ATLAS - GdR MADICS

CALYPSO: systèmes de recommandation pour la publicité en ligne
Partenariat entre le LIG, Kelkoo et Purch

Charlotte Laclau

Equipe dAta analysis, Modeling and mAchine learning (AMA)
Laboratoire d'Informatique de Grenoble (LIG)

Action ATLAS

- **Responsables**: Marianne CLAUSEL (LJK, INSMI) et Massih-Reza AMINI (LIG, INS2i)
- **Contexte**: Apprentissage, opTimisation à Large échelle et cAlcul diStribué.
- **Objectif** : développer de nouveaux outils pour prendre en compte la nature des données analysées
- **4 axes**:
  - ⋆ collecte des données
  - ⋆ apprentissage de représentation
  - ⋆ optimisation pour l'apprentissage
  - ⋆ calcul haute performance
- **Evènement** : l'action ATLAS et CAp
  - ⋆ Session industrielle en partenariat avec Viseo
  - ⋆ Léon Bottou (USA, Facebook), Yurii Nesterov (Belgique, UCL), Yves Grandvallet (UTC Compiègne)
  - ⋆ Sponsors industriels : Purch, Viseo, Kelkoo, Xerox, Coffreo

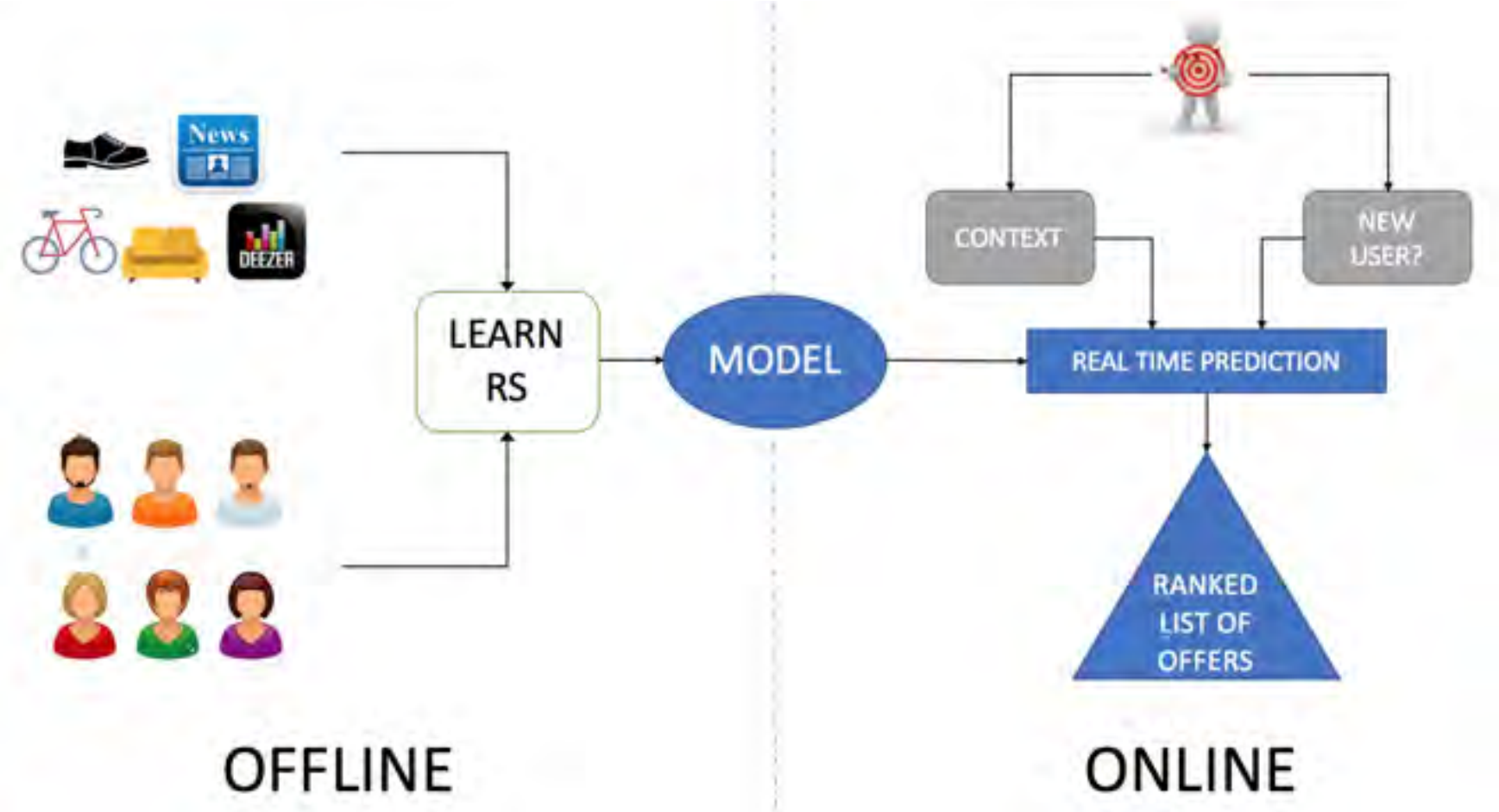
Des données non structurées en abondance

- D'après les prévisions du projet EMC, en 2020 il y aura 40 zetta octets (40×10^{21} octets) de données non-structurées sur la Toile.
- Ces données sont considérées comme le pétrole du XXI^{e} siècle.
- Nécessité de développer de nouveaux outils automatiques pour la recherche et accès à l'information.

Systèmes de recommendation

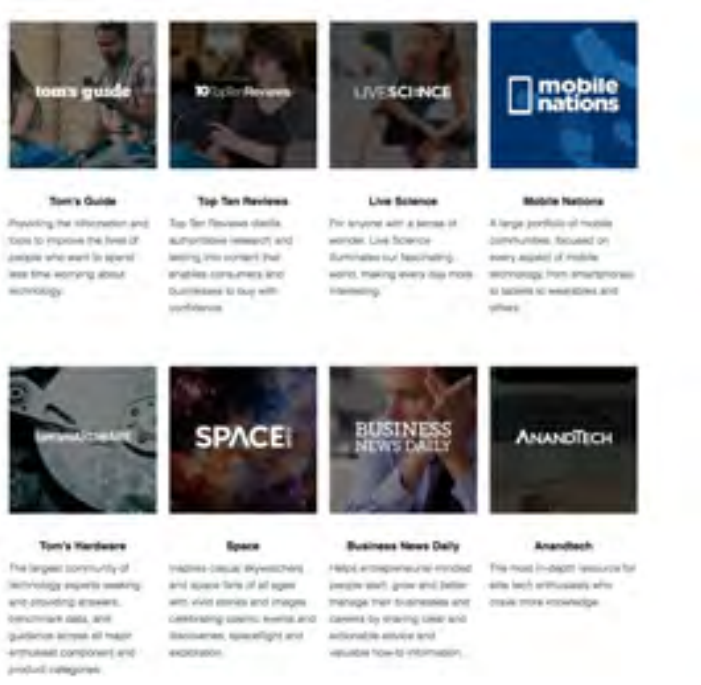
- Un moyen efficace d'analyser l'appétence des utilisateurs
- Bénéfices du côté utilisateurs ...
  - ⋆ faire face à un nombre d'options non évaluable par l'humain
  - ⋆ exemple: Amazon movie propose une sélection de 20 000 films
- et du côté des Industriels
  - ⋆ 60% des films regardés sur Netflix sont des films recommandés
  - ⋆ 35% des ventes sur Amazon sont grâce à la recommandation
  - ⋆ 38% des clicks sur Google sont générés sur des produits recommandés

Systèmes de recommandation: work-flow



Calypso: les partenaires

- **Kelkoo** et **Purch**: deux acteurs du marché de la publicité en ligne



- **LIG**: Massih-Reza Amini, Sumit Sidana, et al.
- **Objectif**: développer une nouvelle génération de système de recommandation pour la publicité en ligne

Les grands défis

- Apprendre à partir de retours implicites: comment inférer la notion de préférence à partir de clics?
- Modèle de recommandation adaptatif: comment prendre en compte les retours des utilisateurs de manière dynamique?
- Recommandation pour de nouveaux utilisateurs/items: que recommender à des nouveaux utilisateurs? A qui recommender de nouveaux items?
- Passage à l'échelle pour la recommendation en ligne

Valorisation des données

Objectifs

- Définir une architecture des données pour le développement d'un modèle
- Extraction des données sur une période définie pour les tests off-line
- Partage des données pour la communauté scientifique

KASANDR (Kelkoo lArge ScAle juNe Data for Recommendation)

- Article publié dans le cadre de SIGIR'17
- Données accessibles sur UCI
- 1 mois de données sur 20 pays

Spécificité des données (1/2)

- Données stockés dans 4 bases: Click, Offer, PageView, Search
- Données contextuelles
  - ⋆ **Utilisateurs**: géographiques
  - ⋆ **Items**: mots-clés, catégorie, prix, titre, marchand
  - ⋆ **Intéraction**: requête saisie, timestamp, filtre
- Quelques chiffres

# of users	# of unique offers	# of offers shown	# of clicks
123,529,420	56,667,919	3,210,050,267	16,107,227

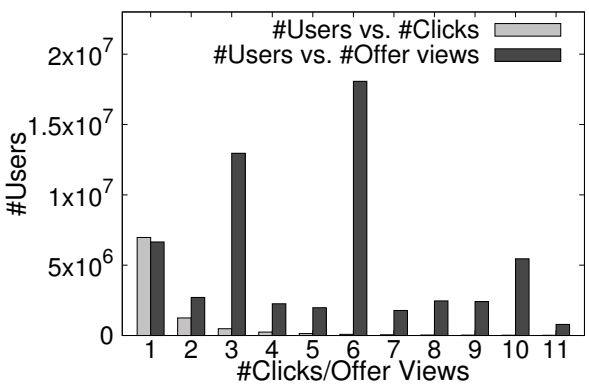
Sparsity	99.9999997848%
Average # of Offers Shown to 1 user	26
Maximum # of clicks done by 1 user	3,722
Minimum # of clicks done by 1 user	0
Average # of clicks done by 1 user	0.13
Average # of clicks done by 1 user (if user did at least one click)	1.71

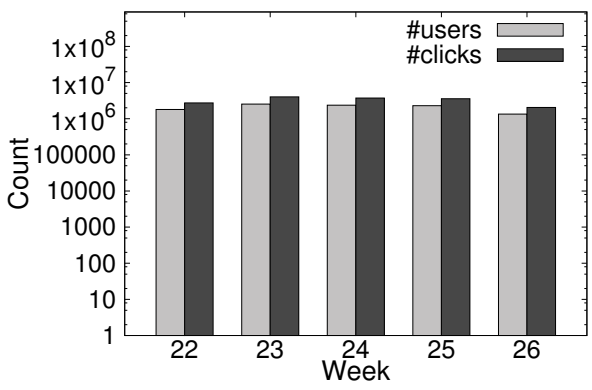
Spécificité des données (2/2)

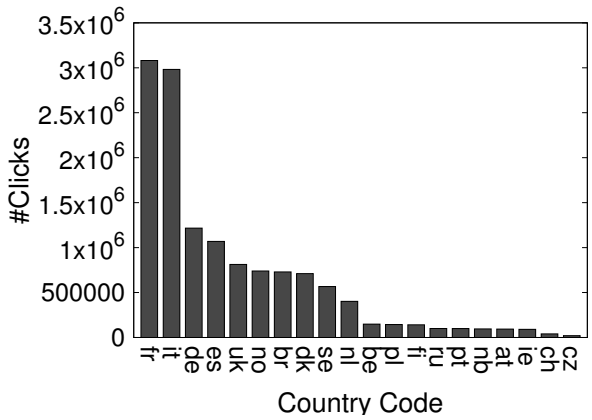
- Problème du cold-start: 200 fois plus de nouveaux utilisateurs que d'anciens

Week Number	# New Users	# Returning Users
23	36,932,009	165,951
24	26,736,201	199,467
25	22,358,876	185,749
26	13,908,242	135,303

- Quelques figures







SOTA sur KASANDR

Approches testées

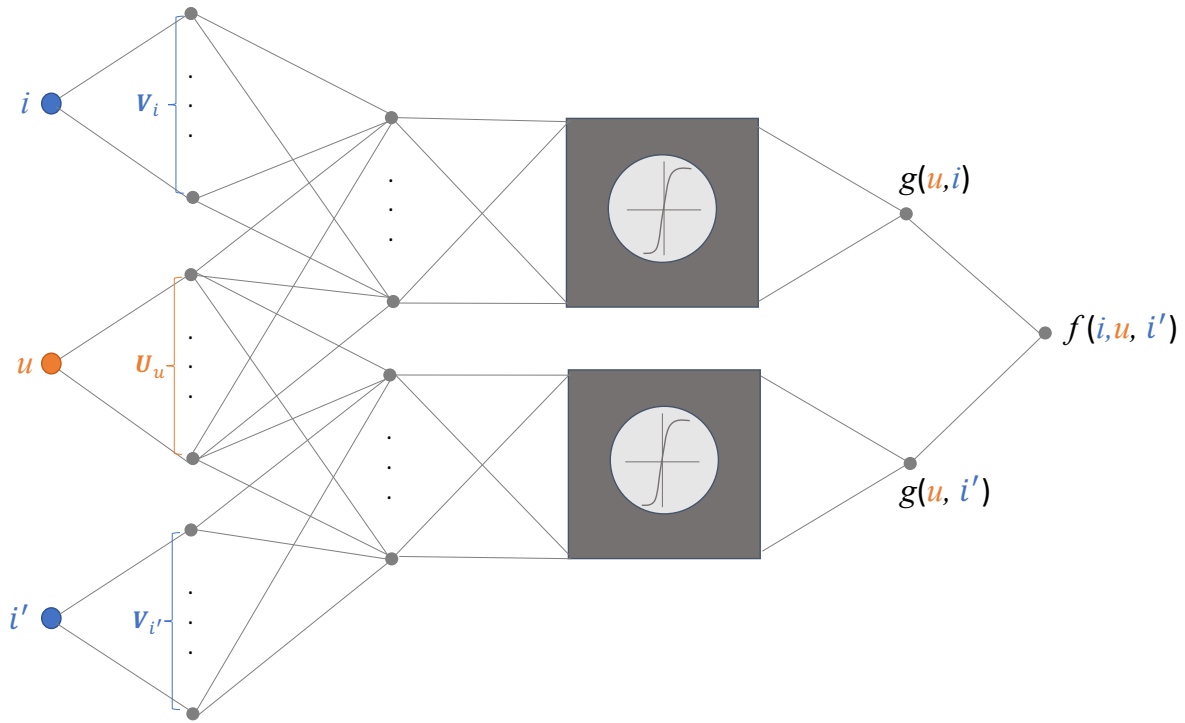
- Non-machine learning: popularité, intéractions passées et aléatoire
- Matrix Factorization (MF)
- Factorization Machines (FM)
- Field-Aware Factorization Machines (FFM)
- Field-Aware Factorization Machines-F

Table: Comparison between all tested methods in terms of Micro and Macro MAP. The best results are in bold.

	Rand		Pop		PastI		MF		FM		FFM		FFM-F	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
MAP@5	2.41E-6	1.54E-005	0.004	0.004	0.017	0.011	0.044	0.037	0.721	0.814	0.732	0.829	**0.760**	**0.861**
MAP@30	4.25E-6	2.33E-005	0.004	0.005	0.017	0.011	0.044	0.037	0.726	0.817	0.736	0.831	**0.764**	**0.862**
MAP@100	5.64E-6	2.996E-005	0.005	0.005	0.016	0.011	0.044	0.037	0.726	0.817	0.735	0.831	**0.763**	**0.862**

Solution proposée par le LIG: RECNET

- Un cadre unifié pour l'apprentissage de représentation et d'ordonnancement
- Repose sur une représentation type one-hot encoding et word2vec
- Optimisation d'une fonction d'ordonnancement par paire



Premiers résultats

- Performances de RecNet sur trois benchmarks
  - ⋆ **RECNET (0,1)** Fonction à minimiser sur la qualité de la représentation
  - ⋆ **RECNET (1,0)** Fonction à minimiser sur la qualité de la sortie du réseau
  - ⋆ **RECNET (1,1)** Combinaison linéaire des fonctions précédentes

	ML-100K			ML-1M			KASANDR		
	MAP@1	MAP@5	MAP@10	MAP@1	MAP@5	MAP@10	MAP@1	MAP@5	MAP@10
BPR-MF	0.753	0.759	0.727	**0.887**	**0.876**	**0.845**	0.159	0.166	0.167
RECNET (0,1)	0.813	**0.856**	**0.863**	0.852	0.842	0.818	**0.951**	**0.959**	**0.958**
RECNET (1,0)	0.814	0.840	0.843	0.869	0.861	0.838	0.899	0.908	0.911
RECNET (1,1)	0.787	0.832	0.836	0.873	0.861	0.838	0.947	0.950	0.953
LightFM	**0.861**	0.815	0.790	0.833	0.826	0.796	0.938	0.938	0.936
CoFactor	0.801	0.781	0.753	0.806	0.806	0.776	0.916	0.912	0.914
Popularity	0.633	0.594	0.553	0.571	0.586	0.553	0.037	0.036	0.038

Travaux en cours

Industriels

- Kelkoo: implémentation de FFM (A/B testing)
  - ⋆ Industrialisation des codes
  - ⋆ Feature engineering
  - ⋆ Tuning des hyper-paramètres et des fenêtres temporelles
  - ⋆ Un peu de tests à la main!
- Purch: récupération et publication de données stables
  - ⋆ Publication sur le même modèle que KASANDR
  - ⋆ Modèles prenant en compte les spécificités de Purch

Optimisation

- Collaboration avec Aleksandra Burashnikova et Marianne Clausel
- RECNET repose sur une optimisation de type SGD (RMSProp)
- Optimisation adaptative et efficace
- Modélise l'arrivée des utilisateurs progressivement
- Modélise l'arrivée des intéraction progressivement

Travaux futurs

Purch

Organisation d'un data challenge sur une extraction de données

Sélection du bon ensemble d'items candidats

- Impossible de calculer toutes les paires de score (users x items)
- Catégorisation des produits
- Maintenir des critères telles que la nouveauté et la diversité

SOTA

- Travaux sur les calculs de similarité entre items
- Travaux sur les approches de type word-embedding pour le clustering d'items (Criteo)

Liens utiles

- Action ATLAS:
  `http://ama.liglab.fr/ATLAS/index.php?title=Main_Page`
- KASANDR Germany :
  `https://archive.ics.uci.edu/ml/datasets/KASANDR`
  - ⋆ 17,764,280 d'utilisateurs, 2,158,859 offres
  - ⋆ Fichier train et test au format csv
- KASANDR `http://ama.liglab.fr/kasandr/`: en cours