LSH One

Yassine Abou Hadid, Clément Préaut, Benjamin Sykes

LSH One
Locality Sensitive Hashing

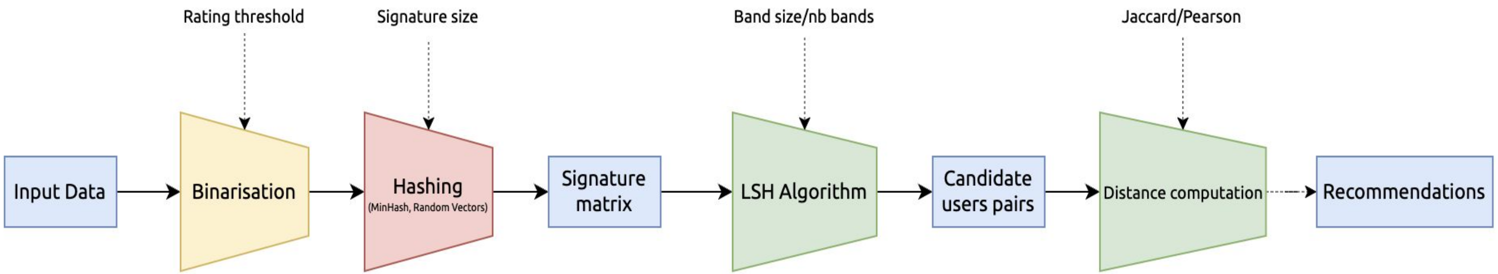
Dauphine | PSL
UNIVERSITÉ PARIS

Recap on LSH : LSH workflow for recommendation

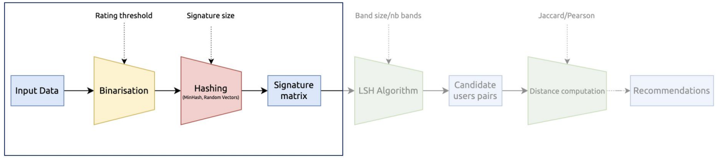
$$R_{user=i,movie=j} = \sum_{k=1}^{N} \mathrm{UserSim}(i,k).R_{user=k,movie=j}$$

<u>Goal</u>: only relevant similarities should be taken into consideration

$$R_{user=i,movie=j} = \sum_{k/\exists h_\alpha, h_\alpha(i)=h_\alpha(k)} \mathrm{UserSim}(i,k).R_{user=k,movie=j}$$



Signature Matrix Generation



Binarization

(threshold=3)

	movie 1	movie 2	movie 3
user 1	**3**	**2**	**4**
user 2	**1**	**2**	**5**
user 3	**4**	**3**	**2**

	movie 1	movie 2	movie 3
user 1	**1**	**0**	**1**
user 2	**0**	**0**	**1**
user 3	**1**	**1**	**0**

V1	-3	4	2
V2	0	-1	2
V3	-1	2	-2

Random Hyperplans

(3 vectors)

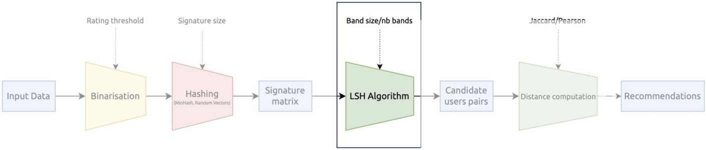
1	**0**	**1**
0	**0**	**1**
1	**1**	**0**

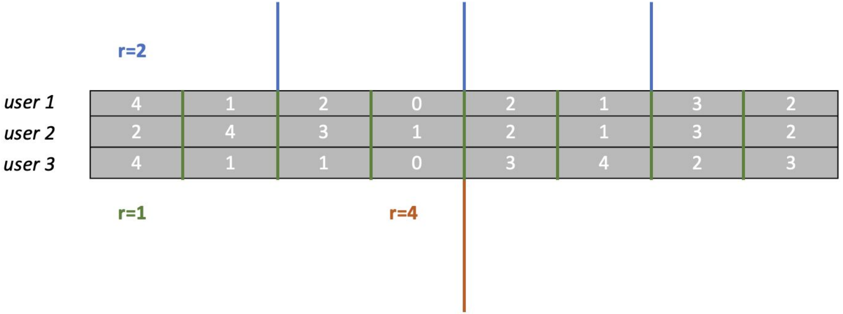
$$H_{V_i}(C_j) = \begin{cases} 1 & if\ V_i . C_j^T > 0 \\ 0 & otherwise \end{cases}$$

0	**1**	**0**
1	**1**	**0**
1	**0**	**1**

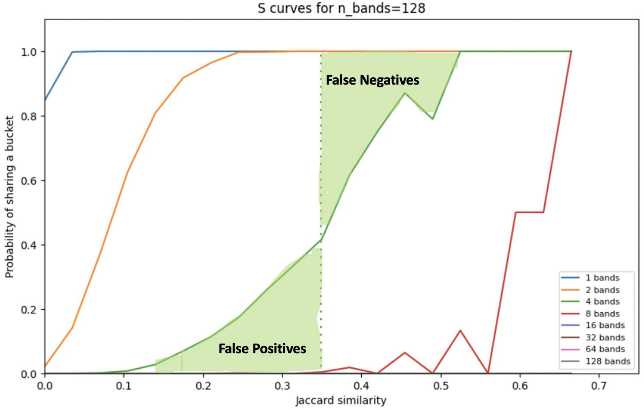
Signature Matrix

Approximating the s-curve

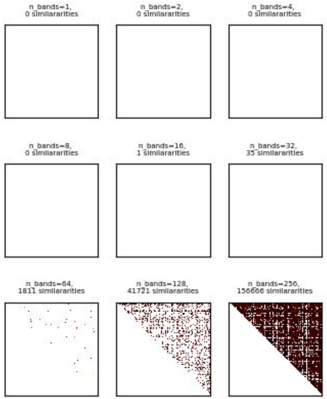




user 1	4	1	2	0	2	1	3	2
user 2	2	4	3	1	2	1	3	2
user 3	4	1	1	0	3	4	2	3



Evolution of computed similarities on full matrix

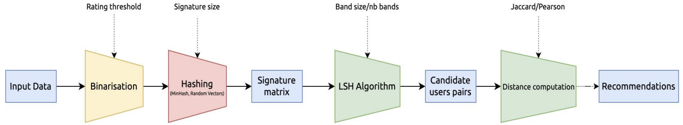


Increasing threshold t →

$$\textit{threshold} = (1/b)^{1/r}$$

Trade-off **speed/false negatives** vs **false positives**

Optimisation → scaling



LSH is made for large data

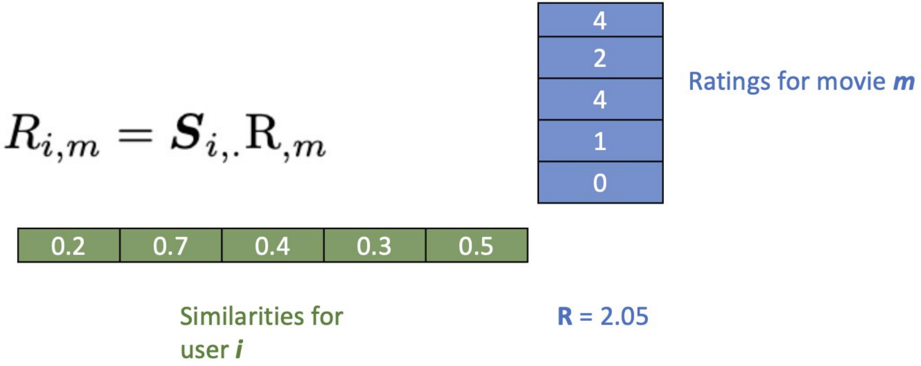
$$\mathcal{O}(n_{\text{users}}.n_{\text{movies}}.8bits) \xrightarrow{\text{hash}} \mathcal{O}(n_{\text{users}}.s_{\text{signature}}.8bits)$$

Precomputation of the signature matrix

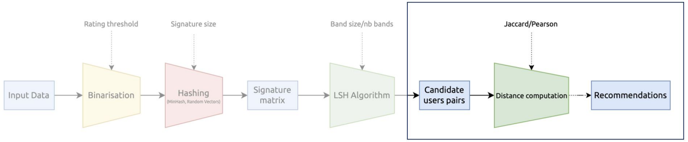
*All the computations for **similarity evaluation** can be then made **offline***

Computation of the sparse similarity matrix (numpy)

$$R_{i,m} = \boldsymbol{S}_{i,.}\mathrm{R}_{,m}$$



Curve interpretation


Rating threshold

Signature size

Band size/nb bands

Jaccard/Pearson

Input Data

Binarisation

Hashing
(MinHash, Random Vectors)

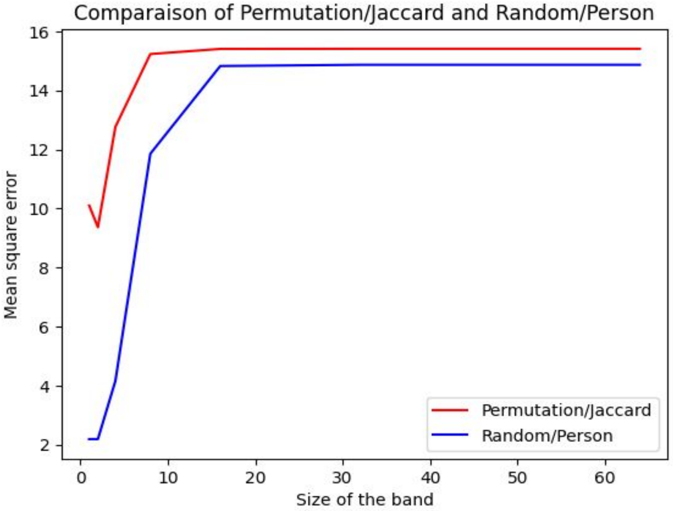
Signature matrix

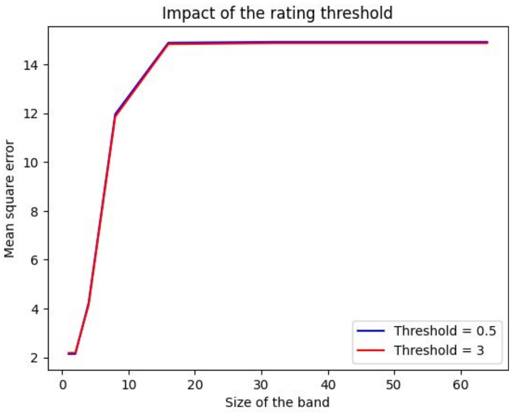
LSH Algorithm

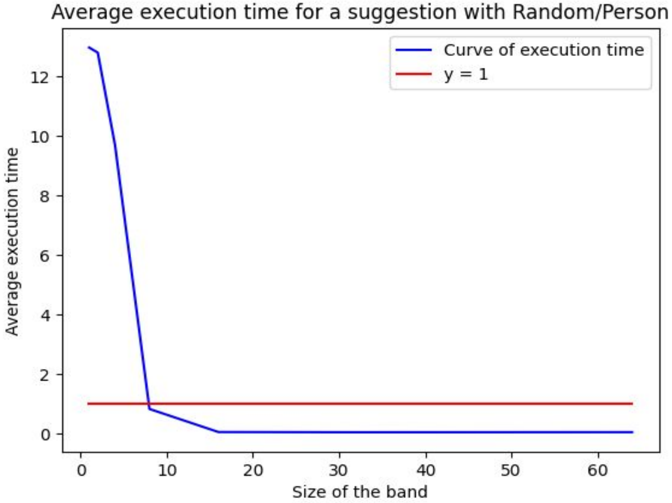
Candidate users pairs

Distance computation

Recommendations


Comparaison of Permutation/Jaccard and Random/Person


Impact of the rating threshold


Average execution time for a suggestion with Random/Person

Thank you for your attention!

LSH One

Locality Sensitive Hashing