Mangaki.fr, système de recommandation de mangas et d'anime

Jill-Jênn Vie

11 mai 2016

Timeline

Novembre 2011 : Eigaki, projet de master

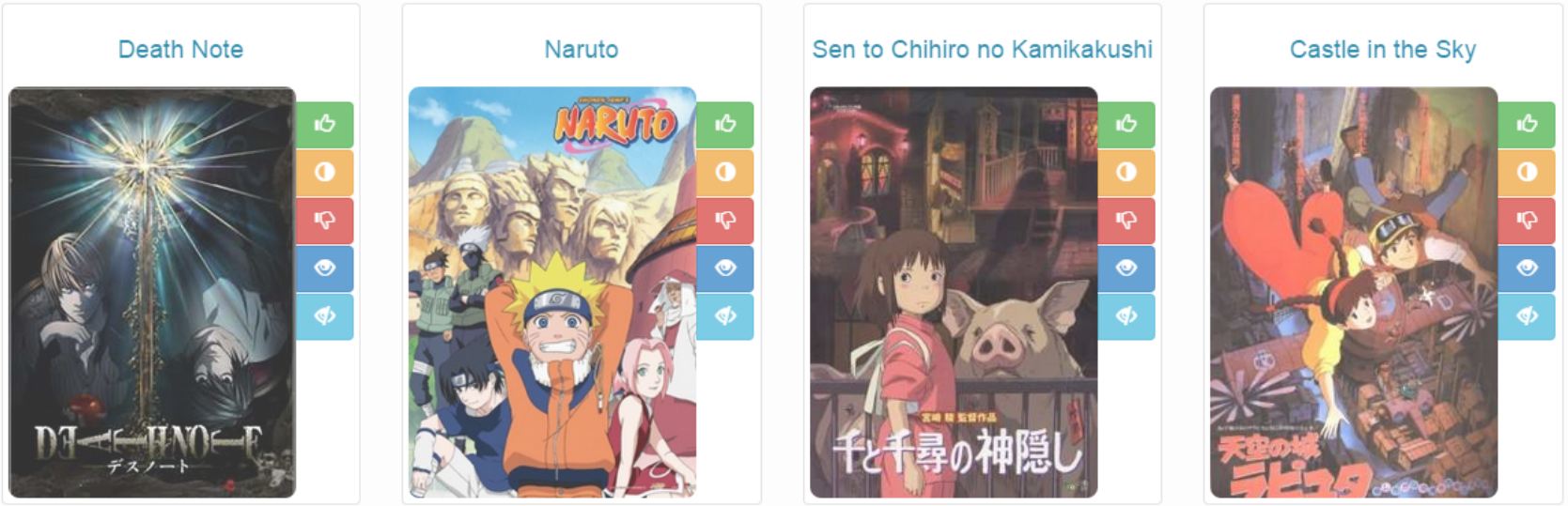
```
$this->db->exec("INSERT INTO eigaki_sim (SELECT $idUser, distances.user_id AS sim_id,
dist/(sqrt(my.norm)*sqrt(users.norm)) AS score FROM (SELECT eigaki_ratings2.user_id,
sum((eigaki_myratings2.rating)*(eigaki_ratings2.rating)) AS dist, count(id) AS nbcommon
FROM eigaki_ratings2, eigaki_myratings2 WHERE eigaki_myratings2.user_id = $idUser AND
eigaki_myratings2.film_id = eigaki_ratings2.film_id GROUP BY user_id HAVING nbcommon > 5)
AS distances, (SELECT eigaki_ratings2.user_id, sum((rating)*(rating)) AS norm FROM
eigaki_ratings2 GROUP BY user_id) AS users, (SELECT SUM((rating)*(rating)) AS norm FROM
eigaki_myratings2 WHERE eigaki_myratings2.user_id = $idUser) AS my WHERE users.user_id =
distances.user_id ORDER BY score DESC LIMIT 30)"); // ZOMG
```

Timeline

- Novembre 2011 : Eigaki
- Février 2014 : thèse au LRI (Akinator pour l'éducation)
- Octobre 2014 : Mangaki
- Octobre 2015 : Student Demo Cup, prix de Microsoft
- Février 2016 : prix Maison de la culture du Japon à Paris

Aujourd'hui : 2k utilisateurs, 14k uvres, 281k ratings.

Un système de recommandation



Principe

- Un utilisateur s'inscrit et rentre ses préférences
- On lui recommande des films susceptibles de lui plaire

Objectifs

- Elles doivent être pertinentes (sinon l'utilisateur s'en va)
- Rapides à calculer (sinon l'utilisateur s'en va)

Filtrage collaboratif

Problème

- On dispose d'utilisateurs $u = 1, \ldots, n$ et d'items à noter $i = 1, \ldots, m$
- Chaque utilisateur u attribue une note à une partie des items (r_{ui} : note de l'utilisateur u sur l'item i)

$\Rightarrow$ Quels nouveaux items recommander à chaque utilisateur ?

Exemple (notes sur 5)

	Death Note	*L'Attaque des titans*	*Naruto*	*Bleach*
Sacha	*****	****	?	?
Ondine	**	?	*	?
Pierre	*	?	****	?

Objets : n vecteurs à m dimensions, éléments de $\{-1, 0, 1\}^m$

k plus proches voisins

Intuition

- On introduit un score de similarité entre utilisateurs
- On détermine les k utilisateurs les plus proches d'un utilisateur u
- On lui recommande ce qu'ils ont aimé qu'il n'a pas vu

Similarité

$\mathcal{R}_u$: le vecteur de notes $(r_{u1}, r_{u2}, \ldots, r_{um})$ $\quad u = 1, \ldots, n$
Le score de similarité entre 2 utilisateurs u et v est donné par :

$$score(u, v) = \mathcal{R}_u \cdot \mathcal{R}_v.$$

Intuition

Les points communs augmentent le score :

	Paprika	*Oldboy*	*Gattaca*	*12 Monkeys*
Alice	1	-1	0	0
Bob	1	1	-1	0
Charles	1	-1	1	-1

$$score(Alice, Bob) = 1 + (-1) = 0$$
$$score(Alice, Charles) = 1 + 1 = 2$$

Alice est plus proche de Charles que de Bob.

Estimation des notes inconnues

$N(u)$: les k plus proches voisins de u, $\quad u = 1, \ldots, n$
notés $\{v_1, \ldots, v_k\}$

$$\widehat{r_{ui}} = \frac{r_{v_1 i} + \ldots + r_{v_k i}}{k}$$

On calcule $\widehat{r_{ui}}$ pour chaque film i non noté $\Rightarrow$ les 10 meilleurs.

Version pondérée : les plus proches ont plus de poids

$$\widehat{r_{ui}} = \frac{\sum_{v \in N(u)} w_v \times r_{vi}}{\sum_{v \in N(u)} w_v} \quad \text{où } w_v = score(u, v)$$

Variantes

- Changer le nombre k de voisins ?
- Calculer une similarité non sur les utilisateurs sur mais sur les films.
- Pour une uvre donnée, considérer les plus proches voisins parmi ceux qui ont effectivement noté l'uvre.

Évaluation du modèle par validation croisée

À quel point j'ai bien recommandé ?

- Je suppose que je connais 80 % des utilisateurs (*train*)
- Je teste les recommandations sur les 20 % restants (*test*)

Pénalité : les moindres carrés

$$RMSE = \frac{1}{N}\sqrt{\sum_{u,i} (\widehat{r_{ui}} - r_{ui})^2}.$$

Une autre méthode : complétion de matrice

Supposons que la matrice M de notes soit de faible rang r :

$$M = \begin{pmatrix} \mathcal{R}_1 \\ \mathcal{R}_2 \\ \vdots \\ \mathcal{R}_n \end{pmatrix} = \boxed{} = \boxed{C} \boxed{P}$$

Chaque ligne $\mathcal{R}_u$ est une combinaison linéaire des lignes de P.

$$M : n \times m \qquad C : n \times r \qquad P : r \times m.$$

$$\mathcal{R}_1 = c_{11} P_1 + c_{12} P_2 + \ldots + c_{1r} P_r \qquad C_1 = (c_{11}, c_{12}, \ldots, c_{1r})$$

Exemple

Si P	P_1 : « aventure »	P_2 : « romance »	P_3 : « plot twist »
Si C_u	0,2	$-0{,}5$	0,6

ça veut dire :

j'aime un peu l'aventure, je n'aime pas la romance,
j'aime beaucoup les plot twists.

Top 30 du premier vecteur propre

Nausicaä of the Valley of the Wind
Princesse Mononoké
Le Château dans le ciel
Le Voyage de Chihiro
Toki wo Kakeru Shoujo
Tengen Toppa Gurren Lagann
Baccano !
Cowboy Bebop
Les Enfants Loups : Ame & Yuki
Mahou Shoujo Madoka Magica
Suzumiya Haruhi no Yuuutsu
Porco Rosso
Summer Wars
Neon Genesis Evangelion
Mon voisin Totoro
Ghost in the Shell
Kiki la petite sorcière
Suzumiya Haruhi no Shoushitsu
Le Château ambulant
Paprika
The Garden of Words
Barakamon
Steins ;Gate
5 centimètres par seconde
Grave of the Fireflies
The Tale of The Princess Kaguya
Akira
Mushishi
Bakemonogatari
Durarara !!

Bottom 30 du premier vecteur propre

Zero no Tsukaima
To LOVE-Ru
Soul Eater
D.Gray-man
Another
Bleach
Rosario to Vampire Capu2
Vampire Knight
High School DxD
Naruto
Black Butler
Dragon Ball GT
Guilty Crown
Akame ga Kill !
Naruto the Movie 2 : Legend of the Stone of Gelel

Mirai Nikki
Tokyo Ghoul
Rosario to Vampire
L'Attaque des Titans
IS : Infinite Stratos
Fairy Tail
Sword Art Online II
Ao no Exorcist
One Piece
Highschool of the Dead
Sword Art Online
Bleach
Naruto
Fairy Tail
Naruto : Shippuuden

Top 30 du deuxième vecteur propre

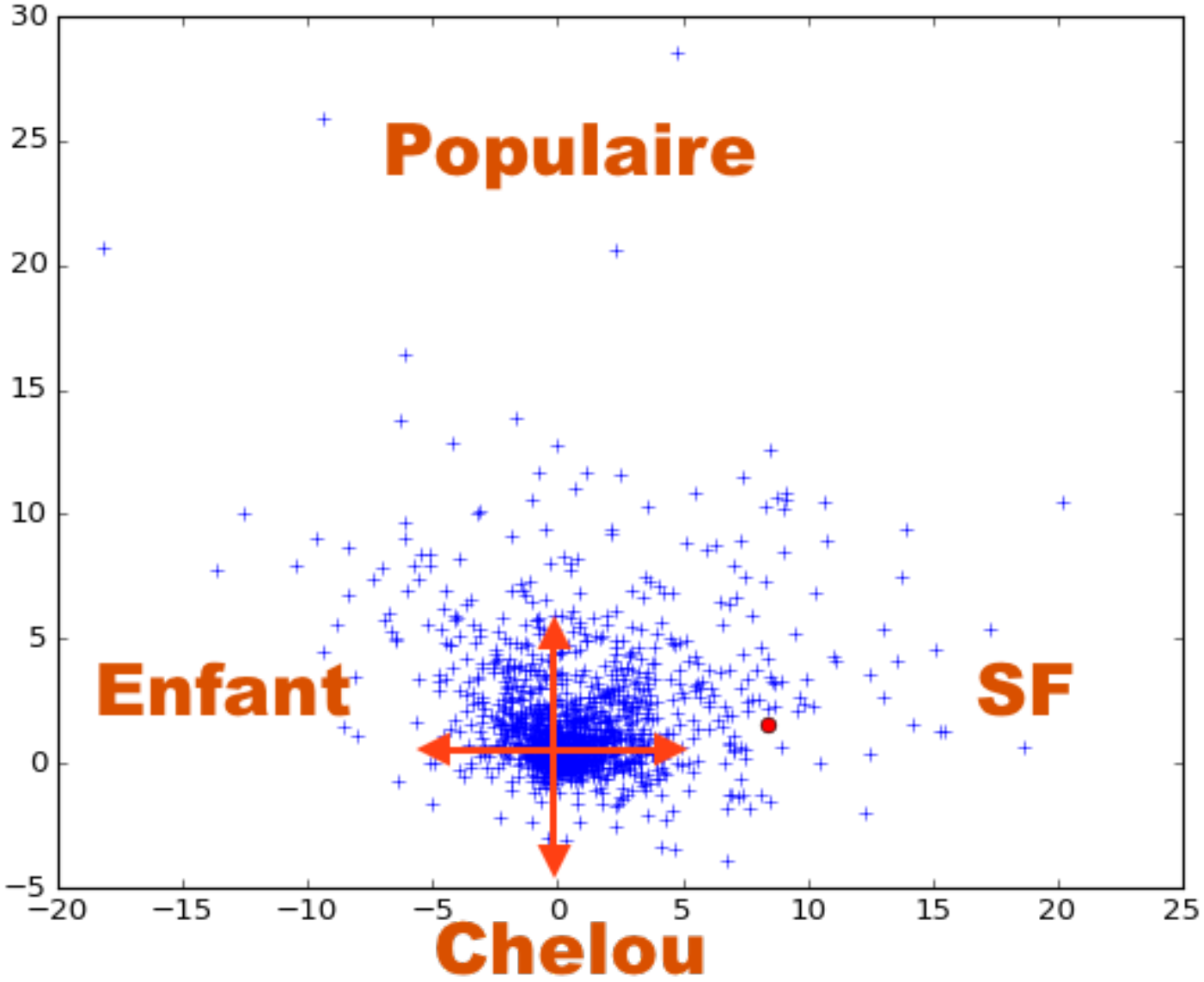
L'Attaque des Titans
Fullmetal Alchemist : Brotherhood
Death Note
Fullmetal Alchemist
Sword Art Online
Le Voyage de Chihiro
Princesse Mononoké
Ao no Exorcist
No Game No Life
Tokyo Ghoul
Mon voisin Totoro
FullMetal Alchemist
Psycho-Pass
Attaque Des Titans (l')
Code Geass : Hangyaku no Lelouch
Naruto
Fate/Zero
Les Enfants Loups : Ame & Yuki
Hunter x Hunter
Fullmetal Alchemist : Brotherhood OVA Collection
Mirai Nikki
Death note
Steins ;Gate
Soul Eater
One Piece
Le Château ambulant
Le Château dans le ciel
Bleach
Durarara !!
Tokyo ghoul

Bottom 30 du deuxième vecteur propre

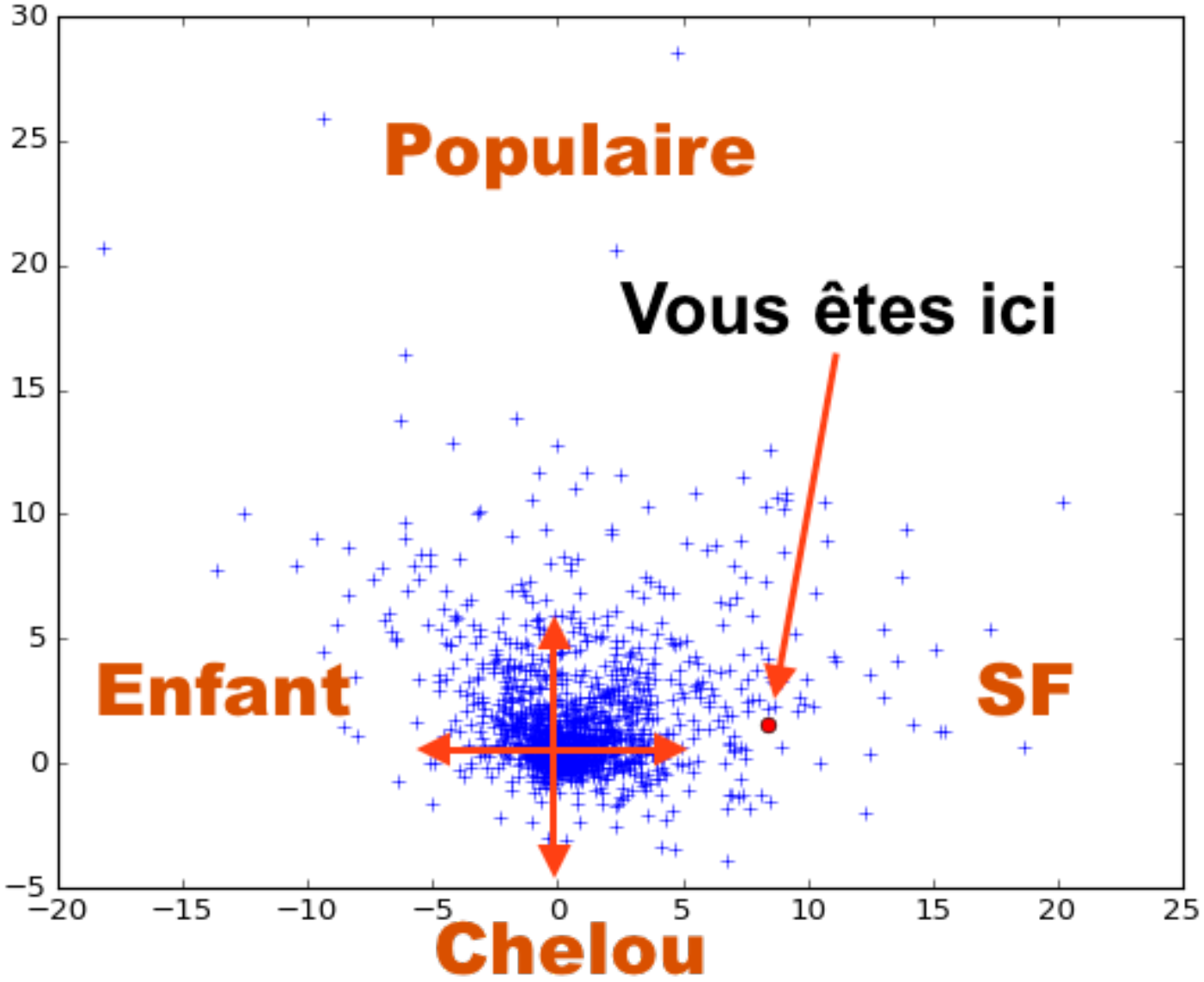
Infinite Stratos 2
IS : Infinite Stratos
Ikkitousen : Dragon Destiny
The Severing Crime Edge
IS : Infinite Stratos Encore - Koi ni Kogareru ...
A Bridge to the Starry Skies
Sailor Moon R
Ikki Tousen
Vividred Operation
School Days : Magical Heart
Kokoro-chan
Papa to Kiss in the Dark
D.C. Da Capo
Rail Wars !
Strawberry Panic
Freezing

To LOVE-Ru
School Days
Tokyo Mew Mew
Haruka Nogizaka's Secret
R-15
Wizard Barristers
Choujigen Game Neptune : The Animation
Yu-Gi-Oh ! GX
Dragon Ball GT
Captain Earth
Astarotte's Toy
Sakura Trick
Girls Bravo : First Season
Kiss x Sis
Dog Days

Map


Populaire
Enfant
SF
Chelou
30
25
20
15
10
5
0
−5
−20
−15
−10
−5
0
5
10
15
20
25

Map


Populaire
Vous êtes ici
Enfant
SF
Chelou
30
25
20
15
10
5
0
−5
−20
−15
−10
−5
0
5
10
15
20
25

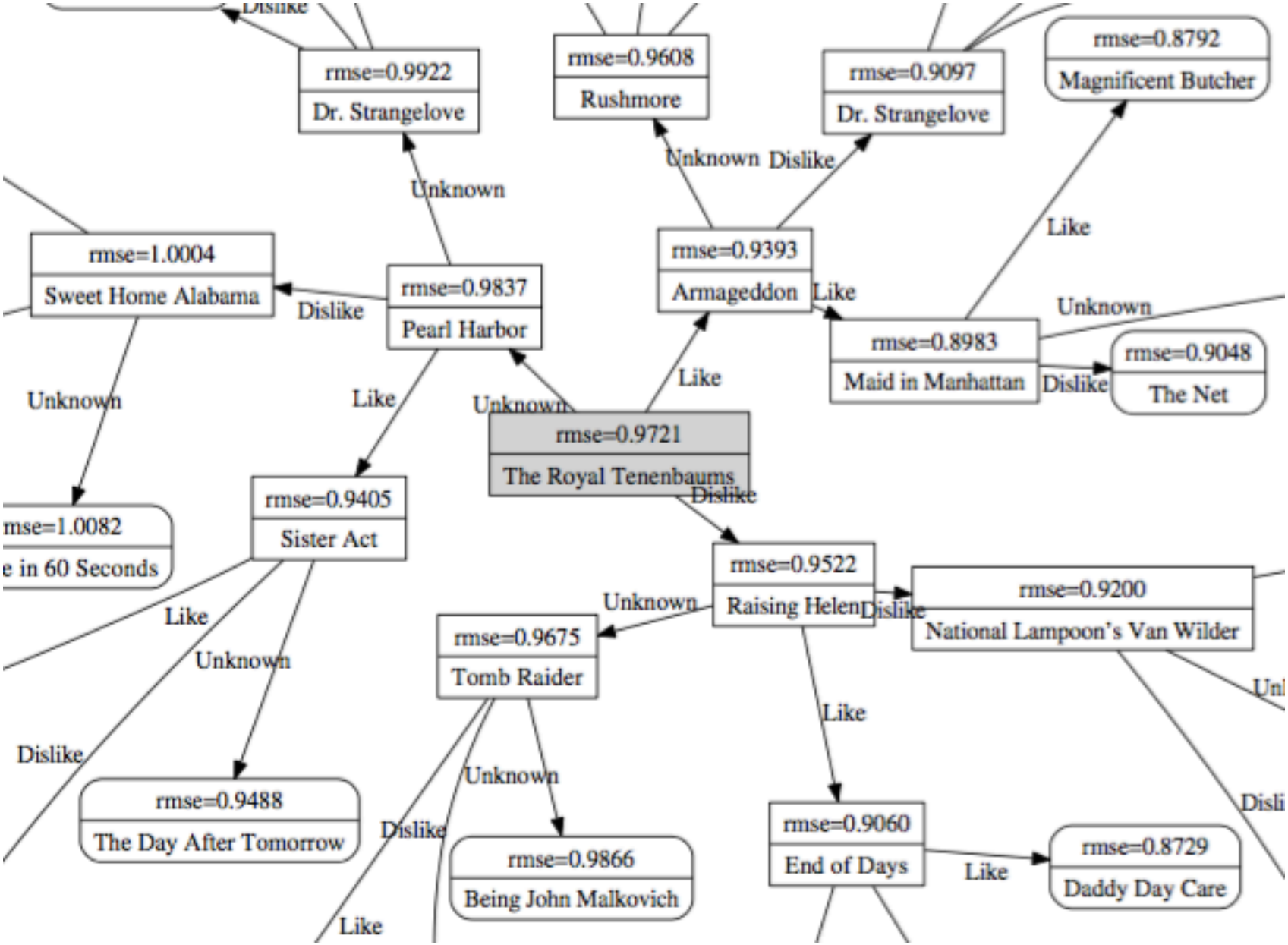
Mon sujet de thèse

Sur quels items vaut-il mieux sonder un nouveau venu pour le profiler efficacement ?

- populaires, pour que l'utilisateur puisse les noter
- controversés, pour que ce soit informatif

(Arbres de décision, tests adaptatifs.)

Les arbres de décision de Yahoo



Merci de votre attention !

- research.mangaki.fr
- jj@mangaki.fr

Fork us
on GitHub!