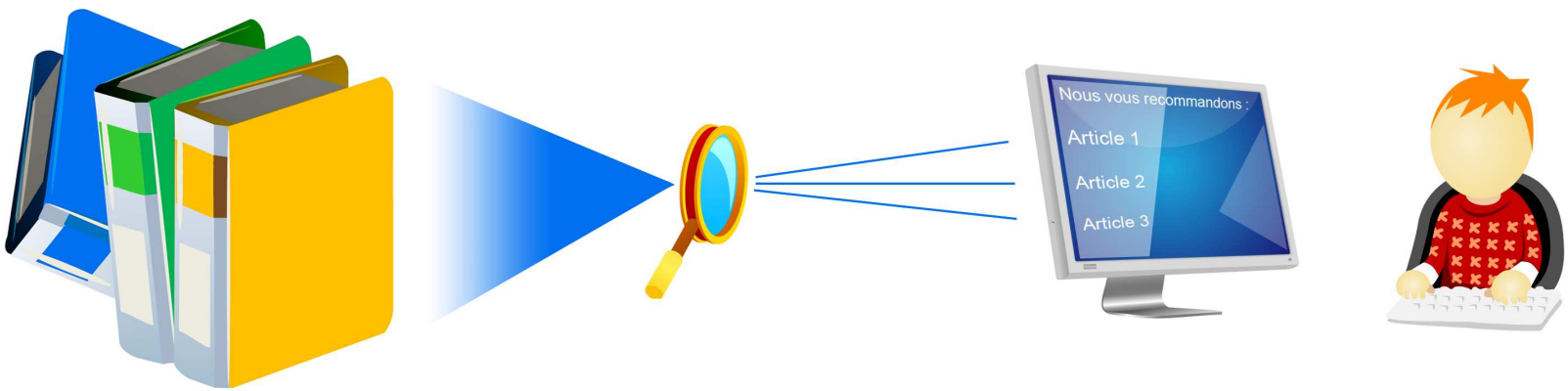
Des Textes Communautaires à la Recommandation

Damien Poirier

Soutenance de thèse pour obtenir le grade de Docteur en Informatique

Orléans, le 11 février 2011

Qu'est-ce que la recommandation automatique ?



Objectif : sélectionner des articles (appelés items) dans un catalogue pour en faciliter la lecture.

Qu'est-ce que la recommandation automatique ?

Deux grands types de recommandations possibles :

Contextuelle :

Consiste à proposer des items en lien avec l'item observé

Personnalisée :

Consiste à proposer des items adaptés à chaque utilisateur

Qu'est-ce que la recommandation automatique ?

Deux grands types de recommandations possibles :

Contextuelle :

Consiste à proposer des items en lien avec l'item observé

Personnalisée :

Consiste à proposer des items adaptés à chaque utilisateur

Exemple d'application : la VOD

Recommandation personnalisée

L'intégration d'un service de recommandation personnalisée nécessite tout d'abord des **informations sur les utilisateurs**...

On peut alors collecter des données du type :

- L'utilisateur a regardé tel ou tel film
- L'utilisateur a aimé/pas aimé tel ou tel film

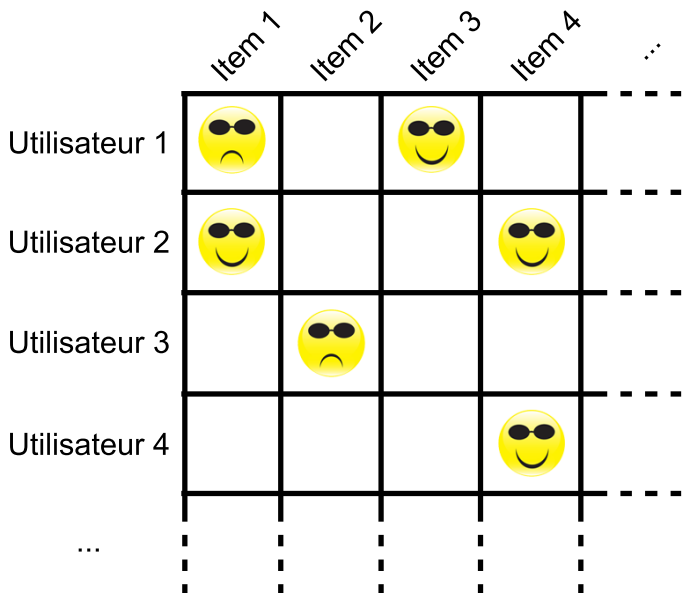
Recommandation personnalisée

L'intégration d'un service de recommandation personnalisée nécessite tout d'abord des **informations sur les utilisateurs**...

On peut alors collecter des données du type :

- L'utilisateur a regardé tel ou tel film
- L'utilisateur a aimé/pas aimé tel ou tel film

$\Longrightarrow$ **Matrice d'usages**



Recommandation personnalisée

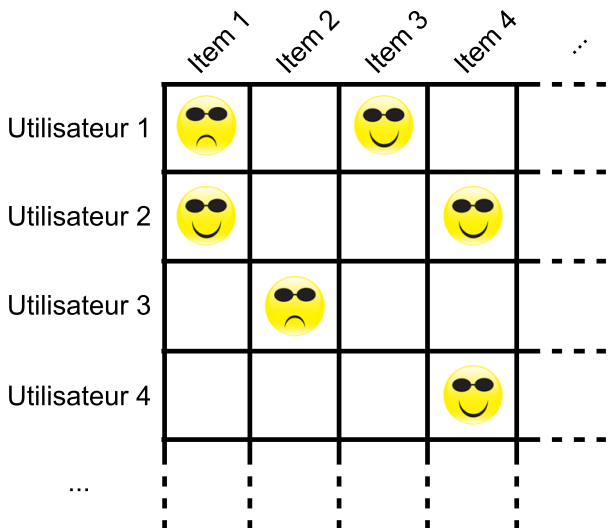
Ainsi que des **informations sur les items**.

Deux grandes approches sont possibles (Nageswara Rao et al. [1]) :

- **Filtrage collaboratif** (basé sur les profils des utilisateurs)
- **Filtrage basé sur le contenu** (basé sur des descripteurs d'items)

1. Application domain and functional classification of recommender systems a survey, *In Desidoc journal of library and information technology, volume 28, pages 17-36, ACM Press, 2008*

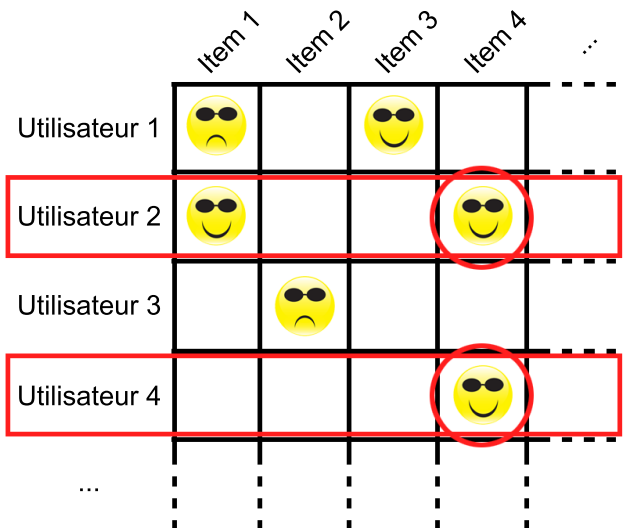
Principe du filtrage collaboratif



- 1ère étape :
  - Calcul des similarités entre utilisateurs
  - ou similarités entre items
- 2ème étape :
  - Prédiction des notes manquantes à l'aide des similarités et des profils utilisateurs
- 3ème étape :
  - Recommandation des meilleures notes prédites

$\Rightarrow$ Très performant mais nécessite une grande quantité de données

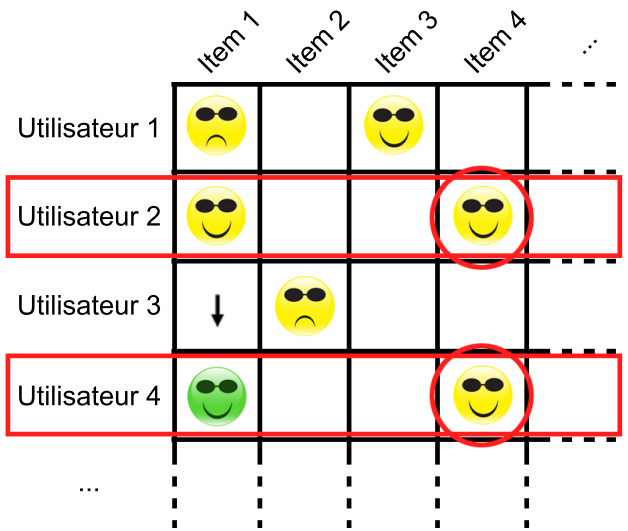
Principe du filtrage collaboratif



- 1ère étape :
  - Calcul des similarités entre utilisateurs
  - ou similarités entre items
- 2ème étape :
  - Prédiction des notes manquantes à l'aide des similarités et des profils utilisateurs
- 3ème étape :
  - Recommandation des meilleures notes prédites

$\Rightarrow$ Très performant mais nécessite une grande quantité de données

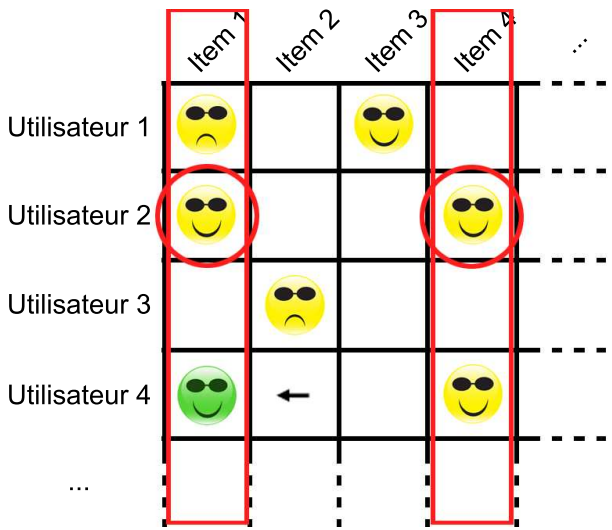
Principe du filtrage collaboratif



- 1ère étape :
  - Calcul des similarités entre utilisateurs
  - ou similarités entre items
- 2ème étape :
  - Prédiction des notes manquantes à l'aide des similarités et des profils utilisateurs
- 3ème étape :
  - Recommandation des meilleures notes prédites

$\Rightarrow$ Très performant mais nécessite une grande quantité de données

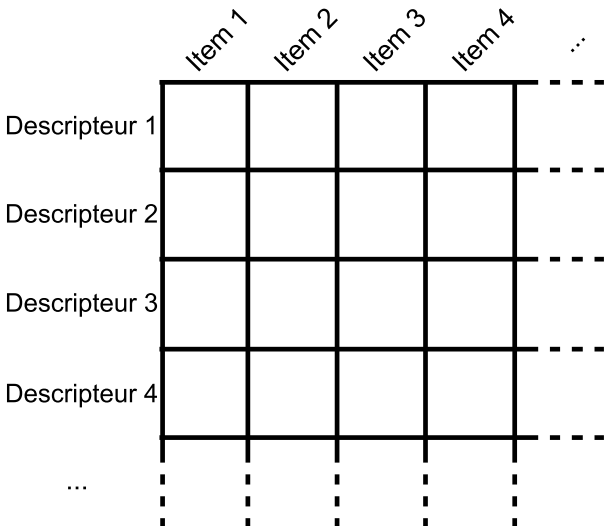
Principe du filtrage collaboratif



- 1ère étape :
  - Calcul des similarités entre utilisateurs
  - ou similarités entre items
- 2ème étape :
  - Prédiction des notes manquantes à l'aide des similarités et des profils utilisateurs
- 3ème étape :
  - Recommandation des meilleures notes prédites

$\Rightarrow$ Très performant mais nécessite une grande quantité de données

Principe du filtrage basé sur le contenu



- Exemples de descripteurs de films :
  - Genre, réalisateur, acteurs principaux, etc.
- 1ème étape :
  - Calcul des similarités entre items
- 2ème étape :
  - Prédiction des notes
- 3ème étape :
  - Recommandation des meilleures notes prédites

$\Rightarrow$ Moins performant que le filtrage collaboratif

Problématique

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

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

Que faire lorsqu'elles ne sont pas disponibles ?

Problématique

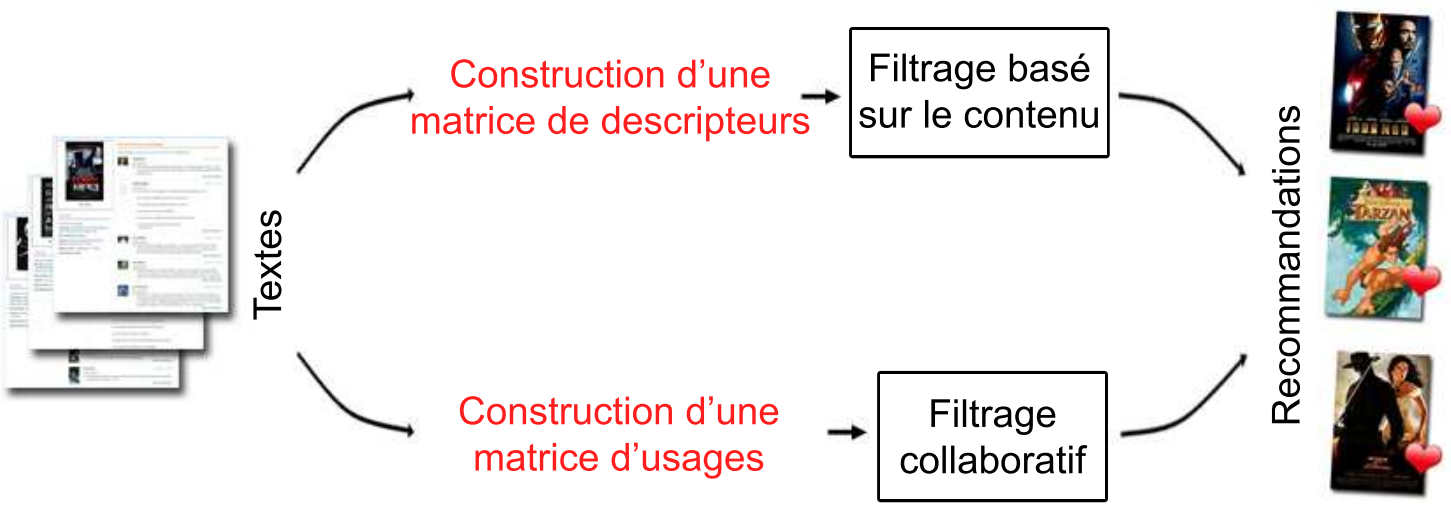
Le Web regorge de ressources, principalement textuelles, qui contiennent un grand nombre d'avis et d'opinions d'utilisateurs sur des sujets de toutes sortes.

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

Problématique

Objectif principal de la thèse

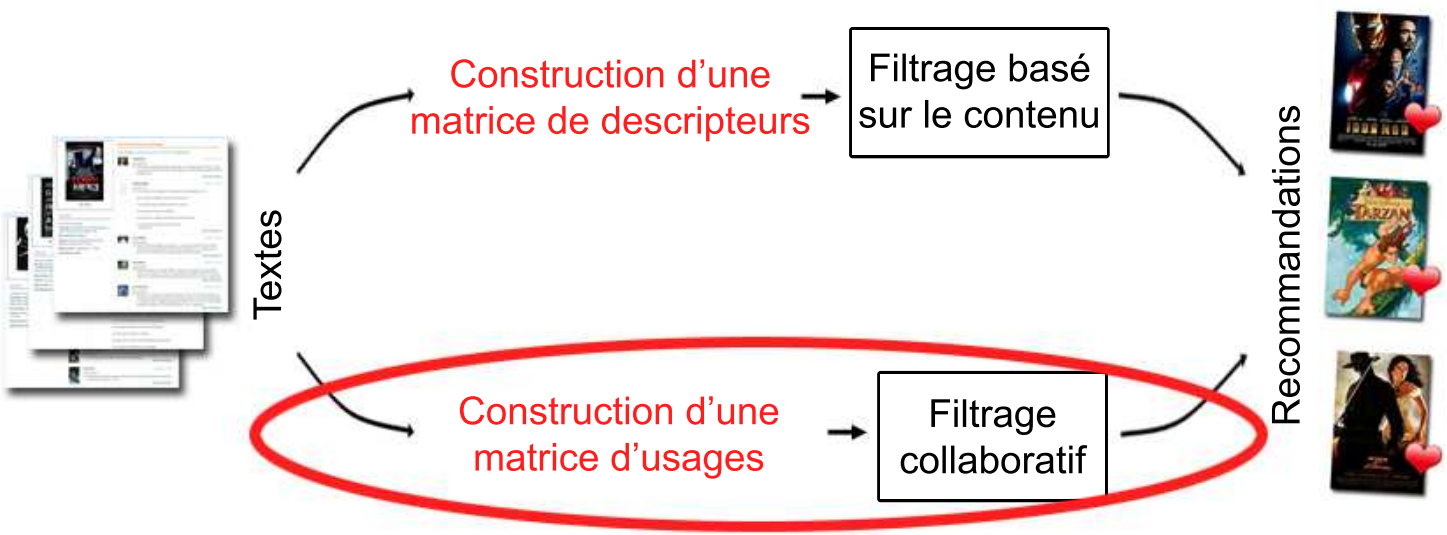
Montrer que l'on peut extraire, à partir de textes communautaires non structurés, des informations utiles aux moteurs de recommandation existants.



Problématique

Objectif principal de la thèse

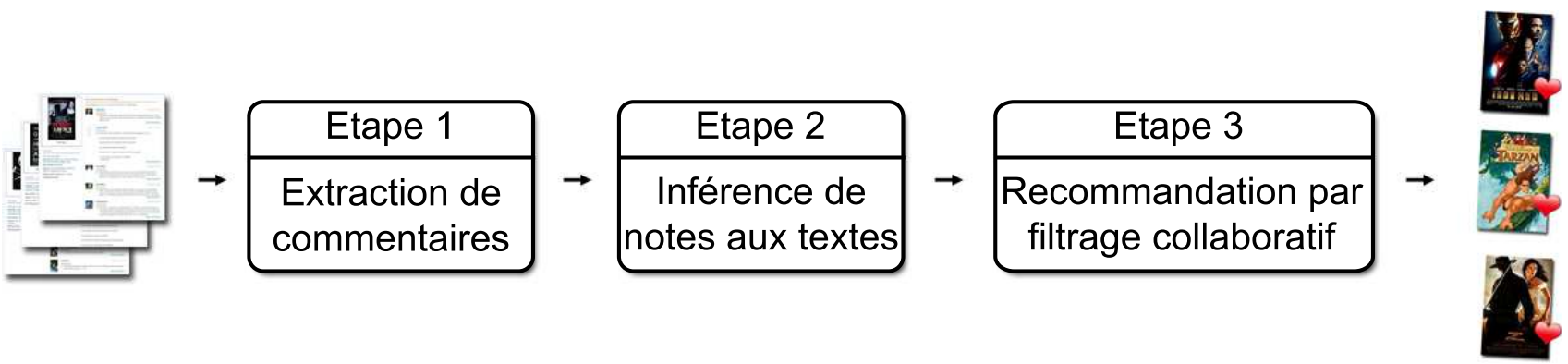
Montrer que l'on peut extraire, à partir de textes communautaires non structurés, des informations utiles aux moteurs de recommandation existants.



Méthodologie

Principale contribution :

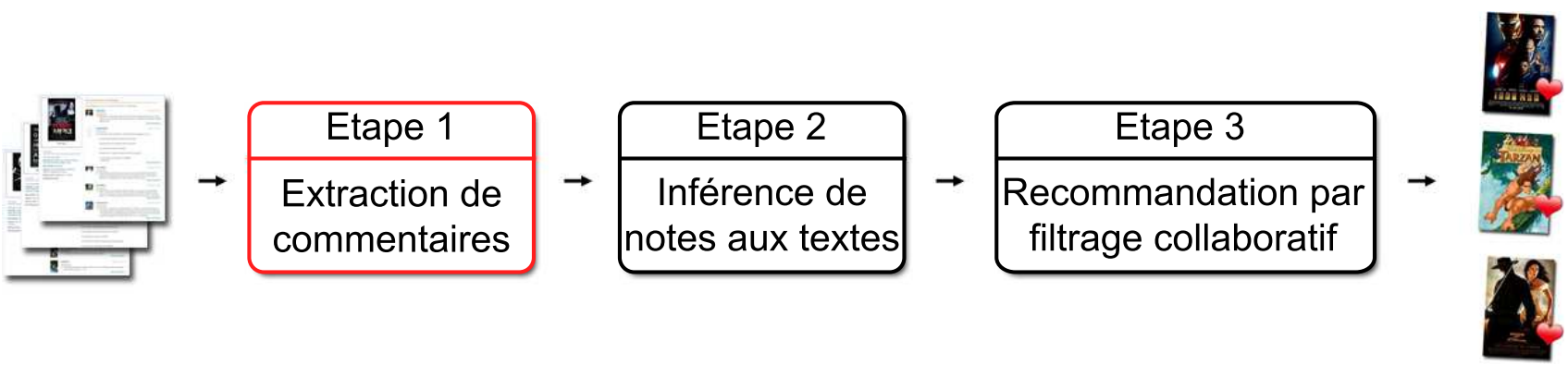
Mise en place d'une méthodologie complète, **automatisée** et en grande partie **reproductible** (étapes 2 et 3)



1ère étape : Extraction des textes

Objectif de la tâche :

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



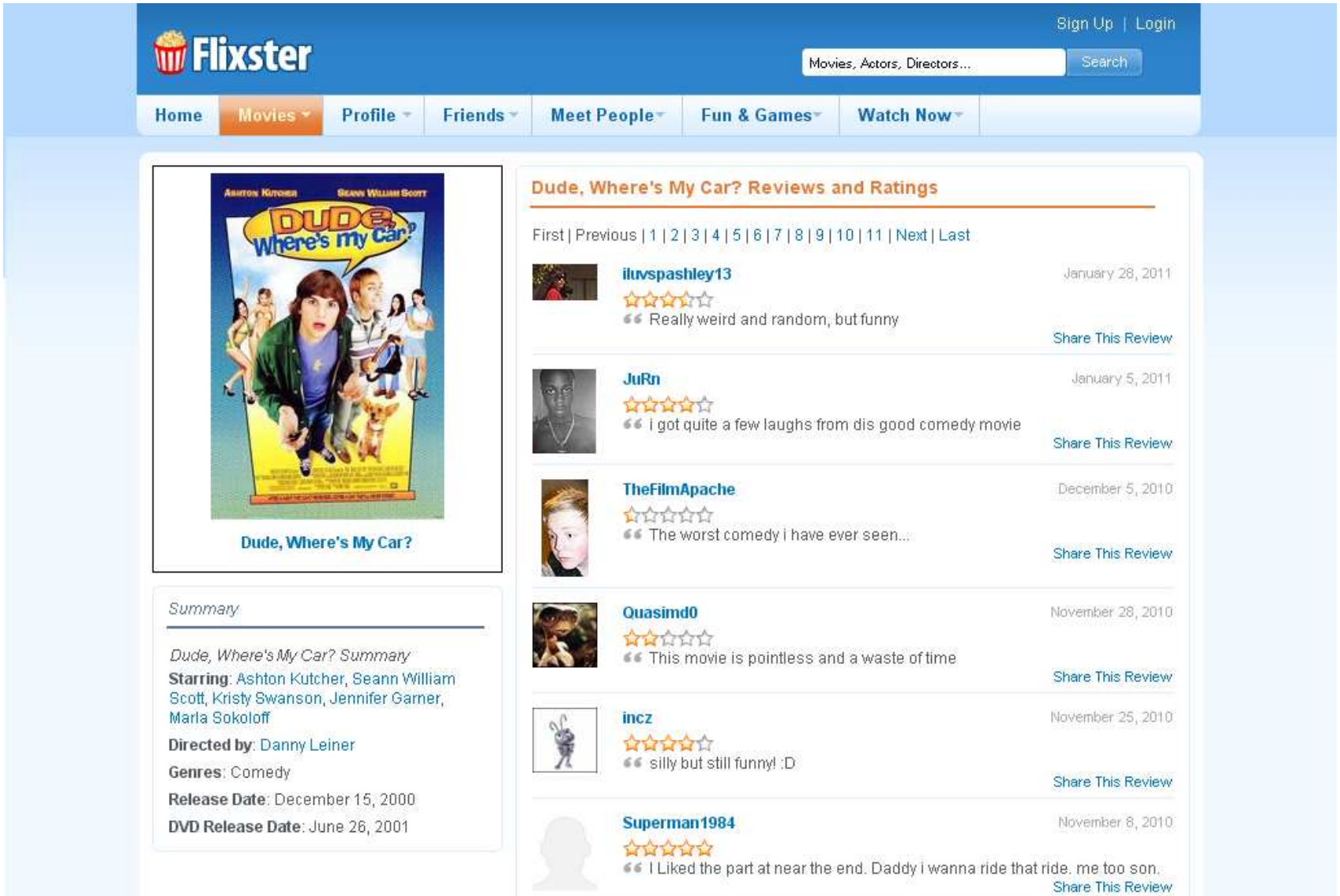
Provenance des commentaires

Domaine d'application : les films

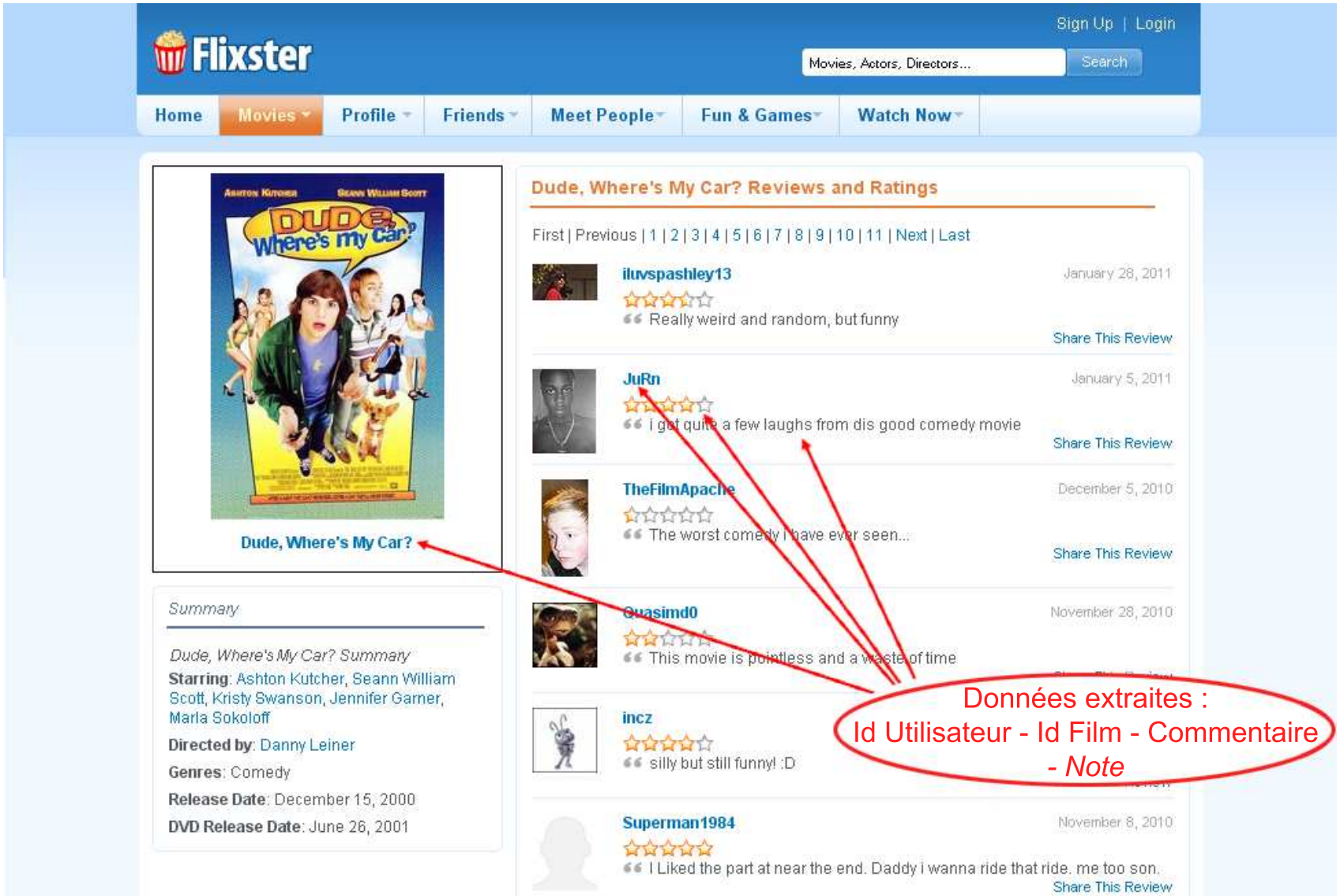
- Construction d'un crawler pour le site communautaire *Flixster*[2]
  1. Recherche les films souhaités à l'aide du moteur de recherche du site
  2. Aspiration des pages contenant les commentaires sur les films
  3. Extraction des commentaires
- Nettoyage des textes (suppression des urls, des balises html, des lignes vides, des caractères inconnus, etc.)

2. www.flixster.com

Provenance des commentaires



Provenance des commentaires

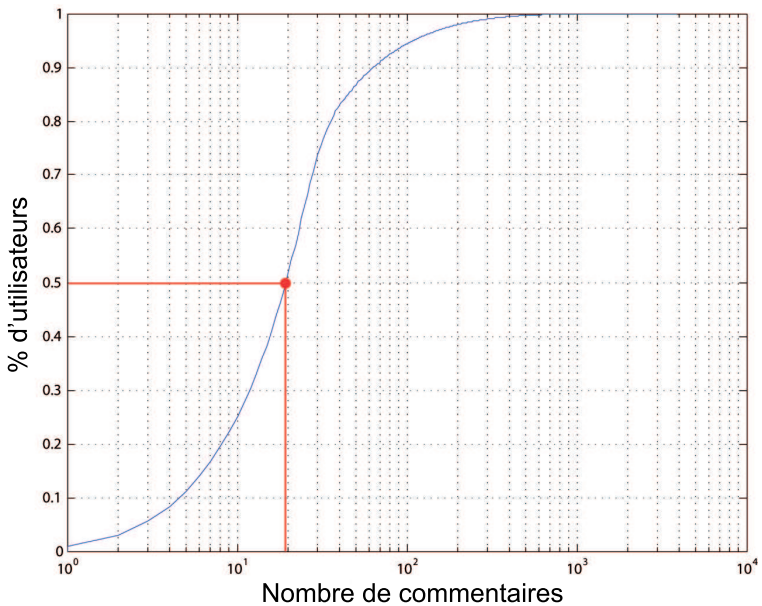


Description du corpus

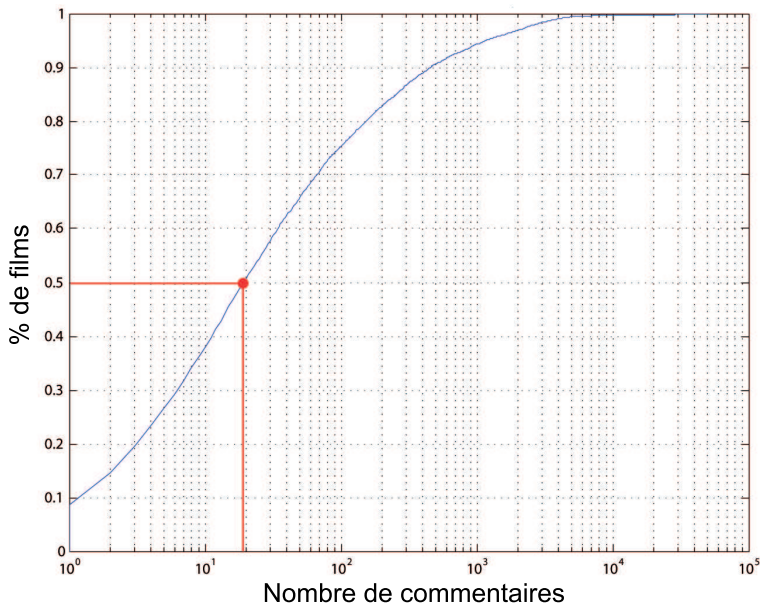
Chiffres :

- Nombre : $>$ 3 millions de commentaires
  (sous la forme de 4-uplets Id utilisateur-Id film-Texte-Note)

- Utilisateurs concernés : $\simeq$ 100 000
  $\Rightarrow$ Moyenne : **34** com./utilisateurs
  $\Rightarrow$ Médiane : **20** com./utilisateurs



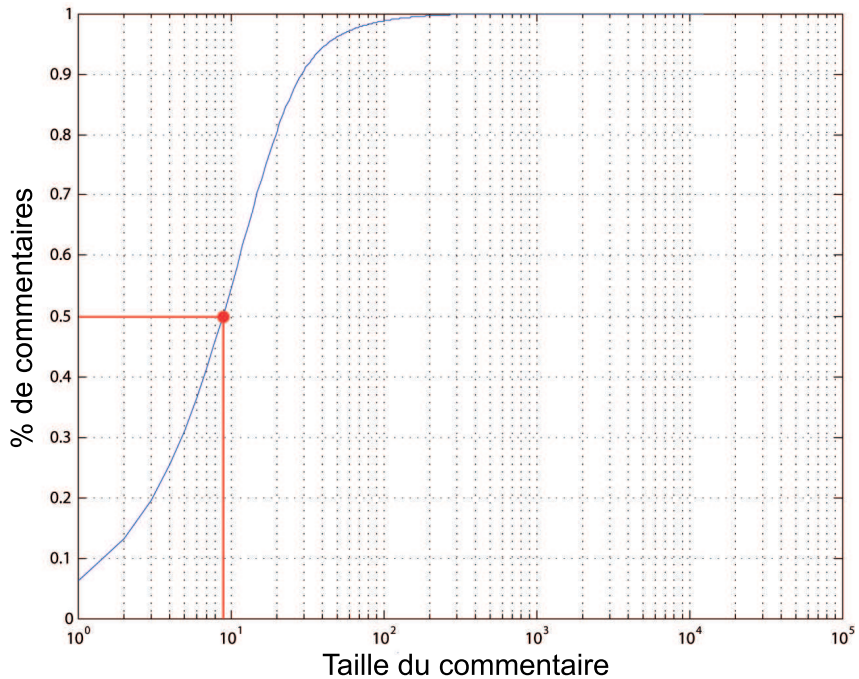
- Films concernés : $\simeq$ 10 000
  $\Rightarrow$ Moyenne : **318** com./films
  $\Rightarrow$ Médiane : **20** com./films



Nature des textes

1ère particularité : la taille

- Nombre moyen d'unités (mots + ponctuations + symboles + chiffres, etc.) par commentaire : 15
- 90% des commentaires font moins de 30 mots
- 50% des commentaires font moins de 9 mots



Nature des textes

2ème particularité : le style d'écriture

$\Longrightarrow$ Emploi du **Discours Électronique Médié** (DEM) (Panckhurst [3])

- Didascalies électroniques
  - Smileys : " ;-) "
  - Abus de majuscules : " NON "
  - Étirement de mots : " cooool "
  - Répétitions de ponctuation : " ? ? ? "
- Fautes d'orthographe
- Néographie
  - Substitutions : " g " (j'ai), " mwa " (moi)
  - Réductions : " dsl ", " jaten "
  - Suppressions : " salut cava "
  - Ajouts : " oki "

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

Nature des textes

Particularités moins visibles :

- Moins de verbes
- Moins de pronoms
- Langage familier
- **Dialecte évolutif**

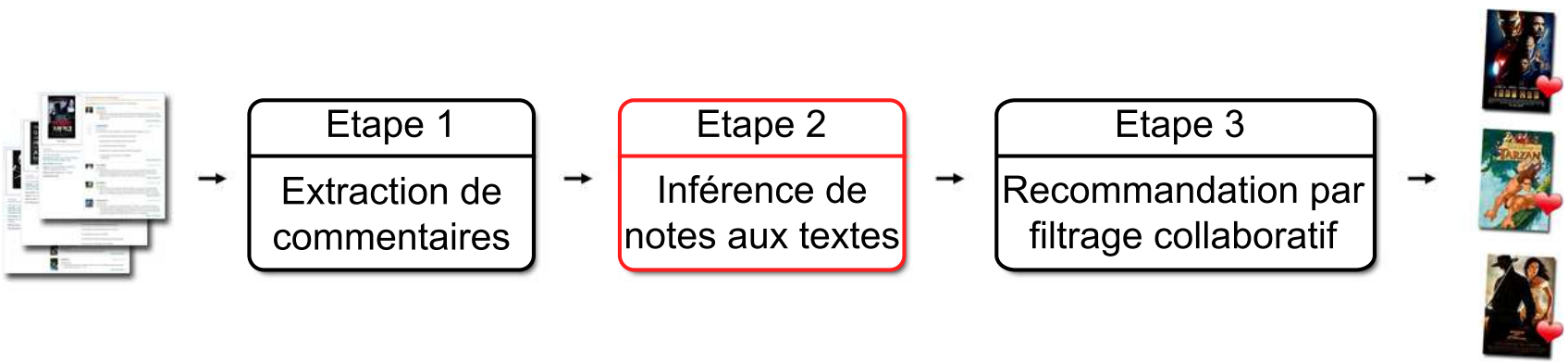
Exemples :

" weeeeeeiiiiiiirrrrrddddddd 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 "

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 ou négatif)

Deux grands types de méthodes (Pang et al.[4]) :

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

4. 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 :

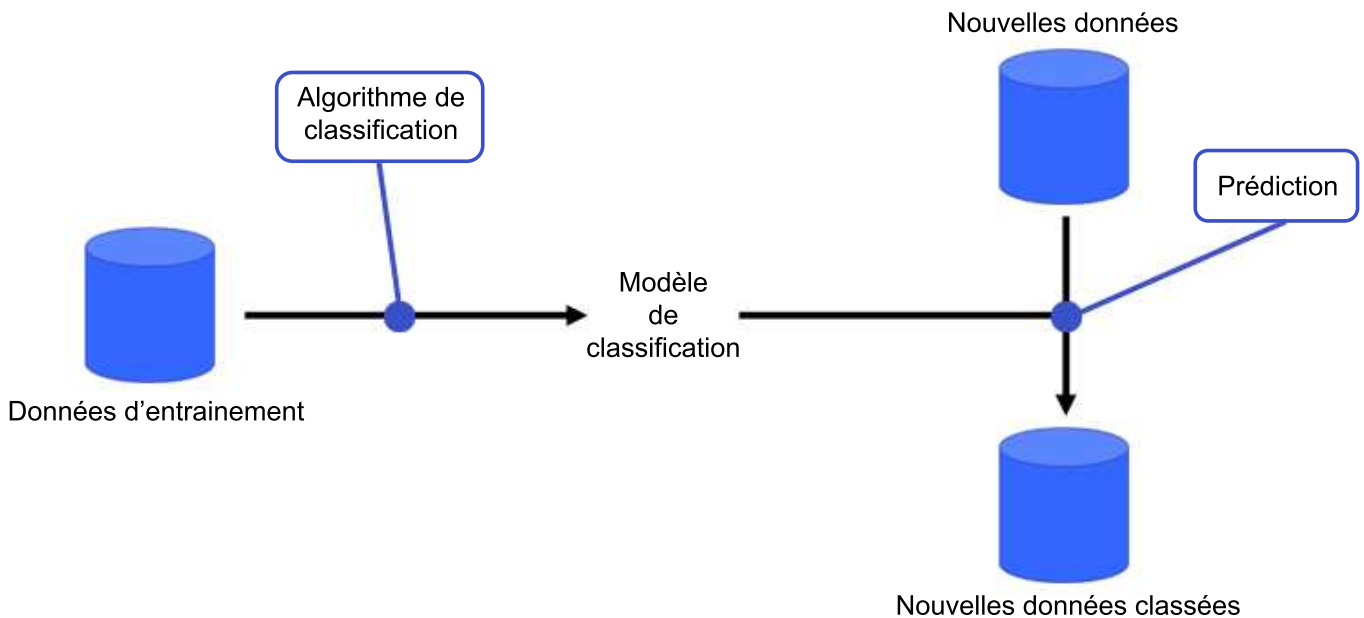
Mots positifs	awesome, great, best, love, ...
Mots négatifs	dislike, awful, hate, bad, ...

$\Longrightarrow$ 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 :

Segmentation	Représentation
Mots	Binaire
N-grammes de mots	Fréquentielle
Unités	Fréquentielle normalisée
N-grammes de caractères	TF-Idf
...	...

Classifieur	Nombre de classes
SVM	2
(S) Naïve Bayes [5]	3
Réseaux de neurones	5
Règles de décision	Échelle continue
...	...

5. SNB : Selective Naïve Bayes

Méthodes basées sur l'apprentissage automatique

Pré-sélection basée sur la littérature :

- Pang et al. : Thumbs up ? : sentiment classification using machine learning techniques. *In Proceedings of EMNLP '02, 2002*
- Wilson et al. : Just how mad are you ? finding strong and weak opinion clauses. *In In Proceedings of AAAI, 2004*
- Nigam et al. : Towards a robust metric of polarity. *In Computing Attitude and Affect in Text : Theory and Applications, 2006*
- Kaiser et al. : Swarm intelligence for analyzing opinions in online communities. *Hawaii International Conference on System Sciences, 2010*
- Plantié et al. : Is a voting approach accurate for opinion mining ? *In Proceedings of DaWaK '08, 2008*
- Maurel et al. : Classification d'opinions par méthodes symbolique, statistique et hybride. *In Actes du 3ème DEfi Fouille de Textes, 2007*
- etc.

Méthodes basées sur l'apprentissage automatique

Segmentation	Représentation
Mots	Binaire
N-grammes de mots	**Fréquentielle**
Unités	**Fréquentielle normalisée**
N-grammes de caractères	**TF-Idf**
...	...

Classifieur	Nombre de classes
SVM [6]	**2**
(S) Naïve Bayes [7]	**3**
Réseaux de neurones	**5**
Règles de décision	Échelle continue
...	...

6. SVM^{light}
7. KHIOPS

Méthodes basées sur l'apprentissage automatique

Nombre initial d'unités (mots + ponctuations...) = **150 000**

Choix concernant les pré-traitements :

- Pré-traitements de base

Pré-traitements	Taille du vocabulaire
Minusculisation	$\simeq$ 60 000
Suppression des mots très peu fréquents	$\simeq$ 20 000

- Pré-traitements additionnels

Pré-traitements	Taille du vocabulaire
Corrections orthographiques et lemmatisation [8]	$\simeq$ 15 000
Suppression des mots vides [9]	$\simeq$ 19 700

8. Utilisation du logiciel TiLT
9. Évaluation de plusieurs listes prises sur Internet

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 :

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

$$Précision = \sum_{i=1}^{n} \frac{Documents\ correctement\ attribués\ à\ la\ classe\ i}{Nombre\ de\ documents\ attribués\ à\ la\ classe\ i}$$

$$Rappel = \sum_{i=1}^{n} \frac{Documents\ correctement\ attribués\ à\ la\ classe\ i}{Nombre\ de\ documents\ appartenant\ à\ la\ classe\ i}$$

Méthodes basées sur l'apprentissage automatique

Exemple d'évaluation :

F_{score} obtenus pour chaque représentation avec les trois classifieurs pour une classification binaire

	Fréquentielle	Fréq. normalisée	TF-IDF
SVM	**0,741**	0,724	0,726
SNB	0,726	0,729	0,728
NB	0,695	0,708	0,708

$\Longrightarrow$ Meilleurs résultats obtenus avec SVM et la représentation fréquentielle

Méthodes basées sur l'apprentissage automatique

Exemple d'évaluation :

F_{score} obtenus avec les pré-traitements minimaux et linguistiques et toutes les représentations possibles avec le classifieur SVM

	Fréquentielle	Fréq. normalisée	TF-IDF
Basiques	**0,741**	0,724	0,726
Anti-dictionnaire	0,720	0,700	0,703
Linguistiques	0,738	0,722	0,724

$\Longrightarrow$ Corrections et lemmatisation n'entraînent pas d'amélioration des résultats

Observations sur l'impact des pré-traitements courants

Suppression de la ponctuation

ATTENTION : La **ponctuation** joue un rôle important

Cas du " ! "		
Fréquence	Négatif	Positif
$n \geq 2$	37%	**63%**

" so funny ! ! ! ! !u have 2 c it ! ! ! ! ! "

Cas du " . "		
Fréquence	Négatif	Positif
$n \geq 1$	**65%**	35%

" so disappointed ..... "

Cas du " ? "		
Fréquence	Négatif	Positif
$n \geq 1$	**74%**	26%

Observations sur l'impact des pré-traitements courants

Suppression des mots vides

ATTENTION : Les **mots vides** jouent un rôle important

Cas de " but "		
Fréquence	Négatif	Positif
$n \geq 1$	**71%**	29%

"*i like science fiction but not this one*"

Cas de " it "		
Fréquence	Négatif	Positif
$n \geq 1$	23%	**77%**

"*it's tooooo funny*"

Cas de " or "		
Fréquence	Négatif	Positif
$n \geq 1$	**72%**	28%

Observations sur l'impact des pré-traitements courants

Lemmatisation ou stemmatisation

ATTENTION : La **conjugaison** des verbes joue également un rôle

Cas de " is "		
Fréquence	Négatif	Positif
$n \geq 1$	41%	**59%**

" I Love It. It is A Great Movie"

Cas de " was "		
Fréquence	Négatif	Positif
$n \geq 1$	**67%**	33%

" the movie was a waste of money"

Cas de " were "		
Fréquence	Négatif	Positif
$n \geq 1$	**69%**	30%

Conclusion des expérimentations

Méthode sélectionnée :

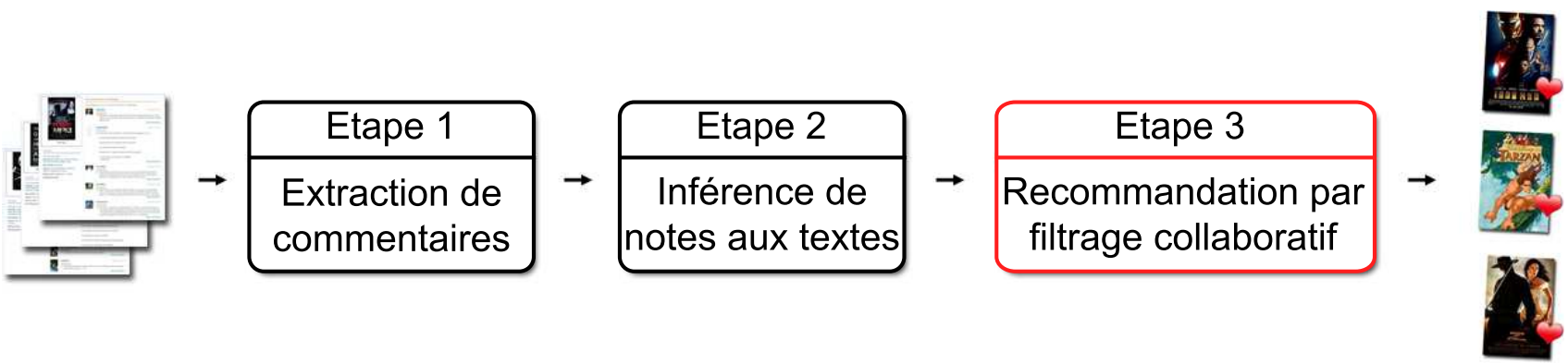
Critère	Choix
Pré-traitements	Basiques
Segmentation	Unités (Mots, ponctuations, nombres...)
Représentation	Fréquentielle
Classifieur	SVM
Nombre de classes	2

$\Longrightarrow$ **Construction de la matrice d'usages**

3ème étape : Recommandations

Objectif de la tâche :

Établir des recommandations à partir de la matrice d'usages



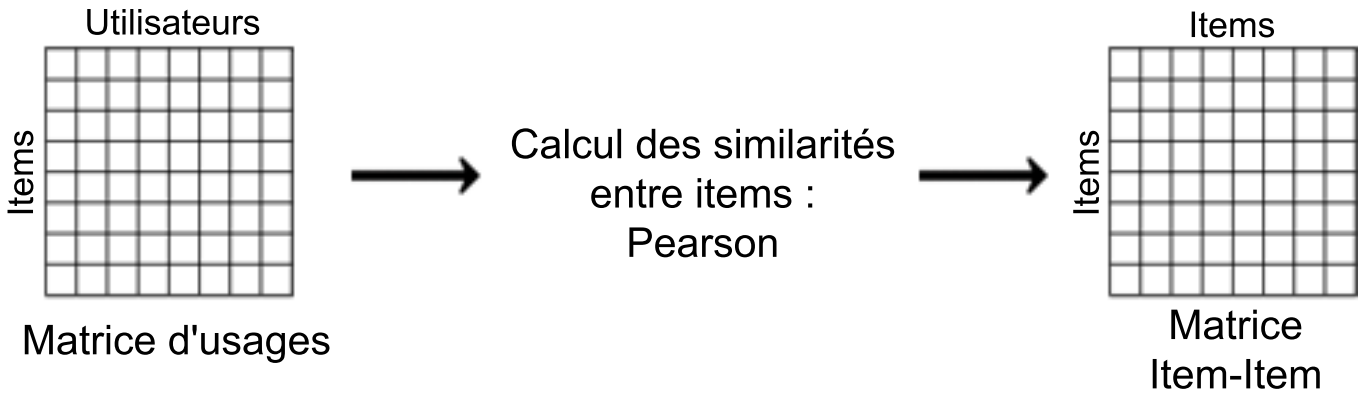
Description du moteur

- Moteur développé chez Orange (Meyer [10])
- Choix issus du contexte industriel
- 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

10. 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

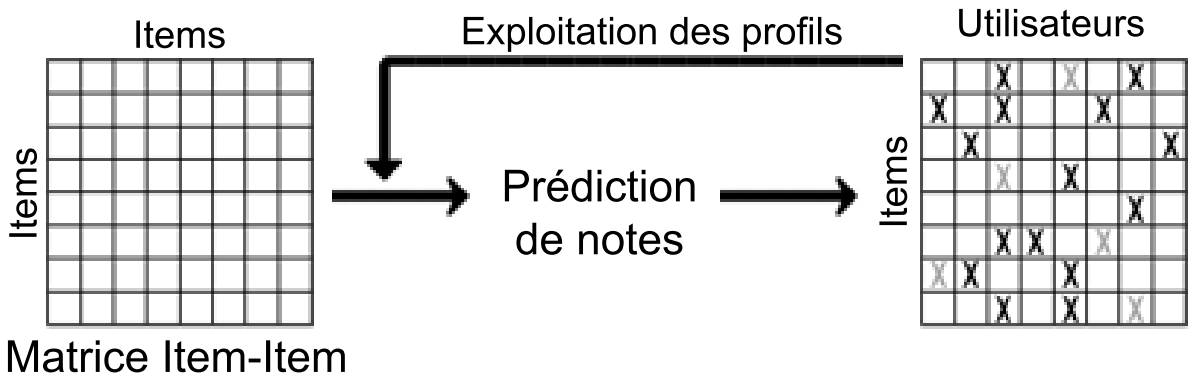


$$Pearson(i,j) = \frac{\sum_{\{u \in S_i \cap S_j\}} (r_{iu} - \overline{r_i}) \times (r_{ju} - \overline{r_j})}{\sqrt{\sum_{\{u \in S_i \cup S_j\}} (r_{iu} - \overline{r_i})^2 \sum_{\{u \in S_i \cup S_j\}} (r_{ju} - \overline{r_j})^2}}$$

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

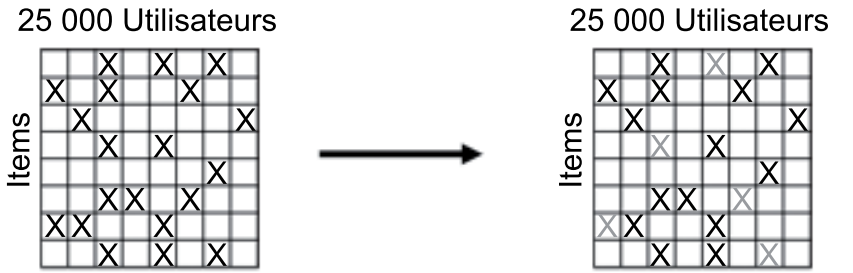


$$p_{ui} = \overline{r_i} + \frac{\sum_{\{j \in S_u\}} Pearson(i,j) \times (r_{uj} - \overline{r_j})}{\sum_{\{j \in S_u\}} Pearson(i,j)}$$

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

Conditions expérimentales

Données de test : extraites du challenge Netflix (Bennett et al.) [11]



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

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

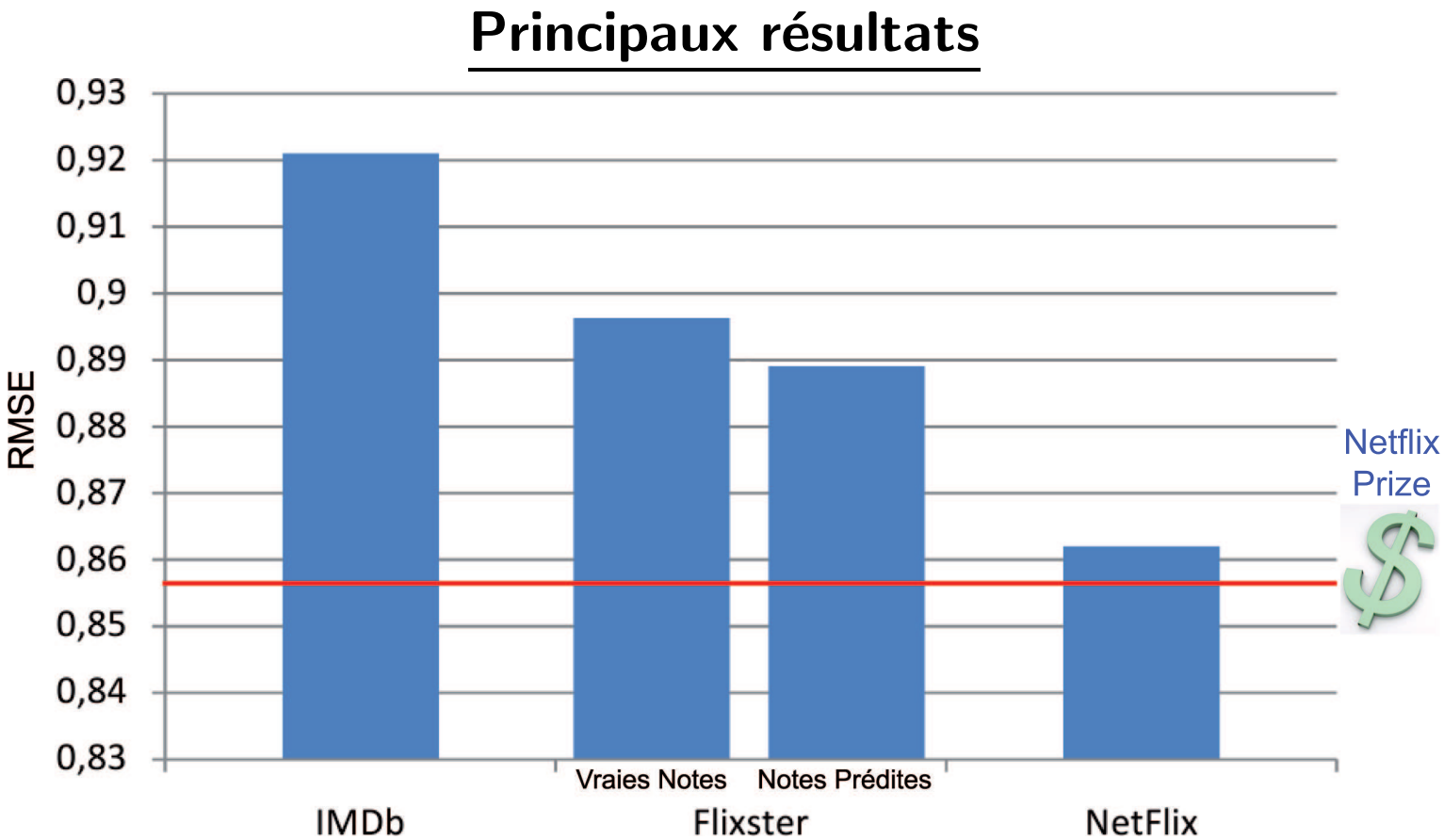
Méthode d'évaluation

Mesure utilisée : Erreur quadratique moyenne

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

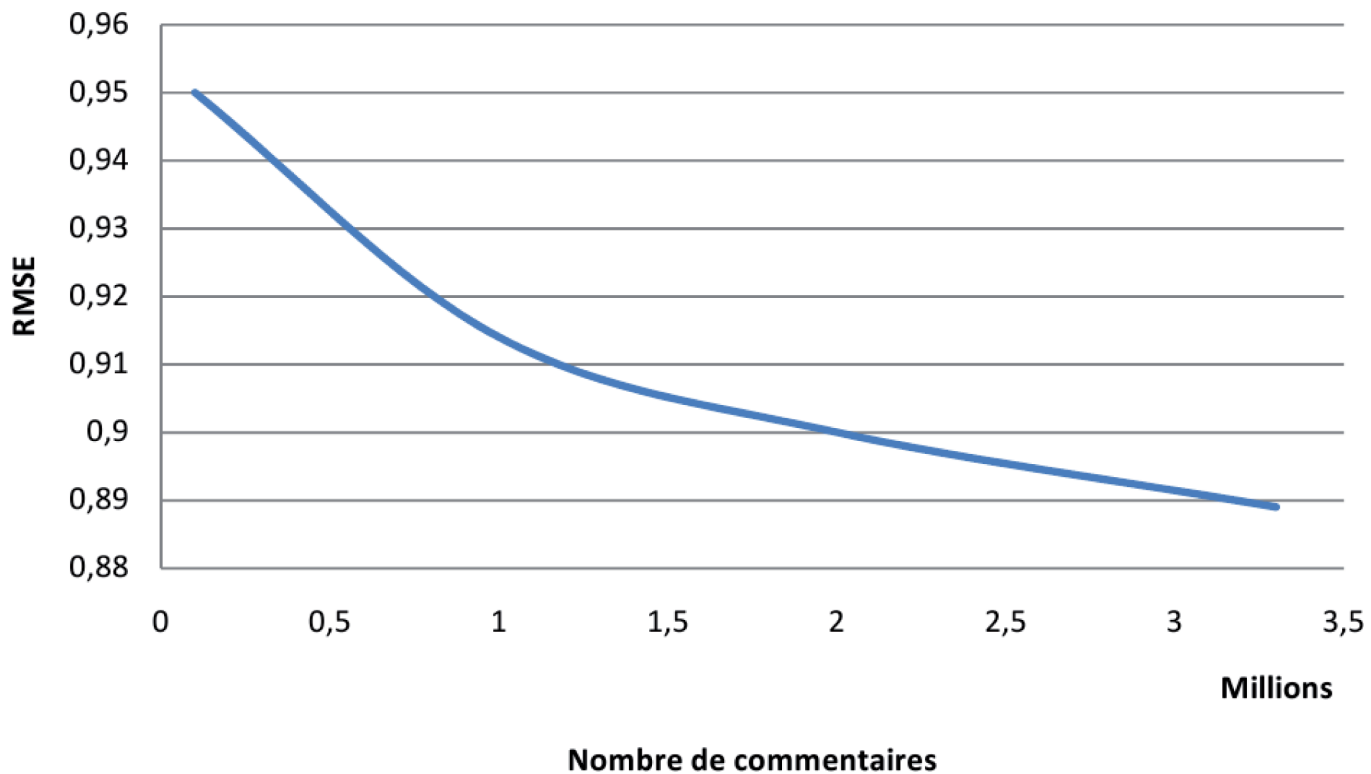
n_{ui} = note réelle donnée par l'utilisateur u sur l'item i
p_{ui} = note prédite par le moteur de recommandation
n = nombre total de notes prédites

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 s'est avéré pertinent
  $\Longrightarrow$ Meilleures performances qu'avec les descripteurs IMDb
- Validation et évaluation de chaque élément de la chaîne
- Étapes 2 (inférence des notes) et 3 (recommandation) automatisables et applicables à d'autres domaines que les films

Perspectives

- À court terme :
  - Évaluation de la chaîne sur d'autres domaines que les films et d'autres langues que l'anglais
- À plus long terme :
  - Automatisation complète de la chaîne
    - Détection automatique des commentaires sur les sites Web
    - Extraction automatique du sujet traité dans le commentaire
  - Extraction d'informations sociales (liens entre utilisateurs)

Merki ! ! !

Des kestions ? :-)