Frequent Itemset Mining and Association Rules

Dmitry I. Ignatov[◇]

[◇]HSE
Computer Science Faculty
Dept. of Data Analysis and Artificial Intelligence

ML&DM 2019

Outline

1. Main part
   - Introduction
   - Applications
   - Formal Concept Analysis
   - Frequent Itemsets and Association Rules
   - Apriori Algorithm
   - FP-growth Algorithm
   - Rules Interestingness Measures
   - Compact representation of frequent itemsets
2. Applied Problems and Experiments
   - Mining of web site audience
   - Recommendation of advertising terms
3. Software tools
4. What to read and watch?

Introduction

KDD & Data Mining

- Data mining is the main step of Knowledge Discovery in Databases
- Association rules and frequent itemset mining are among the key methods of Data Mining
- The original problem is market basket analysis

On the terminology. KDD and Data Mining

Knowledge discovery in Databases (KDD)

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Fayyad, Piatetsky-Shapiro, and Smyth 1996

Data Mining

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.

The same paper.

On the terminology. KDD и Data Mining

KDD scheme

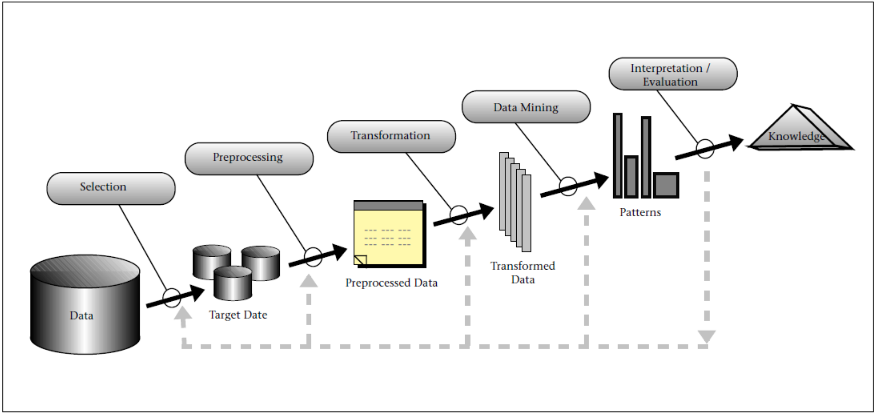


Figure 1. An Overview of the Steps That Compose the KDD Process.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

On the terminology. KDD и Data Mining

[J. Han et al., Data Mining. Concepts and Techniques, 3rd Ed., 2012]

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation
7. Knowledge presentation

Data Mining

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

On the terminology. Machine Learning

[T. Mitchell. The Discipline of Machine Learning,2006]

The main question in Machine Learning

How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

More precisely

To be more precise, we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

Interdisciplinary relations

Hypothesis

Data Mining $\stackrel{?}{=}$ Machine Learning

Related disciplines

- Computer Science
- Artificial Intelligence
- Pattern Recognition
- Information Retrieval
- Social Network Analysis
- Probability Theory and Mathematical Statistics
- Discrete Mathematics (including orders and graphs)
- Optimization

Applications of DM&ML

Applied domains

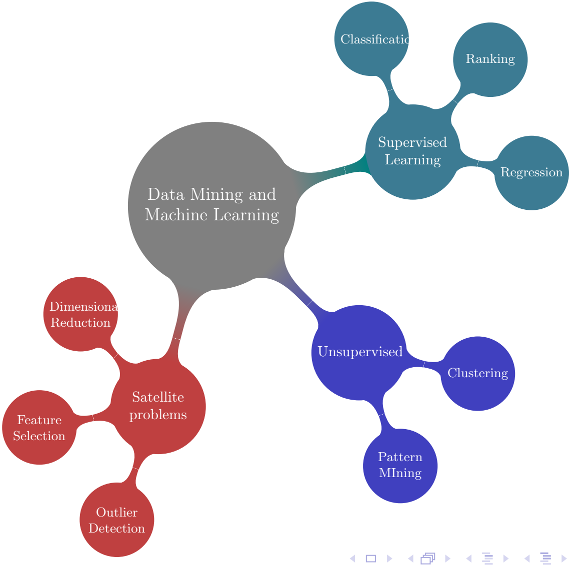
- Business
- Medicine
- Education
- Life sciences
- Internet data
- Banking and finance
- ...

Applied Trends DM&ML

[J. Han et al., 2012]

- Application exploration: e.g., counter-terrorism and mobile (wireless) data mining
- Scalable and interactive data mining methods
- Integration of data mining with search engines, database systems, data warehouse systems, and cloud computing systems
- Mining social and information networks
- Mining spatiotemporal, moving-objects, and cyber-physical system
- Mining multimedia, text, and web data
- Mining biological and biomedical data
- Data mining with software engineering and system engineering
- Visual and audio data mining
- Distributed data mining and real-time data stream mining
- Privacy protection and information security in data mining

Taxonomy of DM&ML



Pattern Mining

Problem Statement

- Pattern mining from data about (shared) usage of different resources, for example, those which are frequently used together.
- Example: $support(\{\text{bread, milk}\}) = 0.7$
- Such dependencies are often expressed as rules:

$$A \longrightarrow B$$

- Example: $\{\text{Student, Age in [16,25]}\} \longrightarrow \{iPhone, iPad\}$

Pattern Mining

The FIMI'03 best implementation award was granted to Gosta Grahne and Jianfei Zhu (on the left). The award consisted of the most frequent itemset: $\{diapers, beer\}$.

Formal Concept Analysis

[Wille, 1982], [Ganter,1999]

- G is a set of objects, M is a set of attributes attributes
- a incidence relation $I \subseteq G \times M$ such that gIm, iff the object g has the attribute m.
- $\mathbb{K} = (G, M, I)$ is called a formal context.

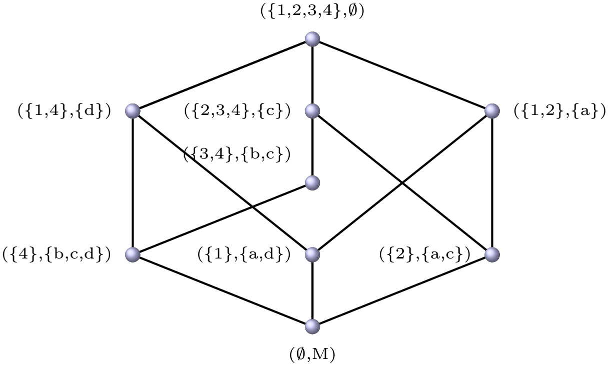
Galois operator (derivation operators): $A \subseteq G$, $B \subseteq M$

$$A' = \{m \in M \mid gIm \text{ for all } g \in A\}, B' = \{g \in G \mid gIm \text{ for all } m \in B\}.$$

A formal concept is a pair (A, B): $A \subseteq G,\ B \subseteq M,\ A' = B,\ \ B' = A.$

- A is called the (formal) extent, and B is the (formal) intent of concept (A, B).
- The concepts, ordered by $(A_1, B_1) \geq (A_2, B_2) \iff A_1 \supseteq A_2$,
forms a complete lattice, which is called the concept lattice $\underline{\mathfrak{B}}(G, M, I)$.
- $(\cdot)''$ is a closure operator (idempotent, monotone, and extensive)

Example of context of geometrical figures and its concept lattice



	G \ M	a	b	c	d
1		×			×
2		×		×	
3			×	×	
4			×	×	×

a – has exactly 3 vertices,

b – has exactly 4 vertices,

c – has a right angle,

d – is equilateral

Implications over sets of attributes

Def.

An implication $A \to B$, where $A, B \subseteq M$, takes place if $A' \subseteq B'$, i.e. each object that has all attributes from A also has all attributes from B.

Def.

Implications fulfills Armstrong rules:

$$\frac{}{X \to X}, \quad \frac{X \to Y}{X \cup Z \to Y}, \quad \frac{X \to Y, Y \cup Z \to W}{X \cup Z \to W}$$

Basic Definitions

Def. 1

Let $\mathbb{K} := (G, M, I)$ be a context, where G is a set of objects (transactions, clients), M is a set of attributes (items), $I \subseteq G \times M$
An association rule of the context $\mathbb{K}$ is defined as a dependency between attribute sets as $A \to B$, where $A, B \subseteq M$.

Often $A \cap B = \emptyset$

Basic Definitions

Def. 2

The Support of an association rule $A \to B$ is defined as follows
$supp(A \to B) = \frac{|(A \cup B)'|}{|G|}$.

The value $supp(A \to B)$ shows which fraction of objects from G contains $A \cup B$. Often this value is given in %.

Basic Definitions

Def. 2

The Support of an association rule $A \to B$ is defined as follows
$supp(A \to B) = \frac{|(A \cup B)'|}{|G|}$.

The value $supp(A \to B)$ shows which fraction of objects from G contains $A \cup B$. Often this value is given in %.

Def. 3

The confidence of an association rule $A \to B$ is defined as $conf(A \to B) = \frac{|(A \cup B)'|}{|A'|}$.

The values $conf(A \to B)$ shows which fraction of objects that have A contains $A \cup B$. This value is often expressed in %.

Basic definitions

Def. 4

A set of attributes $F \subseteq M$ is called frequent (itemset) if $supp(F) \geq min\_supp$.

Example

Object-attribute table of clients' transactions

Clients/Items	Beer	Cookies	Milk	Müesli	Chips
c_1	1	0	0	0	1
c_2	0	1	1	1	0
c_3	1	0	1	1	1
c_4	1	1	1	0	1
c_5	0	1	1	1	1

- $supp(\{\text{Beer, Chips}\}) = 3/5$
- $supp(\{\text{Cookies, Müesli}\} \to \{\text{Milk}\}) =$
  $= \frac{|(\{\text{Cookies, Müesli}\} \cup \{\text{Milk}\})'|}{|G|} = \frac{|\{C2, C5\}|}{5} = 2/5$
- $conf(\{\text{Cookies, Müesli}\} \to \{\text{Milk}\}) =$
  $= \frac{|(\{\text{Cookies, Müesli}\} \cup \{\text{Milk}\})'|}{|\{\text{Cookies, Müesli}\}'|} = \frac{|\{C2, C5\}|}{|\{C2, C5\}|} = 1$

Problem Statement

Searching for association rules, min-confidence and min-support

We need to find all the association rules of an input context such that their support and confidence are higher the constraints, min_supp and min_conf, respectively [Agrawal et al., 1993].

Association rules and implications

- The association rules with $min\_supp = 0\%$ and $min\_conf = 100\%$ are the implications of an input context.
- Sometimes, association rules are given as $A \xrightarrow[s]{c} B$, c and s are the confidence and support of the rule, respectively.

Association rules search

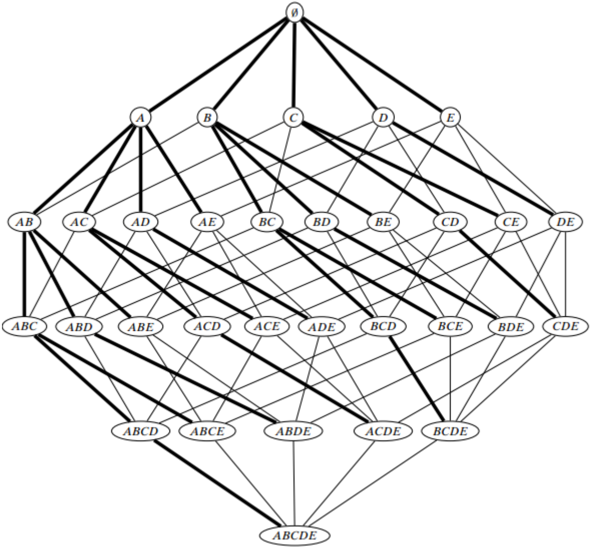
Main steps

1. Frequent itemsets search, i.e. we are looking for attribute sets with their no less than *min_supp*.
2. Generation of association rules based on the found frequent itemsets.

- The first is the most exhaustive, the second step is rather trivial.
- On of the classic algorithms for the first step is Apriori [Agrawal, Srikant, 1994]

Frequent itemset mining

Boolean Lattice Traversing



FCA meets Data Mining

- Agrawal R., RSFDGrC – 2011, Moscow

Antimonotony

Property 1 (antimonotony)

For $\forall A, B \subseteq M$ и $A \subseteq B \Rightarrow supp(B) \leq supp(A)$

- The key property for multi-element frequent itemsets
- The larger the set, the lower its support (or it remains the same)
- The support of any itemset is not greater than the minimal support of every its subset
- If the set of items of size n is frequent, then all its $(n-1)$-element sets are frequent

Apriori Algorithm

Description

It finds all frequent itemsets

Алгоритм 1.1. Apriori($Context, min\_supp$)

input: $Context$ – dataset, min_supp – minimal support
output: all frequent itemsetsI_F

$C_1 \leftarrow \{$1-itemsets$\}$
$i \leftarrow 1$
while $(C_i \neq \emptyset)$

$$\text{do} \begin{cases} SupportCount(C_i) \\ F_i \leftarrow \{f \in C_i \mid f.support \geq min\_supp\} \\ //F - \text{frequent itemsets} \\ C_{i+1} \leftarrow AprioriGen(F_i)//C - \text{candidates} \\ i{+}{+} \end{cases}$$

$I_F \leftarrow \bigcup F_i$
return (I_F)

AprioriGen Procedure

Description

for i-element frequent itemsets it generates all $(i+1)$-supersets and returns only a set of prospective frequent candidates

---

Алгоритм 1.2. AprioriGen(F_i)

---

input: F_i – frequent itemset of length$i$
output: C_{i+1} – prospective frequent candidate itemsets

insert into C_{i+1} // union
select $p[1], p[2], ..., p[i], q[i]$
from $F_i p, F_i q$
where $p[1] = q[1], ..., p[i-1] = q[i-1], p[i] < q[i]$
for each $c \in C_{i+1}$ // removal

$$\text{do} \begin{cases} S \leftarrow i\text{-element subsets} c \\ \text{for each } s \in S \\ \quad \text{do} \begin{cases} \text{if } (s \notin F_i) \\ \quad \text{then } C_{i+1} \leftarrow C_{i+1} \setminus c \end{cases} \end{cases}$$

return (C_{i+1})

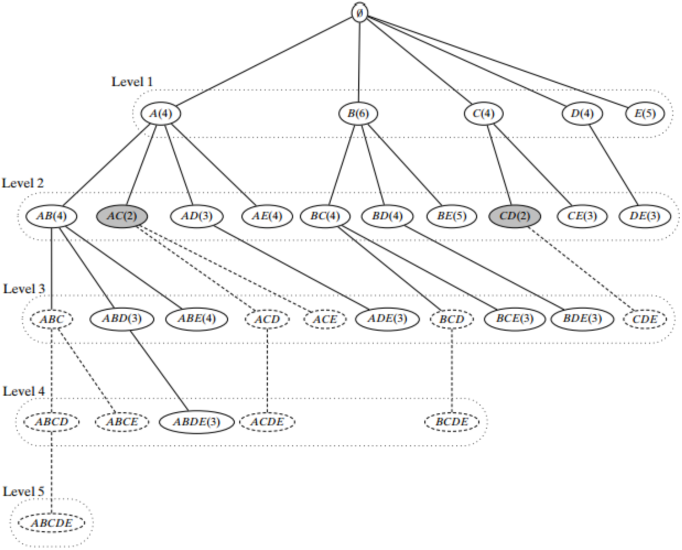
AprioriGen Example

Union and elimination steps

- $F_3 = \{\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{a, c, e\}, \{b, c, d\}\}$
- $C_4 = \{\{a, b, c, d\}, \{a, c, d, e\}\}$ — union
- $C_4 = \{\{a, b, c, d\}\}$, so we should exclude $\{a, c, d, e\}$ since $\{c, d, e\} \notin F_3$ — the removal step

Frequent itemset search

Frequent Itemset Lattice ($minsupp = 3$)



Rules generation

Rules extraction from frequent itemsets

Let F be a frequent itemset. Generate the rule $f \to F \setminus f$ if

$$conf(f \to F \setminus f) = \frac{supp(F)}{supp(f)} \geq min\_conf$$

Rules generation

Property 2

$conf(f \to F \setminus f) = \frac{supp(F)}{supp(f)}$ is maximal when $support(f)$ is maximal.

- The rule confidence is minimal when its premise consists of a single attribute. All the supersets of this attribute have lower (or at least the same) support values and, hence, greater confidence values.
- The rule extraction procedure is recursive. We start with a single-element premise f that fulfils min_conf and min_sup and check all supersets of a given F. We use all attributes from F at each step of the rule construction.

Exercise

1. By means of Apriori build all frequent itemset of the context from Example 1 for $min\_sup = 1/3$

Exercise

1. By means of Apriori build all frequent itemset of the context from Example 1 for $min\_sup = 1/3$
2. Please, say "I ♡ Apriori".

FP-growth Algorithm
[Han et al., 2000]

- Jiawei Han, Jian Pei, Yiwen Yin: Mining Frequent Patterns without Candidate Generation. SIGMOD Conference 2000: 1-12
- Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Min. Knowl. Discov. 8(1): 53-87 (2004)

FP-growth Algorithm

Example data from (Zaki & Meira, 2014)

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

(a) Binary database

t	$\mathbf{i}(t)$
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

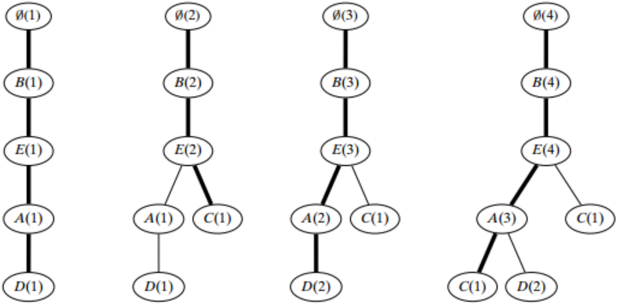
(b) Transaction database

x	A	B	C	D	E
$\mathbf{t}(x)$	1	1	2	1	1
	3	2	4	3	2
	4	3	5	5	3
	5	4	6	6	4
		5			5
		6			

(c) Vertical database

FP-growth Algorithm

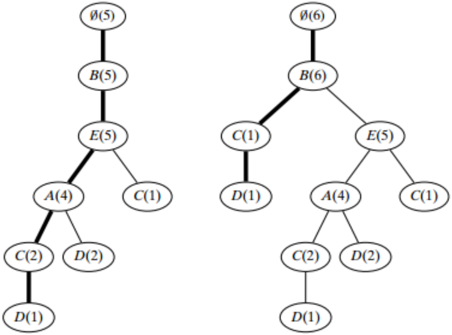
FP-tree: transactions 1-4



(a) $\langle 1, BEAD \rangle$ (b) $\langle 2, BEC \rangle$ (c) $\langle 3, BEAD \rangle$ (d) $\langle 4, BEAC \rangle$

FP-growth Algorithm

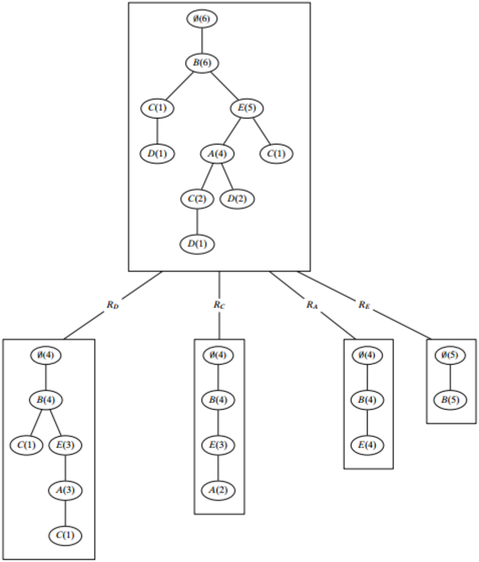
FP-tree: transactions 5–6



(e) ⟨5, *BEACD*⟩ (f) ⟨6, *BCD*⟩

Frequent Itemset Lattice

Projection for D



Rules Interestingness Measures

Zaki & Meira 2014, Chapter 12 “Pattern and Rule Assessment”

Jaquard coefficient

$$Jaquard(A,B) = \frac{|A' \cap B'|}{|A' \cup B'|} = \frac{sup(AB)}{sup(A) + sup(B) - sup(AB)}$$

Lift of $A \rightarrow B$

$$lift(A,B) = \frac{P(AB)}{P(A)P(B)} = \frac{P(A|B)}{P(A)}$$

Lift of $\neg A \rightarrow B$

$$lift(\neg A,B) = \frac{P(\neg AB)}{P(\neg A)P(B)} = \frac{P(\neg A|B)}{P(\neg A)}$$

Compact representation of frequent itemsets

Let $\mathbb{K} := (G, M, I)$ be a context.

Def. 5

An itemset $FC \subseteq M$ is called frequent closed itemset if $supp(FC) \geq min\_supp$ and there is no F such that $F \supset FC$ and $supp(F) = supp(FC)$.

Def. 6

An itemset $MFC \subseteq M$ is called maximal frequent closed itemset if it is frequent and there is no F such that $F \supset FMC$ and $supp(F) \geq min\_supp$.

Compact representation of frequent itemsets

Let $\mathbb{K} := (G, M, I)$ be a context.

Proposition 1

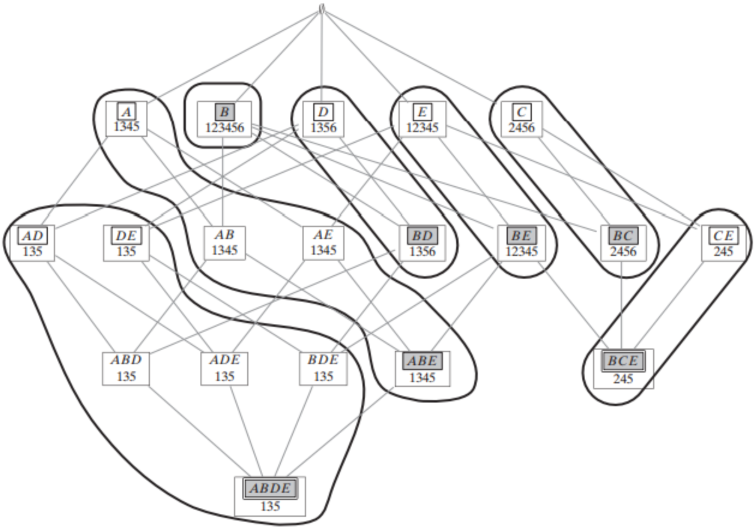
$\mathcal{MFC} \subseteq \mathcal{FC} \subseteq \mathcal{F}$, where $\mathcal{MFC}$ are maximal frequent itemsets of $\mathbb{K}$, $\mathcal{FC}$ are frequent closed itemsets, and $\mathcal{F}$ are frequent itemsets with min_supp.

Proposition 2

The lattice of formal concepts of a context $\mathbb{K}$ is isomorphic to the lattice of its frequent closed itemsets with $min\_supp = 0$.

Frequent Itemset lattice

Maximal and closed sets ($minsupp = 3$)



Outline

1 Main part
- Introduction
- Applications
- Formal Concept Analysis
- Frequent Itemsets and Association Rules
- Apriori Algorithm
- FP-growth Algorithm
- Rules Interestingness Measures
- Compact representation of frequent itemsets

2 Applied Problems and Experiments
- Mining of web site audience
- Recommendation of advertising terms

3 Software tools

4 What to read and watch?

Problem statement

Masterhost company (Spylog → Openstat), 2006-2007

- Having webcounters data, to identify audience tastes
- We proposed an FCA-based model with criteria for relevant concepts selection

Website taxonomies: a model

External taxonomy

$\mathbb{K}_{ex} = (V, S_{ex}, I)$, where
V is the set of all visitors of the target website, S_{ex} is the set of all websites excluding the target one, I is the incidence relation such that vIs, $v \in V$, $s \in S_{ex} \Leftrightarrow$ if the visitor v "went" to the site s.

Internal taxonomy

$\mathbb{K}_{in} = (V, S_{in}, I)$, where
V is the set of all visitors of the target website, S_{in} is the set of all webpages of the target website, I is the incidence relation such that vIs, $v \in V$, $s \in S_{in} \Leftrightarrow$ if v "went" to the site s.

- The concept is a pair (A, B) such that
- $A' = \{$ the sites $s \in S$ that have been visited by $v \in A\} = B$
- $B' = \{$ the visitors $v \in V$ that visited all the sites $s \in B\} = A$.

Relevant concepts criteria

Let $\mathbb{K} = (G, M, I)$ be a formal context, (A, B) be a certain formal concept $\mathbb{K}$.

Stability index

Thestability index σ of (A, B) is defined as

$$\sigma(A, B) = \frac{|\{C \subseteq A | C' = B\}|}{2^{|A|}}.$$

Clearly, $0 \leq \sigma(A, B) \leq 1$.

Iceberg lattice

The support of the intent of (A, B) is defined as $supp(A, B) = \frac{|A|}{|G|}$. Let $minsupp \in [0, 1]$, then an iceberg lattice is a set $\{(A, B) | supp(B) \geq minsupp\}$.

Input data

- a sample of 10000 websites with a flat thematic catalog for 59 categories.
- a university website, household equipment webstore, large bank, car dealer.

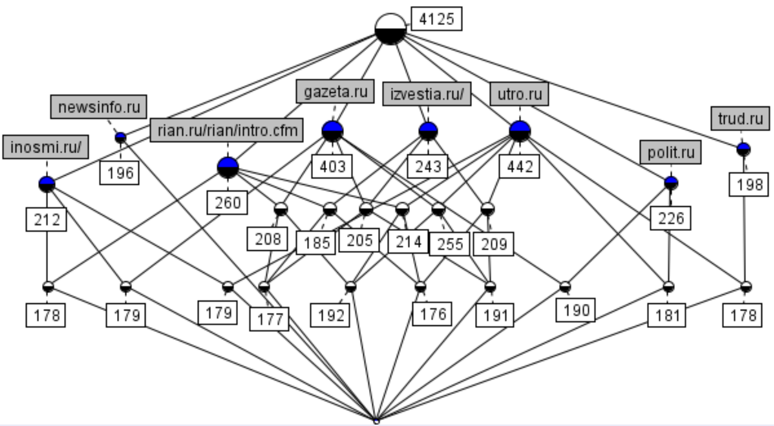
Data description

id; \\user id
first_ts; \\the time of the first visit
last_ts; \\the time of the last visit
num; \\the number of all sessions

External taxonomy building

HSE website in September, 2006 in terms of news resources.

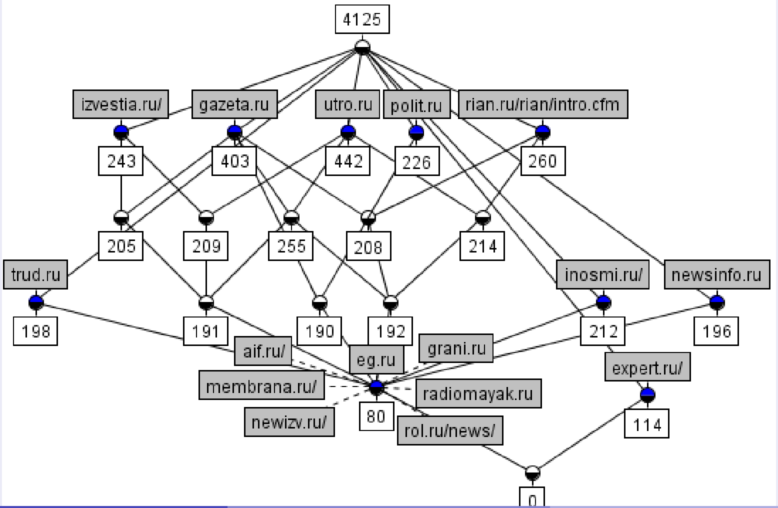
Iceberg lattice for 25 the largest concepts



External taxonomy example

HSE website in September, 2006 in terms of news resources.

The line diagram of partially ordered set of 25 the most stable concepts



Recommendation of advertising terms

1. R&D of algorithms for forming recommendations on Internet data
2. Experimental validation of Data Mining techniques for Internet advertising

Problem Statement

- contextual Internet advertising
- searching for potentially relevant terms (for companies)
- example — Google AdWords

Recommendation of advertising terms

Input data

Data about terms' purchases. A formal context $\mathbb{K}_{FT} = (F, T, I_{FT})$, F is a set of firms, T is a set of advertising terms, fIt means that the firm $f \in F$ bough the term $t \in T$. The context size is 2000×3000.

Problem statement

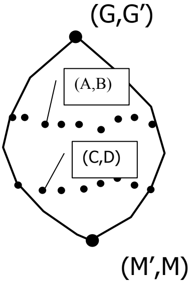
To identify advertising markets to form recommendations

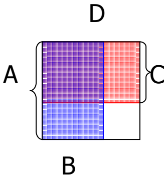
Prospective tools

- FCA: D-miner algorithm
- association rules
- association rules+morphology
- association rules+ontology

Recommendation of advertising terms: FCA

[Besson et al, 2004], D-miner, $O(|G|^2|M||L|)$





Results of D-miner

Min size of extent	Min size of intent	Number of formal concepts
0	0	8 950 740
10	10	3 030 335
15	10	759 963
15	15	150 983
15	20	14 226
20	15	661

Recommendation of advertising terms: D-miner

Web hosting market

{affordable hosting web, business hosting web, cheap hosting, cheap hosting site web, cheap hosting web, company hosting web, cost hosting low web, discount hosting web, domain hosting, hosting internet, hosting page web, hosting service, hosting services web, hosting site web, hosting web}

Hotel business

{ angeles hotel los, atlanta hotel, baltimore hotel, dallas hotel, denver hotel, diego hotel san, francisco hotel san, hotel houston, hotel miami, hotel new orleans, hotel new york, hotel orlando, hotel philadelphia, hotel seattle, hotel vancouver }

Recommendation of advertising terms: association rules

- [Szathmary, 2005]
- Coron system, Zart algorithm, informative base of association rules

Rules' examples

minsupp=30 minconf=0,9

- $\{florist\} \to \{flower\}$ supp=33 [1.65%]; conf=0.92;
- $\{gift\ graduation\} \to \{anniversary\ gift\}$, supp=41 [2.05%]; conf=0.82;

Results of associations' search

min_supp	max_supp	min_conf	max_conf	number of rules
30	86	0,9	1	101 391
30	109	0,8	1	144 043

Recommendation of advertising terms: association rules+morphology

- t — advertising term, $t = \{w_1, w_2, \dots, w_n\}$
- $s_i = stem(w_i)$ — the stem of the word w_i
- $stem(t) = \bigcup\limits_i stem(w_i)$ — the set of the stems of t
- $\mathbb{K}_{TS} = (T, S, I_{TS})$ — a formal context, where T is the set of all terms, S the set of all stems for terms in T, i.e. $S = \bigcup\limits_i stem(t_i)$
- tIs means that the stems of t contain s

Recommendation of advertising terms: association rules+morphology

A context example, $\mathbb{K}_{FT}$, for the "long distance calling" market

firm \ term	call distance long	calling distance long	calling distance long plan	carrier distance long	cheap distance long
f_1	x		x		x
f_2		x	x	x	
f_3				x	x
f_4		x	x		x
f_5	x	x		x	x

Recommendation of advertising terms: association rules+morphology

A context example, $\mathbb{K}_{TS}$, for the "long distance calling" market

phrase \ stem	call	carrier	cheap	distanc	long	plan
call distance long	x			x	x	
calling distance long	x			x	x	
calling distance long plan	x			x	x	x
carrier distance long		x		x	x	
cheap distance long			x	x	x	

Recommendation of advertising terms: association rules+morphology

Examples

- $t \xrightarrow{FT} s_i^{I_{TS}}$

  $\{last\ minute\ vacation\} \rightarrow \{last\ minute\ travel\}$
  Supp= 19 Conf= 0,90
- $t \xrightarrow{FT} \bigcup\limits_i s_i^{I_{TS}}$
- $\{mail\ order\ phentermine\} \rightarrow$
  $\{adipex\ online\ order, adipex\ order, adipex\ phentermine, \ldots,$
  $phentermine\ prescription, phentermine\ purchase, phentermine\ sale\}$
  Supp= 19 Conf= 0,95

Recommendation of advertising terms: association rules+morphology

Examples

- $t \xrightarrow{FT} (\bigcup_i s_i)^{I_{TS}}$
- $\{distance\ long\ phone\} \rightarrow$
  $\{call\ distance\ long\ phone, carrier\ distance\ long\ phone, \ldots,$
  $distance\ long\ phone\ rate, distance\ long\ phone\ service\}$
  Supp= 37   Conf= 0,88
- $t_1 \xrightarrow{FT} t_2$ such that $t_2^{I_{TS}} \subseteq t_1^{I_{TS}}$
- $\{ink\ jet\} \rightarrow \{ink\}$, Supp= 14   Conf= 0,7

Recommendation of advertising terms: association rules+morphology

$min\_conf = 0.5$

Rules assessment

Rule type	Average value of supp	Average value of conf	Number of rules
$t \xrightarrow{FT} s_i^{I_{TS}}$	15	0,64	454
$t \xrightarrow{FT} \bigcup\limits_i s_i^{I_{TS}}$	15	0,63	75
$t \xrightarrow{FT} (\bigcup\limits_i s_i)^{I_{TS}}$	18	0,67	393
$t \xrightarrow{FT} t_i$, где $t_i^{I_{TS}} \subseteq t^{I_{TS}}$	21	0,70	3922
$t \xrightarrow{FT} \bigcup\limits_i t_i$, где $t_i^{I_{TS}} \subseteq t^{I_{TS}}$	20	0,69	673

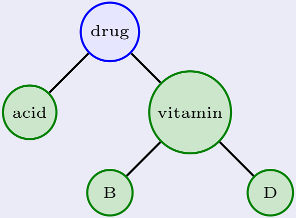
Recommendation of advertising terms: association rules

Crossvalidation for association rules

	Number of rules	Number of rules c sup > 0	average_conf	Number of rules c min_conf=0.5	average_conf (min_conf=0.5)
1	147170	73025	0,77	65556	0,84
2	69028	68709	0,93	68495	0,93
3	89332	89245	0,95	88952	0,95
4	107036	93078	0,84	86144	0,90
5	152455	126275	0,82	113008	0,90
6	117174	114314	0,89	111739	0,91
7	131590	129826	0,95	128951	0,96
8	134728	120987	0,96	106155	0,97
9	101346	67873	0,72	52715	0,92
10	108994	107790	0,93	106155	0,94
average	115885	99112	0,87	92787	0,92

Recommendation of advertising terms: association rules+ontology

Composing an ontology (hierarchical catalog or taxonomy)



Metarules

- сопоставление правилам онтологии ассоциаций
- $t \to g_i(t)$, где $g_i(t)$ — множество понятий онтологии на i уровней выше t
- $t \to n(t)$, где $n(t)$ — множество соседних для t понятий онтологии, имеющих общего предка

Recommendation of advertising terms: association rules+ontologies

Examples of rules

- $t \to g_1(t)$
- {d vitamin} $\to$ {vitamin }, Supp$=$ 19 Conf$=$ 0,90
- $t \to n(t)$
- {b vitamin} $\to$ { b complex vitamin, b12 vitamin, c vitamin, d vitamin, discount vitamin, e vitamin, herb vitamin, mineral vitamin, multi vitamin, supplement vitamin} Supp$=$ 18 Conf$=$ 0,7

Outline

1. Main part
   - Introduction
   - Applications
   - Formal Concept Analysis
   - Frequent Itemsets and Association Rules
   - Apriori Algorithm
   - FP-growth Algorithm
   - Rules Interestingness Measures
   - Compact representation of frequent itemsets
2. Applied Problems and Experiments
   - Mining of web site audience
   - Recommendation of advertising terms
3. Software tools
4. What to read and watch?

Freely available tools

- SPMF – an open-source data mining mining library
- The CORON Data Mining Platform
- Bart Goethals webpage and FIMI repository
- Conexp — concept lattices, implications and association rules
- Orange – contains widgets for frequent itemset mining and association rules (version 2.7)
- Spark ML Lib – frequent itemset mining via FP-growth and association rules
- Frequent Itemset Mining in Python
- Frequent Itemset Mining Implementations Repository

Outline

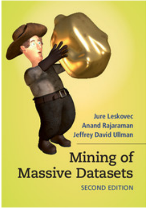
1. Main part
   - Introduction
   - Applications
   - Formal Concept Analysis
   - Frequent Itemsets and Association Rules
   - Apriori Algorithm
   - FP-growth Algorithm
   - Rules Interestingness Measures
   - Compact representation of frequent itemsets
2. Applied Problems and Experiments
   - Mining of web site audience
   - Recommendation of advertising terms
3. Software tools
4. What to read and watch?

Books

- M. Zaki et al. Data Mining and Analysis: Fundamental Concepts and Algorithms, 2014 (free)
- J. Leskovec et al. Mining of Massive Datasets, 2014 (free)
- J. Han et al. Data Mining. Concepts and Techniques, 2012
- Aggarwal, Charu C., Han, Jiawei (eds.) Frequent Pattern Mining
- Barsegian A. et al. Analysis of Data and Processes, 2009 (In Russian)

Coursera: courses and specialisations

http://www.coursera.org/



- Jiawei Han Pattern Discovery in Data Mining (current)
- Jure Leskovec et al. (current)

Specialisations (Paid certificates) feature separate courses (participation is for free)

- Data Mining (current)

ИНТУИТ

http://intuit.ru

- Internet University of Information Technologies
- K. Vorontsov Machine Learning, 2015 (Videos for Yandex Data Analysis School (In Russian))
- I. Chubukova. Data Mining (In Russian), 2006

Community and Data Sources

- FIMI – Frequent Itemset Mining Implementations Repository
- IMLS – The International Machine Learning Society
- Kaggle – Data mining competition platform
- KDD Nuggets – Data Mining Community Top Resource
- Open ML – Machine Learning community portal
- UCI Machine Learning Repository – Data repository

Conferences

- IEEE ICDM – IEEE International Conference on Data Mining
- KDD – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- ECML & PKDD – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- AIST (АИСТ) – International conference on Analysis of Images, Social Networks, and Texts

Just for fun или шутки ради

http://dilbert.com

Questions and contacts

www.hse.ru/staff/dima

Thank you!

dmitrii.ignatov[at]gmail.com