Pattern Set Mining

prof. dr Arno Siebes

Algoritmic Data Analysis

April 29, 2022

The Course

Things to Discuss

- the topic
- how it is done
- how it is examined

More information

- this teams site, obviously
- http://www.cs.uu.nl/docs/vakken/mpsm/

The Course

The Topic

Patterns

At the basis of (almost) all data mining and machine learning techniques you will find

patterns

Patterns in data

- are things that recur in the data

You can see a pattern as a query

- here it occurs, there it doesn't

Patterns are described in some

- *pattern language*

And usually there is a *quality function*

- to distinguish interesting from uninteresting patterns

Frequent Patterns

One way to determine whether or not a pattern is interesting is by its *frequency*

- how often it occurs in the data

For example

- items sold together in a shop
- routes taken by commuters
- characteristics of patients
- and so on and so on

The rationale is:

- things that often occur tell you something about the data
- and may help you, e.g., to classify or cluster, or ...

Posh Food Ltd

.

You own an upmarket supermarket chain, selling mindbogglingly expensive food and drinks to customers with more money than sense

- and you wonder how you can wring even more money out of your customers

You think it might be a good idea to suggest items that go well with what they already plan to buy.

- e.g., that a bottle of Krug Clos d'Ambonnay goes very well with the Iranian Almas Beluga Caviar they just selected.

But, unfortunately, you are not as rich as your clientèle, so you actually don't know

- so, you decide to look for patterns in your data
- sets of items – also known as itemsets – that your customers regularly buy together.

You decide to mine your data for all frequent itemsets

- all sets of items that have been bought more then θ times in your chain.

Your First Idea

You collect all transactions over the past year in your chain

- there turn out to be millions of them, a fact that makes you happy.

Since you only sell expensive stuff

- nothing below a thousand or so; you do sell potatoes, but only "La Bonnotte"

you only have a few thousand different items for sale.

Since, you want to know which sets of items sell well,

- you decide to list simply all sets of items
- and check whether or not they were sold together θ times or more.

And this is a terrible idea!

- as you discover when you break off the computation after a week long run

Why is it Naive?

A set with n elements has 2^n subsets

- $2^{1000} \approx 10^{301}$

The universe is about 14×10^9 years old

- and a year has about 31 million seconds

A modern CPU runs at about 5Ghz $= 5 \times 10^9$ Hz

- which means that the universe is about $14 \times 10^9 \times 31 \times 10^6 \times 5 \times 10^9 = 2,2 \times 10^{25}$ clockticks old

So, if your database fits into the CPU cache

- and you can check one itemset per clocktick
- and you can parallelise the computation perfectly

You would need

- $10^{301}/(2,2 \times 10^{25}) \approx 5 \times 10^{275}$ computers that have been running in parallel since the big bang to finish about now!

The number of elementary particles in the observable universe is, according to Wikipedia, about 10^{97}

- excluding dark matter

A New Idea

Feeling disappointed, you idly query your database.

- how many customers bought your favourite combination?
- Wagyu beef with that beautiful white Italian truffle accompanied by a bottle of Romanée-Conti Grand Cru

And to your surprise you find 0! You search for a reason

- plenty people buy Wagyu or white truffle or Romanée-Conti – actually, they belong to your top sold items
- quite a few people buy Wagyu and Romanée-Conti and the same holds for Wagyu and white truffle
- but no-one buys white truffle and Romanée-Conti
- those Philistines prefer a Chateau Pétrus with their truffle!
  - on second thoughts: not a bad idea

Clearly you cannot buy Wagyu and white truffle and Romanée-Conti more often

- than you buy white truffle and Romanée-Conti!

A Priori

With this idea in mind, you implement the A Priori algorithm

- simple levelwise search
- you only check sets of n elements for which all subsets of $n-1$ elements are frequent

After you finished your implementation

- you throw your data at it
- and a minute or so later you have all your frequent itemsets.

In principle, all subsets could be frequent

- and A Priori would be as useless as the naive idea

But, fortunately, people do not buy that many different items in one transaction

- whatever you seem to observe on your weekly shopping run

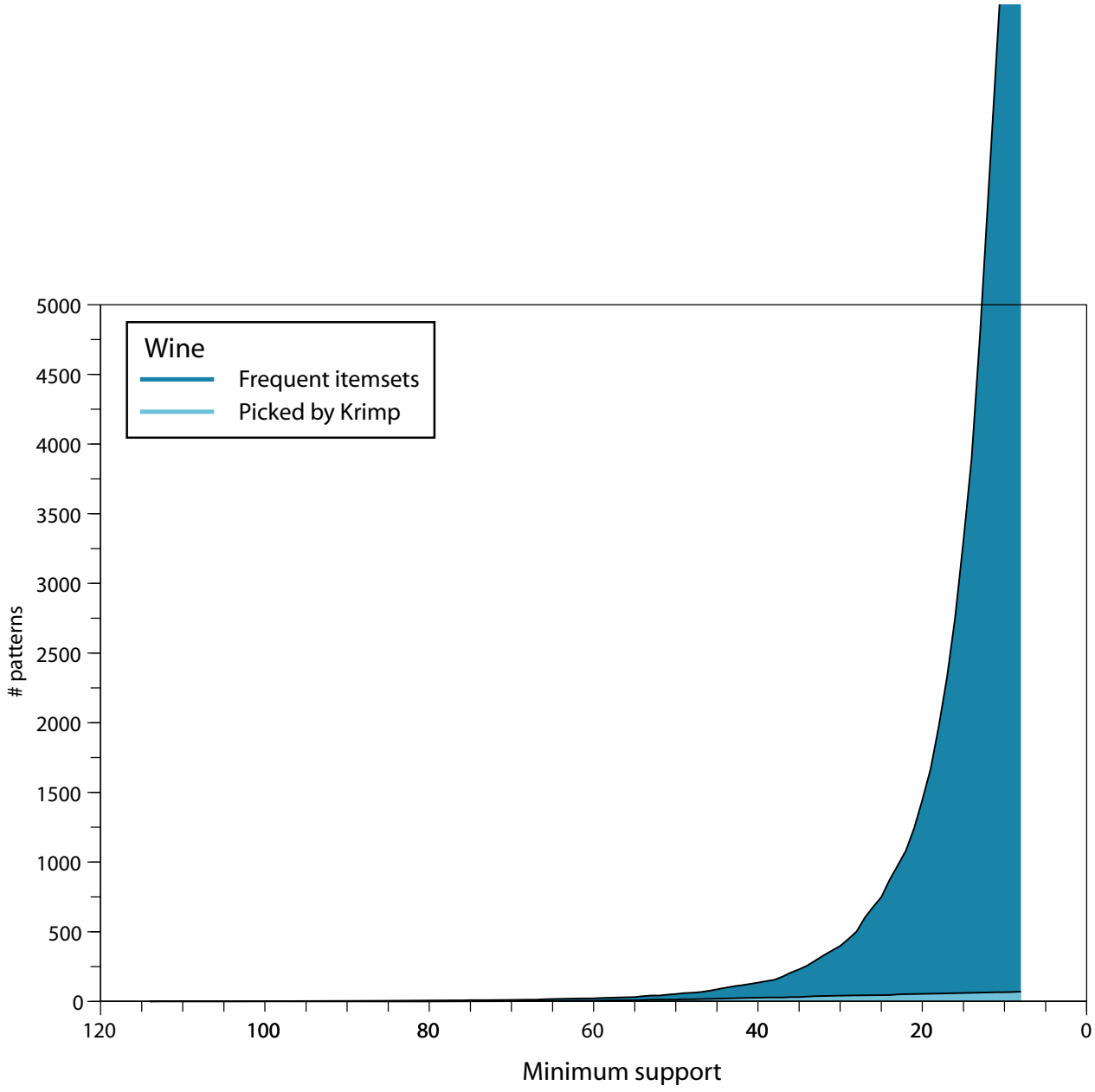
Trouble in Paradise

- If min-sup is high, only a few, well-known, patterns will be found
- If min-sup is low, the number of patterns discovered explodes

All is a lot to ask for!

From data explosion to pattern explosion

A Small Example



How Comes?

There are many possible reasons for this phenomenon, e.g.,

- transactions can fit many patterns
- if p_1 and p_2 are independent patterns that occur in 50% of the transactions, $p_1 \wedge p_2$ occurs in 25%
- if $p_3 = p_1 \wedge p_2$ and p_3 is frequent, so are p_1 and p_2
  - and p_3 may have multiple such decompositions

Hence, the frequent patterns explosion is a real problem

- one that should be solved

And that is the topic of this course

Three Approaches

There are three main approaches to tame the pattern explosion, viz.,

- Constraints
  - if we allow you to formulate more precise what patterns are interesting to you, you'll find fewer
- Condensed Representations
  - Select a "basis" of the set of frequent patterns,
  - i.e., the set of all frequent patterns including their support can be reconstructed from this basis only
- Pattern Set Mining
  - select a small subset of patterns which is collectively interesting.

Constraints

While the decision predicate $q(db, \phi)$ could be anything, it is usually restricted to something like support

- not a very precise way to delineate the patterns you are interested in

So, what if we allow you to specify extra requirements to filter out what patterns are interesting to you?

- note that it isn't really to specify patterns that will surprise you – by definition you do not know them

Computing many patterns can be costly

- hence the interest in constraints that can be pushed into the discovery process
- i.e., only interesting patterns will be discovered

Condensed Representations

The pattern explosion means that the set of all frequent patterns $\mathcal{F}$ is redundant

- it has many patterns that tell you basically the same thing

What if we find a subset $\mathcal{B} \subseteq \mathcal{F}$ such that

- you can reconstruct $\mathcal{F}$ from $\mathcal{B}$ including all supports

This is known as a condensed representation

- the first example are *closed* item sets

An item set I is closed iff

- $\forall A \notin I : supp_{db}(I \cup \{A\}) < supp_{db}(I)$

Exercise: the set of all closed frequent item sets is a condensed representation of the set of all frequent item sets.

Pattern Set Mining

If many patterns describe more or less the same set of objects

- you cannot decide whether or not a pattern is interesting by looking at that pattern only

You have to look at all patterns you want to select

- interestingness is a relative property

For example, by selecting sets of patterns that collectively describe the database well.

The Course

The Organization

Papers

For some topics there are excellent textbooks

- for many others there are none

There are many reasons for this

- it takes a while before new developments can be condensed into a book
- publishers may not see a huge market
- researchers are more interested in new research rather than writing a textbook

In all these cases there is only one thing you can do

- read papers on the topic

And that is what we will do

Reading and Discussing Papers

For every class you are supposed to read 1 - 3 papers

- for the schedule, see
- http://www.cs.uu.nl/docs/vakken/mpsm/

The "Library Access" extension for Chrome will allow you to download the papers we discuss from your home (or any other place you happen to be).

- check the library site and/or the Chrome Store for more information on this extension.

Papers may use math and/or CS terms you are not familiar with

- Google is your friend

No, I'm not sponsored by Google,

- in fact, I don't trust them very much

Reading and **Discussing** Papers

The class meeting is about

your comments on and questions about those papers

I don't give lectures on the papers

- we discuss them together

The less I talk, the better the class meeting

- Socratic dialogues
- as made famous by Plato
  - you all have read Plato I presume?

If you have no comments or questions

- class is over very quickly

You are really in the driving seat

Why Like This?

This way of doing things makes it a tough(er) course. Still I think it is a good way to do it. Some reasons in random order:

- discussing papers is the best way to understand them
- reading scientific papers is something you'll do throughout your career, whatever you become; so why not now?
- I have read the papers we'll discuss and many more already; by telling you what is in those papers I learn nothing new
  - I want to learn from your observations, musings, and thoughts
- the authors know best, so learn from them
  - a paper based MOOC, much older than its internet version.
  - flipping the class room the old fashioned way

If you persevere,

- you will learn a lot

How to Read a Paper

There are many ways to approach this, a way I use often is:

- What do the authors want to achieve (introduction)
  - and how is that related to what you already know (related work)
- Did they achieve this goal (conclusions)
- How do they back up these claims (experiments, theorems)
  - and are these results convincing
- only then start reading the technical details

In our discussions we will follow this approach as well.

Preparing the Meeting

Your questions and comments will come to you while you are reading the paper

- excellent, write them down in a file straight away
- annotated by
  - intro, related work, conclusions, backing up, technical details
  - and the page in the pdf file

For each meeting we'll have a separate channel on the Teams site

- and I'll put a document there that you can edit
- one for each paper we'll read and discuss
- When you have your questions and comments ready
- add them to this document at the appropriate place, i.e.,
  - keep the suggested order
  - cluster with related questions and remarks

Note, it is anonymous!

Discussing

Given that discussions are far better held in person that on-line, the default for the course is on campus. If you are ill, you can follow the discussion on-line, but note that the focus will be on the lecture hall.
We will go through all questions and remarks,

- if you want to say something: do so
  - if two people attempt to speak at the same time, we will resolve it on the spot
- if you read a remark/question by another student on which you want to comment
  - please do so in the file already. Just mark it clearly as such.
  - Don't worry, we will discuss your contribution

Because of your privacy and to ensure that the discussion is as much open and free as possible

- the sessions will *not* be recorded.

I have Another Question?

You may have a question that does not relate directly to the papers discussed in a session

- e.g., about the exam

Simply add these questions at the end of that sessions document

- I will then also discuss and answer those questions

Why not via email?

- often other people will have similar questions
- the answer is often interesting for more people
- your mail may very well get lost in my torrent of unread email.

The Course

The Exam

Is Already Online

See:

http://www.cs.uu.nl/docs/vakken/mpsm/

for all details of the exam.

- You can already start preparing your submission today!

How Do We Test?

To pass the course you'll have to submit an essay

- an essay that proves that you have mastered pattern set mining
- you understand the problems and solutions proposed, you can reflect and choose
- see the next slide

Do I test whether or not you read the assigned papers?

- in principle: no!
  - if no one has a point to discuss, we are done very quickly
  - if you don't want to learn, I'm not going to waste my time on you
- if you want to learn, you'll read the papers
  - you are old enough not to need anyone holding your hand

Essay

Your essay should contain two parts:

1. a brief survey of the content of the course: what is the problem and what are the main directions of solutions we have seen (5 - 6 pages)
   - if you fail this part, you cannot pass.
2. show that you really understand what is going on, by going beyond the material we covered in class (4 - 6 pages), e.g.,
   - survey on one of the major techniques going beyond the papers we already discussed
   - a critical reflection on one of the main approaches: can it work at all? is it the best or worst possible approach?
   - a critical reflection on the whole endeavour: is pattern set mining doomed from the start?

The first part decides whether or not you pass, the second how well you do

- an empty or incoherent or nonsensical second part will make you obviously still fail.

Next Time

The next session will be on Wednesday, May 4.

We'll discuss three papers at that meeting

- Mining Association Rules between Sets of Items in Large Databases, by Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Sigmod 1993.
- Fast Algorithms for Mining Association Rules, by Rakesh Agrawal and Ramakrishnan Srikant, VLDB 94
- Discovery of Frequent Episodes in Event Sequences, by Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo, KDD96 and Data Mining and Knowledge Discovery 97

Check

- http://www.cs.uu.nl/docs/vakken/mpsm/

regularly!