Fouille et classement d'ensembles fermés dans des données transactionnelles de grande échelle

Thèse préparée au sein de l'équipe SLIDE du LIG
pour le titre de Docteur de l'Université de Grenoble, présentée par

Martin Kirchgessner

Sous la direction de Sihem Amer-Yahia et Vincent Leroy

Le projet Datalyse

Collaboration avec le Groupe Les Mousquetaires, acteur majeur de la grande distribution en France et en Europe.

Objectif : étudier les habitudes d'achat.

Méthode : fouille d'associations entre items.

Exemples :

$$\text{Algues nori, Wasabi, Sauce Soja} \rightarrow \quad \text{Riz à Suhis}$$

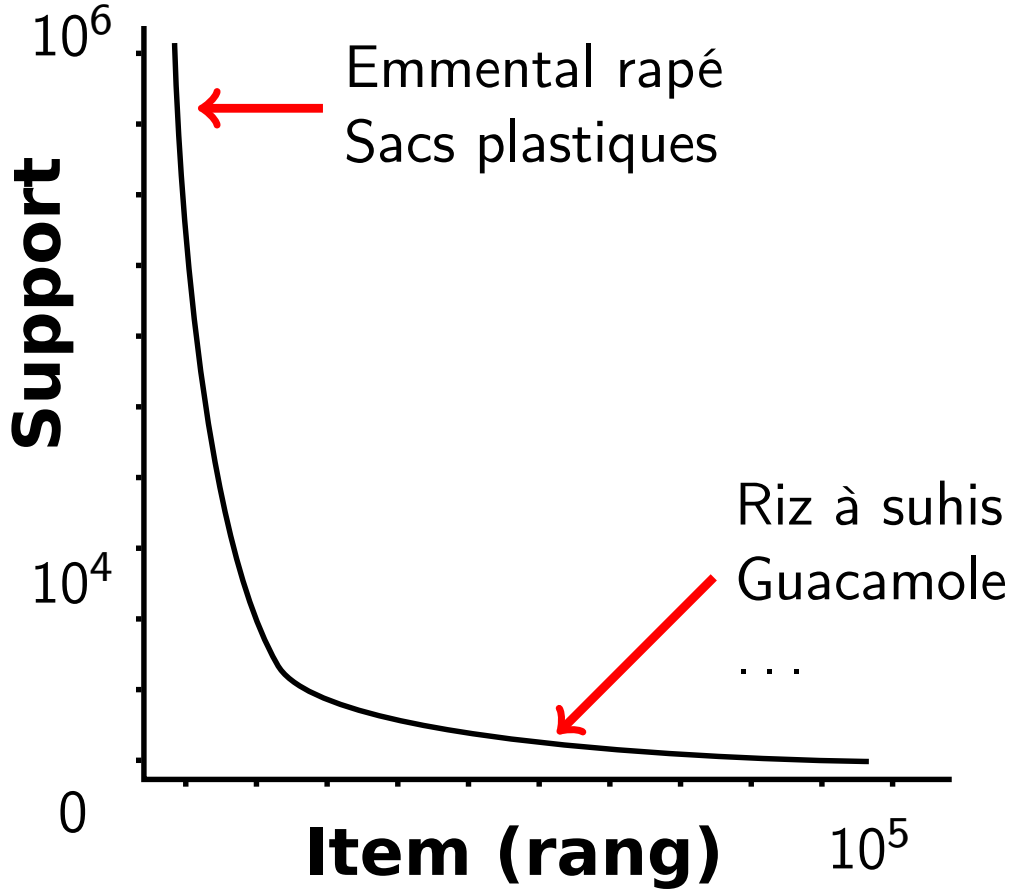
$$\textit{Nord}, < 35\ \textit{ans}, \textit{Homme} \rightarrow \quad \textit{Sodas}$$

"de grande échelle"

Sur l'année 2013:

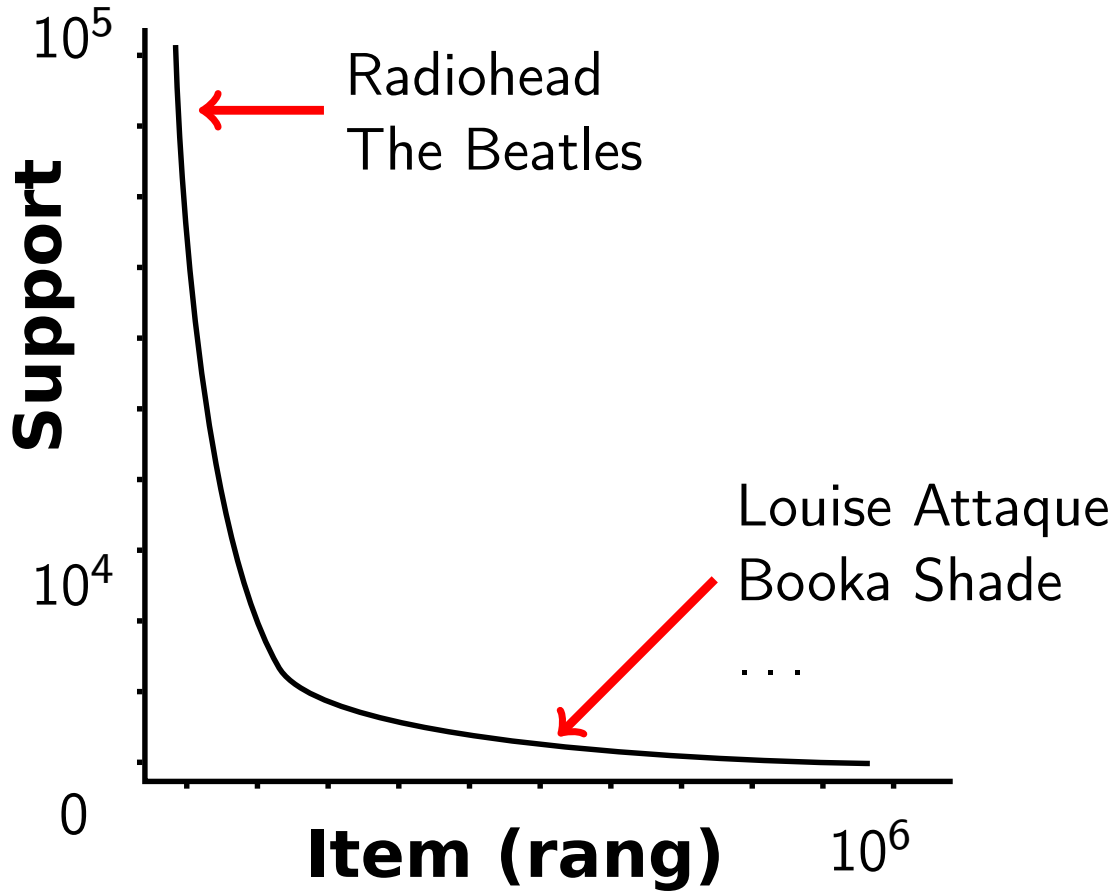
- 290,734,163 tickets
- 222,228 produits
- 9,267,961 clients
- 1884 magasins à travers la France

Distribution en longue traine



Support(*item*) = Nombre de transactions contenant cet item.

Distribution en longue traine



Anatomy of the long tail: ordinary people with extraordinary tastes,
Goel, Broder, Gabrilovich, Pang @ WSDM'10

Problematiques de la recherche d'associations significatives

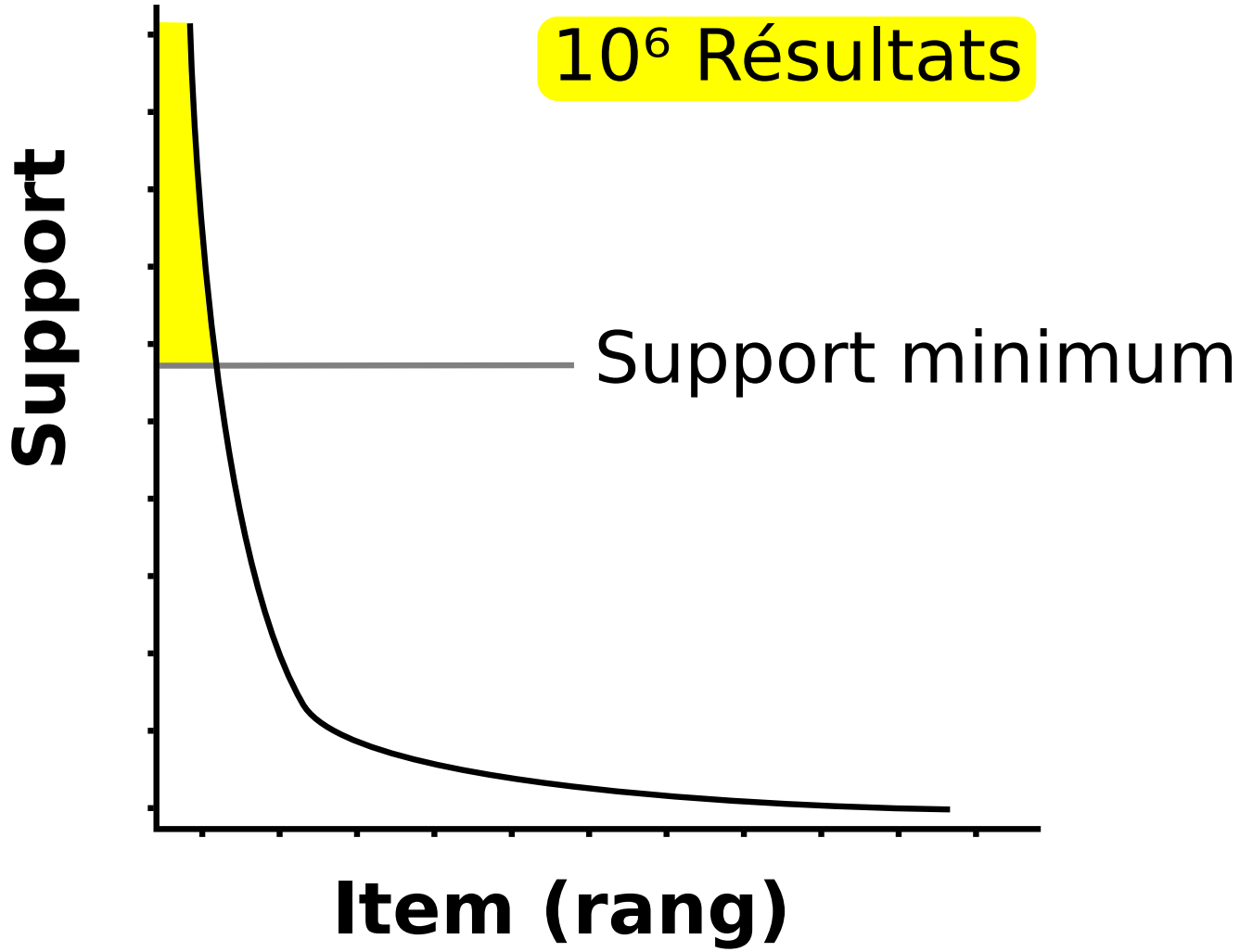
- Couverture
  - Etudier les associations concernant *n'importe quel* item
- Passage à l'échelle
  - Analyser des millions de transactions
- Qualité
  - Indiquer les associations remarquables

Contributions

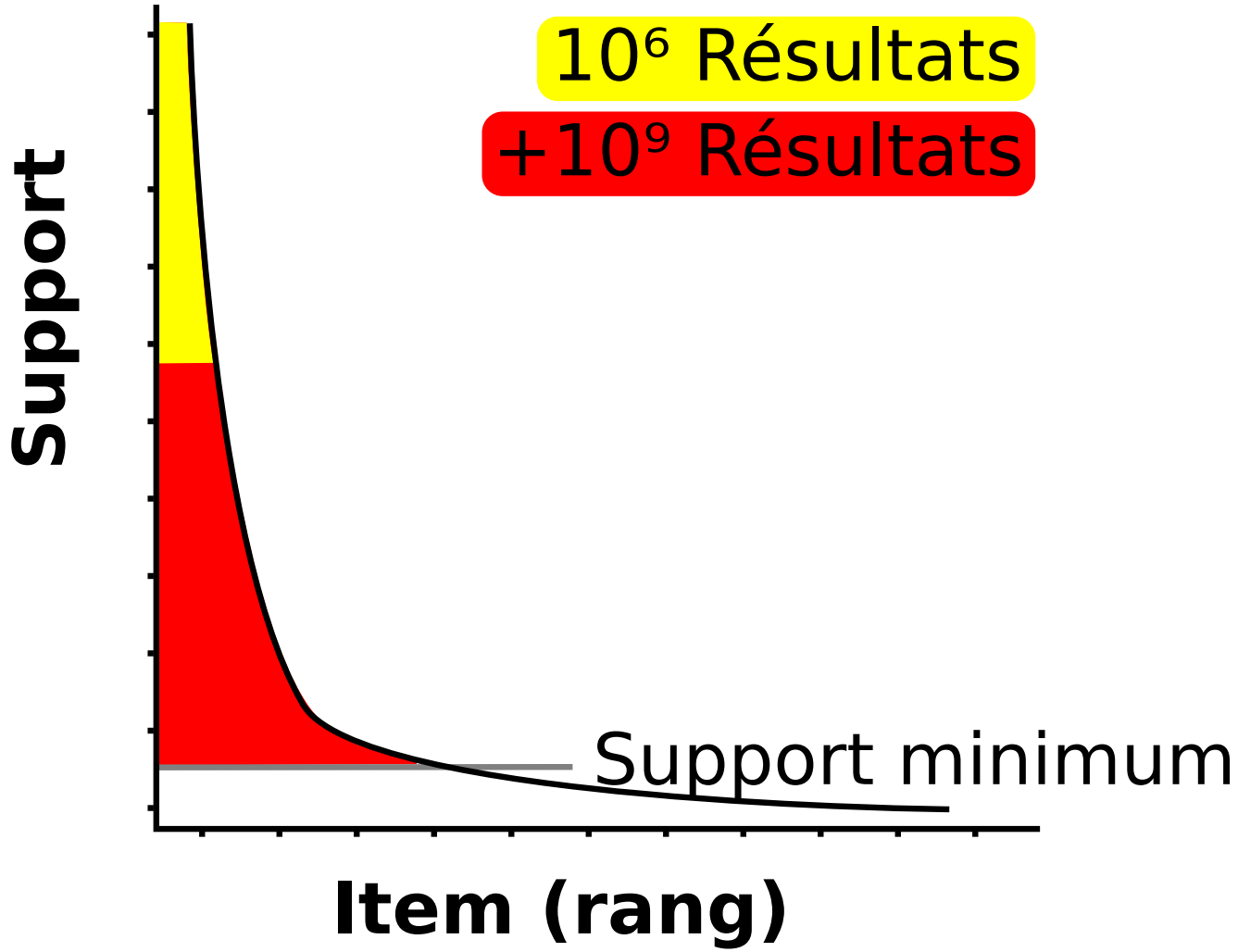
- Couverture et passage à l'échelle
  - Fouille item-centrée avec TopPI
- Qualité
  - Comparaison des mesures de qualité avec CAPA

Fouille item-centrée avec TopPI

Itemsets fréquents et longue traine



Itemsets fréquents et longue traine



Itemsets fréquents et longue traine

Support(P)	P
861 304	Emmental rapé, Crème fraîche 30%
793 310	Emmental rapé, 10 Oeufs
747 539	Sucre en poudre, Farine
652 493	Emmental rapé, Beurre
616 696	Sucre en morceaux, Sucre en poudre
597 144	Emmental rapé, Lardons fumés
549 742	Emmental rapé, Jambon
542 979	Emmental rapé, Sucre en morceaux
508 593	Emmental rapé, Soda 1.5L
481 942	Emmental rapé, Huile de tournesol
	...

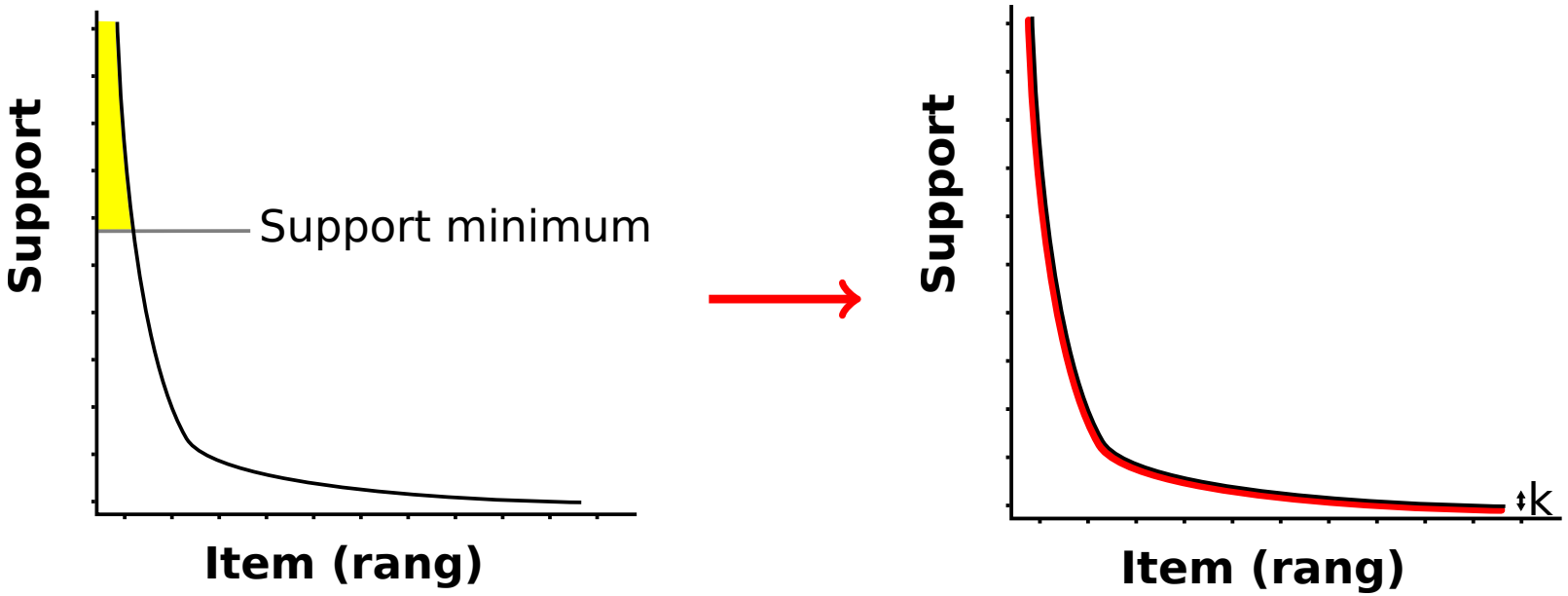
Fouille item-centrée avec TopPI

1. Fouille item-centrée ?
2. Etat de l'art
3. L'algorithme
4. Expériences
5. Distribuer TopPI sur MapReduce

TopPI: An Efficient Algorithm for Item-Centric Mining.
Kirchgessner, Leroy, Termier, Amer-Yahia, Rousset @ DaWaK'16 p.19-33

TopPI: An Efficient Algorithm for Item-Centric Mining.
Leroy, Kirchgessner, Termier, Amer-Yahia — à paraître dans *Information Systems*.

Fouille item-centrée



Une nouvelle sémantique pour la fouille d'itemsets

t_1 : *Camembert, Saucisses, Compté, Chèvre long, Salade mélangée, Sucre en morceaux*
t_2 : *Brandade de morue, Boite 6 oeufs, Désodorisant, Lessive, Emmental rapé*
t_3 : *Haricot vert, Boxer, Nettoyant à moquette, Mouchoirs, Salade mélangée*
...

Une nouvelle sémantique pour la fouille d'itemsets

t_1 : *Camembert, Saucisses, Compté, Chèvre long, Salade mélangée, Sucre en morceaux*
t_2 : *Brandade de morue, Boite 6 oeufs, Désodorisant, Lessive, Emmental rapé*
t_3 : *Haricot vert, Boxer, Nettoyant à moquette, Mouchoirs, Salade mélangée*
...
Résultats :

top(Emmental rapé)

Support	Itemset
9 395 643	Emmental rapé
861 304	Emmental rapé, Crème fraîche
793 310	Emmental rapé, 10 Oeufs
652 493	Emmental rapé, Beurre
597 144	Emmental rapé, Lardons fumés

top(Crème choc.)

Support	Itemset
581042	Crème choc.
58569	Crème choc., Crème à la vanille
32701	Crème choc., Emmental rapé 200g
30451	Crème choc., Cola 1.5L
29671	Crème choc., Beurre doux

top(Riz à sushis)

Support	Itemset
14887	Riz à sushis
5935	Riz à sushis, Algues nori
3669	Riz à sushis, Vinaigre de riz
1843	Riz à sushis, Algues nori, Vinaigre de riz
1762	Riz à sushis, Wasabi

...

Méthodes existantes

Méthodes existantes

Par post-processing

1. Fouiller tous les itemsets fréquents ($minsup = 2$),
2. Insérer chaque itemset dans les $top(i)$ concernés.

Méthodes existantes

Par post-processing

1. Fouiller tous les itemsets fréquents ($minsup = 2$),
2. Insérer chaque itemset dans les $top(i)$ concernés.

Par pre-processing (méthode de référence)

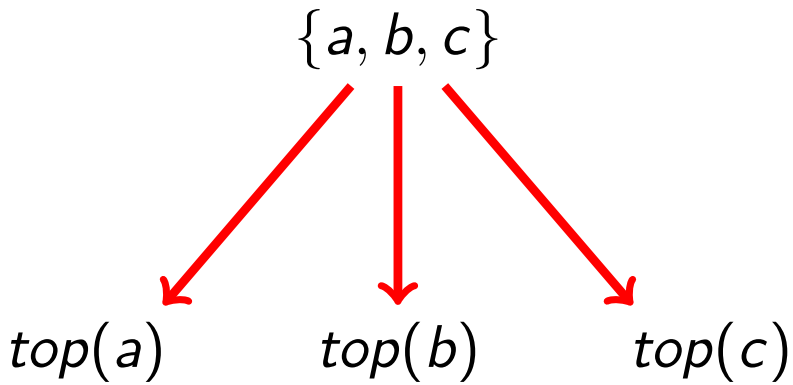
Pour chaque item i:

1. Instanciation de $\mathcal{D}[i] = \{t \in \mathcal{D} | i \in t\}$
2. Exécution de TFP sur $\mathcal{D}[i]$, qui produit directement $top(i)$.

Mining top-k frequent closed patterns without minimum support.
Han, Wang, Lu, Tzvetkov @ ICDM'02

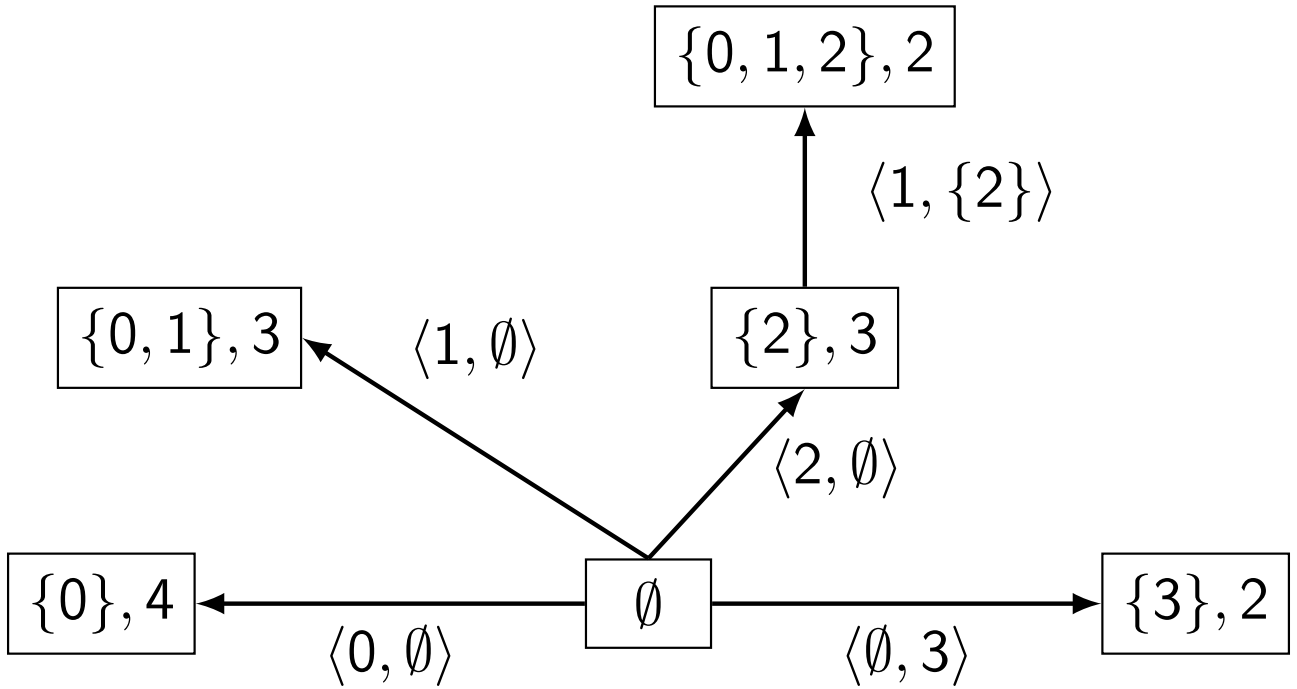
Notre approche : TopPI

1. Obtenir tous les $top(i)$ en une exécution



Notre approche : TopPI

1. Obtenir tous les $top(i)$ en une exécution
2. Un parcours intelligent du treillis des itemsets

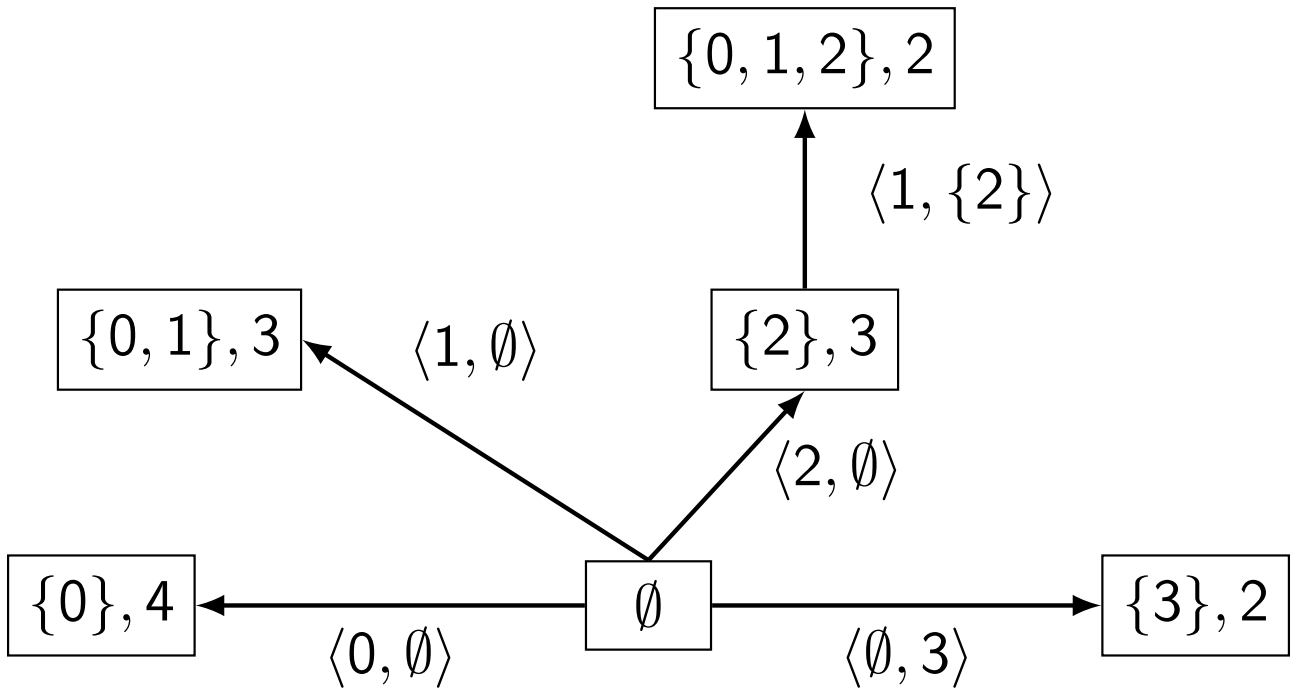


LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets.

Uno, Kiyomi, Arimura @ FIMI'04

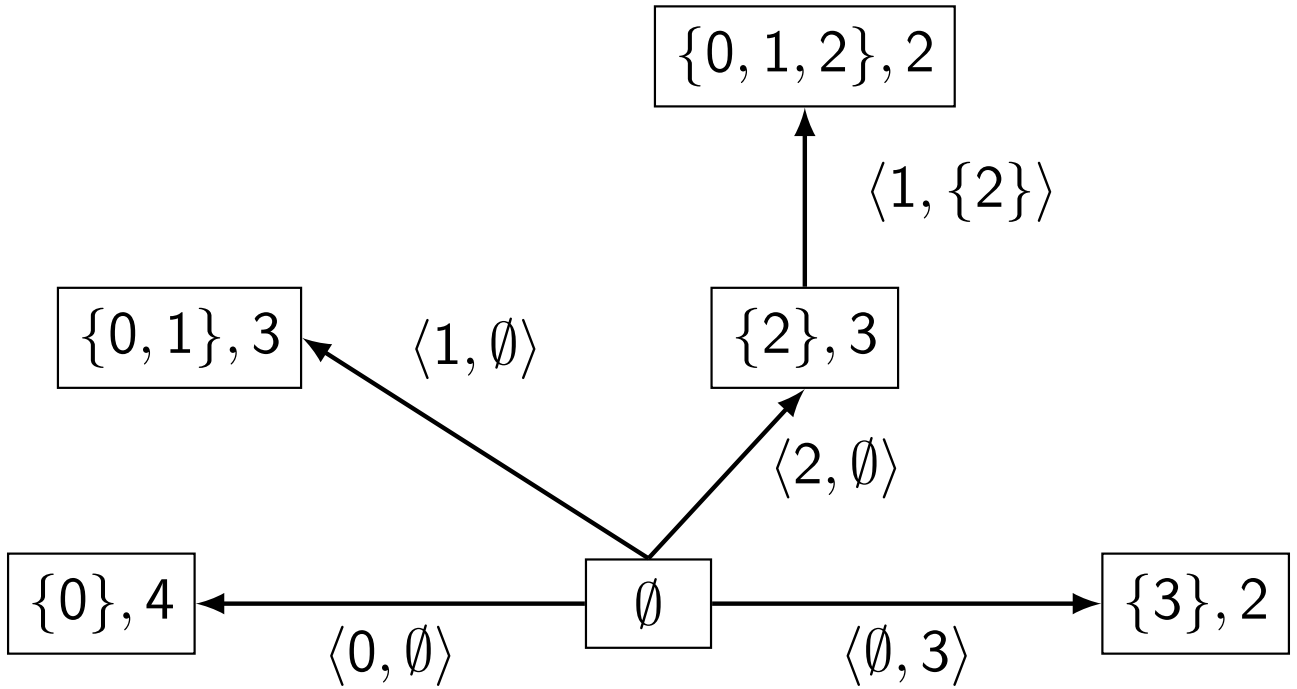
Notre approche : TopPI

1. Obtenir tous les $top(i)$ en une exécution
2. Un parcours intelligent du treillis des itemsets
3. Limite la fouille aux itemsets potentiellement dans un $top(i)$



Notre approche : TopPI

1. Obtenir tous les $top(i)$ en une exécution
2. Un parcours intelligent du treillis des itemsets
3. Limite la fouille aux itemsets potentiellement dans un $top(i)$
4. Répartition des branches entre les threads



Discovering closed frequent itemsets on multicore.

Négrevergne, Termier, Méhaut, Uno @ HPCS'10

TopPI, un algorithme rapide

Temps d'exécution ($k = 50$)

Données	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{D}\rvert$	**Taille**	**Temps**
Tickets	$222,228$	$290,734,163$	24GB	4 min.
Tickets, par client	$222,228$	$9,267,961$	13.3GB	11 min.
LastFM	$1,206,195$	$1,218,831$	277MB	2 min.
WebDocs	$5,267,656$	$1,692,082$	1.4GB	8 heures

Environnement expérimental :

- 32 threads en parallèle
- 128 GB de RAM

TopPI, distribué sur cluster MapReduce

Plus de CPUs pour l'énumération d'itemsets.

TopPI, distribué sur cluster MapReduce

Plus de CPUs pour l'énumération d'itemsets.

Pas de calculs redondants

- L'ensemble des items $\mathcal{I}$ est partitionné en *groupes* : un par machine.
- Chaque machine produit une partition des résultats

TopPI, distribué sur cluster MapReduce

Plus de CPUs pour l'énumération d'itemsets.

Pas de calculs redondants

- L'ensemble des items $\mathcal{I}$ est partitionné en *groupes* : un par machine.
- Chaque machine produit une partition des résultats

Pas de calculs inutiles

- Complétion des $top(i)$ en deux étapes
- Conserve l'efficacité de l'élagage

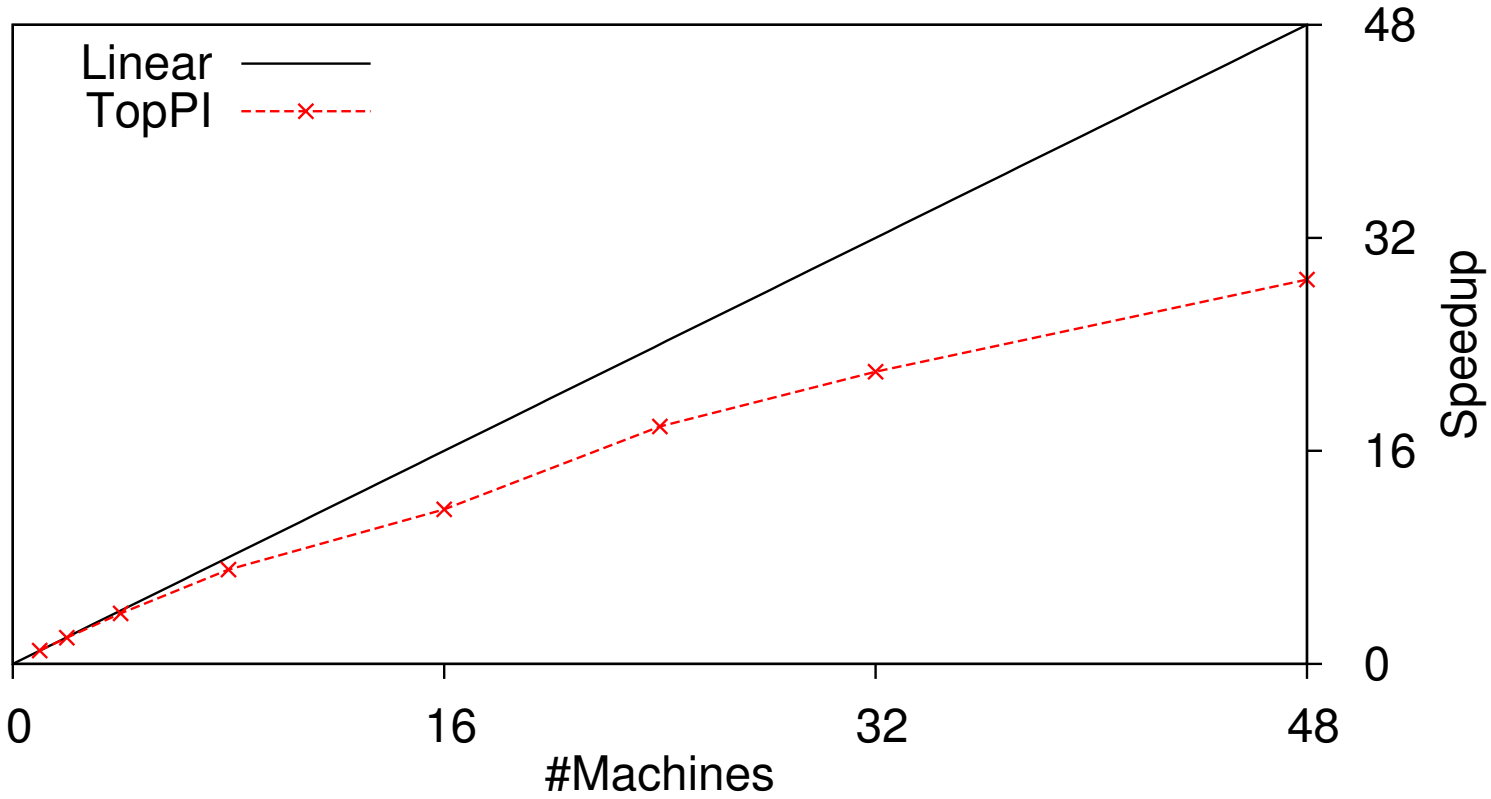
TopPI sur cluster MapReduce

Sur le cluster “edel” de Grid5000.

WebDocs, $k = 10$:

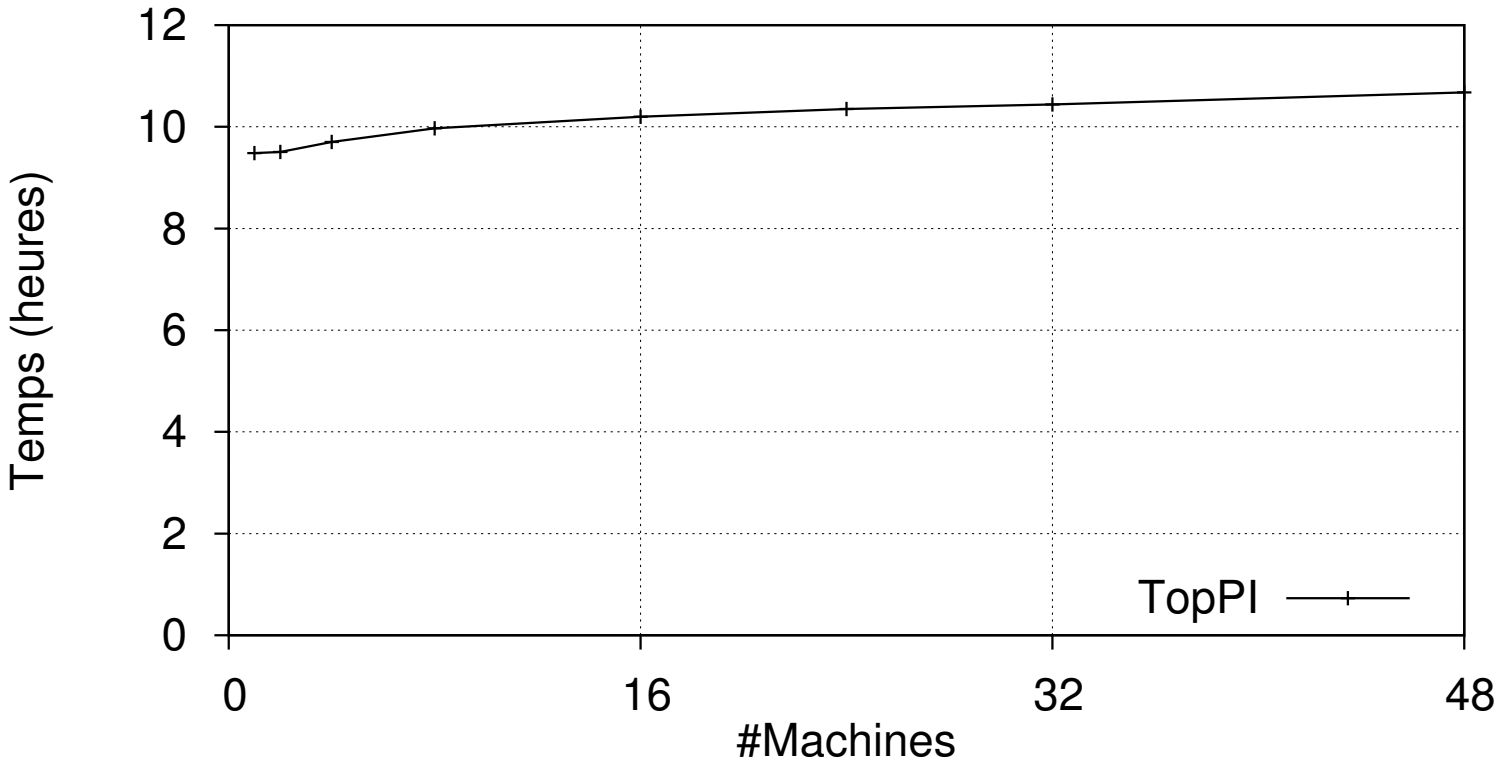
- 4570s avec 32 machines (8 threads chacune)
- 2641s avec 64

TopPI sur cluster MapReduce



Supermarket, $k = 1000$.

TopPI sur cluster MapReduce



Supermarket, $k = 1000$, temps CPU cumulé passé à la fouille

TopPI

- Une sémantique adaptée aux longues traines
  - Un (unique) paramètre, k
  - Résultats complets et organisés intuitivement (par item)
  - Applicable dans différents domaines (Web, grande distribution, ...)
- Un algorithme qui passe à l'échelle
  - Sur serveur multi-coeurs
  - Sur cluster MapReduce
  - Elaguage très efficace de l'espace des solutions
- Utilisation industrielle
  - `https://github.com/slide-lig/TopPI`

CAPA : Comparative Analysis of PAtterns

Le bruit du tri par fréquence

Sur 290 millions de tickets :

k	$support(P)$	P
1	581042	Crème au chocolat
2	58569	Crème au chocolat, Crème à la vanille
3	32701	Crème au chocolat, Emmental rapé 200g
4	30451	Crème au chocolat, Cola 1.5L
5	29671	Crème au chocolat, Beurre doux
6	29376	Crème au chocolat, Emmental rapé 3x70g
7	24869	Crème au chocolat, Emmental rapé 200g Marque B
8	23032	Crème au chocolat, Lait 1/2 écrémé
9	19929	Crème au chocolat, Lait 6x1L
10	16547	Crème au chocolat, Pâte feuilletée

Trier des règles d'association

39 mesures de qualité dans la littérature.

- Sont-elles vraiment différentes ?
- Laquelle choisir ? Pour la grande distribution ?
- Comment simplifier l'évaluation par des experts ?

Interestingness Measures for Data Mining: A Survey.
Geng, Hamilton, *ACM Computer Surveys*, 2006

Association Rule Interestingness Measures: Experimental and Theoretical Studies.
Lenca, Vaillant, Meyer, Lallich, *Quality Measures in Data Mining*, 2007

Beyond Support and Confidence: Exploring Interestingness Measures for Rule-Based Specification Mining. Le, Lo @ SANER'15

La méthode CAPA

Comparative Analysis of PAtterns

1. Comparaison automatique des mesures
   $\rightarrow$ distingue des familles de mesures donnant des classements similaires.
2. Validation empirique, où les chargées d'étude marketing comparent les familles de classements.

Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns.
Kirchgessner, Leroy, Amer-Yahia, Mishra @ DSAA'16

Comparaison automatique des mesures

1. Définition de cibles d'étude

Contraintes sur les ensembles $A \rightarrow B$

Comparaison automatique des mesures

1. Définition de cibles d'étude
2. Fouille des associations correspondantes

$A \rightarrow B$	$support(A)$	$support(A \cup B)$
$\{35-49\} \rightarrow$ *Patisserie indus.*	66 811 806	22 270 602
$\{35-49\} \rightarrow$ *Boissons*	66 811 806	16 513 795
$\{35-49, F\} \rightarrow$ *Patisserie indus.*	45 267 831	15 089 277
$\{F\} \rightarrow$ *Epicerie sucrée*	205 330 640	112 931 852
$\{> 64, F,$ Rhône-Alpes $\} \rightarrow$ *Crèmerie*	6 649 289	4 255 545
...		

Comparaison automatique des mesures

1. Définition de cibles d'étude
2. Fouille des associations correspondantes
3. Classement d'après les 39 mesures

Confiance
$\{> 65, F, \text{Aube}\} \rightarrow \textit{Dairy}$
$\{> 65, F, \text{Aveyron}\} \rightarrow \textit{Dairy}$
$\{> 65, F, \text{Val de Marne}\} \rightarrow \textit{Dairy}$
$\{> 65, F, \text{Seine S}^t \text{ Denis}\} \rightarrow \textit{Dairy}$
$\{> 65, F, \text{Haute Saone}\} \rightarrow \textit{Dairy}$
$\{> 65, F, \text{Meuse}\} \rightarrow \textit{Dairy}$
$\{> 65, *, \text{Aube}\} \rightarrow \textit{Dairy}$
$\{> 65, F, \text{Haute Vienne}\} \rightarrow \textit{Dairy}$
$\{> 65, F, \text{Maine et Loire}\} \rightarrow \textit{Dairy}$
$\{> 65, *, \text{Val de Marne}\} \rightarrow \textit{Dairy}$

Piatetsky-Shapiro
$\{*, *, \text{Nord}\} \rightarrow \textit{Liquids}$
$\{*, *, \text{Nord}\} \rightarrow \textit{Soft drinks}$
$\{*, *, \text{Nord}\} \rightarrow \textit{Beers}$
$\{*, *, \text{Nord}\} \rightarrow \textit{Spreads}$
$\{*, F, \text{Nord}\} \rightarrow \textit{Soft drinks}$
$\{*, *, \text{Nord}\} \rightarrow \textit{Imported beers}$
$\{*, F, \text{Nord}\} \rightarrow \textit{Liquids}$
$\{*, F, \text{Nord}\} \rightarrow \textit{Beers}$
$\{*, *, \text{Finistere}\} \rightarrow \textit{Butters}$
$\{*, F, \text{Garonne}\} \rightarrow \textit{Drugstore}$

$\{*, *,$
$\{*, F,$
$\{> 65, *,$ M
$\{> 65, *,$
$\{*, *,$
$\{*, *,$ Cotes
$\{> 65, F,$ M
$\{*$
$\{*$
$\{*, *$

Comparaison automatique des mesures

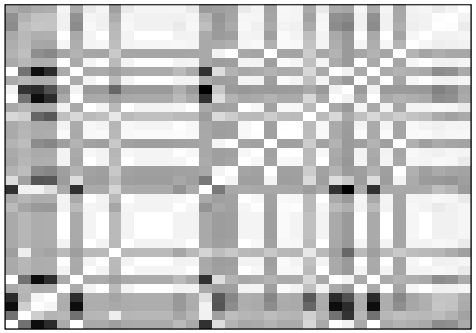
1. Définition de cibles d'étude
2. Fouille des associations correspondantes
3. Classement d'après les 39 mesures
4. Comparaison des classements avec 4 distances
   - Coefficient de Spearman
   - τ de Kendall
   - Overlap@20
   - NDCC : *Normalized Discounted Correlation Coefficient*

NDCC

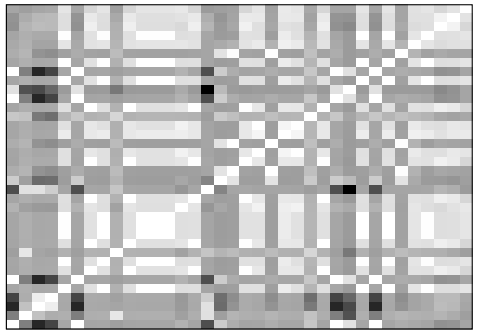
- Calcul inspiré de NDCG
- Favorise les classements similaires en tête

Comparaison automatique des mesures

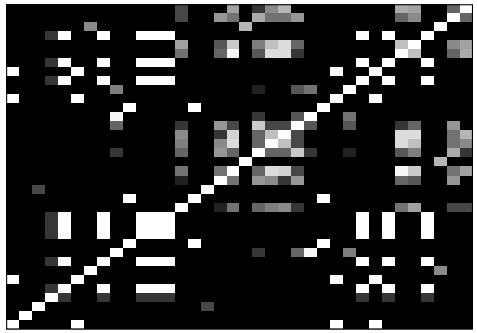
1. Définition de cibles d'étude
2. Fouille des associations correspondantes
3. Classement d'après les 39 mesures
4. Comparaison des classements avec 4 distances



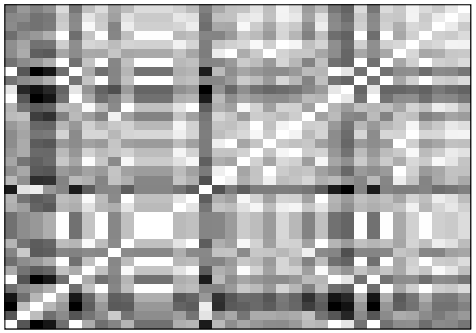
Coefficient de Spearman



τ de Kendall



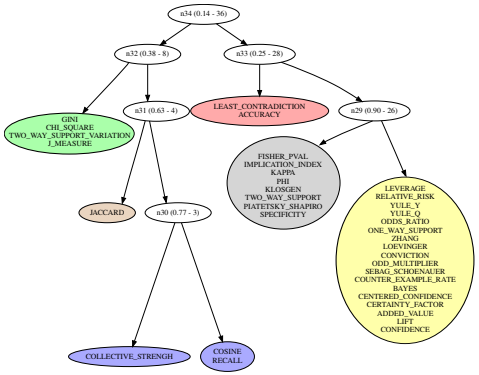
Overlap@20



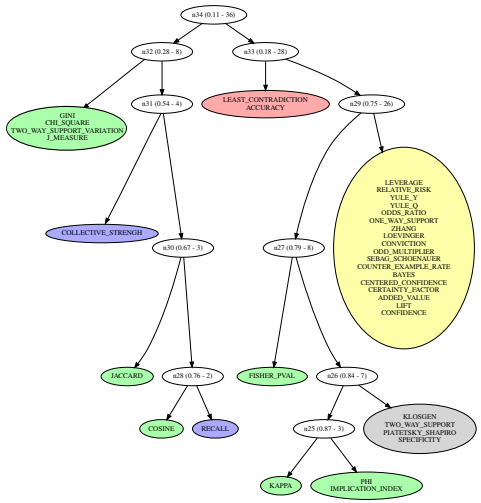
NDCC

Comparaison automatique des mesures

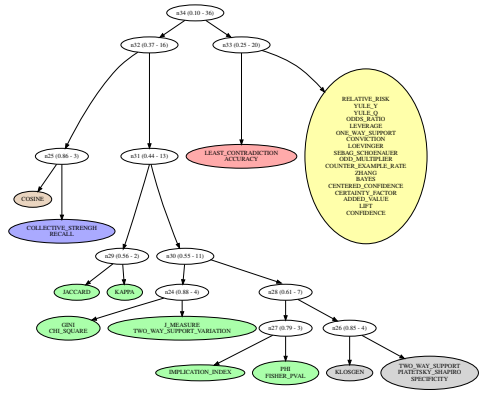
1. Définition de cibles d'étude
2. Fouille des associations correspondantes
3. Classement d'après les 39 mesures
4. Comparaison des classements avec 4 distances
5. Regroupement hiérarchique



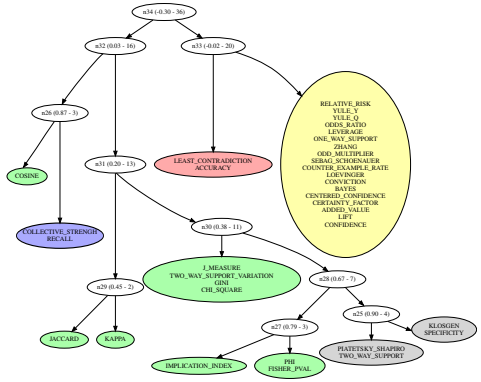
Coefficient de Spearman



τ de Kendall

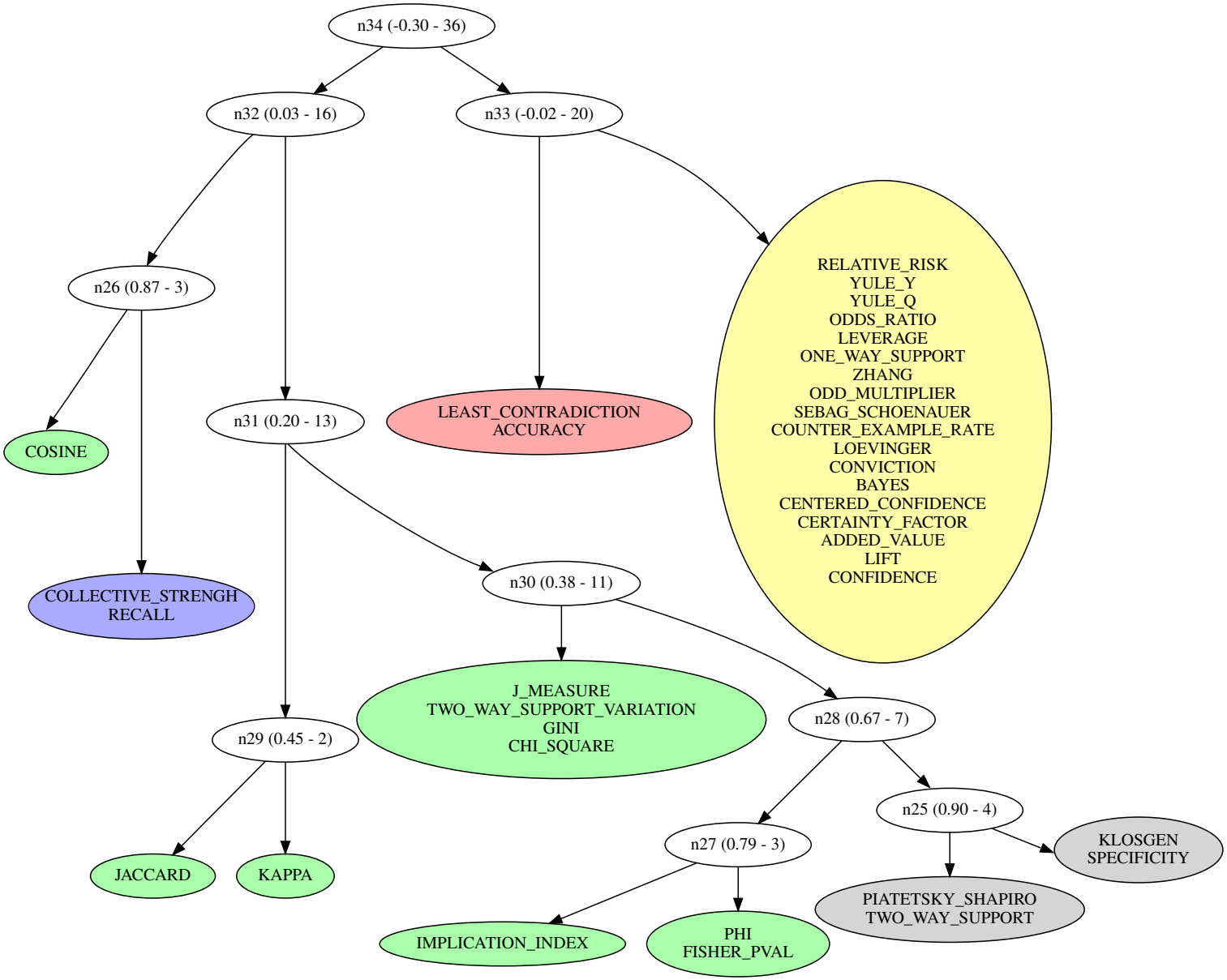


Overlap@20

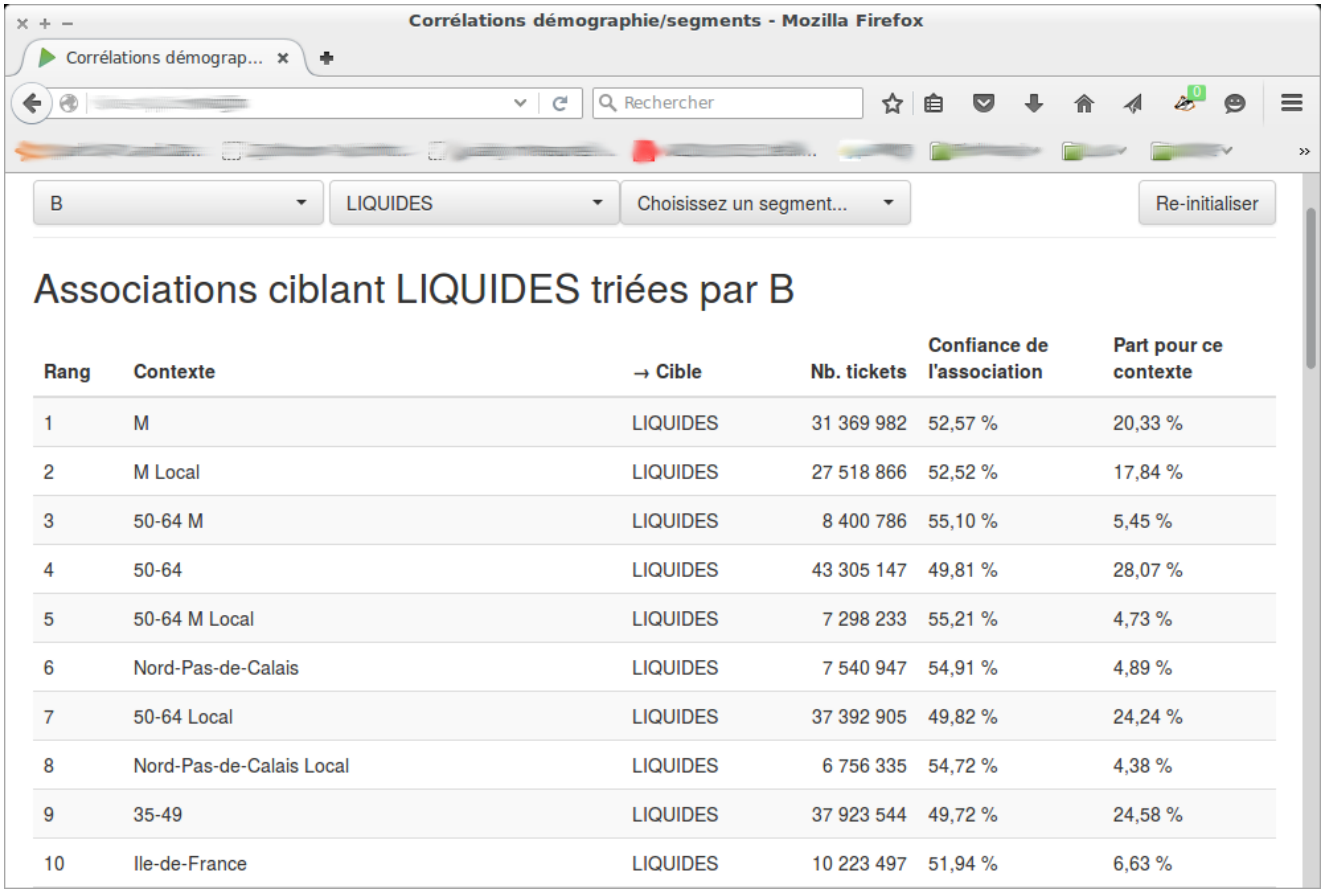


NDCC

Résultat de la comparaison automatique : 5 familles



Etude empirique



Associations ciblant LIQUIDES triées par B

Rang	Contexte	→ Cible	Nb. tickets	Confiance de l'association	Part pour ce contexte
1	M	LIQUIDES	31 369 982	52,57 %	20,33 %
2	M Local	LIQUIDES	27 518 866	52,52 %	17,84 %
3	50-64 M	LIQUIDES	8 400 786	55,10 %	5,45 %
4	50-64	LIQUIDES	43 305 147	49,81 %	28,07 %
5	50-64 M Local	LIQUIDES	7 298 233	55,21 %	4,73 %
6	Nord-Pas-de-Calais	LIQUIDES	7 540 947	54,91 %	4,89 %
7	50-64 Local	LIQUIDES	37 392 905	49,82 %	24,24 %
8	Nord-Pas-de-Calais Local	LIQUIDES	6 756 335	54,72 %	4,38 %
9	35-49	LIQUIDES	37 923 544	49,72 %	24,58 %
10	Ile-de-France	LIQUIDES	10 223 497	51,94 %	6,63 %

Etude empirique

Capacité d'attention

- 10 à 20 premiers résultats
- les derniers, aussi !

Etude empirique

Capacité d'attention

- 10 à 20 premiers résultats
- les derniers, aussi !

Préfèrences :

- Généralement, G_1 et G_2 (mesures favorisant la confiance)
- Pour une cible donnée, G_3 et G_4

L'important est de différencier
{crème vanille, emmental rapé}$\rightarrow$ *crème au chocolat* (conf. 32%)
et
{crème vanille}$\rightarrow$ *crème au chocolat* (conf. 31%)

CAPA : généralisation

- 39 mesures de qualité, mais des classements similaires
- Première étude du genre pour la grande distribution
  - Le classement par confiance reste plebiscité
  - La mesure de Piatetsky-Shapiro retire bien le bruit des “stars”
- Méthode applicable à d’autres domaines

Conclusion

Conclusion

TopPI

- Un paramètre et des résultats intuitifs
- Analyse 300 millions de tickets en quelques minutes
  - 47000 par Agrawal & Srikant en 1994 avec APriori
- Speedup linéaire, version distribuée

CAPA

- Bonne présentation des résultats
- Permet de choisir un post-traitement

TopPI: An Efficient Algorithm for Item-Centric Mining.
Kirchgessner, Leroy, Termier, Amer-Yahia, Rousset @ DaWaK'16 p.19-33

TopPI: An Efficient Algorithm for Item-Centric Mining.
Leroy, Kirchgessner, Termier, Amer-Yahia — à paraître dans *Information Systems*.

Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns.
Kirchgessner, Leroy, Amer-Yahia, Mishra @ DSAA'16

Classement en deux phases pour TopPI

Classement en deux phases pour TopPI

1. Production des top-par-item avec TopPI, $k \in [100; 500]$

$top(i)$	Support
i	1000
a,i	800
i,j	600
a,b,i	500
a,g,i	400
a,g,z	200

Classement en deux phases pour TopPI

1. Production des top-par-item avec TopPI, $k \in [100; 500]$
2. Transformation en règle d'association

$A \rightarrow$	B	$support(A)$	$support(A \cup B)$
a	i	240000	800
j	i	2000	600
a,b	i	220000	500
a,g	i	800	400

Classement en deux phases pour TopPI

1. Production des top-par-item avec TopPI, $k \in [100; 500]$
2. Transformation en règle d'association
3. Re-classement (par confiance ou Piatetsky-Shapiro)

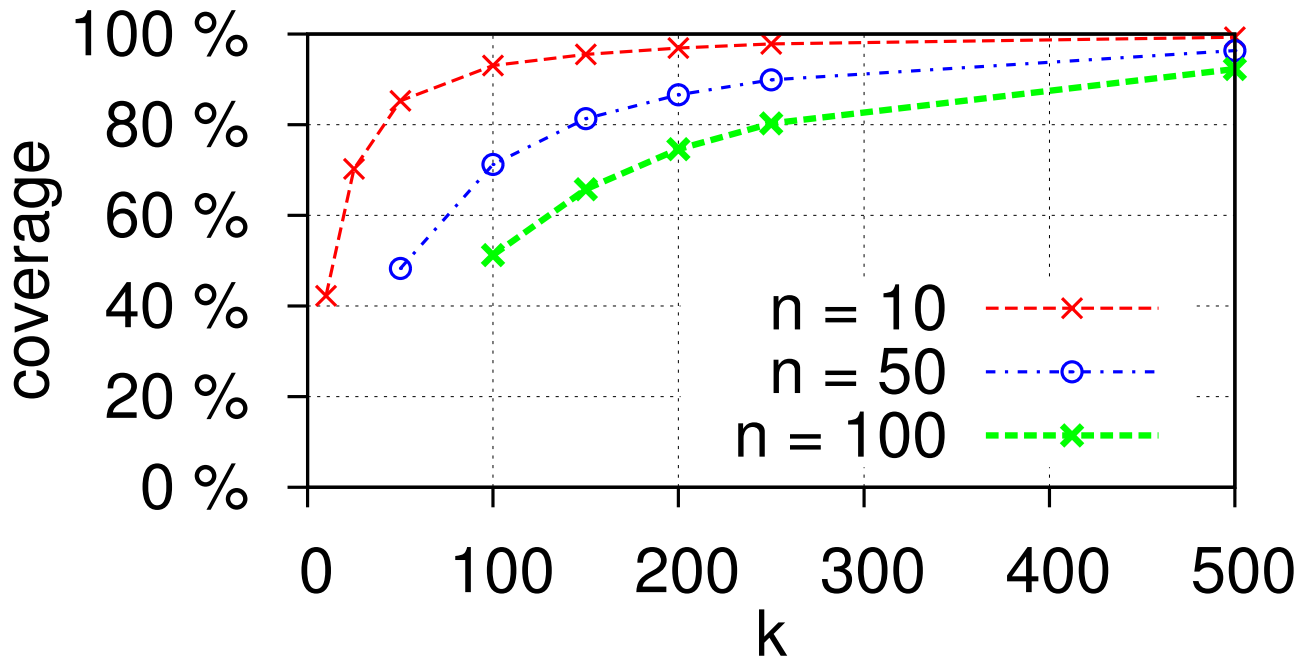
$A \rightarrow$	B	$support(A)$	$support(A \cup B)$	Confiance
a,g	i	800	400	50%
j	i	2000	600	30%
a	i	240000	800	0.3%
a,b	i	220000	500	0.2%

Classement en deux phases pour TopPI

1. Production des top-par-item avec TopPI, $k \in [100; 500]$
2. Transformation en règle d'association
3. Re-classement (par confiance ou Piatetsky-Shapiro)
4. Affichage des top-n, $n \in [10; 50]$

$A \rightarrow$	B	$support(A)$	$support(A \cup B)$	Confiance
a,g	i	800	400	50%
j	i	2000	600	30%

Classement en deux phases pour TopPI



Couverture de TopPI des top-n par p-value sur *LastFM*

Vers l'analyse en ligne

Systèmes d'analyse en mémoire : production de $top(i)$ à la demande.

- 128GB dans un seul serveur
- Apache Spark

Exploitation des retours utilisateur

- Apprentissage automatique des classements

Merci pour votre attention !