Algorithmic Gems in the Data Miner’s Cave

Paolo Boldi

Dipartimento di Informatica
Università degli Studi di Milano
via Comelico 39/41, 20135 Milano, Italy

Abstract. When I was younger and spent most of my time playing in the field of (more) theoretical computer science, I used to think of data mining as an uninteresting kind of game: I thought that area was a wild jungle of ad hoc techniques with no flesh to seek my teeth into. The truth is, I immediately become kind-of skeptical when I see a lot of money flying around: my communist nature pops out and I start seeing flaws everywhere.

I was an idealist, back then, which is good. But in that specific case, I was simply wrong. You may say that I am trying to convince myself just because my soul has been sold already (and they didn’t even give me the thirty pieces of silver they promised, btw). Nonetheless, I will try to offer you evidences that there are some gems, out there in the data miner’s cave, that you yourself may appreciate.

Who knows? Maybe you will decide to sell your soul to the devil too, after all.

1 Welcome to the Dungeon

Data mining is the activity of drawing out patterns and trends from data; this evocative expression started being used in the 1990s, but the idea itself is much older and does not necessarily involve computers. As suggested by many, one early example of successful data mining is related to the 1854 outbreak of cholera in London. At that time it was widely (and wrongly) believed that cholera was a “miasmal disease” that was transmitted by some sort of lethal vapor; the actual cause of the disease, a bacterium usually found in poisoned waters, would have been discovered later by Filippo Pacini and Robert Koch. John Snow was a private physician working in London who was deeply convinced that the killing agent entered the body via ingestion, due to contaminated food or water. In late August 1854, when the outbreak started in Soho, one of the poorest neighborhoods of the city, Snow began his investigation to obtain evidences of what was the real cause behind the disease.

Through an accurate and deep investigation that put together ideas from different disciplines, and by means of an extensive analysis of the factual data

1 Filippo Pacini in fact published his results right in 1854, but his discoveries were largely ignored until thirty years later, when Robert Koch independently published his works on the Vibrio cholerae (now officially called Vibrio cholerae Pacini 1854).
he collected, he was able to find the source of the epidemic in a specific infected water pump, located in Broad Street. His reasoning was so convincing that he was able to persuade the local authorities to shut down the pump, hence causing the outbreak to end and saving thousands of lives. John Snow is now remembered as a pioneering epidemiologist, but we should also take his investigation as an early example of data mining. (I strongly suggest those of you who like this story to read the wonderful *The ghost map* by Steven Johnson [1], that albeit being an essay is as entertaining as a fiction novel).

In the 160 years that have passed since John Snow’s intuitions, data mining have come to infect almost every area of our lives. From retail sales to marketing, from genetics to medical and biomedical applications, from insurance companies to search engines, there is virtually no space left in our world that is not heavily scrutinized by data miners who extract patterns, customs, anomalies, forecast future trends, behaviours, and predict the success or failures of a new business or project.

While the activity of data miners is certainly lucrative, it is at the same time made more and more difficult by an emerging matter of size. If big data are all around, it is precisely here were data are bigger, more noisy, and less clearly structured. The data miner’s cave overflows, and not even the most apparently trivial of all data-mining actions can be taken lightheartedly.

If this fact is somehow hindering miners’ activity, it makes the same activity more interesting for those people (like me) who are less fascinated in the actual mining and more captivated by the tools and methods the miners use to have their work done. This is what I call data-mining algorithmics, which may take different names depending on the kind of data that are concretely being processed (web algorithmics, social-network algorithmics, etc.). In many cases, the algorithmic problems I am referring to are not specific to data mining; they may be entirely new or they may even have been considered before, in other areas and for other applications. Data mining is just a reason that makes those methods more appealing, interesting or urgent.

In this brief paper, I want to provide some examples of this kind of techniques: the overall aim is to convince a skeptical hardcore theoretician of algorithms that data mining can be a fruitful area, and that it can be a fertile playground to find stimuli and ideas. I will provide virtually no details, but rather try to give a general idea of the kind of taste these techniques have. The readers who are already convinced may hopefully find my selection still interesting, although I will follow no special criterion other than personal taste, experience and involvement. So welcome to the miner’s dungeon, and let’s get started.

## 2 Please, Crawl as you Enter the Cave

One of the first activities a data miner must unavoidably face is harvesting, that is, collecting the dataset(s) on which the mining activity will take place. The real implications and difficulties of this phase depend strongly on the data that are being considered and on the specific situation at hand. A classical example