Job Recommender Systems: challenges and pitfalls

Presenters: Guillaume Bied, Solal Nathan

22nd of February, 2023

Summary

Summary

Introduction

- Context:
  - Unemployment in France, partly due to imperfect information (Belot et al., 2019)
  - Recommender systems (RS) help users in find their way in large databases (*e.g.* all posted job ads), leveraging past user behavior

- The VADORE Project:
  - Partnership between researchers (economics, CS) and the French Public Employment Service (PES)
  - Goal: build recommender systems using the PES's data, and assess their impact on the labor market

- Goals of the talk:
  - Provide an overview of challenges in job recommendations and of the proposed RS
  - Discuss the (many) pitfalls in job recommendation and perspectives for further work

The VADORE Team

- CREST-ENSAE: Economics
  B. Crépon, C. Gaillac (now at Oxford University), E. Perennes (now at *Pôle emploi*)
- LISN (Université Paris Saclay), INRIA TAU (AO) Team: Computer Science
  P. Caillou, M. Sebag, G. Bied, S. Nathan
- French PES (*Pôle emploi*):
  C. Nouveau (DSEE), H. Caillol (DSEE); P. Beurnier (DSEE), C. Vessereau (DSEE); Y. De Coster (DSEE); S. Robidou (DSI/DER); R. Mir (DSI/DER); S. Semmar (DSAE)

Summary

Context: the matching problem

We have two sets $A = \{a_i\}_{i \in [[0,N]]}$ and $B = \{b_i\}_{i \in [[0,N]]}$ which have preferences on each other (rankings). We want to match them in order to maximise individual preferences and such as it is stable (no swap would yield a higher sum of preference).

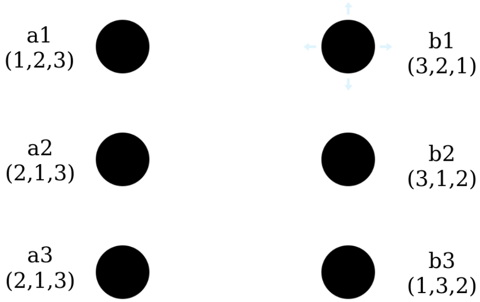


Figure: A matching game of size 3

Context: the matching problem

- Stable Marriage Problem (SMP);
- The hospital-resident assignment problem (HR);
- The student-allocation problem (SA);
- The stable roommates problem (SR);

Naive and historic solutions

- first-come, first-served;
- Gale-Shapley (1962) to solve SMP;
- Roth & Peranson (1999) to solve HR/ HR with couples/ SA / SR;

These algorithms (or variant) are still used (or where until recently): The Match (USA), APB, Parcoursup.

But not everywhere: dating sites (Tinder, etc) does not work like that.

```
algorithm stable_matching is
    Initialize m ∈ M and w ∈ W to free
    while ∃ free man m who has a woman w to propose to do
        w := first woman on m's list to whom m has not yet proposed
        if ∃ some pair (m', w) then
            if w prefers m to m' then
                m' becomes free
                (m, w) become engaged
            end if
        else
            (m, w) become engaged
        end if
    repeat
```

Shortcomings

Gale-Shapley:

- both sets have to be of the same size
- asymmetry (will favor men or women creating the best/worst possible match out of all the matching satisfying stability)
- must rank every item in the other set
- strictly heterosexual relationship[1]

Roth & Peranson:

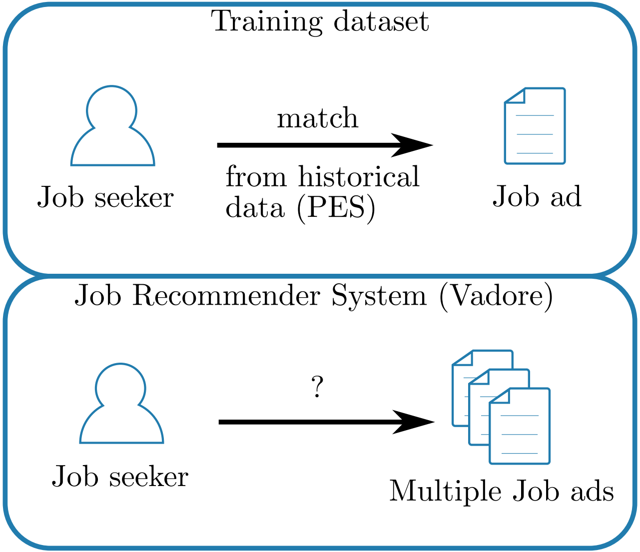
- do not need to have preferences on everything
- has to match everything at once
- participants must have a match at the end
- does not take into account past hires, similarity between users
- no serendipity

In both cases, people **have to** follow the match (assignment), we only want to make a recommendation.

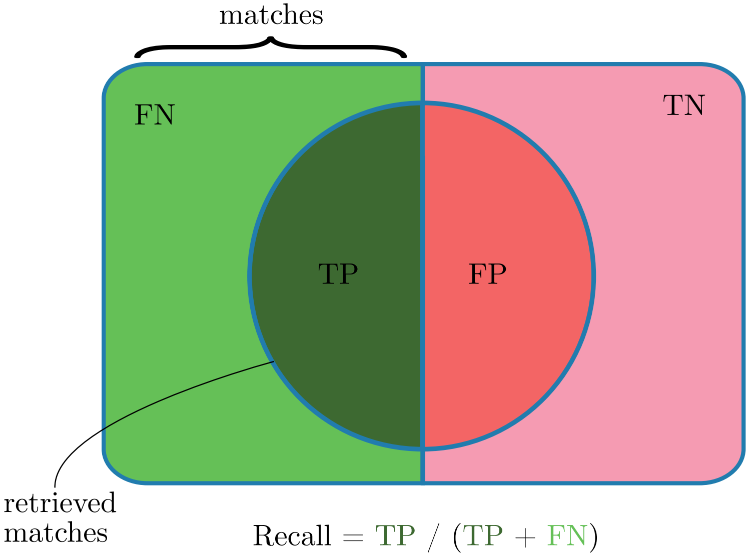
[1] with bisexual relationship/polygamous it becomes NP hard. But it is not relevant to matching jobs.

Summary

Job Recommender Systems



Metric: the recall



$$\text{Recall} = \text{TP} \;/\; (\text{TP} + \text{FN})$$

Collaborative filtering

- Flavor of most classic recommender systems (*e.g.* Amazon, Netflix): "show you items like those you have previously seen", "show you items that people like you have liked"
- For the sake of illustration: memory-based, item-based collaborative filtering
- Interactions between n users and p items summarized in a collaborative filtering matrix $\mathcal{M}$.
  - $\mathcal{M}_{ij} = 1$ if user i clicked on job ad j, 0 otherwise.
  - Denote $\mathcal{M}_{.,j}$ column j of $\mathcal{M}$.
- Recommend to i items **similar to those seen previously**, ranking them by score:

$$\hat{\mathcal{M}}_{ij} = \sum_{k=1}^{p} \mathrm{sim}(\mathcal{M}_{.,j}, \mathcal{M}_{.,k})\mathcal{M}_{ik}$$

  For instance, cosine similarity defined for $u, v \in \mathbb{R}^n$ by

$$\mathrm{sim}(u, v) = \frac{\langle u, v \rangle}{||u||_2||v||_2}$$

- Many other approaches:
  - Recommend what users like you have seen
  - Factorize $\mathcal{M}$ by SVD

Hybrid recommender systems

- **Key issue: cold start.** This setup fails when no previous interactions exist for users / items
  - Here: $\mathcal{M}_{ij}$ is based on hires. Cold start is the norm, not the exception.
- Solution: leverage contextual data on users and items (*e.g.* text, job types, skills)
- Hybrid recommender systems use both contextual data and interactions, for instance:
  - Use features of job ad x_i and job ad y_j to predict if a hire or click took place in the past (standard binary classification):

$$\hat{\mathcal{M}}_{ij} = f_\theta(x_i, y_j)$$

    - Think of f_θ as a neural network with a logistic loss
  - Sort job ads by decreasing $\hat{\mathcal{M}}_{ij}$

Selected design issues

- Incorporating domain knowledge in the architecture of f_θ
  - In VADORE, parts of the architecture account for geography, skills & occupations, all other information
- Downsampling: can't train on all i, j pairs (billions)
  - Negative sampling ...but can we do better than uniformly at random?
- Scalability:
  - When issuing recommendations, $f_\theta(x_i, .)$ needs to be evaluated for all job ads (tens of thousands) for a given i
  - Solution: at least for initial pair selection, "two-tower" structure:
    - Separate embeddings $f_1(x_i)$ for job seekers and $f_2(y_j)$ for job ads
    - $\hat{\mathcal{M}}_{ij} := \langle f_1(x_i), f_2(y_j) \rangle$
    - Maximum inner product search strategies (with approximations) scale to billions of items
  - In VADORE: two-tower structure to select 1,000 top job ads for i; rerank them with a feedforward neural net in a second stage
- Leveraging different types of interactions: hires, applications, clicks ...

Overview of the VADORE architecture

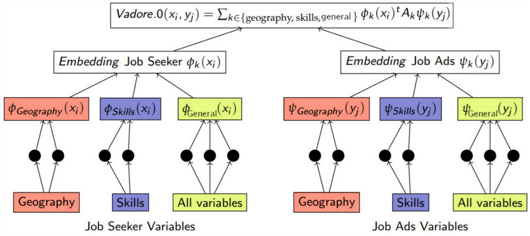


Figure: Architecture of VADORE.0

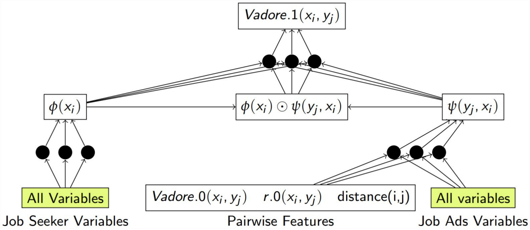
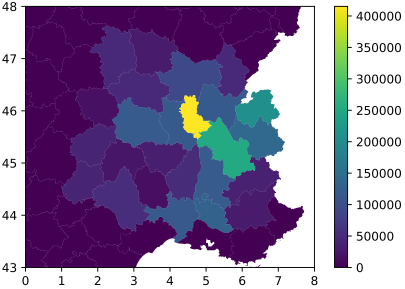


Figure: Architecture of VADORE.1

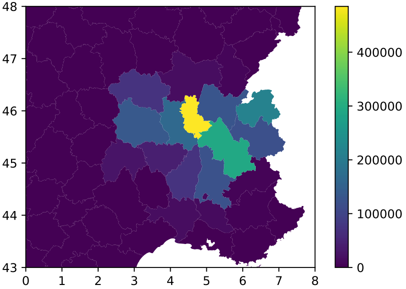
Experimental settings

- Auvergne-Rhône-Alpes region; 2019-mid 2022
- Training is done on hires
- 1.2M job seekers, 2.2M job ads, 285k hires
- Job seeker characteristics $x_i \in \mathbb{R}^{483}$ (after pre-processing):
  - Job search criteria (occupation, wage, location, commuting)
  - Labor market profile (experience, skills, education, text)
  - Administrative data / link with the PES
- Job ads $y_j \in \mathbb{R}^{469}$ (after pre-processing):
  - Wage, location, education / skills requirements, contract, working hours, text
- 85% / 15% train test split on a weekly basis

Experimental settings: geography

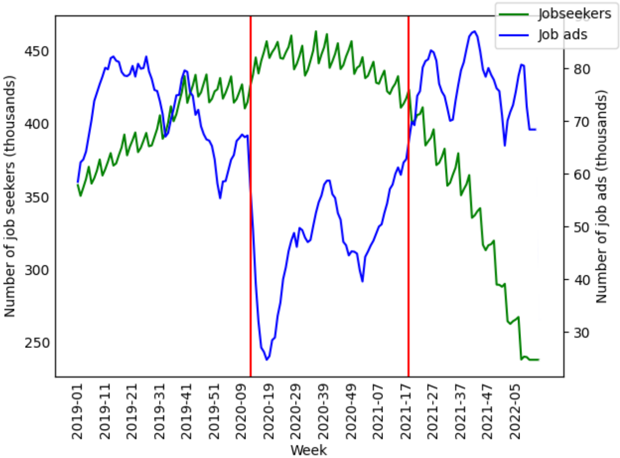


(a) Choropleth map of job ads



(b) Choropleth map of job seekers

Experimental settings: non-stationarity, especially during the COVID pandemic



Number of job seekers and job ads
Vertical red lines: 1st and 3rd French lockdowns

Performance and computational cost

Recall@k	XGBoost	Vadore
10	**26.83**	25.96
20	35.59	**35.65**
50	48.75	**49.07**
100	**58.88**	58.67
Training time (hours)	10	7.6
Recom. time per person (seconds)	7	0.005

- For 26% of hires, the job seeker's future job was among Vadore's top-10 recommendations
- Baseline: ensemble of boosted trees based on Recsys 2017 winners (Volkov et al., 2017)
  - Accuracy: Vadore on par with XGBoost
  - Recommendation time: Vadore faster by three orders of magnitude

Summary

Public benchmarks and privacy

- Public benchmarks (*e.g.* ImageNet) have been crucial to the development of machine learning by enabling the fair comparison of proposed techniques, favoring cumulative research
- No such datasets are available for job recommendation
- Linked to an understandable concern for privacy (and private interests)
- Perspectives: defining public benchmarks and enabling sharing of data / code / results between researchers while respecting privacy
  - Synthetic data generation (*e.g.* GANs) while ensuring differential privacy?

Congestion

- Job ads are **rival goods**: only one person may be hired on a job
  - Thus, recommending the same job to everyone would be sub-optimal
- Congestion exists in real life:
  - Rhone-Alpes, 2019-2022, 2,2M offers, 1.2M job seekers and 1.75M applications in PES data
  - Only 25% of the job ads receive one (or more) applications
- Congestion worsens with recommendations:
  - On a subset of our data (110k job seekers, 14k job ads), only 6% of the job ads are in the Top-1 of some job seeker; only 17% are in the Top-10
- Perspectives: address congestion as a post-processing step (constrained optimization) or during learning

Fit to job seeker preferences / value alignment

VADORE *is learned on past hires. Do recommendations fit with job seeker's search criteria? Are some recommendations irrelevant / inappropriate?*

- **SDR**: a weighted sum of sub scores, measuring fit between js' and recruiters' criteria

Profile (CV)	**Search criteria**
Skills	Occupation
Diploma	Working hours
Languages	Reservation wage
Driving license	Geographic mobility
Years of experience	Duration and type of contract

SDR Rank of VADORE Top-1

Rk SDR	≥101	51-100	26-50	11-25	1-10
Nb Reco.	64%	8%	6%	7%	15%

VADORE **Rank of SDR Top-1**

Rk VADORE	≥101	51-100	26-50	11-25	1-10
Nb Reco	54%	10%	9%	10%	18%

- Perspectives:
  - Understanding what job seekers expect from recommendations
  - Are hires the right labels to train on?
  - Mixtures between VADORE and SDR work surprisingly well, in recall and in practice (A/B tests)

Fairness

- Algorithms trained on real-world data may learn job seekers' and recruiters' biases (*e.g.* w.r.t. gender)
- Many possible questions:
  - Are recommendations equally relevant for different genders?
  - Are different job ads recommended to men and women *because of their gender*?
    - Simply removing gender from x_i is not enough to prevent this
    - But women and men may value different aspects in a job, *e.g.* different wage-commute trade-offs (Le Barbanchon et al., 2021)
  - If job seekers followed recommendations, what would be the impact in terms of labor market inequalities (gender wage gap, occupation segregation, etc.)? Especially w.r.t. the current situation?
- Perspectives (ongoing, with M. Hoffmann):
  - Develop audit methodologies and appropriate metrics
  - Develop algorithmic tools to enforce the respect of fairness metrics, and assess the trade-offs between recommendation performance and fairness

Summary

Conclusion and perspectives

- Job recommendation is a key domain of AI for Good
- Compared to usual RS settings (*e.g.* Amazon, Netflix), it requires careful handling of cold start, and benefits from unusually rich contextual data
- We presented an overview of VADORE, a two-tiered hybrid recommender system built in cooperation with the French PES
- Job recommendation comes with pitfalls subject to ongoing research, including:
  - The design of public benchmarks while respecting privacy
  - Congestion
  - Value alignment
  - Fairness considerations
- Beyond those challenges, next steps for the VADORE project include:
  - Evaluation of the designed recommender system in real life (A/B tests)
  - Recommendation of job seekers to recruiters