Articles de départ

- Information Extraction Based on Pattern Discovery
- A Fast and Robust Method for Web Page Template Detection and Removal

Problématique

Détecter de manière automatique des patterns commun dans une collection de page web

Deux approches

- **Tree Mapping** : construction d'un mapping entre le DOM de plusieurs pages web et extrait les sous-arbres qui sont commun à ces documents
- **Patricia Tree** : construction d'une structure arborescente de recherche (preprocessing)

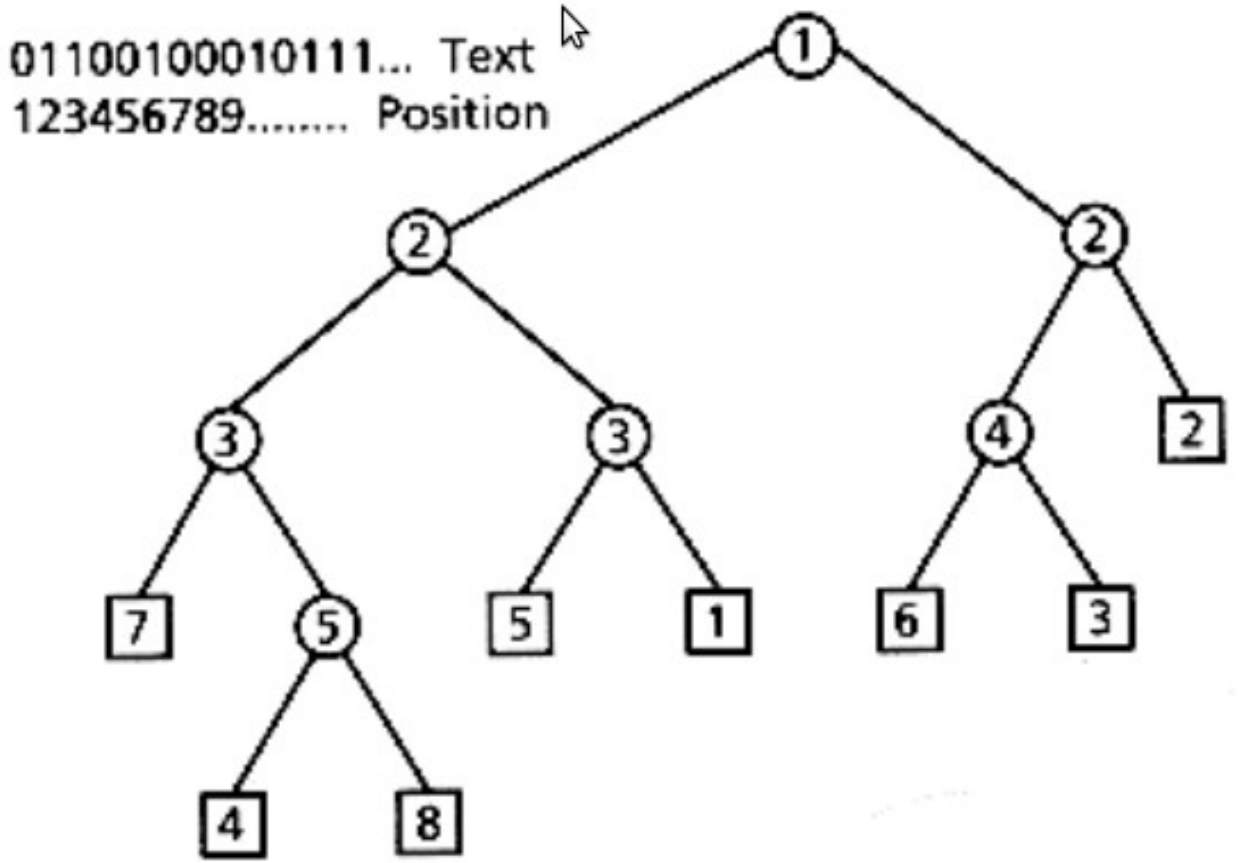
Patricia Tree

- Nouveau modèle de texte → texte = chaîne de caractère unique. Découper en semi-infinite Strings

Example:

```
Text          Once upon a time, in a far away land . . .

sistring 1    Once upon a time . . .

sistring 2    nce upon a time . . .

sistring 8    on a time, in a . . .

sistring 11   a time, in a far . . .

sistring 22   a far away land . . .
```

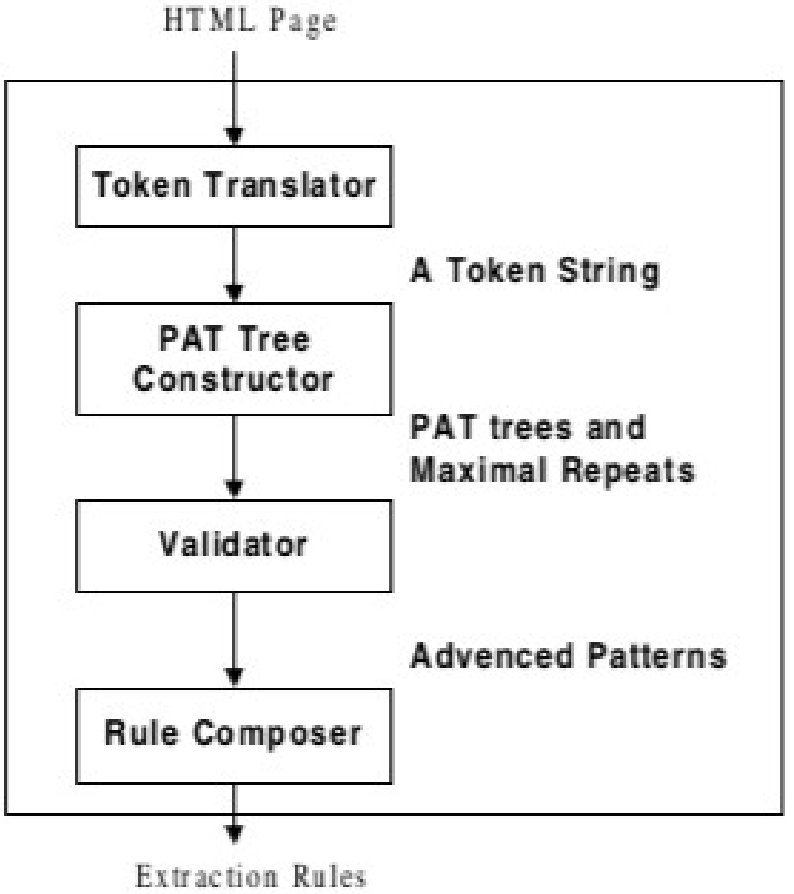
Patricia Tree



Patricia Tree

- Plusieurs type de recherche :
  - Préfixe
  - **Proximité**
  - Range
  - Répétition (plus grande nombre de caractère)
  - Répétition la plus fréquente
  - Expression régulière

Patricia Tree



Tree Mapping

- Distance entre deux arbres : nombre d'opération pour transformer T1 vers T2
- Plusieurs restriction de Mapping :
  - Distance de modification
  - Distance d'alignement
  - Distance de sous-arbres isolés
  - Distance descendante

Tree Mapping

- Apport à notre problématique :
  - Problématique : Dans les pages web, diversité des structures ayant la même fonction
  - Hypothèse : Les structures similaires ont la même fonction
  - Solution : Calculer un indice de similarité entre structures avec le mapping d'arbre