *George,*

As promised here is the "Redlined" agreement.
Let us know what you can and can't live with in here. Thanks,
Jonathan Kinsbery
Effective Solutions eTrade Technology, Inc.
123-456-7890-Wireless
198-765-4321- Fax 654-312-7119- Office
"http://www.etrade.com/web/main"
eTrade Effective Solutions

From: Olga Ginsberger
Sent: Tuesday, August 03, 2009 11:47 AM
To: Jonathan Kinsbery Cc: David Arnold
Subject: KelwinMag Reseller Agreement
Jonathan –

Enclosed please find the redlined KelwinMag Reseller Agreement.
Would you please forward it to your contact at KelwinMag for their review. Many thanks.
Olga Ginsberger
Sr. Contracts Administrator
eTrade Technology, inc.
12531 DullesDrive, Mail Stop #11
Herndon, VA 12131
Tel: 713-914-3042, Fax: 705-981-8362
ginsberger@etrade.com

Subject *FW: KelwinMag Reseller Agreement*
Attachment FileName *KelwinMag Reseller Agreement 03-1-2009.doc*
AttachmentType *Doc*
EMailSender
m_sEmail *jkinsbery@etrade.com*
Person
FirstName *Jonathan*
LastName *Kinsbery*
EMailReceiver
m_sEmail *georget@kelwinmag.com*
EMailDomain
m_sEMailDomain *kelwinmag.com*
m_sEMailDomain *etrade.com*

What Are Main Problems of Knowledge Discovery from Object Data Base?

Conventional challenges of Knowledge Discovery from Data (KDD):

How to manage *huge data size and dimensionality*? How to *select* useful *features*?

How *to fuse interconnected heterogeneous* data and information that can be represented as texts, strings, numbers, partially ordered symbols…and have complex structures?

How to deal with features *assigned many particular values* in the same instance?

- (e.g. *Person phone numbers*: 123-4567890, 198-7654321, 54 -3127119, 713-9143042, 705-9818362)

Novel challenge for context-based data mining (context knowledge discovery from data - CKDD):

How to *extract knowledge* from *object data* if each learning sample instance is presented in *terms of different concepts and attributes* structured differently from instance to instance reflecting in this way particular context?

The talk objective is to outline developed generic technology intended to cope with or, at least, mitigate these challenges

Technology Components

Generic context-driven data mining and knowledge discovery technology is focused on

- context extraction and representation
- Feature selection for decision making support;
- Causal data analysis and causal feature search
- Personalization of knowledge extracted (in regard to each class specification)

2. Methodology of context-driven data mining.

Phase1:

Transformation of (relational) data sample to object DB form

Commonly known statement:

Data mining and machine learning quality depends on the volume of domain knowledge and context involved in data mining procedure and on how effectively it is used in the resulting knowledge base.

Transformation of (relational) data sample to object DB form

Transformation of source (training) data set to the object form:

(1) *Development*, by domain experts, of the *domain ontology* thus enriching data sample with domain context and semantics

(2) *Transformation training data* set to object DB structure

Result:

- Initial *training sample* is represented by the *set of objects' instances* in object DB , that is by the set of the relational DB tables with the *ontology on top of it*. Ontology plays the role of domain meta knowledge intended to provide object-oriented view of the relational data.

  Note: There exists standard middleware that is capable to *in-fly transform data sample* represented in relational DB with *ontology* on top of relational DB *as a data meta model to the object form*. Therefore, *getting an object form* of a relational data given ontology *is a feasible* task.

- *Each instance* of training data sample is assigned the label ω_k of a class it belongs to:

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$$

E-mail Instance

EMailItem

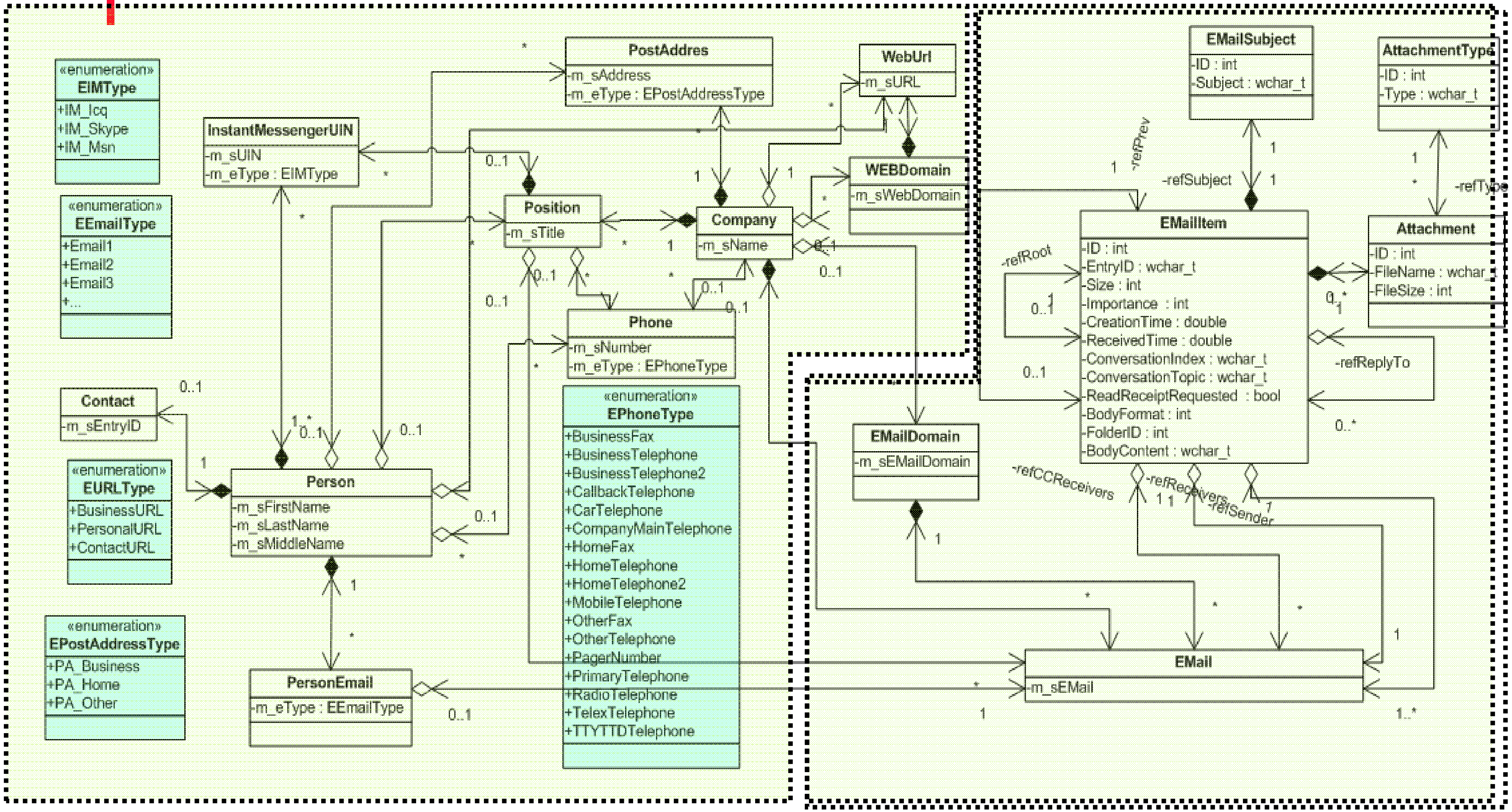
EntryID *000000003465C1F6148B1C40*
932E6C94E9F0490224D52100
Size *477097*
Importance 1
BodyFormat 2
Conversation Topic KelwinMag Reseller Agreemen
Read Receipt Request false
ReceivedTime *38567.903020833335*
CreationTime *38567.905411597225*
FolderID *240*
Body Content *George,*
As promised here is the "Redlined" agreement.
Let us know what you can and can't live with in here. Thanks,
Jonathan Kinsbery
Effective Solutions eTrade Technology, Inc.
123-456-7890-Wireless
198-765-4321- Fax 654-312-7119- Office
"http://www.etrade.com/web/main"
eTrade Effective Solutions

From: Olga Ginsberger
Sent: Tuesday, August 03, 2009 11:47 AM
To: Jonathan Kinsbery Cc: David Arnold
Subject: KelwinMag Reseller Agreement
Jonathan –

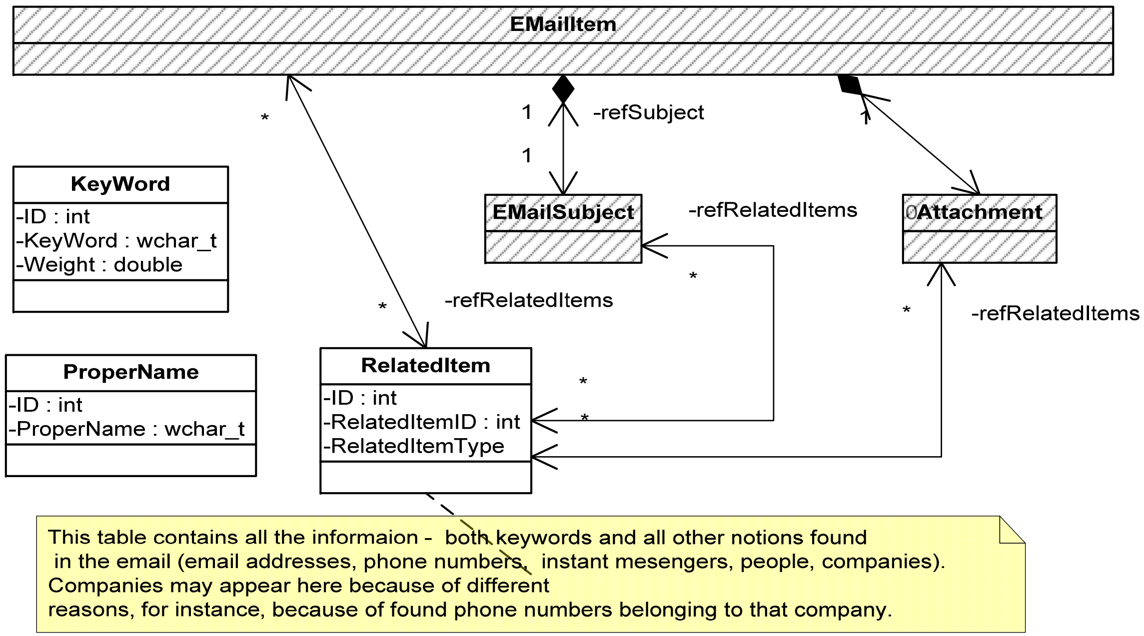
Enclosed please find the redlined KelwinMag Reseller Agreement.
Would you please forward it to your contact at KelwinMag for their review. Many thanks.
Olga Ginsberger
Sr. Contracts Administrator
eTrade Technology, inc.
12531 DullesDrive, Mail Stop #11
Herndon, VA 12131
Tel: 713-914-3042, Fax: 705-981-8362
ginsberger@etrade.com

Subject *FW: KelwinMag Reseller Agreement*
Attachment FileName *KelwinMag Reseller Agreement 03-1-2009.doc*
AttachmentType *Doc*
EMailSender
m_sEmail *jkinsbery@etrade.com*
Person
FirstName *Jonathan*
LastName *Kinsbery*
EMailReceiver
m_sEmail *georget@kelwinmag.com*
EMailDomain
m_sEMailDomain *kelwinmag.com*
m_sEMailDomain *etrade.com*

Application Example: Personalized Outlook E-mail Assistant Ontology



Ontology of Secondary Features



Text Analysis and Mining Technology Components

E-mail body mining

Objective: To extract and interpret e–mail context *in order to automatically use it for instantiation of the primary and secondary features* appearing in the e-mail body

Tools used: IBM Language Resource Ware (LRW) and IBM Ontological Network Miner (ONM) (available at IBM's alphaworks site).

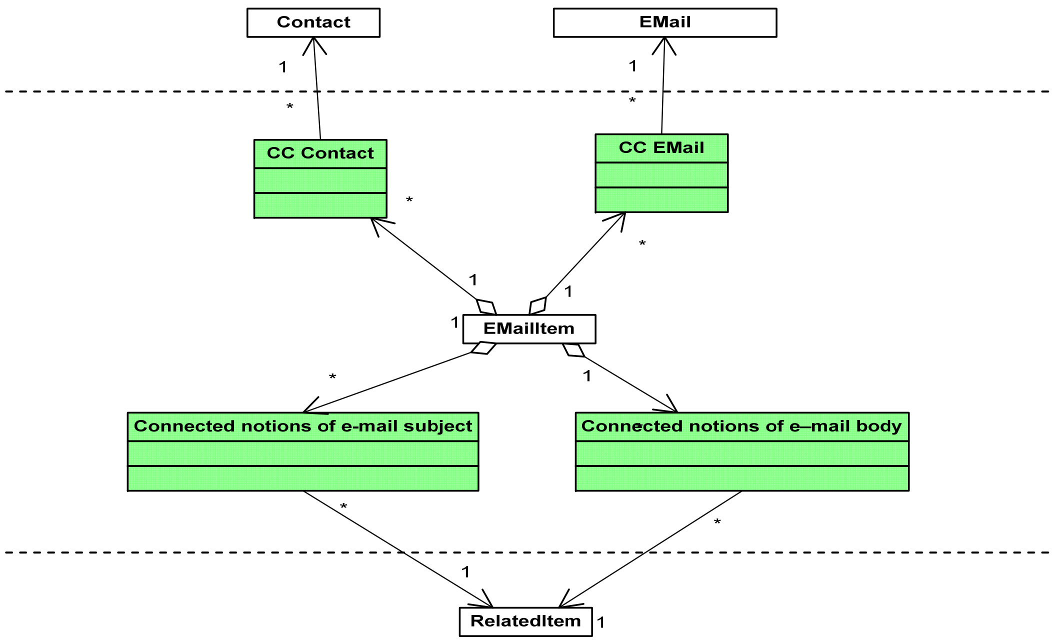
Pattern mined

- *Regular expressions* (e-mail and web addresses);
- *LRW capabilities* are used (1) for annotation, i.e. *Dictionary–Based search* using dictionaries of *people* and *company* names and (2) for *Rule based annotation (text segmentation→ segments of rule-based interpretations*

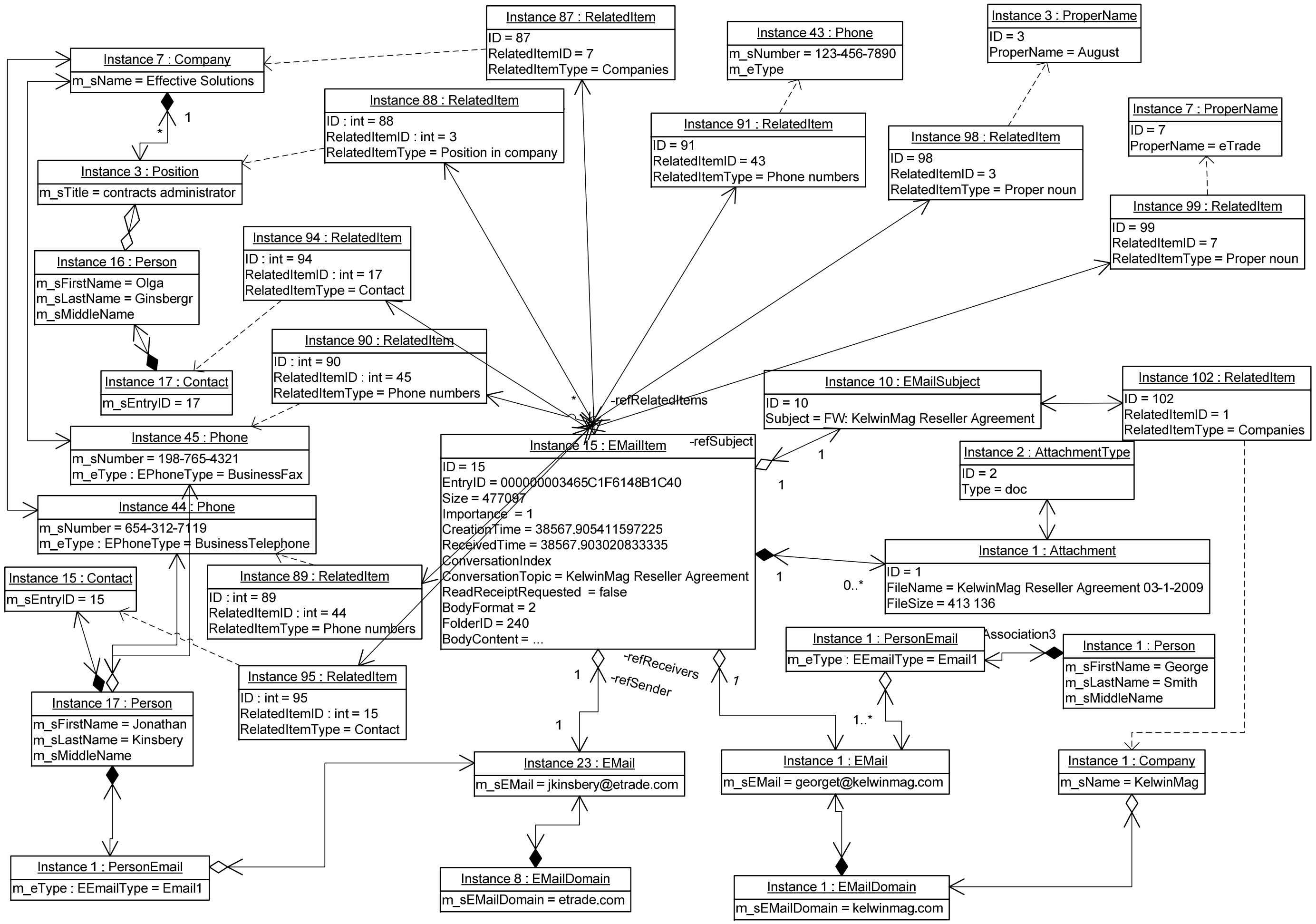
  //E.g., for segment <Barry White >→ <Barry> , <White>→ {<Barry>→ FirstName> (using Dictionary), <White>→Word→ <Barry> + <White> + rule {"if FirstName with subsequent proper noun then they form *FullPersonName"}→<Barry White>→ FullPersonName. //*
- *ONM is used to extract key concepts from the text* (*text focus*), even those that are not presented explicitly in the e-mail body.
- *Ontology for text analysis* and mining is to be developed by expert and specified in XML while using *categories* and *concepts* of an external *ontology*

“Star”-structured Multi–dimensional (1–0..*)

While set of features is completed, customer’s historical data are transformed to the form of *“star”–structured tables*, in which columns of fact tables correspond to elements of the designated expanded feature set with one row in kernel table per every instance.



An Example of E-mail Instance Representation in Object Data Base



3. Methodology of context-driven data mining

Phase 2:

Expert–driven Feature Generation and Corresponding Object Data Sample Transformation

Expert–driven Feature Generation and Corresponding Object Data Sample Transformation

Objective:

- *Feature* generation: domain experts is responsible for generation of potentially useful, clearly understood and simply interpretable *features in terms of ontology concepts* or/and in terms of concepts *attributes* with *no care about feature space dimensionality*;
- *Transformation* of data sample data to new feature space .

Examples of features for E-mail assistant case study:

Formal features		Secondary features	
Feature	**Measurement scale**	**Feature**	**Measurement scale**
E–mail size	Real	CC contact	categorical
E–mail sender	categorical	Connected notions of e-mail subject	Pair–wise of any measurement.
Sender contact	categorical	Connected notions of e–mail body	Pair–wise of any measurement.
CC E-mail	categorical	Attachment format	categorical/ Boolean

Secondary features type: *pair-wise* of the structure <*Notion: {Set of values}* >

Examples of secondary features

- Connected notions of e-mail subject:
  (Company name: *KelwinMag*)
- Connected notions of e-mail body:
  (Position in company: *contracts administrator*)
  (Companies: *effective solutions*)
  (Phone numbers: 123-4567890, 198-7654321, 654 -3127119, 713-9143042, 705-9818362),
  (E-mail address: *ginsberger@etrade.com*),
  (Web address: *http://www.etrade.com/web/main*),
  (Proper noun: *August*), (Proper noun: *eTrade*),
  (Contact: *Jonathan Kinsbery, Olga Ginsbergr).*

4. Methodology of context-driven data mining

Phase 3:

Feature aggregation

Principles of Feature Selection “Philosophy”

Reminder: Every feature should be expressed in terms of *ontology concept and/or their attributes* thus providing for a *well understandable semantics*. For feature that is not explicitly contained in ontology it is necessary *to establishing its relations to the existing ontology concepts*.

1. *Feature as Classifier: There is no semantic difference between the concepts “feature” and “classifier”.* Every feature X_i can be thought of as a (simple) classifier like that:

$$if\ P(X_i \in X_i^{(k)})\ then\ \omega_k\ ,$$

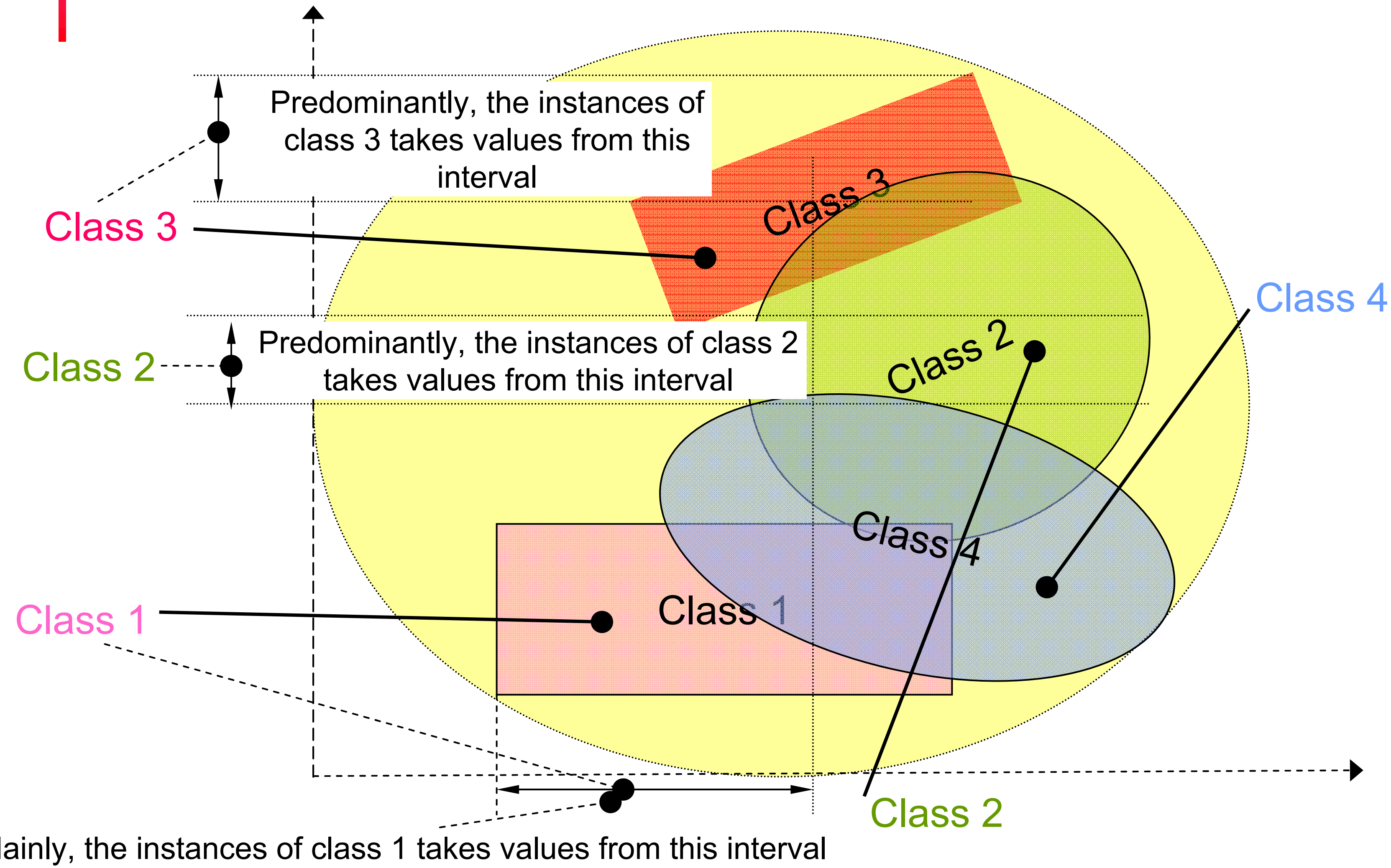
   and, vice versa, every classifier can be considered as a (complex) feature.

2. *Good and bad features*: Slightly re-formulating the Condorcet theorem one can say that a *classifier is "good" if its accuracy is strictly more than 0.5.* Otherwise a classifier is useless. Analogously, one can say that *a feature is "good" if a one-variable classifier using this feature is "good".*

3. *Main receipts*: The "*recipes*" against huge scale of both data size and dimensionality are *feature aggregation, filtering and causality discovery*.

4. *Personalization: Feature selection* procedure should be *class–targeted*, i.e., a *specific* set of *features* are generated for *each class* of object instances.

Feature as Classifier :
One-feature Naive Bayes Classifiers



Feature aggregation: One-feature-Naïve-Bayes classifier case

Let X_i, $i = 1,..., n$, is a feature with discrete domain $\boldsymbol{X}_i$.

Let $x_s^{(i)} \in \boldsymbol{X}_i$ -- a particular value of the feature X_i .

Let us compute disjoint sets $\boldsymbol{X}_i^{(k)} \subset \boldsymbol{X}_i$ in the following way:

For any value $x_s^{(i)} \in \boldsymbol{X}_i$ of the feature X_i this value $x_s^{(i)} \in X_i^{(k)}$

If and only if

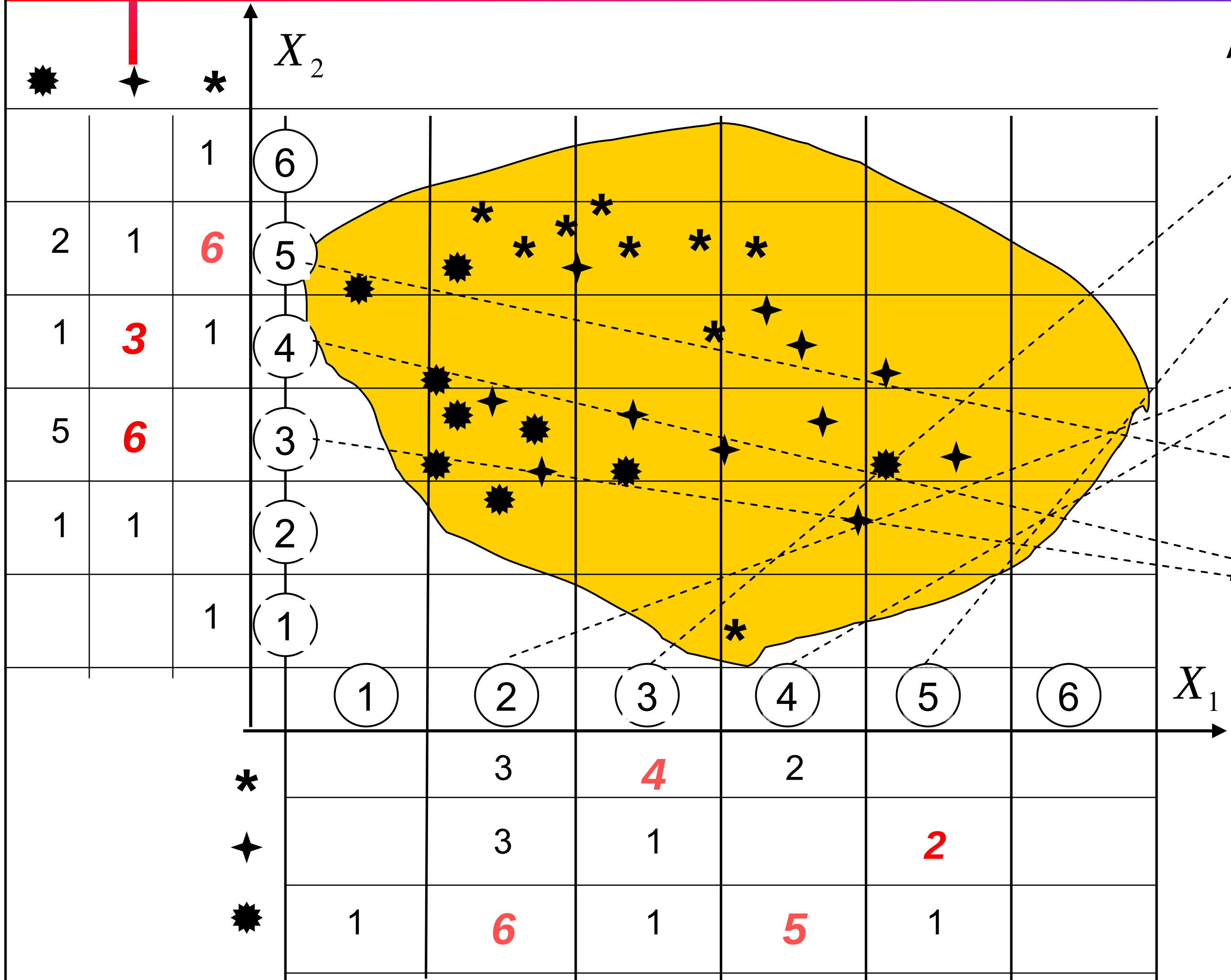
$$p(\omega_k / x_s^{(i)}) > p(\omega_v / x_s^{(i)}) + \delta_i \quad for \;\; any \;\; v \neq k,$$

where $p(\omega_k / x_s^{(i)})$ and $p(\omega_v / x_s^{(i)})$ is conditional probabilities of classes ω_k, ω_v , $\omega_v,\ \omega_k \in \Omega,\ k = 1,...,m,$ respectively.

One-feature Naïve Bayes classifier

If $x_s^{(i)} \in \boldsymbol{X}_i^{(k)}$, then ω_k

A toy example: Feature aggregation



Aggregated features

$$X_1^{(*)} = \{x_3^{(1)}\},$$

$$X_1^{(+)} = \{x_5^{(1)}\},$$

$$X_1^{(\mathbf{o})} = \{x_2^{(1)}, x_4^{(1)}\},$$

$$X_2^{(*)} = \{x_5^{(2)}\},$$

$$X_2^{(+)} = \{x_3^{(2)}, x_4^{(4)}\},$$

$$X_2^{(\mathbf{o})} = \{\text{Ø}\}$$

Aggregated Unary Predicate Search – Target of the Third Phase of the Methodology

For each aggregate $\boldsymbol{X}_i^{(k)}$ the aggregated unary predicate $F_i^{(k)}$ is introduced in the following way: if $x_r^{(k)} \in X_i^{(k)}$ then $L[F_i^{(k)}(x_r^{(i)})] = true$

Therefore $\boldsymbol{X}_i^{(k)}$ is the truth domain for unary predicates $F_i^{(k)}$.

Using training data sample, each aggregated unary predicate can be mapped with conditional probability $p(\omega_k / L[F_i^k] = true) = p(\omega_k / F_i^k),$

$$\sum_k p(\omega_k / F_i^k) = 1$$

Search for *aggregated unary predicates* $F_i^{(k)}$ assigned with conditional probabilities $p(\omega_k / F_i^k)$ for all features X_i , $i = 1,..., n,$ and all classes $\omega_k \in \Omega, k = 1,...,m,$ - *goal of the third phase of the context-driven DM methodology*

Negative Feature Aggregation: One-feature-Naïve-Bayes classifier case

Let X_i , $i = 1,\ldots, n$, is a feature with discrete domain $\boldsymbol{X}_i$.

Let $x_s^{(i)} \in \boldsymbol{X}_i$ -- a particular value of the feature X_i .

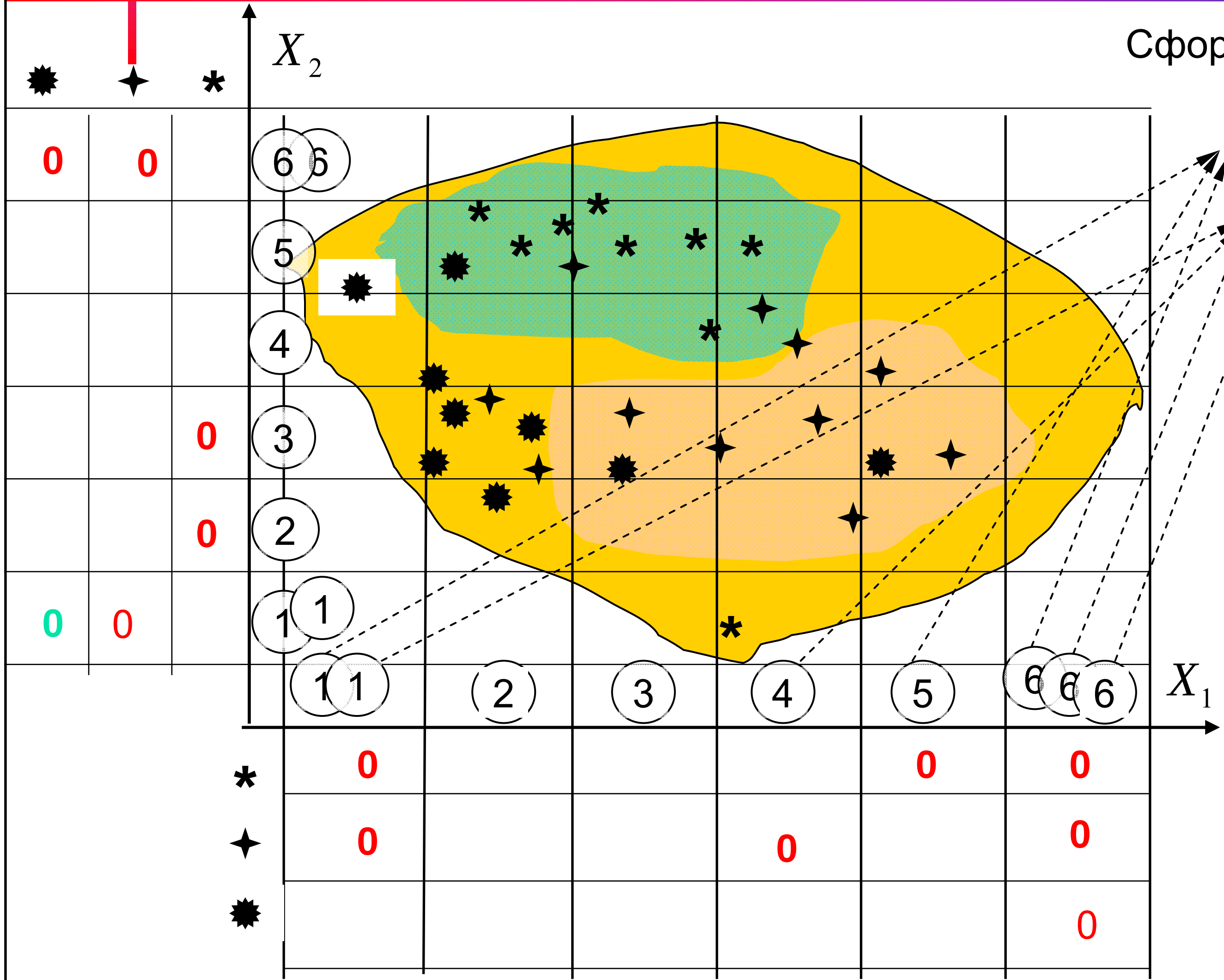
Let us compute disjoint sets $Y_i^{(k)} \subset \boldsymbol{X}_i$ in the following way:

For any value $x_s^{(i)} \in \boldsymbol{X}_i$ of the feature X_i this value $x_s^{(i)} \in \boldsymbol{Y}_i^{(k)}$ if and only if this value $x_s^{(i)}$ is not met in any instance of class ω_k .

One-feature Naïve Bayes classifier

If $x_s^{(i)} \in \boldsymbol{Y}_i^{(k)}$, then $\overline{\omega_k}$

A toy example: Negative Feature Aggregation



$$Y_1^{(*)} = \{x_1^{(1)}, x_5^{(1)}, x_6^{(1)}\},$$
$$Y_1^{(+)} = \{x_1^{(1)}, x_4^{(1)}, x_6^{(1)}\},$$
$$Y_1^{(\mathbf{o})} = \{x_6^{(1)}\},$$
$$Y_2^{(*)} = \{x_2^{(2)}, x_3^{(2)}\},$$
$$Y_2^{(+)} = \{x_1^{(2)}, x_6^{(4)}\},$$
$$Y_2^{(\mathbf{o})} = \{x_1^{(2)}, x_6^{(2)}\}$$

$X_i^{(k)}$

Aggregated Negative Unary Predicate Search – Target of the Third Phase of the Methodology

For any aggregated negative feature $\boldsymbol{Y}_i^{(k)} \subseteq \boldsymbol{X}_i$ the unary predicate $\boldsymbol{G}_i^{(k)}, k \in \{1,\ldots,m\},$ is introduced in the following way:

If $x_s^{(i)} \in \boldsymbol{Y}_i^{(k)}$ then $L[\boldsymbol{G}_i^{(k)}] = true$

Therefore $\boldsymbol{Y}_i^{(k)}$ is the truth domain for unary predicates $\boldsymbol{G}_i^{(k)}$.

Predicate $\boldsymbol{G}_i^{(k)}$ is called negative unary predicate for class $\overline{\omega}_k$

Negative unary predicate $\boldsymbol{G}_i^{(k)}$ is a one feature classifier stating the fact that

If $\boldsymbol{G}_i^{(k)}$ then $\overline{\omega}_k$.

Final Aggregated Feature Set = Union of Aggregated Positive and Negative Unary Predicate Sets

The result of the 3d phase of the context-driven data mining is

- the united set of positive and negative *class-targeted* feature set $\boldsymbol{F}_i^{(k)} \cup \boldsymbol{G}_i^{(k)}$ for every class $\omega_k \in \Omega$;
- the training (and testing) data set transformed to the new aggregated feature set $\boldsymbol{F}_i^{(k)} \cup \boldsymbol{G}_i^{(k)}$;
- all features are of Boolean type, i.e. the feature set is homogeneous.

5. Methodology of context-driven data mining

Phase 4:

Feature filtering

Feature Filtering

Starting point for feature filtering: - The *sets of aggregated unary predicates* assigned with conditional probabilities:

$$\omega_1: \; p(\omega_1 / F_{i_1}^{(1)}), \quad p(\omega_1 / F_{i_2}^{(1)}), \ldots, \quad p(\omega_1 / F_{i_r}^{(1)})$$

$$\omega_2: \; p(\omega_2 / F_{j_1}^{(2)}), \quad p(\omega_2 / F_{j_2}^{(2)}), \ldots, \quad p(\omega_2 / F_{j_s}^{(2)})$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\omega_m: \; p(\omega_m / F_{v_1}^{(m)}), \quad p(\omega_m / F_{v_2}^{(m)}), \ldots, \quad p(\omega_m / F_{v_t}^{(m)})$$

Every unary predicate $F_i^{(k)}$ can be considered as *one-feature classifier* for which *contingency matrix can be computed* using training data sample with conditional probabilities $p(\omega_k / F_i^k)$ and $p(\overline{\omega}_k / \neg F_i^k)$ on their diagonals.

Let us note that all probabilities are computed *using testing data set and cross-validation*

Filtering rule:

Aggregated unary predicate $F_i^{(k)}$ remains in the feature list *if it is a "good feature|:*

$$p(\omega_k / F_i^k) + p(\overline{\omega}_k / \neg F_i^k) > 0,5$$

otherwise it is filtered (In accordance with the Condorset Theorem).

Negative Feature Filtering

The same as for positive features:

$$\omega_1: \quad p(\overline{\omega}_1 / G_1^{(1)}),\ p(\overline{\omega}_1 / G_2^{(1)}),\ldots,\ p(\overline{\omega}_1 / G_{s_1}^{(1)})$$
$$\omega_2: \quad p(\overline{\omega}_2 / G_1^{(2)}), p(\overline{\omega}_2 / G_2^{(2)}),\ldots.,\ p(\overline{\omega}_2 / G_{s_2}^{(2)})$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\omega_m: \quad p(\overline{\omega}_m / G_1^{(m)}), p(\overline{\omega}_m / G_2^{(m)}),\ldots.,\ p(\overline{\omega}_m / G_{s_m}^{(m)})$$

Important note: Each class is specified in terms of its own feature space

Filtering rule:

Aggregated unary predicate $G_i^{(k)}$ remains in the feature list if

$$p(\overline{\omega}_k / G_i^k) + p(\omega_k / \neg G_i^k) > 0{,}5$$

otherwise it is filtered.

Peculiatities of Feature Set Formed

1. The set $\{F_R \cup G_R\}$ of predicates (positive and negative) successfully passed through filtering forms the *context – dependent feature space*.

2. All features of the set $\{F_R \cup G_R\}$, independently of their initial measurement scales, are *finally measured in Boolean* measurement *scale* thus forming *homogeneous feature space*;

3. An important peculiarity of the feature set is that *each class is specified in its specific feature space* and each feature has its own competence domain.

4. Each *feature can be interpreted as a classifier* that can be used in various decision making schemas (voting, ensemble classifier rule, etc.). It also can be *interpreted as feature that can be further aggregated, transformed, etc*.
E.g., in the developed technology the next step is feature causal analysis and second phase filtering.

5. An important *feature* property is that they *represented in terms of unary predicates* of the first order logic but not in terms of propositional variables as it usually take place.

6. Feature causality analysis

Causal Analysis: Problem Statement

Given:

1. Set of Boolean context-based features (subjected to two step filtering)

$$\boldsymbol{F_R} = \{F_R^{(1)}, F_R^{(2)}, \ldots, F_R^{(m)}\}$$

2. Set of Boolean context-based negative features (subjected to two step filtering)

$$\boldsymbol{G_R} = \{G_R^{(1)}, G_R^{(2)}, \ldots, G_R^{(m)}\}$$

3. Data sample needed to compute statistical estimations of probabilities

To find: (according to *associative classification* idea accepted in this work),

Causality-based filtered rule set based of "positive" features:

$$F_{i_1}^{(1)} \to \omega_1, \quad F_{i_2}^{(1)} \to \omega_1, \ldots, \quad F_{i_r}^{(1)} \to \omega_1.$$

$$F_{j_1}^{(2)} \to \omega_2, \quad F_{j_2}^{(2)} \to \omega_2, \ldots, \quad F_{j_s}^{(2)} \to \omega_2.$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$F_{v_1}^{(m)} \to \omega_m, \quad F_{v_2}^{(m)} \to \omega_m, \ldots, \quad F_{v_t}^{(m)} \to \omega_m.$$

Causality-based filtered rule set based on negative features:

$$G_{z_1}^{(1)} \to \overline{\omega}_1, \quad G_{z_2}^{(1)} \to \overline{\omega}_1, \ldots, \quad G_{z_r}^{(1)} \to \overline{\omega}_1.$$

$$G_{w_1}^{(2)} \to \overline{\omega}_2, \quad G_{w_2}^{(2)} \to \overline{\omega}_2, \ldots, \quad G_{w_s}^{(2)} \to \overline{\omega}_2.$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$G_{g_1}^{(m)} \to \overline{\omega}_m, \quad G_{g_2}^{(m)} \to \overline{\omega}_m, \ldots, \quad G_{g_t}^{(m)} \to \overline{\omega}_m.$$

Causality Measure and Filtering Condition

Regression coefficient of two random events (not variables!!!) A and B is defined as follows:

$$|R(A,B)| = |p(B/A) - p(B/\overline{A})| = \\ = p(A)p(B) - p(A,B)/\{p(A)[1-p(A)]\}$$

Causality-based filtering conditions:

$$for \quad \forall i, \forall k: \; R(F_i^{(k)}, \omega_k) = p(\omega_k / F_{i_1}^{(1)}) - p(\omega_k / \overline{F}_{i_1}^{(1)})$$

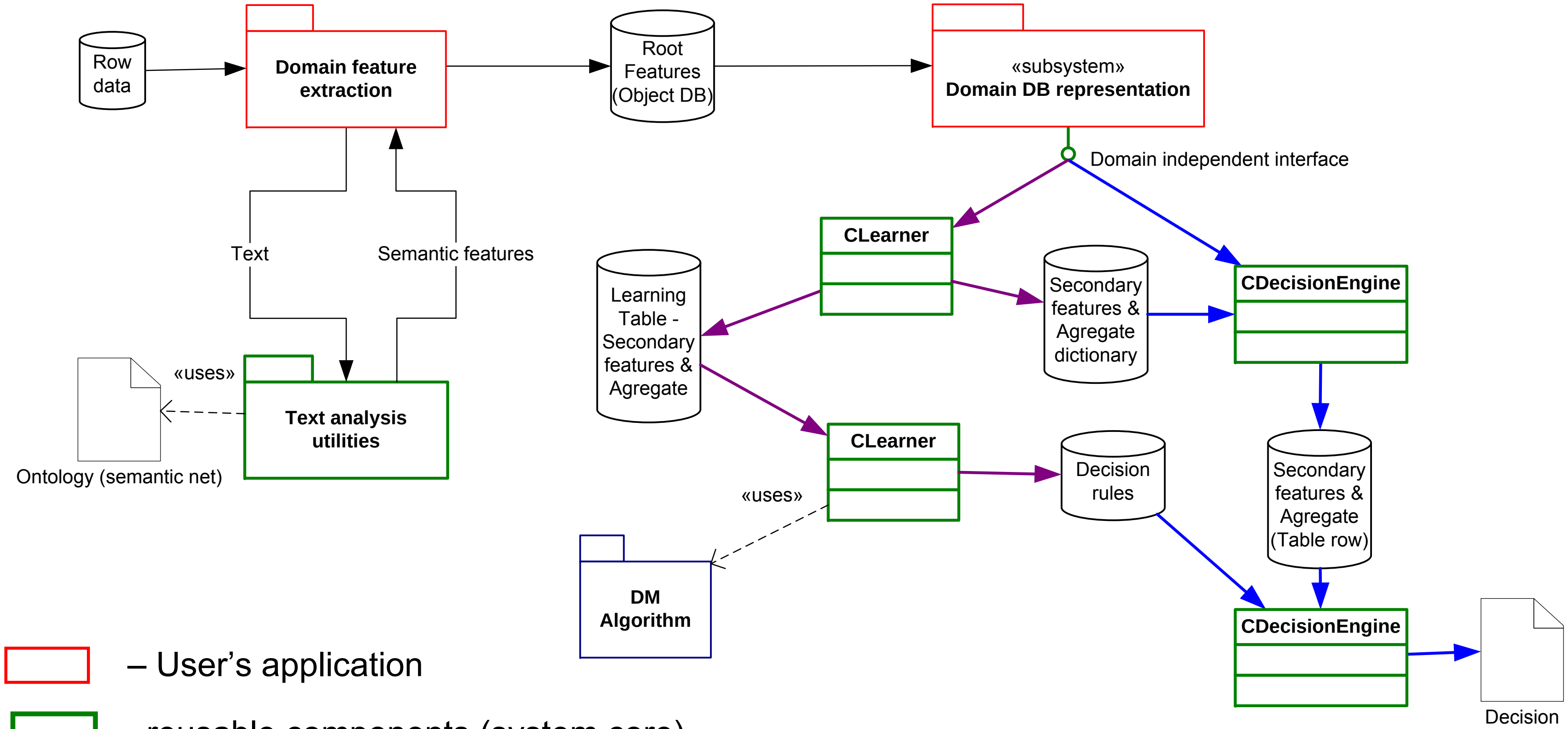
$$|R(F_i^{(k)}, \omega_k)| \geq \Delta_k .$$

$$for \quad \forall i, \forall k: \; R(G_i^{(k)}, \overline{\omega}_k) = p(\overline{\omega}_k / G_{i_1}^{(1)}) - p(\overline{\omega}_k / \overline{G}_{i_1}^{(1)})$$

$$|R(G_i^{(k)}, \overline{\omega}_k)| \geq \Delta_k .$$

7. Software implementation and reusability.

Software Implementation and Reusable Components



– User's application

–reusable components (system core)

–application-dependent components (to be developed ad-hoc)

– Decision making flow

– Data mining data flow

8. Some experimental results

Personal Outlook E-mail Assistant

Experimental settings

1. The numerical features the modified Quinlan information gain measure was used with splitting numerical domains into 10 equally probable intervals.

2. Expert generated features were subjected to two-step filtering (Naïve Bayes-based and causal filtering). For each class, 30 best features (rules) were selected.

3. Several algorithm including weighted voting algorithm were used for decision making.

Testing results (using testing sample)

1. **Accuracy averaged over all classe4s (folders)**			
	Coverage	False Alarm	Refusal
Probability	0,75	0,0833	0,167
Number of e-mails	9	1	2
Accuracy for every particular class (folder)			
Folder number			
4	1	0	0
5	1	0	0
7	0,5	0	0,5
12	1	0	0
13	0	0	1
14	1	0	0
20	0,5	0,5	0

Discussion of Experimental Results

Real user mail box (structured folders and e-mails contained in them) were used for training and testing - *without any simplifications.*

In general, the results *confirm feasibility, efficiency* and high *quality* of the technology proposed. Training and testing jointly take about 10-15 minutes.

Bad results concerning with the folders #5 and #20 result from very *limited* training data *sample* size. Actually, the *ontology needs a refinement* and *thresholds* needs more *experimentations*.

9. Conclusion: New results and perspectives

Conclusion: New Results and Technology Perspectives

- A context-driven data mining technology is proposed. It uses expert-based enrichment of the learning sample with domain ontology
- Technology is oriented to expert-based generation and selection of context-dependent features. As a result, 1) each class of decision is provided with particular set of features that can significantly differs from the sets extracted for other classes; 2) as applied to recommendation systems, the technology provides for user-personalized decision making.
- Technology proposed is applicable to the mining of large scale heterogeneous data that also can contain texts on a natural language.
- An important advantage of the technology is that feature transformation and selection mechanism proposed results in homogeneous feature space independently of types of data in initial data sample. At that, feature are represented in Boolean measurement scale in terms of unary predicates of the first order logics.
- A new technology component is causal analysis that uses new metrics to measure the strength of causal dependence between the variables which is used for effective and efficient filtering of potential set of the features. In contrast with the existing approaches, the proposed measure does require to compute explicitly neither support, nor confidence. Therefore it makes it possible to search for rare and negative causal rules.
- An experimental experience proved that the proposed approach is capable to cope with very "heavy" applications when training data set is of of terra bite size.
- Further research will be oriented for application-based verification and further modification of the technology with the more focus on social network mining and recommendation systems including web-based and mobile application.

Questions…?

Contact

E-mail: gor@iias.spb.su

http://practical-reasoning.com/

http//space.iias.spb.su/ai