Comment exploiter les commentaires d'internautes pour la recommandation automatique

Damien Poirier

Paris, le 11 juin 2012

Contexte et problématique

Description de la méthodologie

- Extraction des textes
- Inférence de notes sur les textes
- Recommandation par filtrage collaboratif

Évaluation

Conclusion

Contexte

Mise en œuvre d'un système de recommandation personnalisée

Deux choix : filtrage collaboratif ou basé sur le contenu.

Contexte

Quelle que soit l'approche utilisée, des données sont nécessaires :

- Données d'usages pour le filtrage collaboratif
- Données descriptives pour le filtrage basé sur le contenu

Que faire lorsqu'elles ne sont pas disponibles ?

Contexte

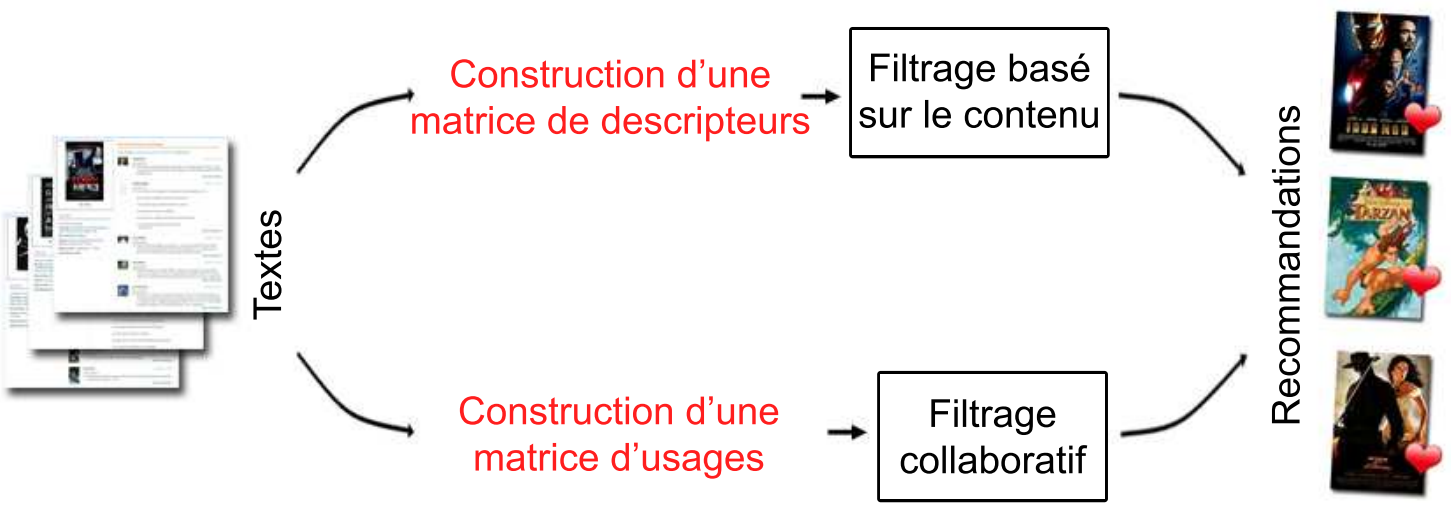
Le Web regorge de ressources, principalement textuelles, qui contiennent un grand nombre d'avis et d'opinions d'utilisateurs sur des sujets de toutes sortes.

(blogs, forums, boîtes à réactions, etc.)

Problématique

Objectif principal

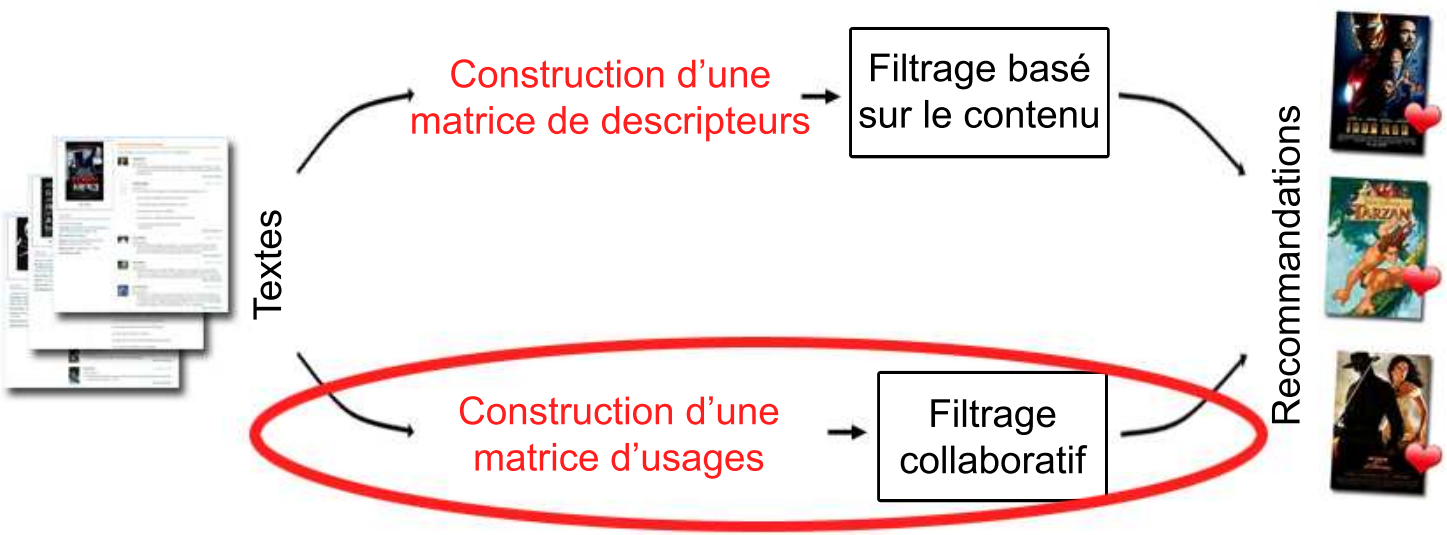
Montrer que l'on peut extraire, à partir de textes communautaires non structurés, des informations utiles aux moteurs de recommandation existants.



Problématique

Objectif principal

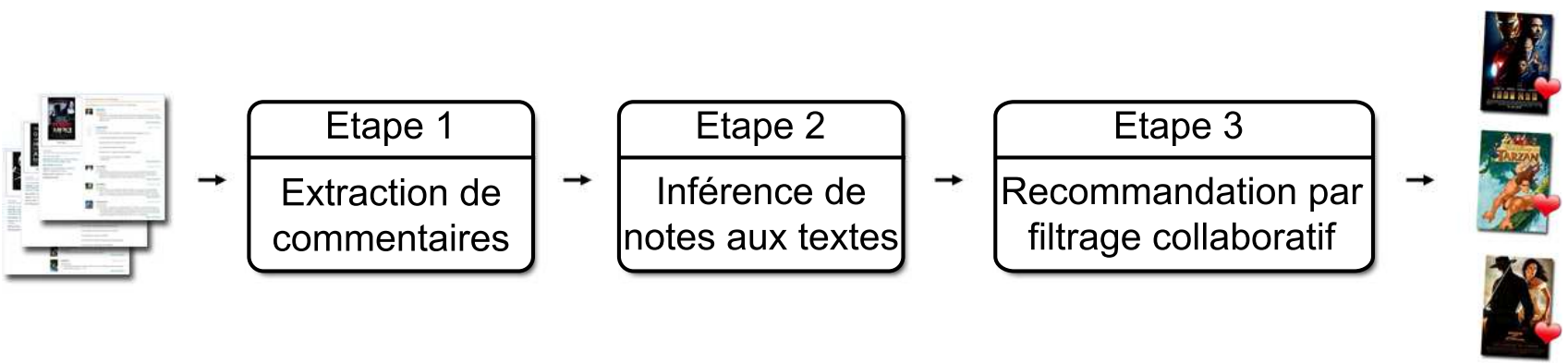
Montrer que l'on peut extraire, à partir de textes communautaires non structurés, des informations utiles aux moteurs de recommandation existants.



Méthodologie

Principale contribution :

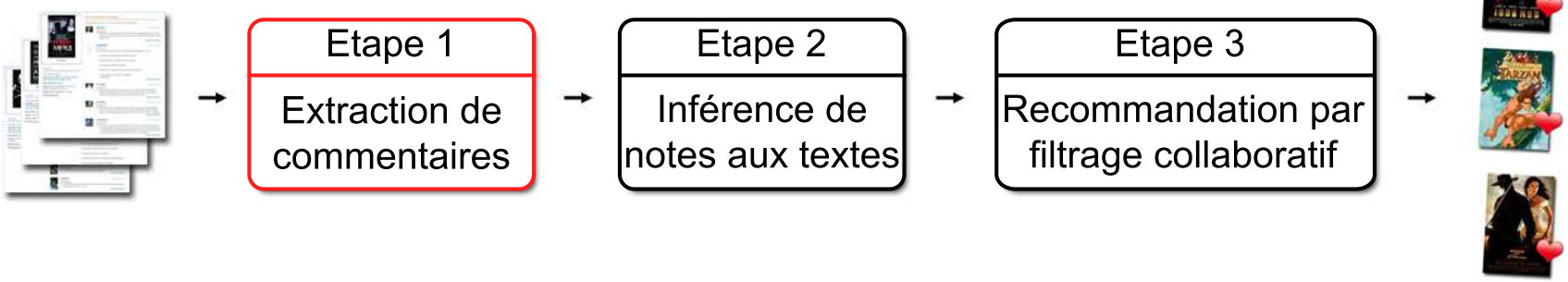
Mise en place d'une méthodologie complète, **automatisée** et en grande partie **reproductible** (étapes 2 et 3)



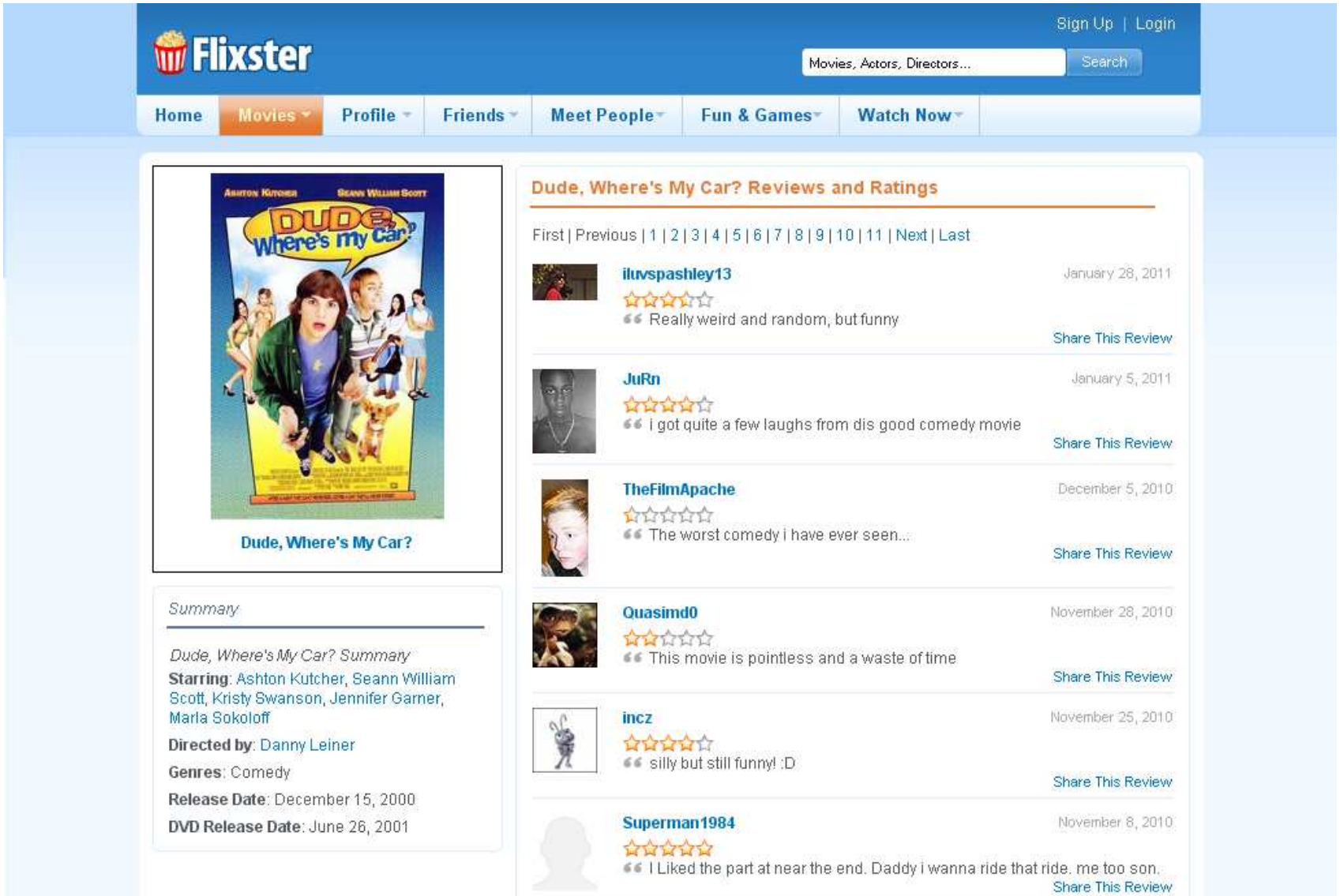
1ère étape : Extraction des textes

Objectif de la tâche :

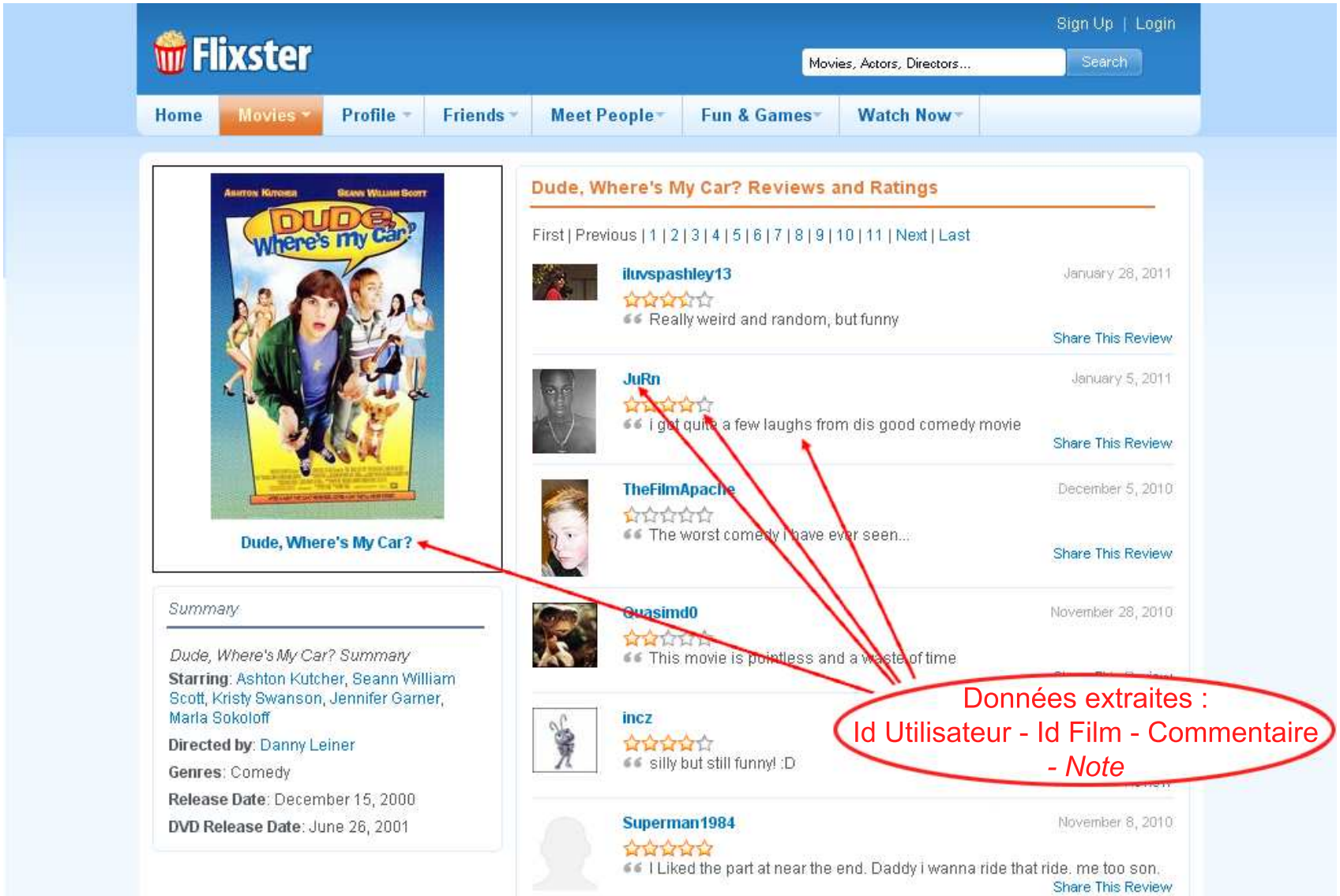
Acquérir des textes exprimant des opinions sur les produits de notre catalogue



Provenance des commentaires



Provenance des commentaires



Description du corpus

Chiffres :

- Nombre d'entrée : $>$ 3 millions de commentaires
  (sous la forme de 4-uplets Id utilisateur-Id film-Texte-Note)
- Utilisateurs concernés : $\simeq$ 100 000
- Films concernés : $\simeq$ 10 000

Particularités des textes :

- Nombre moyen d'unités (mots + ponctuations + symboles + chiffres, etc.) par commentaire : 15
- Emploi du **Discours électronique Médié** (DEM) (Panckhurst [1])

Exemples :

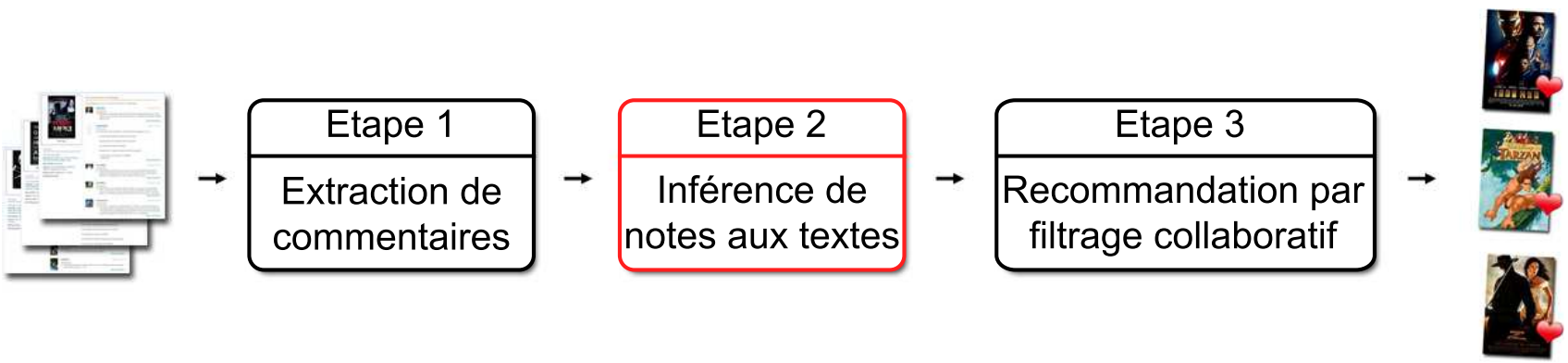
" weeeeeeiiiiiiirrrrrdddddd movie "
" i liked this cuz this shit can really happen. "
" I see it @ work everyday so this movie sucks ! "
" OMG I LOVED THIS MOVIE ! ! ! ! ! ! ! ! ! ! ! ! ! <3 "

1. Le discours électronique médié : bilan et perspectives, *2006*

2ème étape : Inférence des notes

Objectif de la tâche :

Attribuer des notes aux textes afin d'obtenir une matrice d'usages



Classification d'opinion

Objectif de la classification d'opinion

Classer les textes selon l'opinion exprimée par l'auteur (positif, négatif ou neutre)

Deux grands types de méthodes (Pang et al.[2]) :

- Les méthodes basées sur les lexiques
- Les méthodes basées sur l'apprentissage automatique

2. Opinion mining and sentiment analysis, *Found. Trends Inf. Retr., 2008*

Classification d'opinion : 1ère approche

Méthodes basées sur les lexiques

Objectif : Construire des lexiques contenant des mots liés à l'expression de l'opinion et classer les textes en fonction de la présence de ces mots.

Exemples :

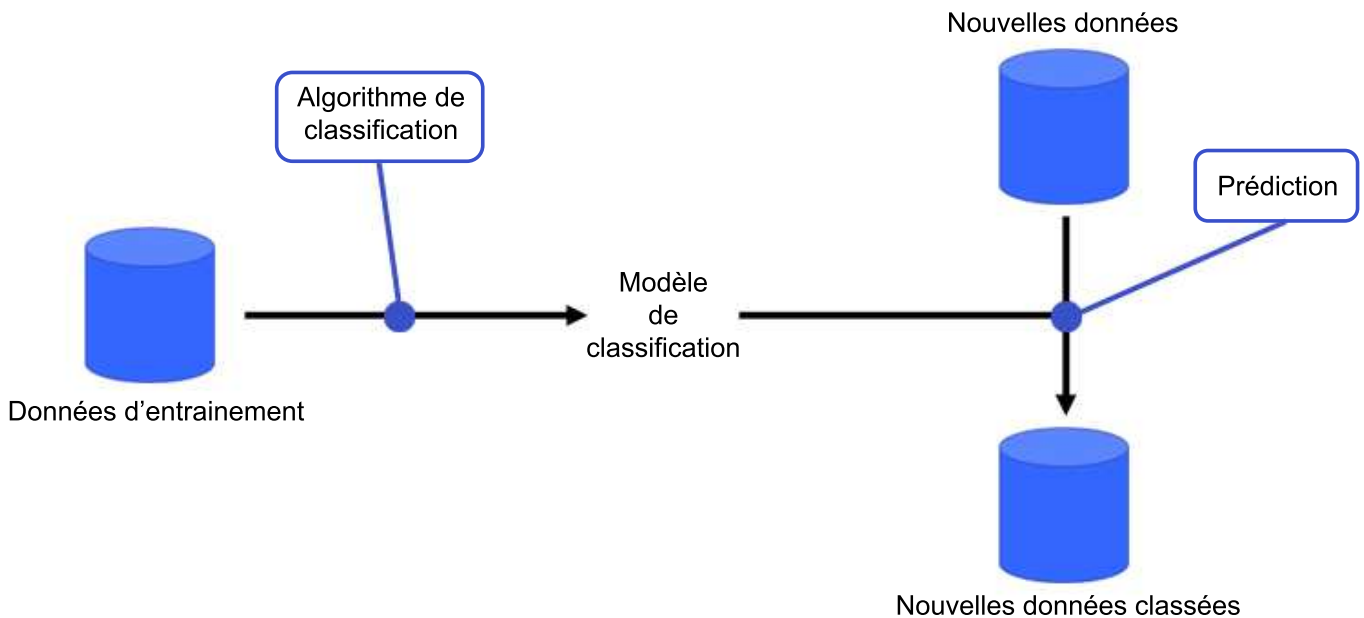
Mots positifs	awesome, great, best, love, ...
Mots négatifs	dislike, awful, hate, bad, ...

$\Longrightarrow$ Trop de contraintes manuelles pour l'automatisation des traitements

Classification d'opinion : 2ème approche

Méthodes basées sur l'apprentissage automatique

Objectif de la classification supervisée : apprendre des modèles à partir d'exemples et classer les nouveaux textes à l'aide de ces modèles.



Méthodes basées sur l'apprentissage automatique

Plusieurs choix à plusieurs niveaux :

Pré-traitements	Représentation
Minusculisation	Binaire
Lemmatisation	Fréquentielle
Stemmatisation	Fréquentielle normalisée
Correction	TF-Idf
...	...

Classifieur	Nombre de classes
SVM	2
Naïve Bayes	3
Réseaux de neurones	5
Règles de décision	Échelle continue
...	...

Méthodes basées sur l'apprentissage automatique

Conditions expérimentales :

- Construction d'un corpus d'apprentissage :
  - 175 000 nouveaux commentaires déjà classés
  - Classes équilibrées
  - Aucune intersection au niveau des films et des utilisateurs
  - Pas plus de 10 commentaires rédigés par le même auteur
- Mesure utilisée :

$$F_{score} = 2 * \frac{Précision * Rappel}{Précision + Rappel}$$

- Meilleur résultat obtenu :

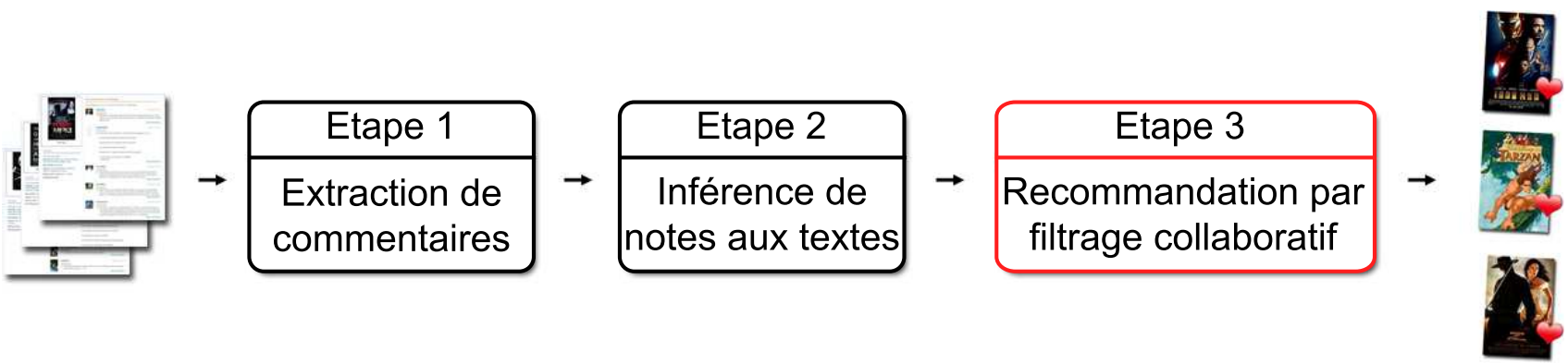
(Pré-traitements minimaux, représentation fréquentielle, SVM, classification binaire $\Longrightarrow$

$$F_{score} = 0,74$$

3ème étape : Recommandations

Objectif de la tâche :

Établir des recommandations à partir de la matrice d'usages



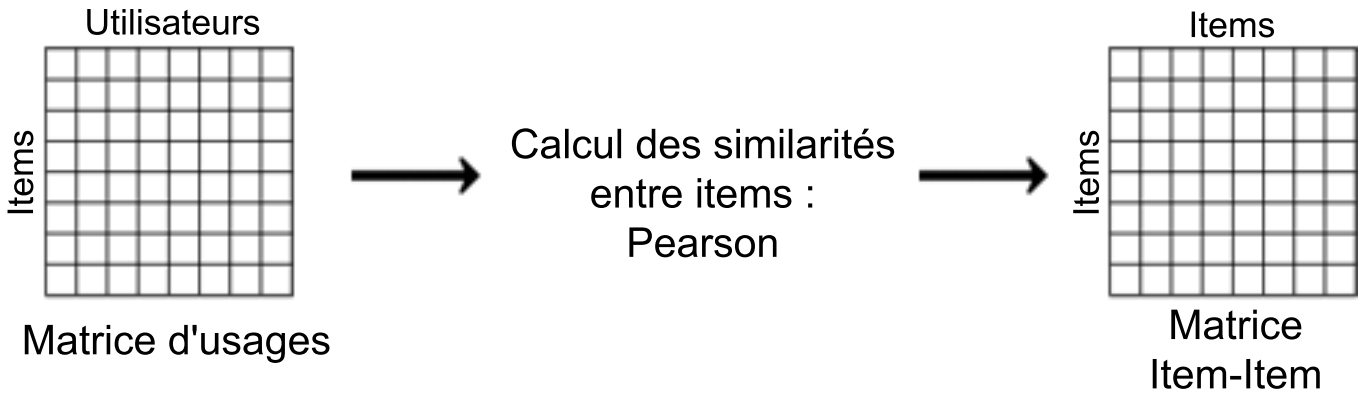
Description du moteur

- Moteur développé chez Orange Labs (Meyer [3])
- Permet le filtrage collaboratif et basé sur le contenu
- Permet la recommandation contextuelle et personnalisée
- Performances au niveau de l'état de l'art

$\Longrightarrow$ Fonctionnement en 2 étapes

3. Apport des données thématiques dans les systèmes de recommandation : hybridation et démarrage à froid, *In Extraction et gestion des connaissances (EGC'2011), Brest, France, 2011*

Construction de la matrice item-item

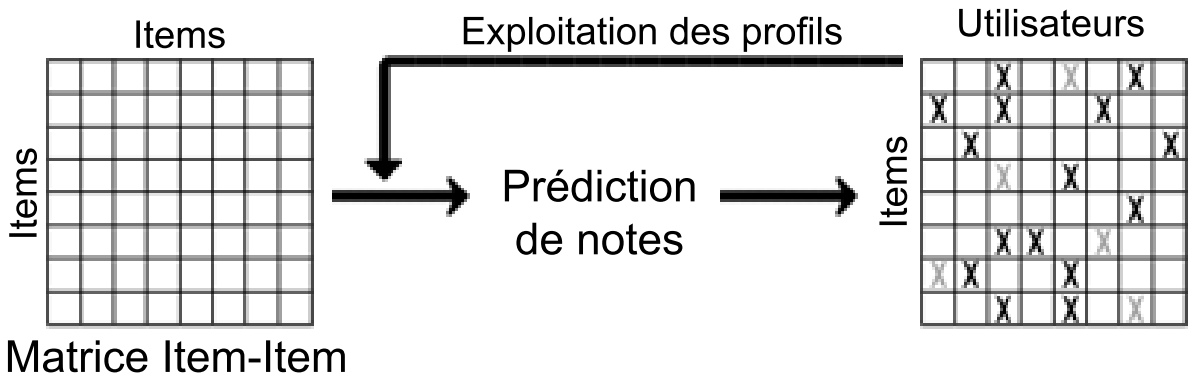


$$Pearson(i,j) = \frac{\sum_{\{u \in S_i \cap S_j\}} (r_{iu} - \overline{r_i}) \times (r_{ju} - \overline{r_j})}{\sqrt{\sum_{\{u \in S_i \cup S_j\}} (r_{iu} - \overline{r_i})^2 \sum_{\{u \in S_i \cup S_j\}} (r_{ju} - \overline{r_j})^2}}$$

S_i (resp. S_j) : l'ensemble des utilisateurs ayant noté l'item i (resp. j)
r_{iu} (resp. r_{ju}) : la note donnée par l'utilisateur u sur l'item i (resp. j)

$\overline{r_i}$ (resp. $\overline{r_j}$) la moyenne des notes obtenues par i (resp. j)

Prédiction des notes

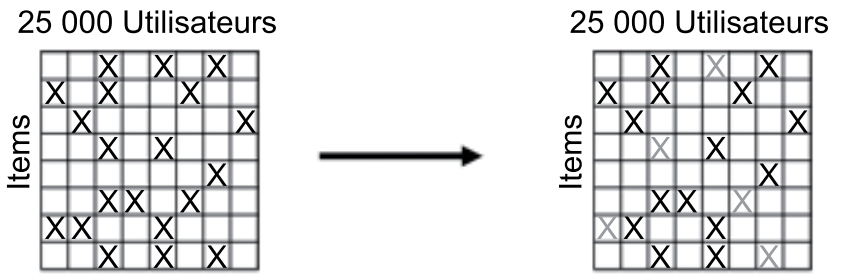


$$p_{ui} = \overline{r_i} + \frac{\sum_{\{j \in S_u\}} Pearson(i,j) \times (r_{uj} - \overline{r_j})}{\sum_{\{j \in S_u\}} Pearson(i,j)}$$

S_u : l'ensemble des notes connues données par l'utilisateur u

Conditions expérimentales

Données de test : extraites du challenge Netflix (Bennett et al.) [4]



X : 90% des notes dédiées aux profils utillisateurs
X : 10% des notes dédiées à la validation des notes prédites

Données d'entrée :

- Données d'usages obtenue à partir des commentaires **Flixster**
  - Notes issues de la classification d'opinion
  - Vraies notes
- Descripteurs **IMDb** pour le filtrage basé sur le contenu
- Données d'usages **Netflix** (profils) pour le filtrage collaboratif

4. The netflix prize, *San Jose, California, ACM, 2007*

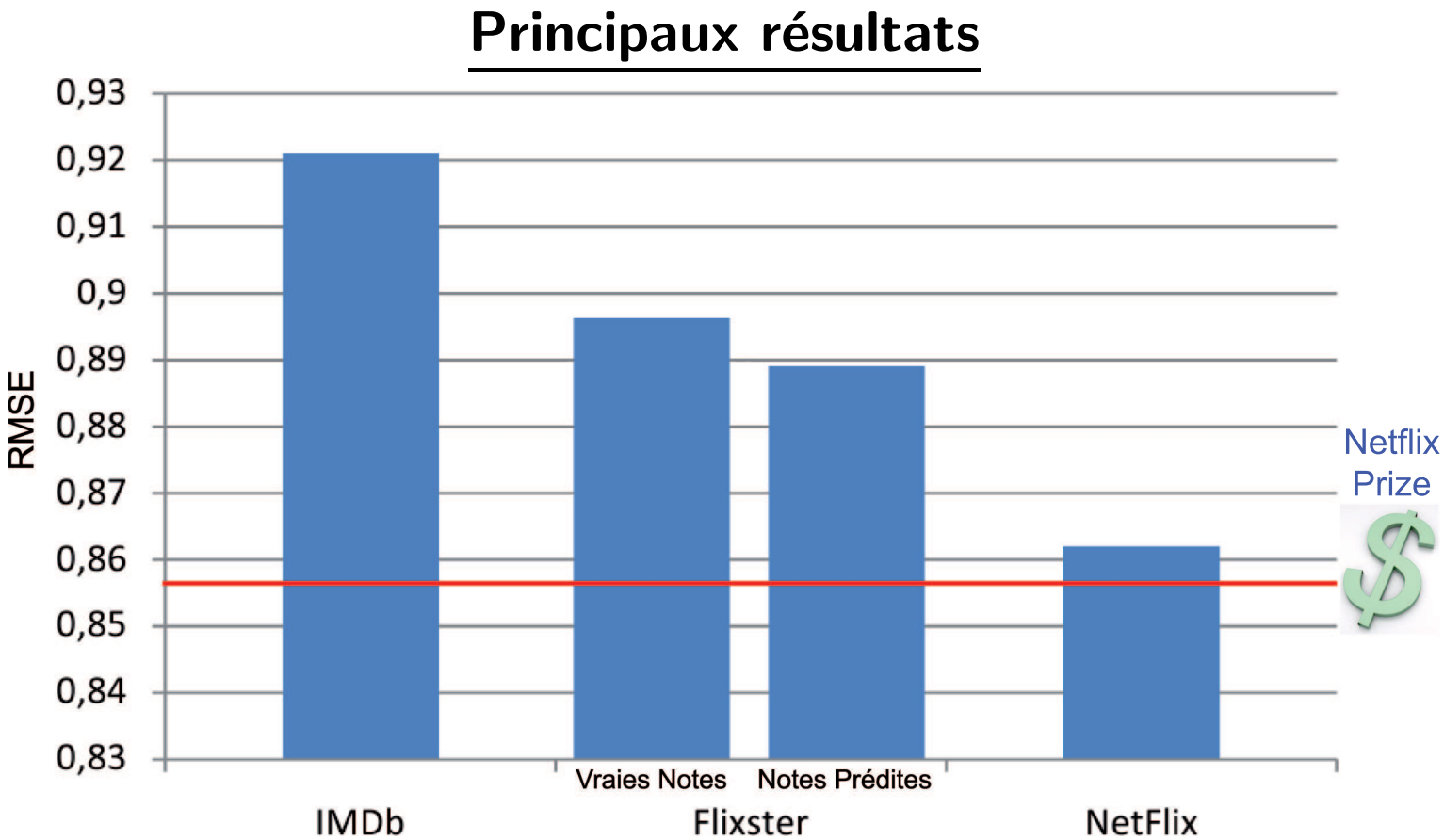
Méthode d'évaluation

Mesure utilisée : Erreur quadratique moyenne

$$RMSE = \sqrt{\frac{\sum_{u,i} (p_{ui} - n_{ui})^2}{n}}$$

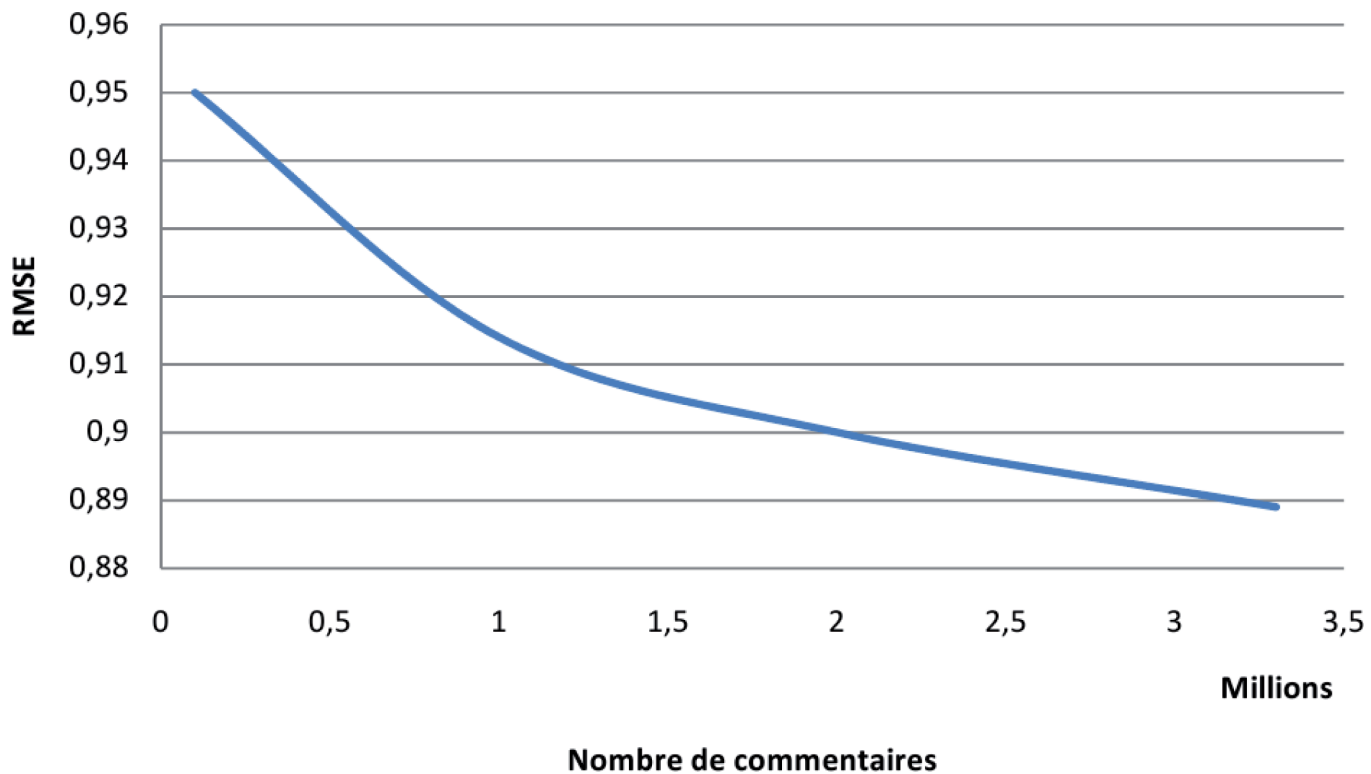
n_{ui} = note réelle donnée par l'utilisateur u sur l'item i
p_{ui} = note prédite par le moteur de recommandation
n = nombre total de notes prédites

Résultats


Principaux résultats
RMSE
0,93
0,92
0,91
0,9
0,89
0,88
0,87
0,86
0,85
0,84
0,83
Vraies Notes
Notes Prédites
IMDb
Flixster
NetFlix
Netflix Prize

Résultats

Une amélioration possible et simple :
Augmenter le nombre de commentaires présents dans la base



Bilan

Objectif atteint !

Démonstration faite que les textes non-structurés issus de blogs peuvent pallier le manque de données en recommandation

- Le choix de transformer les textes en données d'usages (classification d'opinion) s'est avéré pertinent :
  $\Longrightarrow$ Meilleures performances qu'avec les descripteurs IMDb

Merci pour votre attention