Data mining, social networks and ethical implications

Toon Calders – Departement Wiskunde-Informatica

University of Antwerp

Data Mining

Data mining is the use of automatic techniques to "discover *knowledge*"

- Data driven discovery
- Making implicit knowledge explicit

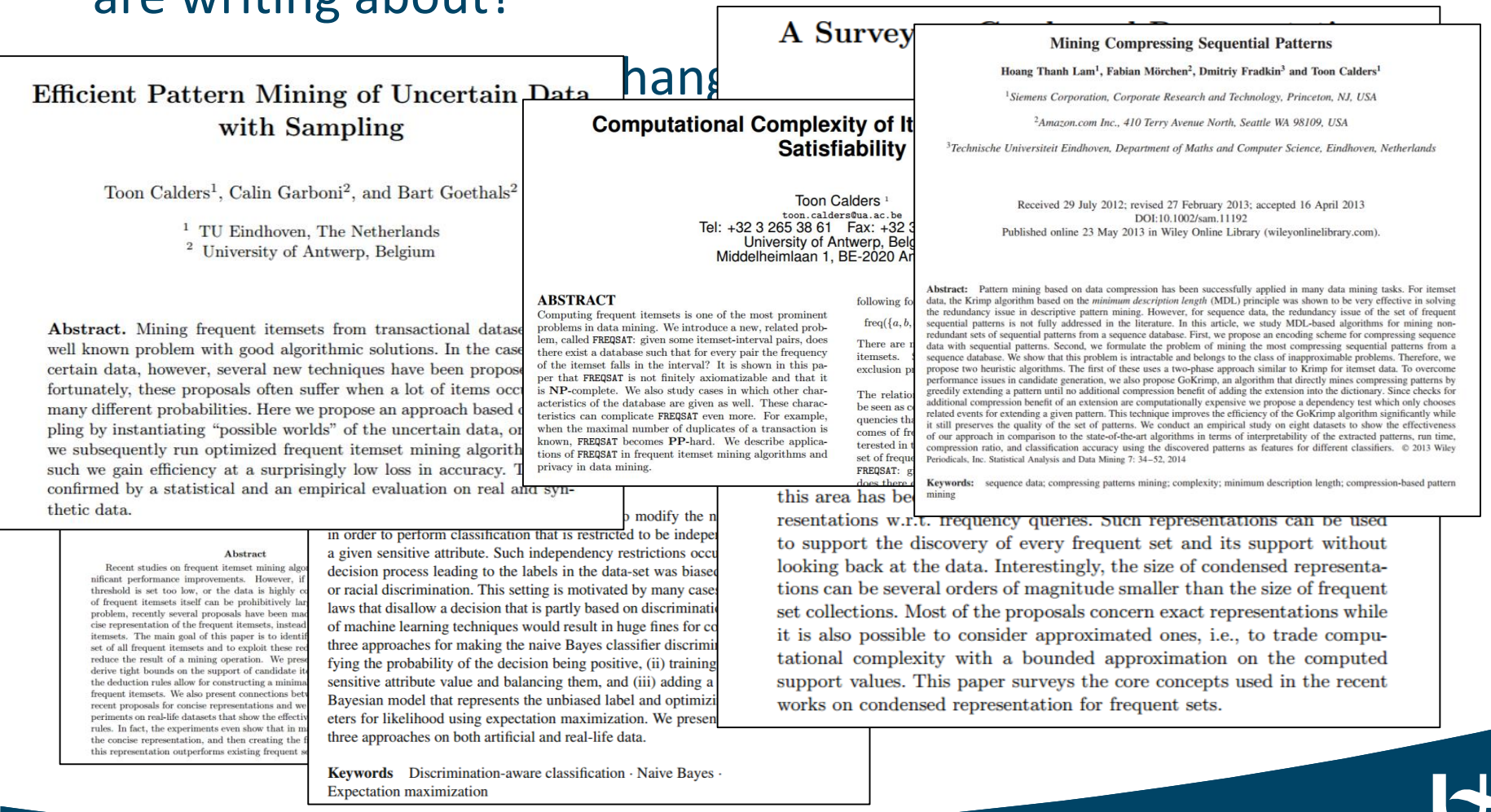
Example: Research papers

Can we automatically detect topics one or more researchers are writing about?

Are there remarkable changes in the topics they write about?

Example: Research papers

Can we automatically detect topics one or more researchers are writing about?



Efficient Pattern Mining of Uncertain Data with Sampling

Toon Calders[1], Calin Garboni[2], and Bart Goethals[2]

[1] TU Eindhoven, The Netherlands
[2] University of Antwerp, Belgium

Abstract. Mining frequent itemsets from transactional datase well known problem with good algorithmic solutions. In the case certain data, however, several new techniques have been propos fortunately, these proposals often suffer when a lot of items occ many different probabilities. Here we propose an approach based pling by instantiating "possible worlds" of the uncertain data, o we subsequently run optimized frequent itemset mining algorith such we gain efficiency at a surprisingly low loss in accuracy. confirmed by a statistical and an empirical evaluation on real and syn- thetic data.

Computational Complexity of It Satisfiability

Toon Calders [1]
toon.calders@ua.ac.be
Tel: +32 3 265 38 61 Fax: +32 3
University of Antwerp, Belg
Middelheimlaan 1, BE-2020 Ar

ABSTRACT
Computing frequent itemsets is one of the most prominent problems in data mining. We introduce a new, related problem, called FREQSAT: given some itemset-interval pairs, does there exist a database such that for every pair the frequency of the itemset falls in the interval? It is shown in this paper that FREQSAT is not finitely axiomatizable and that it is NP-complete. We also study cases in which other characteristics of the database are given as well. These characteristics can complicate FREQSAT even more. For example, when the maximal number of duplicates of a transaction is known, FREQSAT becomes PP-hard. We describe applications of FREQSAT in frequent itemset mining algorithms and privacy in data mining.

Mining Compressing Sequential Patterns

Hoang Thanh Lam[1], Fabian Mörchen[2], Dmitriy Fradkin[3] and Toon Calders[1]

[1]*Siemens Corporation, Corporate Research and Technology, Princeton, NJ, USA*

[2]*Amazon.com Inc., 410 Terry Avenue North, Seattle WA 98109, USA*

[3]*Technische Universiteit Eindhoven, Department of Maths and Computer Science, Eindhoven, Netherlands*

Received 29 July 2012; revised 27 February 2013; accepted 16 April 2013
DOI:10.1002/sam.11192
Published online 23 May 2013 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Pattern mining based on data compression has been successfully applied in many data mining tasks. For itemset data, the Krimp algorithm based on the *minimum description length* (MDL) principle was shown to be very effective in solving the redundancy issue in descriptive pattern mining. However, for sequence data, the redundancy issue of the set of frequent sequential patterns is not fully addressed in the literature. In this article, we study MDL-based algorithms for mining non-redundant sets of sequential patterns from a sequence database. First, we propose an encoding scheme for compressing sequence data with sequential patterns. Second, we formulate the problem of mining the most compressing sequential patterns from a sequence database. We show that this problem is intractable and belongs to the class of inapproximable problems. Therefore, we propose two heuristic algorithms. The first of these uses a two-phase approach similar to Krimp for itemset data. To overcome performance issues in candidate generation, we also propose GoKrimp, an algorithm that directly mines compressing patterns by greedily extending a pattern until no additional compression benefit of adding the extension into the dictionary. Since checks for additional compression benefit of an extension are computationally expensive we propose a dependency test which only chooses related events for extending a given pattern. This technique improves the efficiency of the GoKrimp algorithm significantly while it still preserves the quality of the set of patterns. We conduct an empirical study on eight datasets to show the effectiveness of our approach in comparison to the state-of-the-art algorithms in terms of interpretability of the extracted patterns, run time, compression ratio, and classification accuracy using the discovered patterns as features for different classifiers. © 2013 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 7: 34–52, 2014

Keywords: sequence data; compressing patterns mining; complexity; minimum description length; compression-based pattern mining

A Survey

this area has be
resentations w.r.t. frequency queries. Such representations can be used to support the discovery of every frequent set and its support without looking back at the data. Interestingly, the size of condensed representations can be several orders of magnitude smaller than the size of frequent set collections. Most of the proposals concern exact representations while it is also possible to consider approximated ones, i.e., to trade computational complexity with a bounded approximation on the computed support values. This paper surveys the core concepts used in the recent works on condensed representation for frequent sets.

Abstract

Recent studies on frequent itemset mining algo
nificant performance improvements. However, if
threshold is set too low, or the data is highly c
of frequent itemsets itself can be prohibitively lar
problem, recently several proposals have been ma
cise representation of the frequent itemsets, instead
itemsets. The main goal of this paper is to identif
set of all frequent itemsets and to exploit these re
reduce the result of a mining operation. We pres
derive tight bounds on the support of candidate it
the deduction rules allow for constructing a minima
frequent itemsets. We also present connections bet
recent proposals for concise representations and we
periments on real-life datasets that show the effecti
rules. In fact, the experiments even show that in m
the concise representation, and then creating the f
this representation outperforms existing frequent s

hang

o modify the n
in order to perform classification that is restricted to be indepe
a given sensitive attribute. Such independency restrictions occu
decision process leading to the labels in the data-set was biase
or racial discrimination. This setting is motivated by many case
laws that disallow a decision that is partly based on discriminati
of machine learning techniques would result in huge fines for c
three approaches for making the naive Bayes classifier discrimi
fying the probability of the decision being positive, (ii) training
sensitive attribute value and balancing them, and (iii) adding a
Bayesian model that represents the unbiased label and optimizi
eters for likelihood using expectation maximization. We presen
three approaches on both artificial and real-life data.

Keywords Discrimination-aware classification · Naive Bayes · Expectation maximization

Example: Research papers

Can we automatically detect topics researchers are writing about?

Harder than “just counting” frequent subsequences.

Pattern	Support	Pattern	Support
algorithm algorithm	0.376	method method	0.250
learn learn	0.362	algorithm result	0.247
learn algorithm	0.356	Data set	0.244
algorithm learn	0.288	learn learn learn	0.241
data data	0.284	learn problem	0.239
learn data	0.263	learn method	0.229
model model	0.260	algorithm data	0.229
problem problem	0.258	learn set	0.228
learn result	0.255	problem learn	0.227
problem algorithm	0.251	algorithm algorithm algorithm	0.222

Figure 1: The 20 most frequent non-singleton closed sequential patterns from the JMLR abstracts datasets. This set, despite containing some meaningful patterns, is very redundant.

Example: Research papers

Can we automatically detect topics researchers are writing about?

Harder than “just counting” frequent subsequences.

Minimal description length principle

A good model learns us a lot about the data and allows to represent it compactly. We hence measure model quality as:

$$\text{Length}(M) + \text{Length}(D|M)$$

Better Model = More Compression

Biased die

1/2

0

1/4

10

1/8

110

1/16

1110

1/32

11110

1/32

11111

Expected code length:

$$\frac{1}{2}+\frac{1}{2}+\frac{3}{8}+\frac{4}{16}+\frac{5}{32}+\frac{5}{32}$$

$$= 1.84 \ldots$$

Versus without information:

$$\log(6) = 2.58 \ldots$$

Minimal Description Length

Pattern	Support	Pattern	Support
algorithm algorithm	0.376	method method	0.250
learn learn	0.362	algorithm result	0.247
learn algorithm	0.356	Data set	0.244
algorithm learn	0.288	learn learn learn	0.241
data data	0.284	learn problem	0.239
learn data	0.263	learn method	0.229
model model	0.260	algorithm data	0.229
problem problem	0.258	learn set	0.228
learn result	0.255	problem learn	0.227
problem algorithm	0.251	algorithm algorithm algorithm	0.222

$$\sum_{w \in D} \left(\log \frac{F_C}{f_C(w)} * f_C(w) + g_C(w) \right)$$

Example: Research papers

Can we automatically detect topics researchers are writing about?

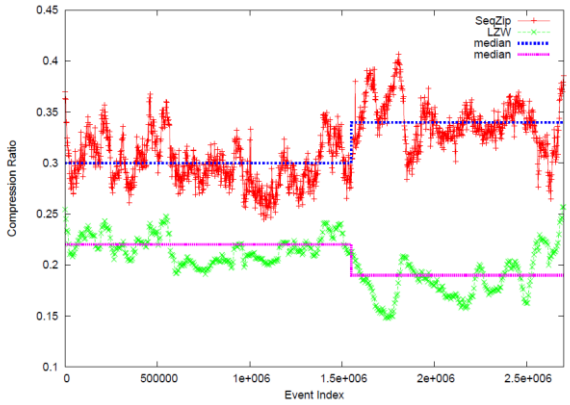
Harder than “just counting” frequent subsequences.

Based on Minimal Description Length principle:

support vector machin	larg scale	featur select	sampl size
machin learn	nearest neighbor	graphic model	learn algorithm
state art	decis tree	real world	princip compon analysi
data set	neural network	high dimension	logist regress
bayesian network	cross valid	mutual inform	model select

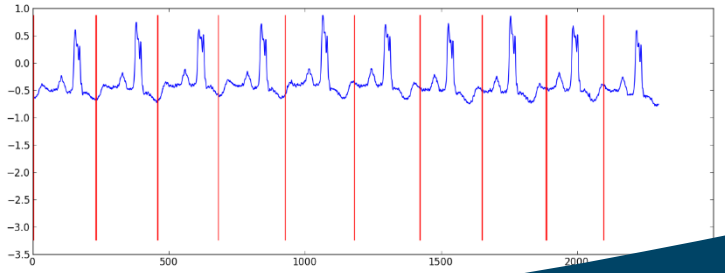
Applications of MDL Pattern Selection

Data summarization



Anomaly and drift detection

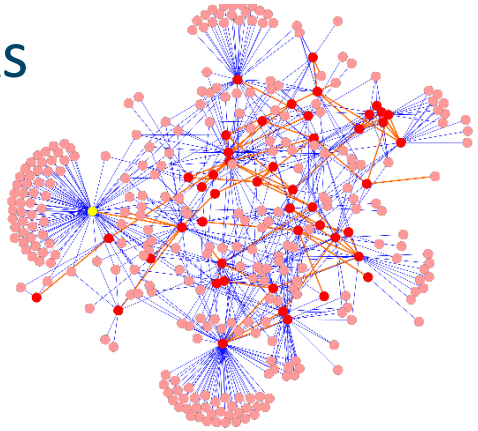
Creating features representing regularity of data



Social Network Data

Another great source of data: social networks

- How do people interact?
- Who is influencing whom?
- How is information spreading?

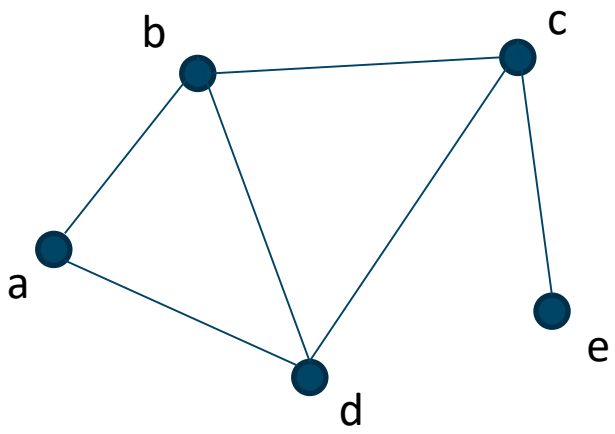


However: due to data size even simple questions become challanging to answer in reasonable time

Neighborhood Function

Count the number of pairs of nodes at distance 1, 2, 3, …

1: 6
2: 3
3: 1



source	dest
a	b
b	c
c	e
d	a
d	c
b	d

Allows to compute average degree, diameter, effective diameter.

Neighborhood Function

Straightforward algorithm:

Input: G(V,E)

- Let $N_i(v)$ be the nodes at distance *i* of *v*;
- Exploit $N_{i+1}(v) = \mathrm{U}_{w:\{v,w\}\in E}\, N_i(w)$

Time: $O(\delta\ |E|\ |V|)$

Space: $O(|V|^2)$

Tool: Sketch
(Main idea behind Flajolet-Martin sketches)

How many people
watch my presentation?

Do not keep track of all people, only remember the lowest number

Tool: Sketch
(Main idea behind Flajolet-Martin sketches)

How low will this lowest number be?

- The more people there are, the lower we expect this number to be (1/(n+1) to be exact)
- We can use this number to estimate the number of people that saw the presentation: 1/min – 1

What about the variance?

- Give every person a tuple of *k* numbers
- Keep track of $\min(num_1)$, $\min(num_2)$, ..., $\min(num_k)$
- Take median of the estimations → large error only if there is a large error in 50% of the estimates

Neighborhood Function

Based on the sketches of two sets, it is straightforward to create a sketch of their union

Using sketches for $|N_i(v)|$ instead of $N_i(v)$ itself brings down complexity to time $O(\delta\ |E|)$ and space $O(|V|)$

- Next to sketching, also sampling, distributed computing
- Solving several social network questions
  - Who influences whom (retweets, repost messages, repeat actions)
  - How does information spread through the network

Ethical Considerations

"Men between the ages of 16 and 25 are much more likely to be involved in accidents, or be cited for traffic violations."
Sam Belden, Insurance.com VP

"Taking the gender of the insured individual into account as a risk factor in insurance contracts constitutes discrimination."
March 1, 2011 EU Court of Justice

Research Question: Can we still build profiles taking legal restrictions into account?

Conclusion

Data Mining for finding patterns in large data collections

Several challenges:

- Finding relevant and orthogonal patterns
  - Minimal Description Length principle
- Data size; e.g., social network analysis
  - Sketching, sampling, approximation
  - Distributed computing
- Ethical aspects
  - What if our classifiers are not “politically correct”. Is this our problem?

University
of Antwerp