Tableau de Bits Indexé (TBI) pour la Recherche de Séquences Fréquentes - Application à l'enquête MENAGE

Lionel Savary, Karine Zeitouni

PRiSM - Versailles

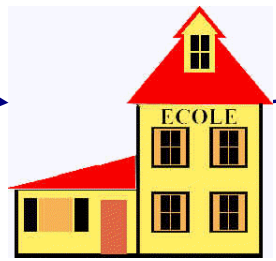
PLAN

- Objectifs et spécificités du problème.
- Proposition.
  - Principales caractéristiques.
  - Représentation en mémoire.
  - Génération de valeurs et de candidats.
- Conclusion.

Objectifs

➔ **Déterminer les séquences d'activités principales D'une population urbaine de Lille.**

Maison



École

Travail

Super Marché

Spécificités du problème

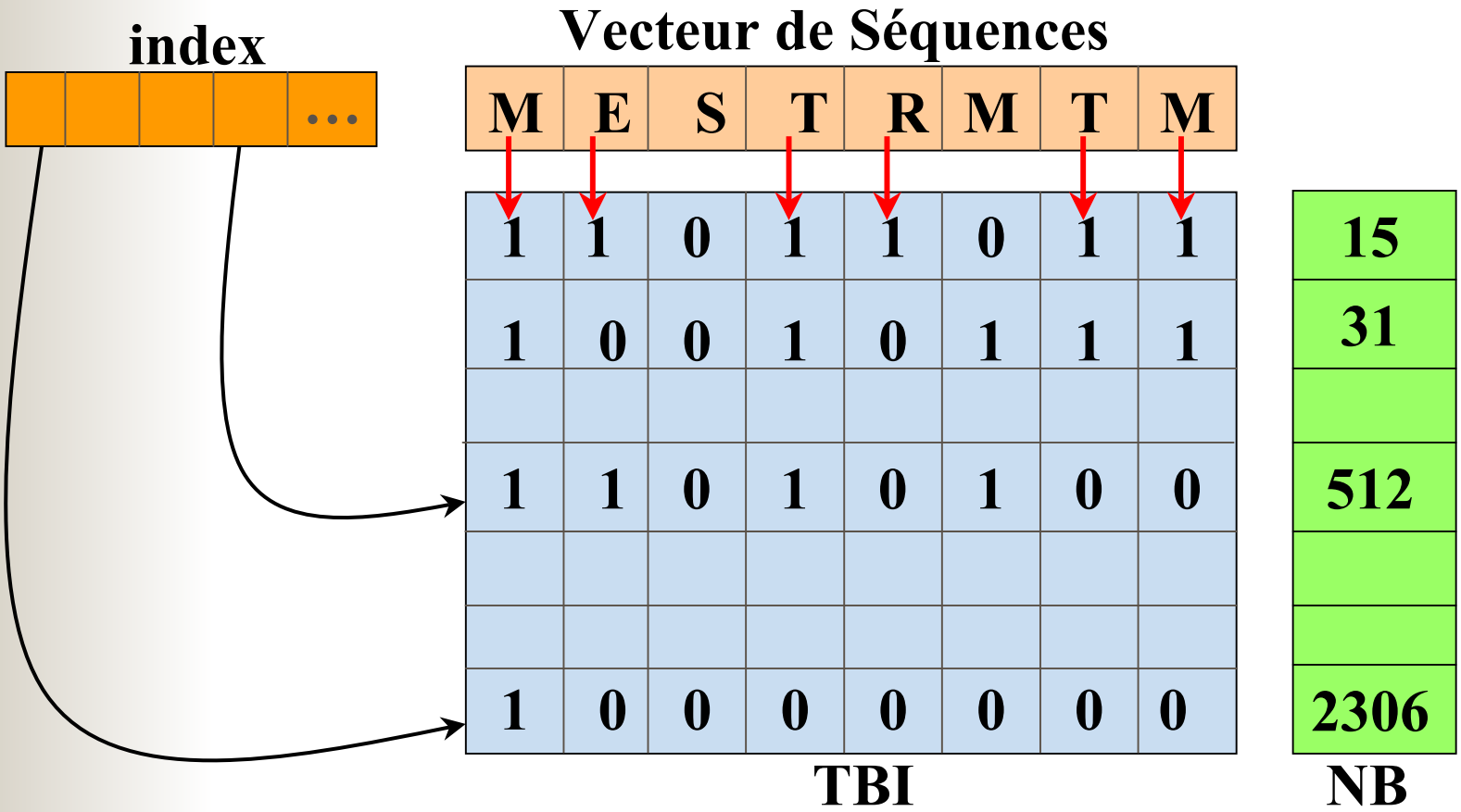
(M,E,T,S,M) → programme d'activité (PA) d'un individu:

- **Des articles peuvent se répéter.**
- **L'ordre est important: (M,E,T,S,M) ≠ (M,T,E,S,M)**
- **le calcul du support ne doit pas prendre en compte la répétition d'articles dans un PA:**
  - **(M,T,R,T,S,T,M)**
  - **(M,E,M)**
  - **(M,S,S,M)**
- **Fixe MinSup = 0.5 → (T) fréquent (FAUX)**

Proposition

- 1 passe :
  - nombre total de séquences,
  - les fréquences de chaque séquence,
  - le nombre de séquences par taille.
- Vecteur de séquences.
- Séquences distinctes codées en mémoire.
- Séquences regroupées par taille.
- Index.

Représentation Mémoire



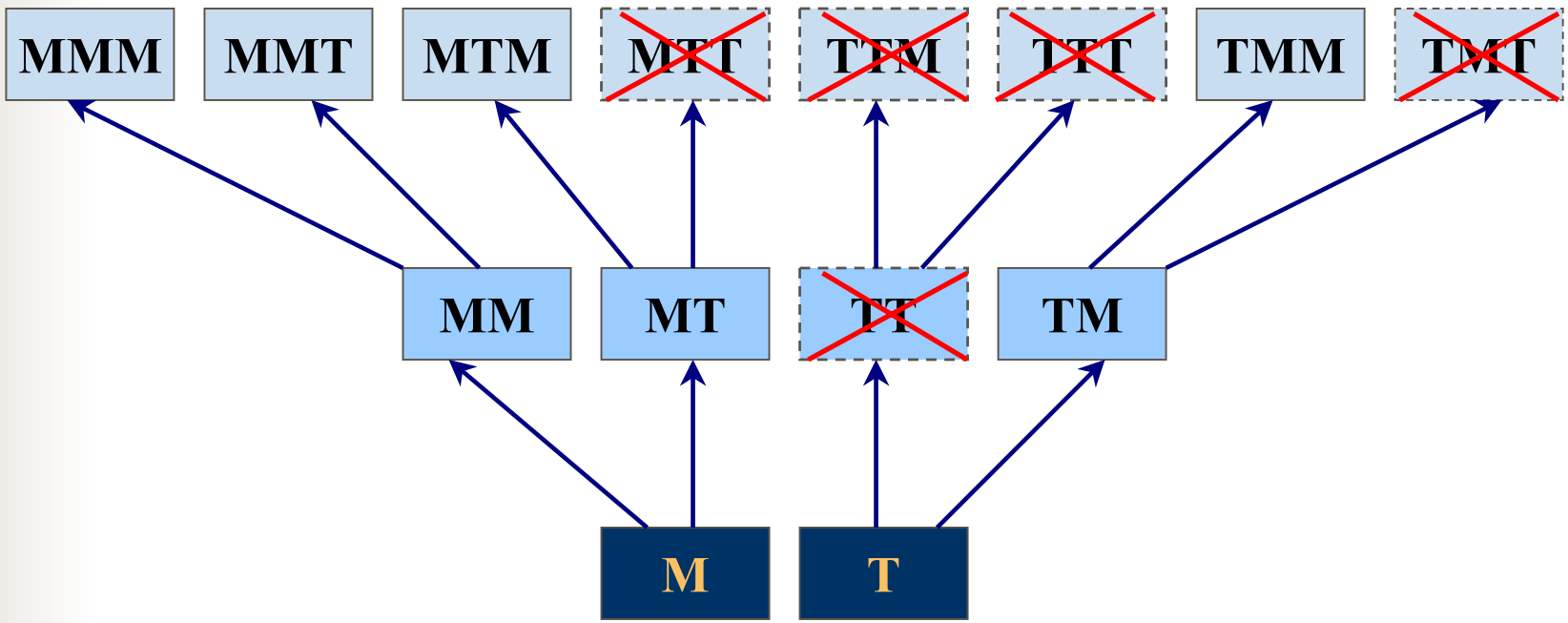
Exemple : Génération du Vecteur de Séquences (VS)

1. (METRTM)
2. (MTMTM)
3. (TMTM)
4. (M)

M = Maison
E = Ecole
T = Travail
R = Restaurant

VS = ∅
VS = (METRTM)
VS = (METRTMTM)
VS = (METRTMTM)
VS = (METRTMTM)

Exemple : Génération de Candidats



Conclusion

- La base de données n'est accédée qu'une seule fois.
- Représentation des données peu gourmande en mémoire:
  - Séquences codées en format binaire.
  - Seules les séquences distinctes sont représentées.
- Comparaisons efficaces:
  - Regroupe les séquences de mêmes tailles, utilisation d'un tableau d'indexes.
  - Comparaison des séquences avec opérateurs binaires.