Vers des solutions adaptatives et génériques pour l'extraction de motifs intéressants dans les données

Soutenance de thèse de Frédéric Flouvat

Directeurs de thèse: Fabien De Marchi et Jean-Marc Petit

8 décembre 2006

Plan de l'exposé

Contexte

- Explosion de la quantité de données stockées
- Beaucoup d'informations "cachées" dans ces données

⇨ Comment extraire ces informations a priori inconnues ?

⇨Fouille de données : fournir des outils pour extraire des connaissances de ces "masses" de données

- grande diversité de problèmes et d'applications
- nombre important de contributions tant au niveau algorithmique que logiciel

Problématique : outils existants très spécialisés ne permettant de traiter qu'une petite partie des applications possibles

Contributions

Famille de problèmes étudiée

Extraction de motifs sous contrainte (anti-)monotone

Mise en place d'un algorithme adaptatif et générique

A partir de l'étude des jeux de données, d'une nouvelle caractérisation et classification des données

Développement d'un outil logiciel générique

Implémentation d'un cadre théorique existant sous forme d'une librairie de composants C++

Contributions

Famille de problèmes étudiée

Extraction de motifs sous contrainte (anti-)monotone

Mise en place d'un algorithme adaptatif et générique

A partir de l'étude des jeux de données, d'une nouvelle caractérisation et classification des données

Développement d'un outil logiciel générique

Implémentation d'un cadre théorique existant sous forme d'une librairie de composants C++

Contributions

Famille de problèmes étudiée

Extraction de motifs sous contrainte (anti-)monotone

Mise en place d'un algorithme adaptatif et générique

A partir de l'étude des jeux de données, d'une nouvelle caractérisation et classification des données

Développement d'un outil logiciel générique

Implémentation d'un cadre théorique existant sous forme d'une librairie de composants C++

Outline

Cadre de travail

Cadre théorique des problèmes d'extraction de motifs intéressants (Mannila et Toivonen, DMKD 1997)

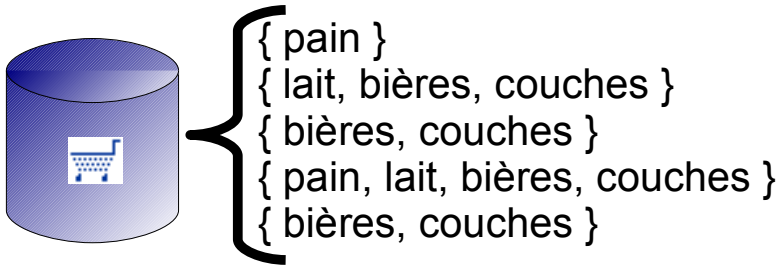
- un **langage** pour exprimer des motifs
- une **relation d'ordre partiel** sur les motifs
- une **base de données**
- un **prédicat**, une contrainte (anti-)monotone sur les données évaluant si un motif est intéressant

Hypothèse : problèmes représentables par des ensembles

Des applications multiples et variées

- extraction des clés d'une base de données relationnelle
- certaines phases d'un processus de réécriture de requêtes (Jaudoin et al., DL 2005)
- extraction de motifs fréquents d'une base de données de transactions

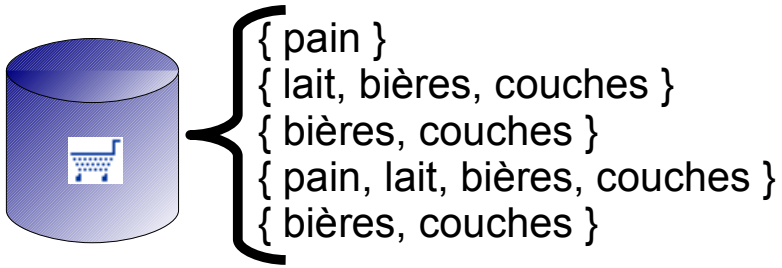
Exemple de l'extraction d'ensembles fréquents



Base de données des achats effectués dans un supermarché

⇨Quels sont les ensembles de produits achetés par plus de 40% des clients ?

Exemple de l'extraction d'ensembles fréquents



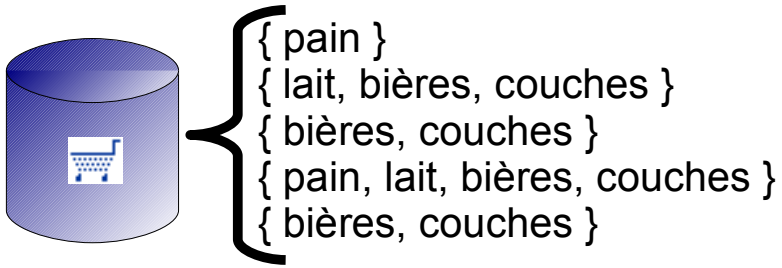
Base de données des achats effectués dans un supermarché

⇨Quels sont les ensembles de produits achetés par plus de 40% des clients ?

Motif	Support
{∅}	5
{pain}	2
{lait}	2
{bières}	4
{couches}	4
{pain,lait}	1
{pain,bières}	1
{pain,couches}	1

Motif	Support
{lait,bières}	2
{lait,couches}	2
{bières,couches}	4
{pain,lait,bières}	1
{pain,lait,couches}	1
{pain,bières,couches}	1
{lait,bières,couches}	2
{pain,lait,bières,couches}	1

Exemple de l'extraction d'ensembles fréquents



Base de données des achats effectués dans un supermarché

⇨Quels sont les ensembles de produits achetés par plus de 40% des clients ?

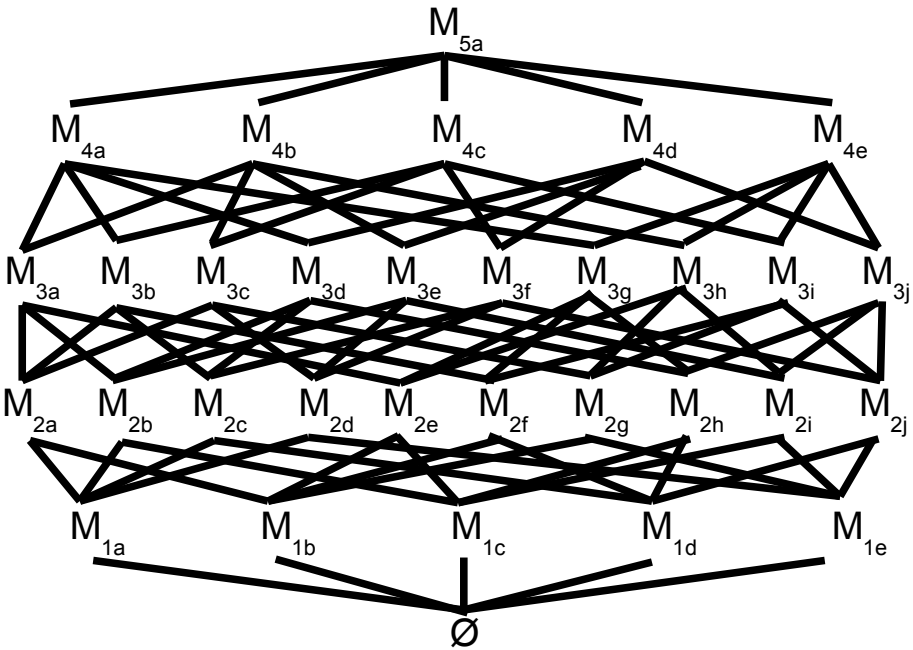
Motif	Support
{∅}	5
{pain}	2
{lait}	2
{bières}	4
{couches}	4
{pain,lait}	1
{pain,bières}	1
{pain,couches}	1

Motif	Support
{lait,bières}	2
{lait,couches}	2
{bières,couches}	4
{pain,lait,bières}	1
{pain,lait,couches}	1
{pain,bières,couches}	1
{lait,bières,couches}	2
{pain,lait,bières,couches}	1

motifs fréquents (support $\geq$ seuil)

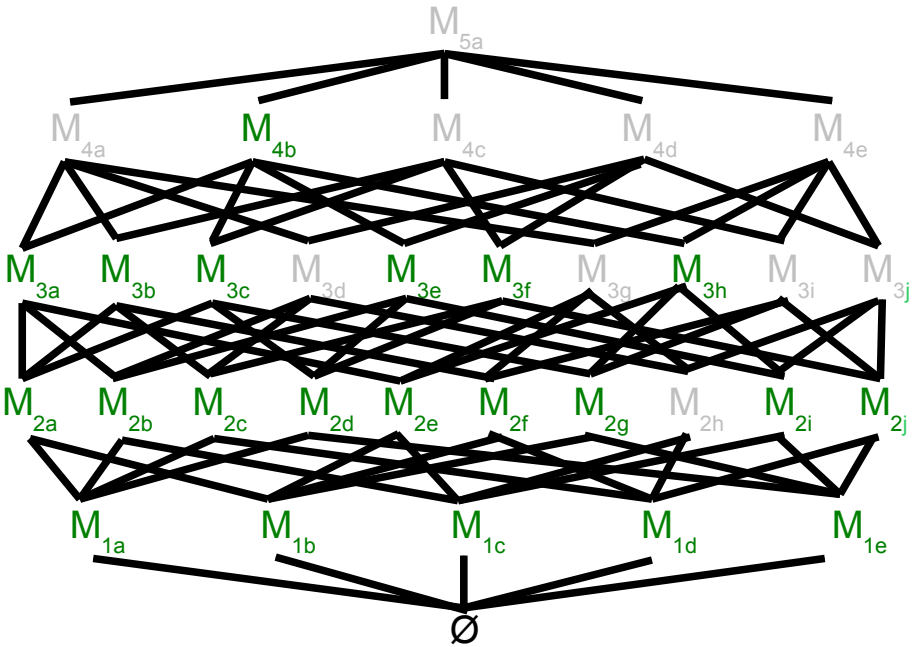
Espace de recherche

Hypothèse du cadre : espace de recherche représentable par des ensembles, i.e. isomorphisme avec un treillis des parties



Espace de recherche

Hypothèse du cadre : espace de recherche représentable par des ensembles, i.e. isomorphisme avec un treillis des parties

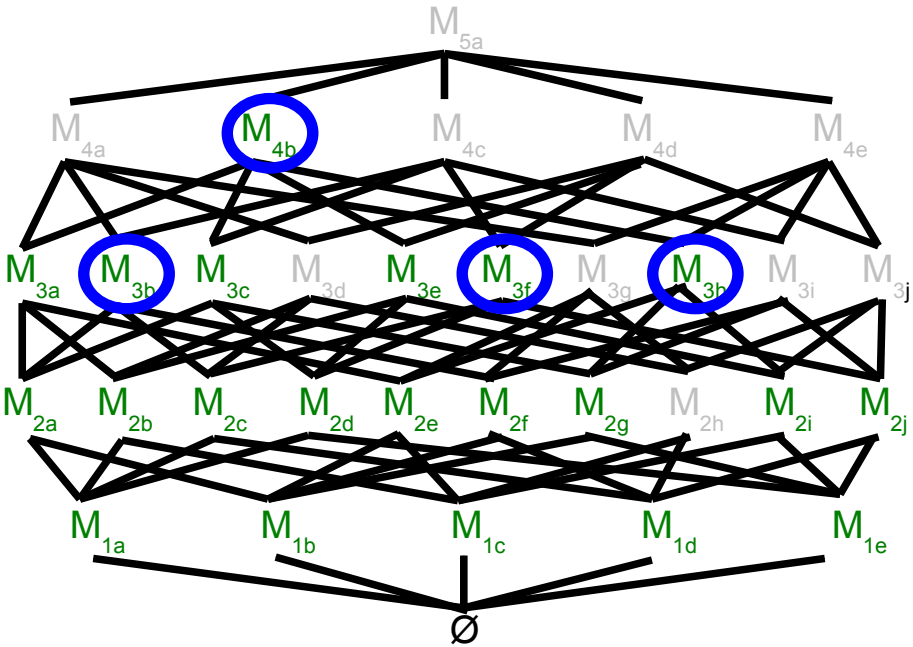


Les solutions de l'extraction

- la **théorie** : l'ensemble des motifs intéressants

Espace de recherche

Hypothèse du cadre : espace de recherche représentable par des ensembles, i.e. isomorphisme avec un treillis des parties

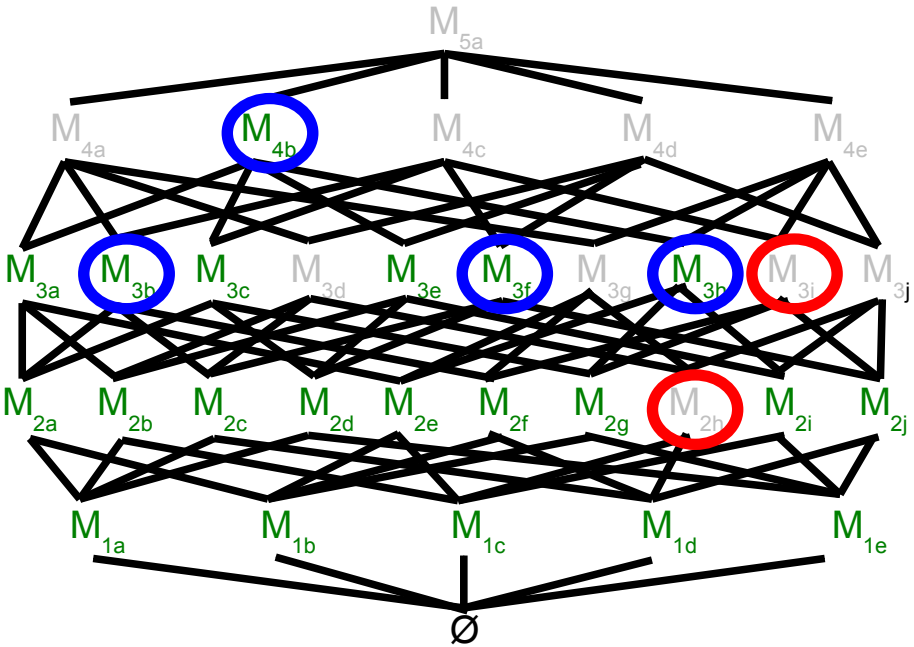


Les solutions de l'extraction

- la **théorie** : l'ensemble des motifs intéressants
- la **bordure positive** Bd^+ : plus "grands" motifs intéressants

Espace de recherche

Hypothèse du cadre : espace de recherche représentable par des ensembles, i.e. isomorphisme avec un treillis des parties

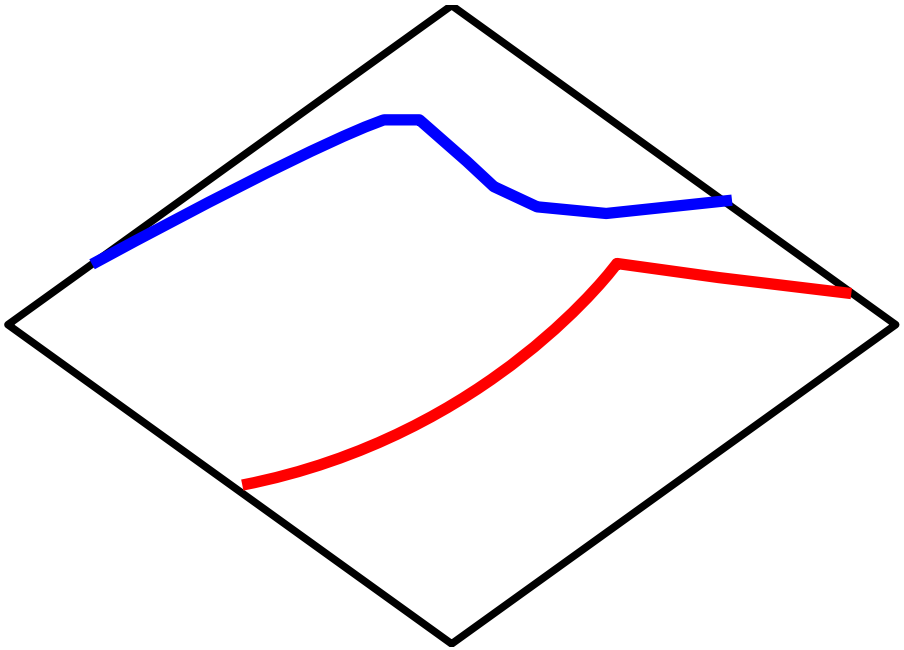


Les solutions de l'extraction

- la **théorie** : l'ensemble des motifs intéressants
- la **bordure positive** Bd^+ : plus "grands" motifs intéressants
- la **bordure négative** Bd^- : plus "petits" motifs non intéressants

Espace de recherche

Hypothèse du cadre : espace de recherche représentable par des ensembles, i.e. isomorphisme avec un treillis des parties

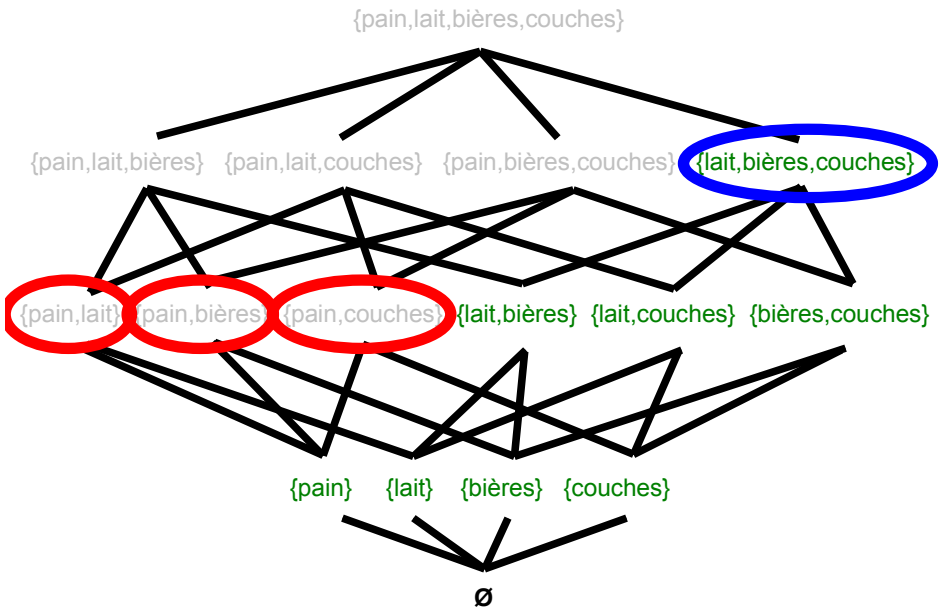


Les solutions de l'extraction

- la **théorie** : l'ensemble des motifs intéressants
- la **bordure positive** Bd^+ : plus "grands" motifs intéressants
- la **bordure négative** Bd^- : plus "petits" motifs non intéressants

Espace de recherche

Hypothèse du cadre : espace de recherche représentable par des ensembles, i.e. isomorphisme avec un treillis des parties

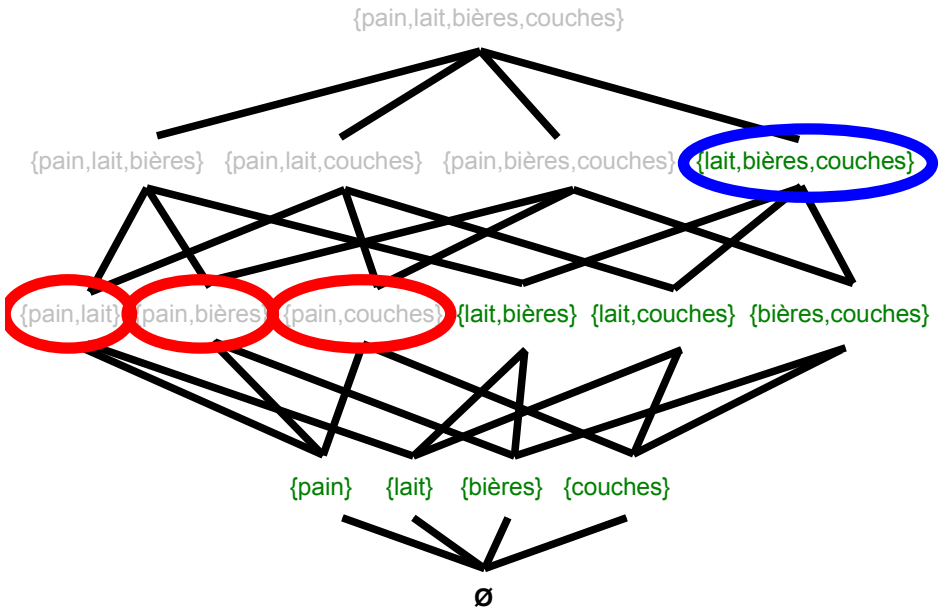


Les solutions de l'extraction

- la **théorie** : l'ensemble des motifs intéressants
- la **bordure positive** Bd^+ : plus "grands" motifs intéressants
- la **bordure négative** Bd^- : plus "petits" motifs non intéressants

Espace de recherche

Hypothèse du cadre : espace de recherche représentable par des ensembles, i.e. isomorphisme avec un treillis des parties



Les solutions de l'extraction

- la **théorie** : l'ensemble des motifs intéressants
- la **bordure positive** Bd^+ : plus "grands" motifs intéressants
- la **bordure négative** Bd^- : plus "petits" motifs non intéressants

⇨Des problèmes complexes : espace de recherche exponentiel !

⇨Importance de la stratégie d'exploration et d'élagage

Les algorithmes d'extraction de motifs

3 algorithmes proposés dans ce cadre : stratégie par niveaux, Guess And Correct et Dualize And Advance

- génériques mais efficacité dépend fortement du problème, des données

Beaucoup d'algorithmes dédiés à la résolution d'un problème d'extraction de motifs

- souvent efficaces mais difficiles à appliquer à un autre problème du cadre

Objectif : Proposer un algorithme efficace et générique pour les problèmes d'extraction de motifs intéressants, représentables par des ensembles

Les algorithmes d'extraction de motifs

3 algorithmes proposés dans ce cadre : stratégie par niveaux, Guess And Correct et Dualize And Advance

- génériques mais efficacité dépend fortement du problème, des données

Beaucoup d'algorithmes dédiés à la résolution d'un problème d'extraction de motifs

- souvent efficaces mais difficiles à appliquer à un autre problème du cadre

Objectif : Proposer un algorithme efficace et générique pour les problèmes d'extraction de motifs intéressants, représentables par des ensembles

Outline

Application d'un algorithme de découverte des DI à l'extraction d'ensembles fréquents (1/2)

Cadre théorique
Algorithmes

⇨ Utilisation d'un de ces algorithmes pour résoudre différents problèmes du cadre ?

Algorithme *ZigZag* (De Marchi et Petit, ICDM 2003)

- extraction des plus grandes DI satisfaites dans une base de données relationnelle
- très performant dans ce contexte

⇨Application à l'extraction d'ensembles fréquents

- moins efficace que les algorithmes existants
- efficacité dépend du jeu de données étudié

Application d'un algorithme de découverte des DI à l'extraction d'ensembles fréquents (2/2)

Principales causes :

- plus grande variété de jeux de données pour l'extraction d'ensembles fréquents
- bonnes performances pour les DI liées à une propriété du problème

➪Nécessité d'adapter la stratégie de l'algorithme en fonction du problème, i.e. en fonction des données

Prérequis : Etudier les caractéristiques des données influençant les algorithmes

Etude expérimentale des jeux de données pour les ensembles fréquents

Objectif : Etudier l'influence des données sur les principales stratégies d'exploration de l'espace de recherche, pour ensuite pouvoir développer un algorithme adaptatif

➪Etude des jeux de données pour le problème de l'extraction des ensembles fréquents

- problème d'extraction de motifs le plus étudié
- grande variété de jeux de données
- implémentations des principaux algorithmes disponibles
- existence de bancs d'essais comparant les performances des implémentations

ateliers FIMI 2003-2004 (*Frequent Itemset Mining Implementations*)

Etude expérimentale des jeux de données pour les ensembles fréquents

Objectif : Etudier l'influence des données sur les principales stratégies d'exploration de l'espace de recherche, pour ensuite pouvoir développer un algorithme adaptatif

➪Etude des jeux de données pour le problème de l'extraction des ensembles fréquents

- problème d'extraction de motifs le plus étudié
- grande variété de jeux de données
- implémentations des principaux algorithmes disponibles
- existence de bancs d'essais comparant les performances des implémentations

ateliers FIMI 2003-2004 (*Frequent Itemset Mining Implementations*)

Caractéristiques des données étudiées dans la littérature

Caractéristiques de la base de données

- nombre/taille des transactions et nombre d'articles

⇨ influence les algorithmes, mais insuffisant face à la grande diversité des données

Densité (Gouda et Zaki, ICDM 2001)

- longs ensembles fréquents même pour des seuils élevés de support

⇨ influence les stratégies d'exploration

Etude expérimentale de la densité (Gouda et Zaki)

Leur proposition : étudier la distribution de la bordure positive

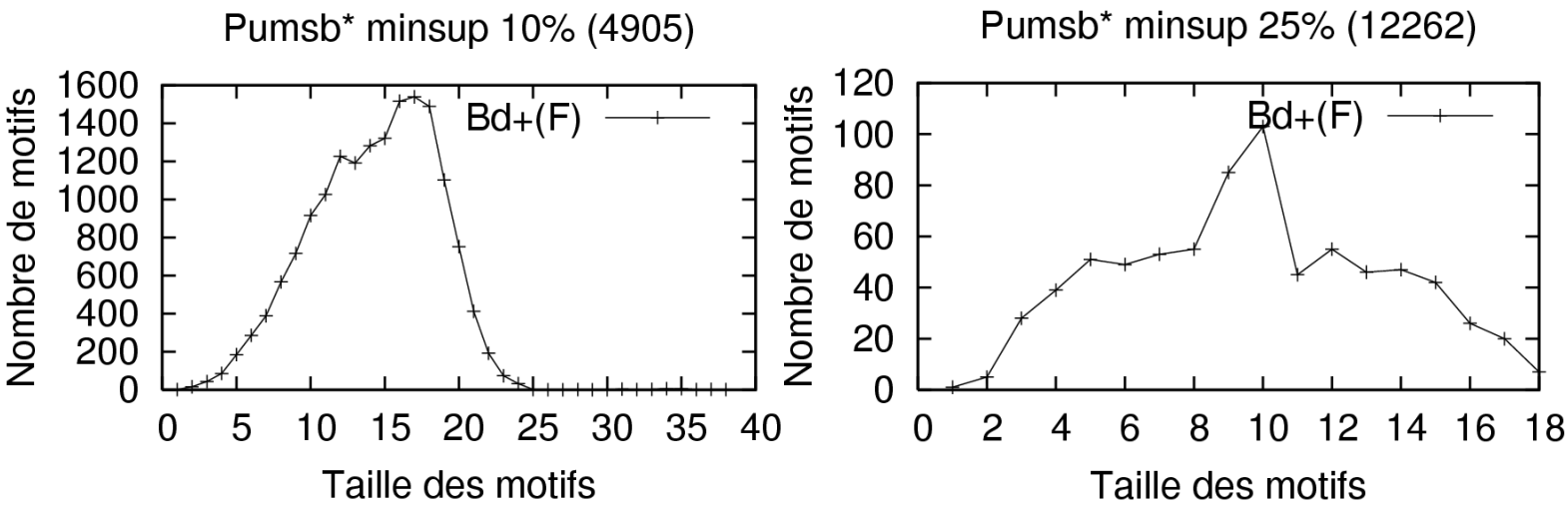
- distribution des ensembles fréquents maximaux par rapport à leur taille

Expérimentations faites à partir de 5 jeux de données réels et 2 synthétiques issus de FIMI

➪Classification des jeux de données en 4 types en fonction de la densité

Limites de la densité et de la classification proposée (1/2)

Pas stable par rapport au changement de seuil de support



➪La densité change selon le seuil de support étudié, i.e. la classification change

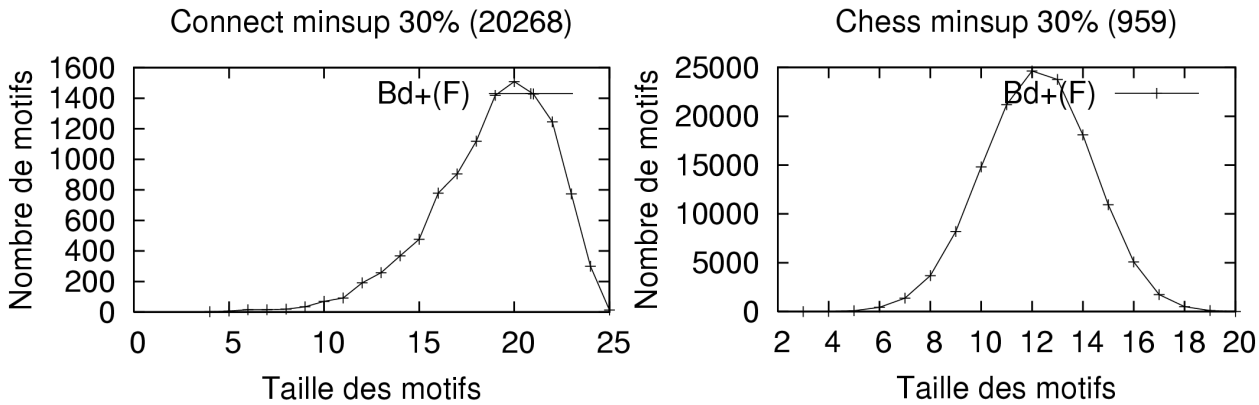
Limites de la densité et de la classification proposée (2/2)

Ne suffit pas à expliquer les performances des algorithmes

Exemple pour les jeux de données *Connect* et *Chess*

- caractéristiques des jeux de données
  Connect · plus de transactions et plus longues que *Chess*
  · plus d'articles

- densité : *Connect* plus dense que *Chess*



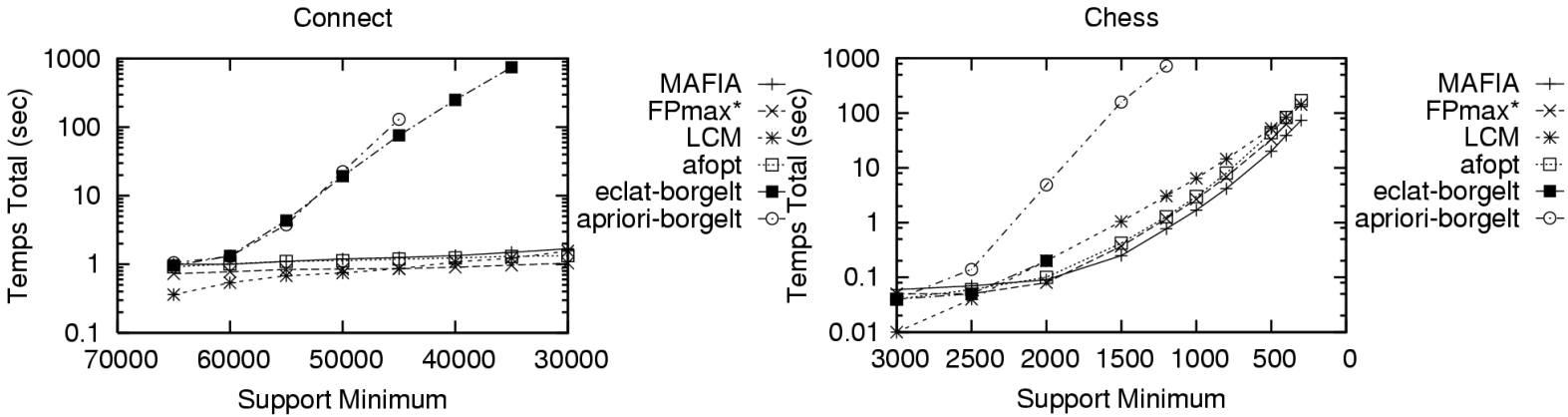
⇨Algorithmes moins efficaces pour *Connect* que pour *Chess*

Limites de la densité et de la classification proposée (2/2)

Ne suffit pas à expliquer les performances des algorithmes

Exemple pour les jeux de données *Connect* et *Chess*

- résultats des bancs d'essais de FIMI



⇨Algorithmes plus efficaces pour *Connect* que pour *Chess*

⇨En contradiction avec les caractéristiques des données et la densité

Intérêt de la bordure positive et négative

Bd^+ insuffisante pour expliquer le comportement des algorithmes

⇨Idée : étudier la distribution des bordures positive **et négative**

- Bd^- utilisée pour élaguer l'espace de recherche
- existence d'une propriété théorique liant les deux bordures
  - Bd^+ équivalent à Bd^- à un calcul de transversaux minimaux près (et inversement)

Protocole expérimental :

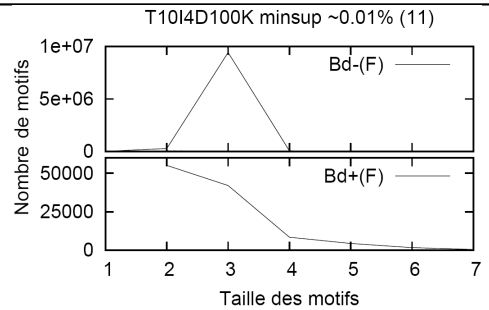
- 14 jeux de données de FIMI
- pour chaque jeux de données, étudier plusieurs seuils de support représentatifs

Distribution des bordures observée

➪Classification des jeux de données en 3 types en fonction de la distribution des bordures :

petits motifs dans Bd^+ et Bd^-

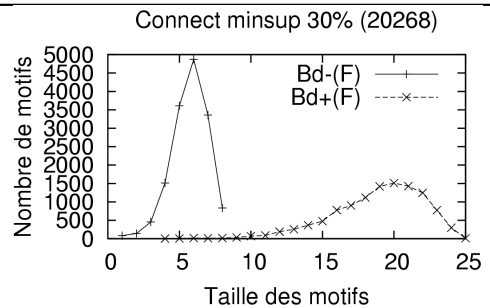
Type I



longs motifs dans Bd^+

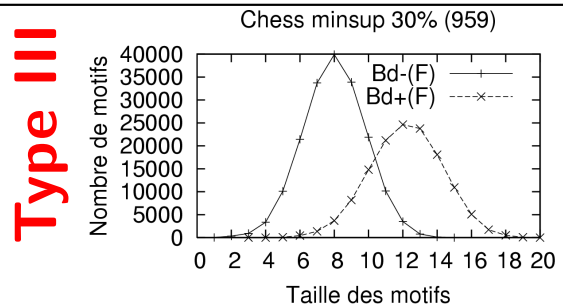
petits motifs dans Bd^-

Type II



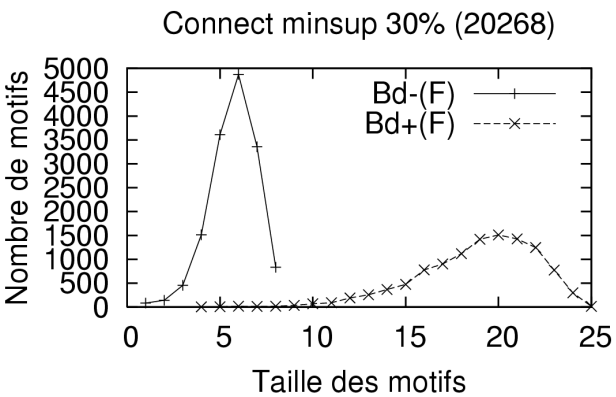
longs motifs dans Bd^-

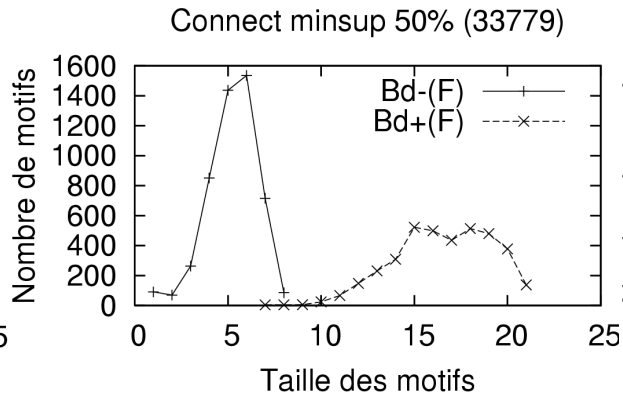
Type III

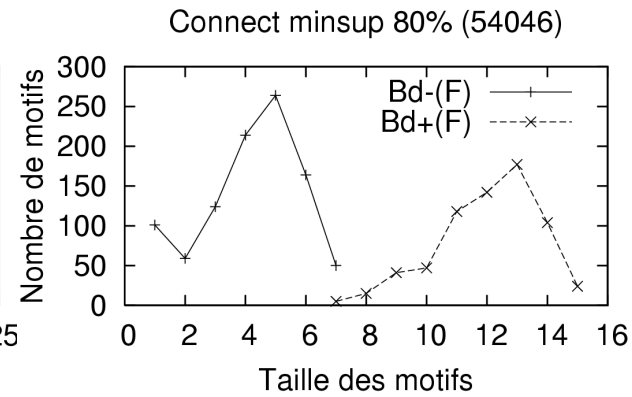


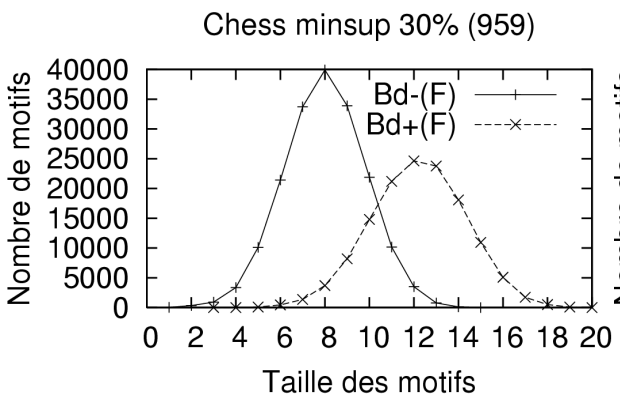
Stabilité du critère et de la classification

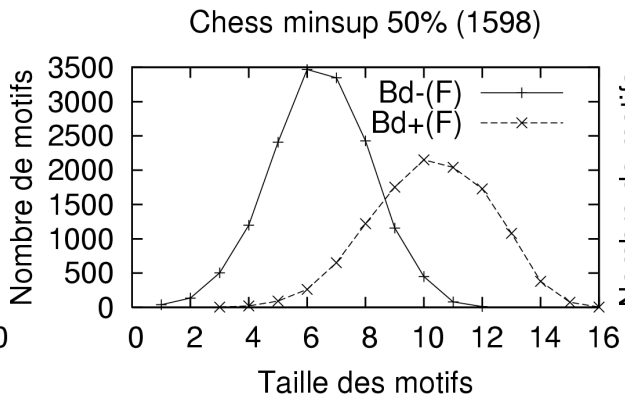
Stable par rapport au changement de seuil de support

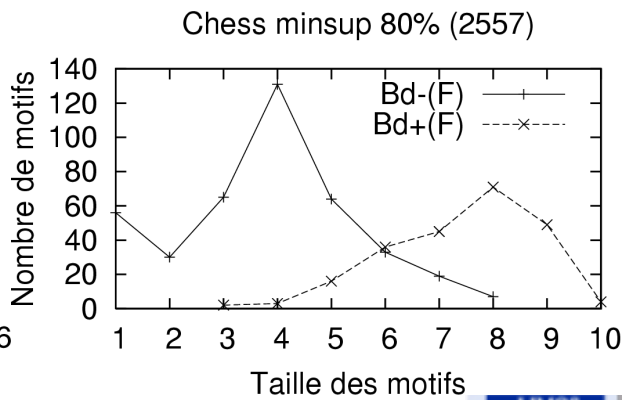










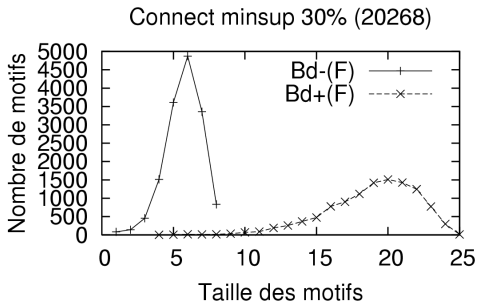


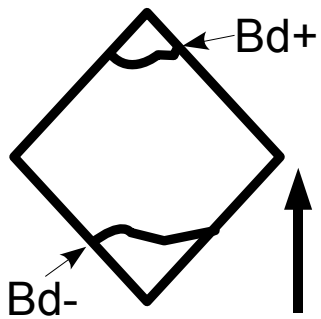
Classification et performances des algorithmes (1/2)

En accord avec les performances des algorithmes

longs motifs dans Bd^+
petits motifs dans Bd^-

Type II

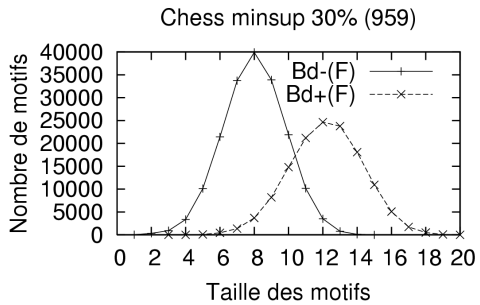


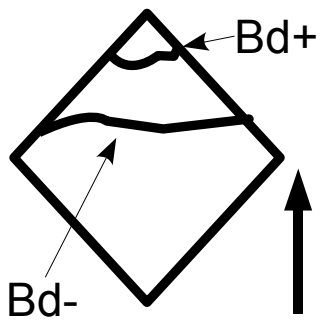


➪élagage et caractérisation de l'espace de recherche rapides

longs motifs dans Bd^+
longs motifs dans Bd^-

Type III



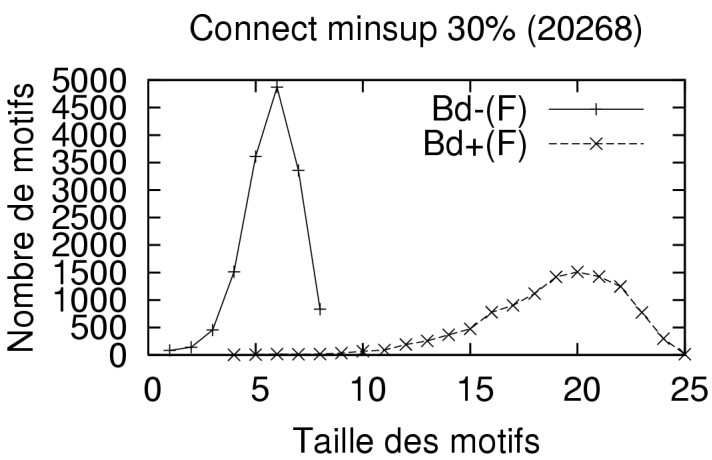


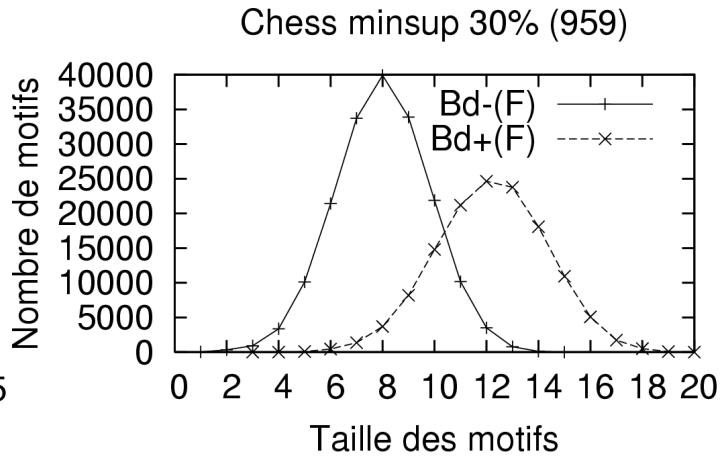
➪élagage et caractérisation de l'espace de recherche progressifs

Classification et performances des algorithmes (2/2)

Exemple pour les jeux de données *Connect* et *Chess*

- Distribution des bordures positive et négative :
  Connect plus grande "distance" entre les bordures que *Chess*





⇨Elagage et caractérisation de l'espace de recherche rapides

⇨Algorithmes plus efficaces pour *Connect* que pour *Chess*

⇨En accord avec les performances observées lors de FIMI

Synthèse de l'étude des jeux de données pour les ensembles fréquents

Etude de la distribution des bordures positive et négative (Flouvat, De Marchi et Petit, ICDM 2005)

- nouvelle caractéristique importante des données
  - stable
  - contribue à expliquer les performances des algorithmes
- conduit à une nouvelle classification des jeux de données

Résultat applicable à tous les problèmes du cadre

- ex : extraction des DI
  - propriété théorique du problème $\longrightarrow$ grande distance entre les bordures $\longrightarrow$ efficacité de l'algorithme *ZigZag*

➪Utilisation de ces résultats pour développer un algorithme adaptatif et générique

Vers un algorithme adaptatif et générique pour l'extraction de motifs intéressants

Objectifs :

- proposer dans le cadre de Mannila et Toivonen un algorithme adaptant sa stratégie en fonction des données, i.e. en fonction du problème
- extraire les **bordures** des motifs intéressants

Problèmes :

- Quelles stratégies choisir ?
- Quel(s) critère(s) utilisé(s) pour décider automatiquement du changement de stratégies ?

Idée de base de notre approche

Utiliser les observations faites sur le comportement des stratégies par rapport à la classification

• stratégie par niveaux	efficace :	"petits" motifs dans Bd^+ (Type I)
	problème :	"longs" motifs dans Bd^+ (Type II et Type III)
• stratégie fondée sur le concept de dualisation	efficace :	"petits" motifs dans Bd^- "longs" motifs dans Bd^+ (Type II)

⇨Idée : combiner ces deux stratégies pour être efficace pour les problèmes/données de type I et II

Notre proposition

Algorithme *ABS* (*Adaptive Borders Search*)

Principe :

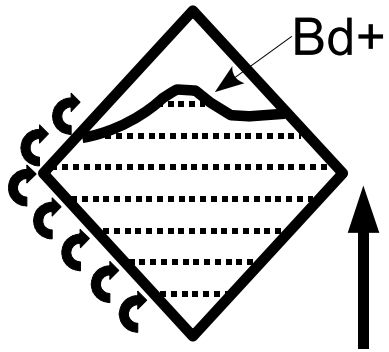
- rechercher les "petits" motifs des bordures par une stratégie par niveaux
- rechercher les "longs" motifs de la bordure positive par dualisation

⇨Stratégie d'*ABS* :

- commencer par une stratégie par niveaux
- puis, si données de type II, alterner les dualisations entre les deux bordures

Stratégie par niveaux

Stratégie de parcours : parcours par niveaux de l'espace de recherche

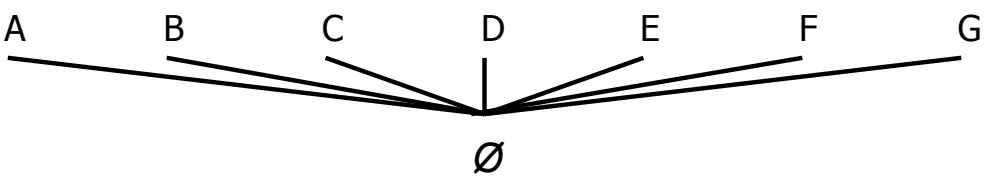


Stratégie d'élagage : exploiter les motifs non intéressants de Bd^- et la propriété de (anti-)monotonie du prédicat pour élaguer l'espace de recherche

Stratégie par niveaux : exemple

(motifs : ensembles)

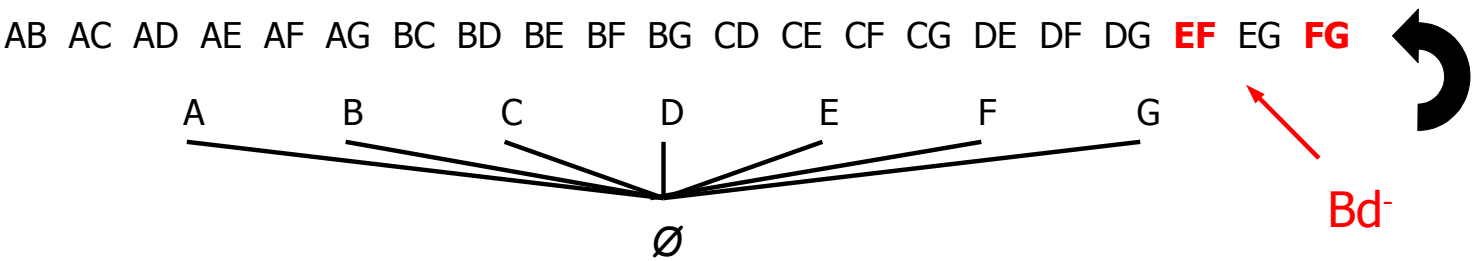
- génération des motifs à partir des motifs intéressants découverts le niveau précedent
- élagage à partir des motifs non intéressants découverts le niveau précedent



Stratégie par niveaux : exemple

(motifs : ensembles)

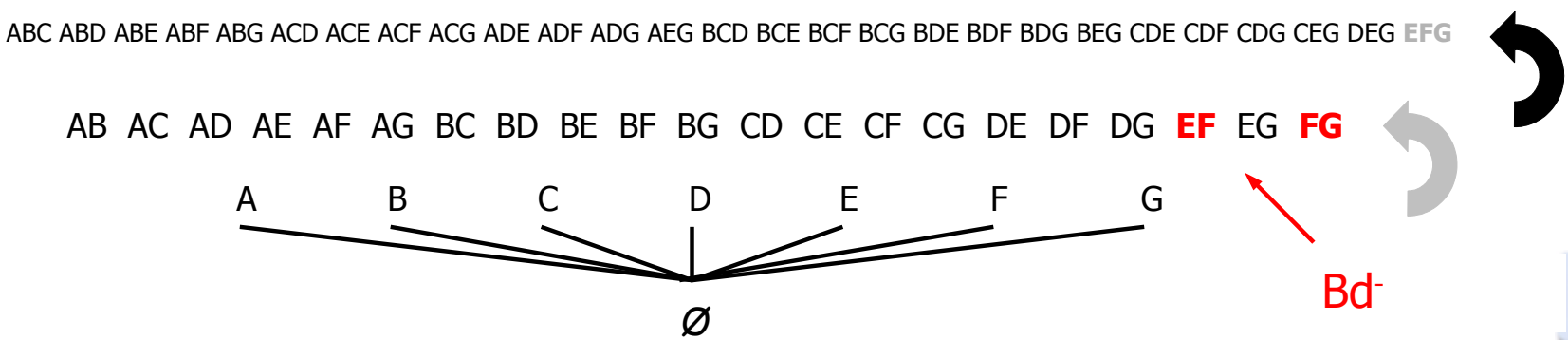
- génération des motifs à partir des motifs intéressants découverts le niveau précedent
- élagage à partir des motifs non intéressants découverts le niveau précedent



Stratégie par niveaux : exemple

(motifs : ensembles)

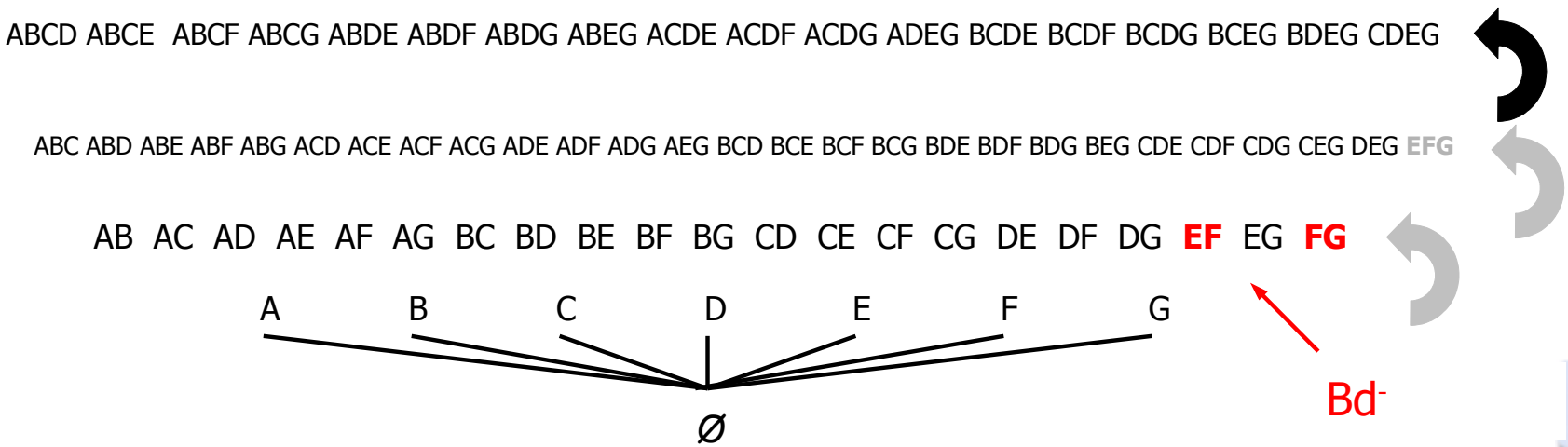
- génération des motifs à partir des motifs intéressants découverts le niveau précedent
- élagage à partir des motifs non intéressants découverts le niveau précedent



Stratégie par niveaux : exemple

(motifs : ensembles)

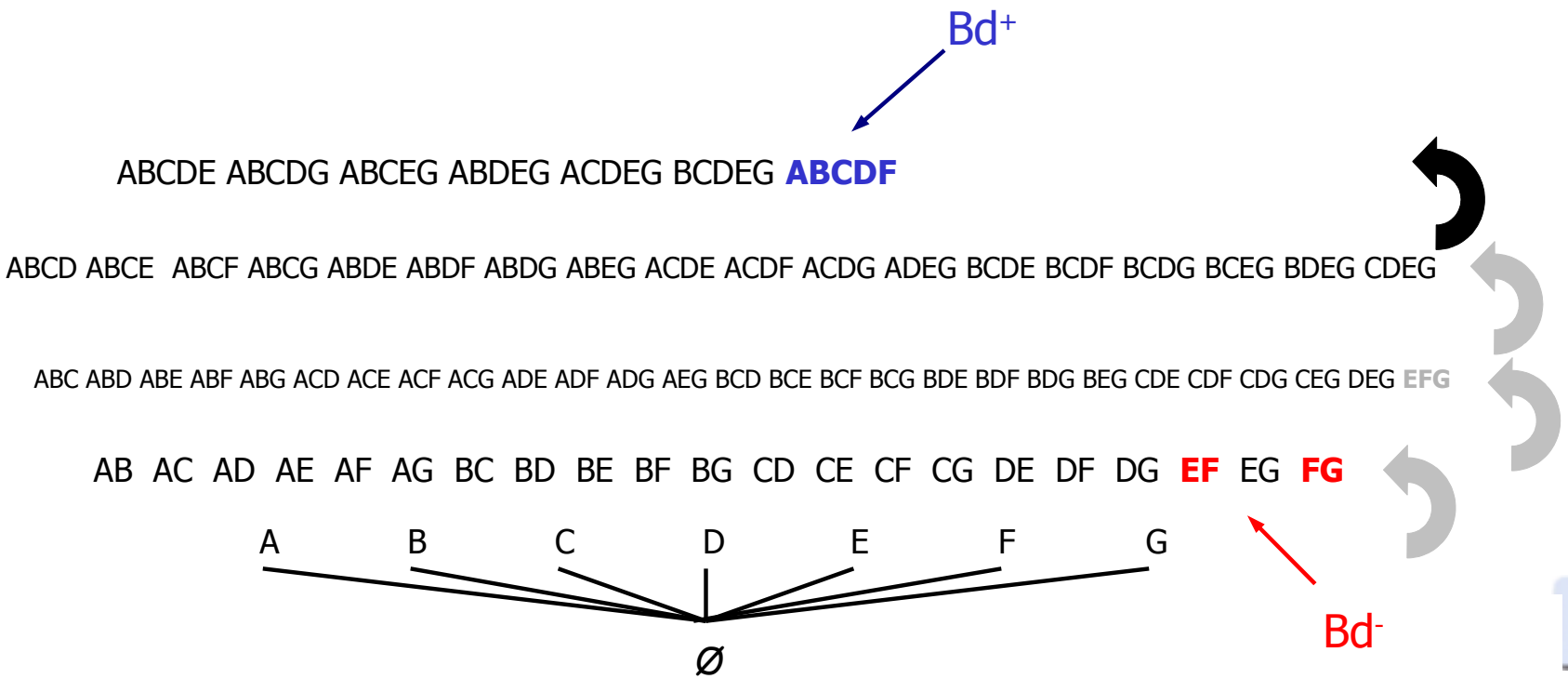
- génération des motifs à partir des motifs intéressants découverts le niveau précedent
- élagage à partir des motifs non intéressants découverts le niveau précedent



Stratégie par niveaux : exemple

(motifs : ensembles)

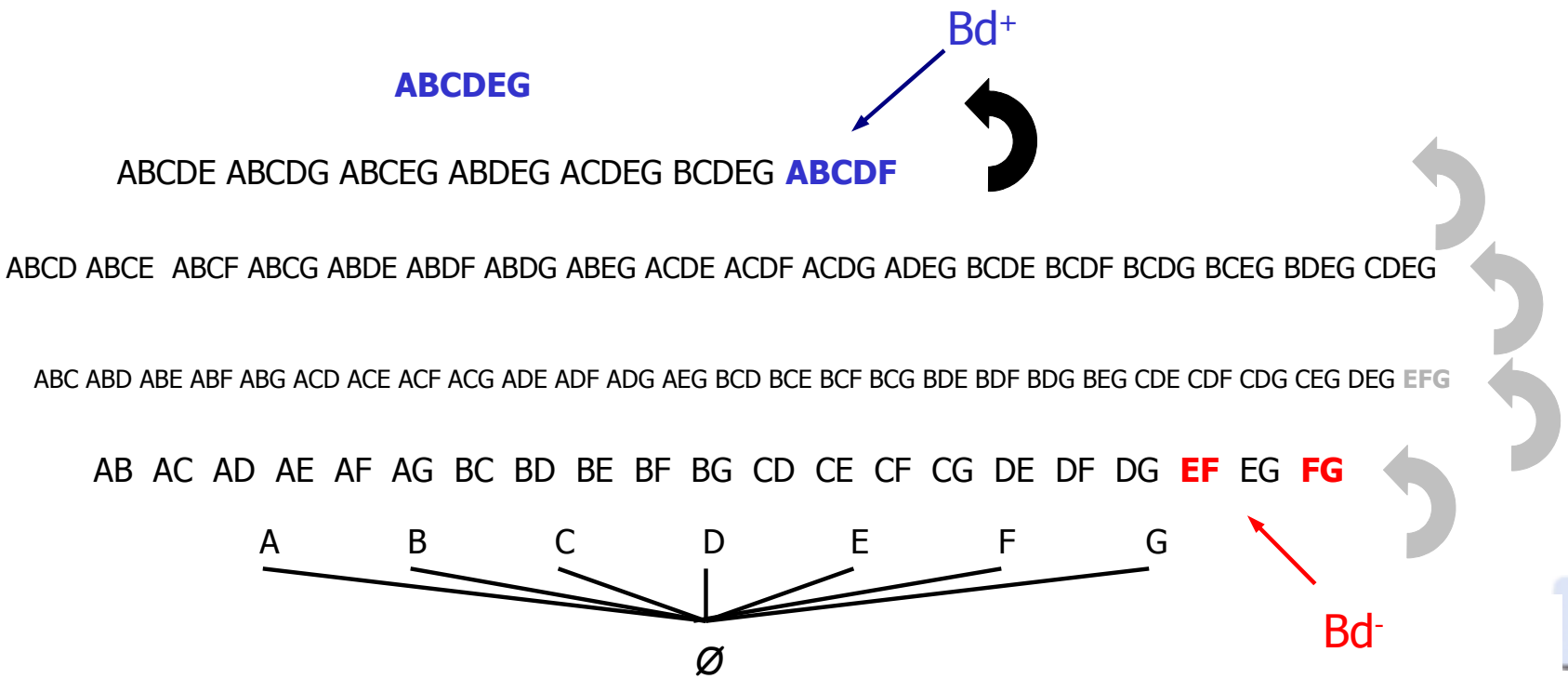
- génération des motifs à partir des motifs intéressants découverts le niveau précedent
- élagage à partir des motifs non intéressants découverts le niveau précedent



Stratégie par niveaux : exemple

(motifs : ensembles)

- génération des motifs à partir des motifs intéressants découverts le niveau précedent
- élagage à partir des motifs non intéressants découverts le niveau précedent



Dualisation

Principe : exploiter les motifs déjà découverts d'une des deux bordures pour estimer l'autre

Stratégie de parcours : équivalent à faire un "saut" dans l'espace de recherche

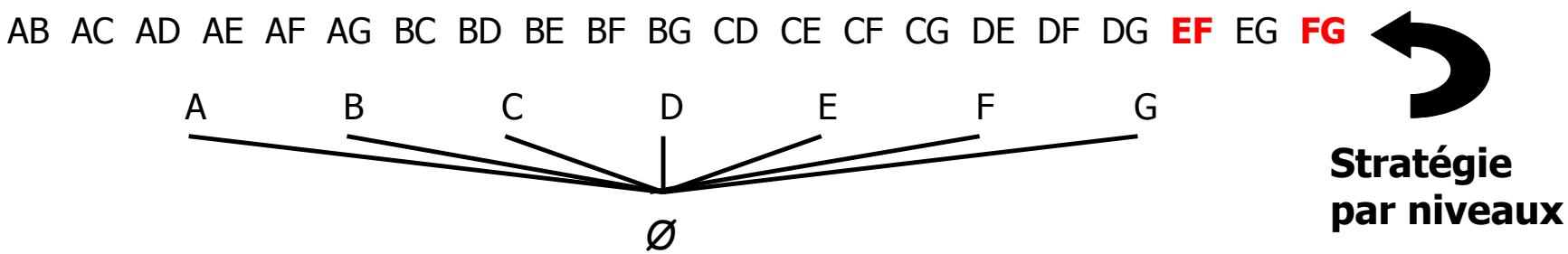
Stratégie d'élagage : ne teste pas tout motif caractérisé par un motif des bordures déjà découvert

Dualisation = calcul des transversaux minimaux d'un hypergraphe

Intérêt : fondée sur une propriété théorique des problèmes du cadre
Prix à payer : le coût de la dualisation

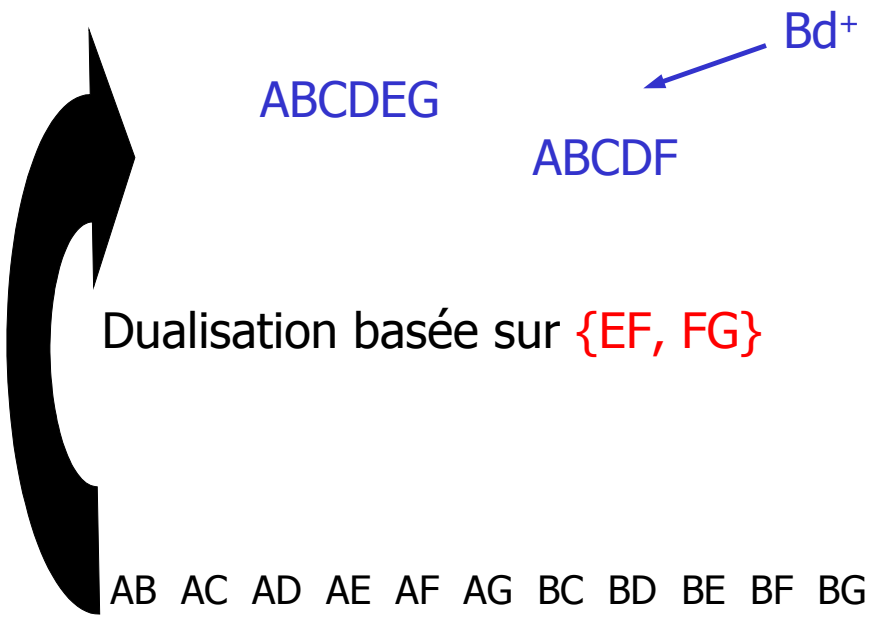
Dualisation : exemple

(motifs : ensembles)



Dualisation : exemple

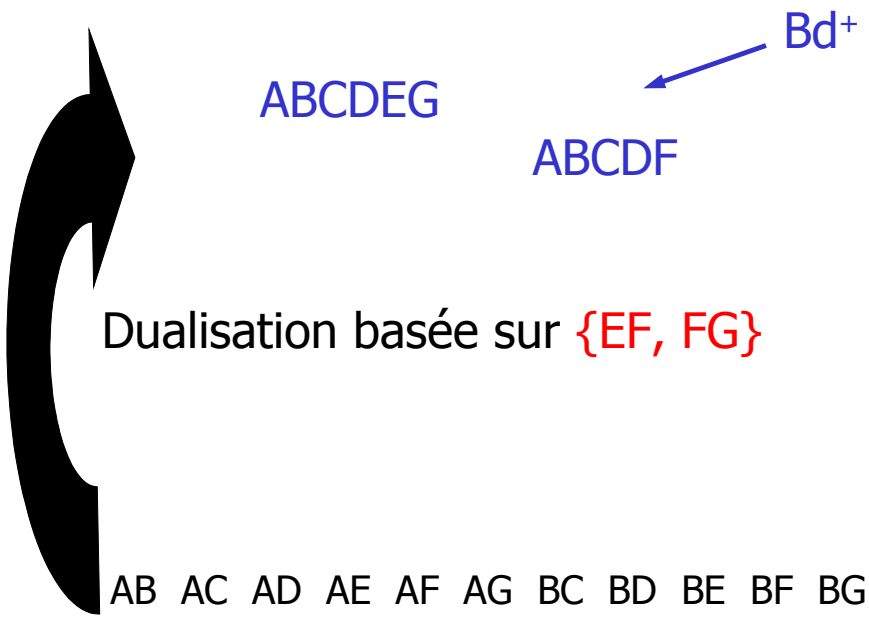
(motifs : ensembles)



Stratégie par niveaux

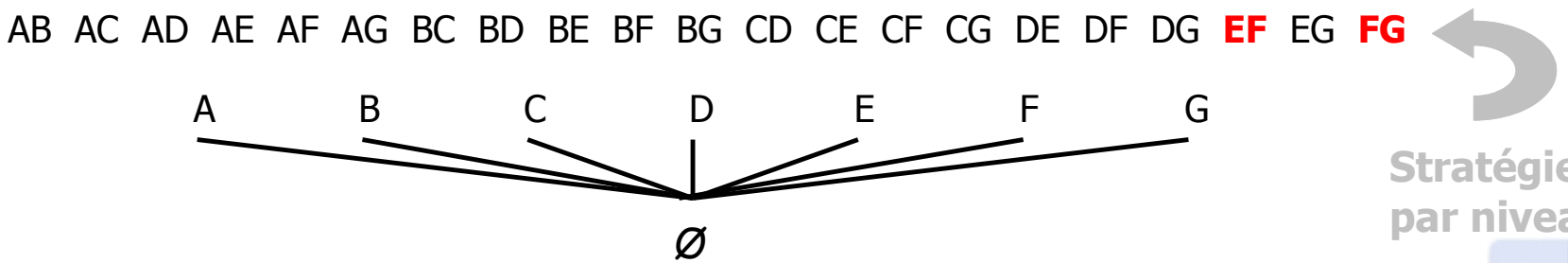
Dualisation : exemple

(motifs : ensembles)



=> caractérisation exacte

=> 1 dualisation à la place de 4 itérations avec la stratégie par niveaux



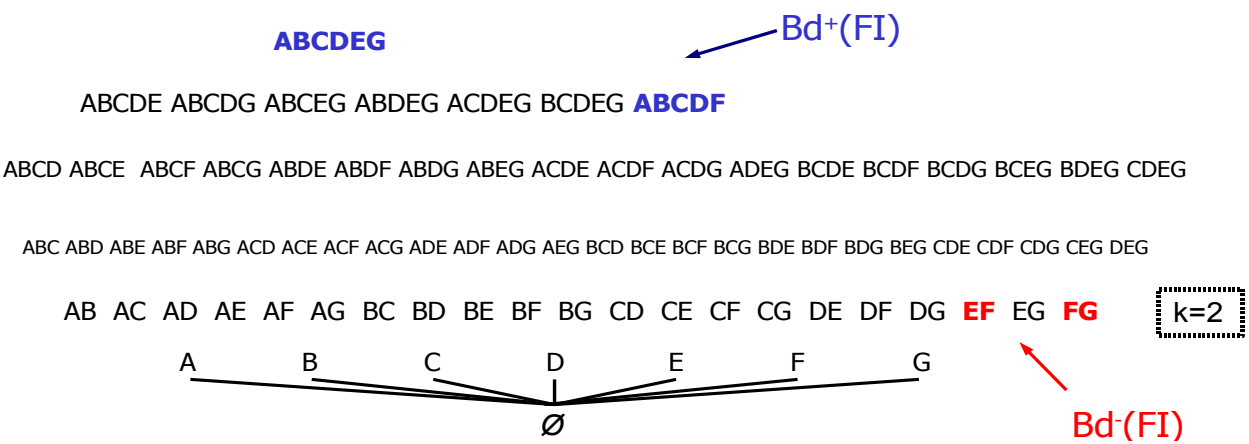
Stratégie par niveaux

Comportement adaptatif

Objectif : détecter dynamiquement le type de problème/données

Constatation :

grande "distance" entre les distributions des bordures ⇨ à partir d'un certain niveau, stratégie par niveaux ne découvre plus de motifs non intéressants



tout motif candidat X généré, $|X| > 2$
⇨X motif intéressant

Comportement adaptatif

Objectif : détecter dynamiquement le type de problème/données

Constatation :

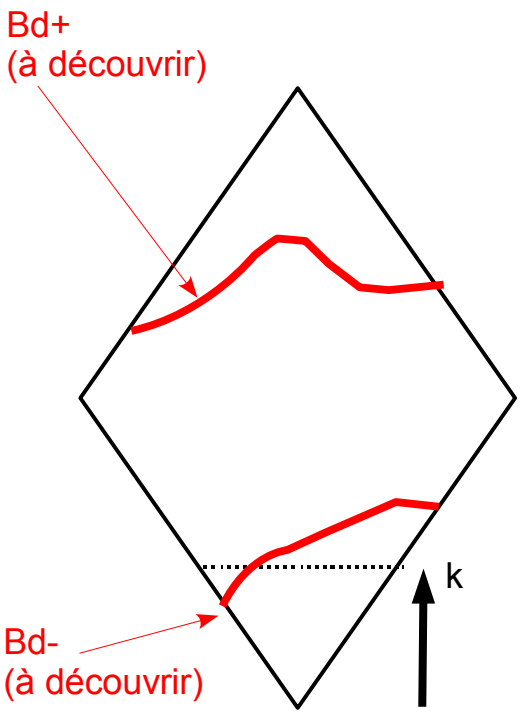
grande "distance" entre les distributions des bordures ⇨ à partir d'un certain niveau, stratégie par niveaux ne découvre plus de motifs non intéressants

⇨Principal paramètre au niveau k pour déterminer le changement de stratégie :

- ratio $|I_k|/|C_k|$:

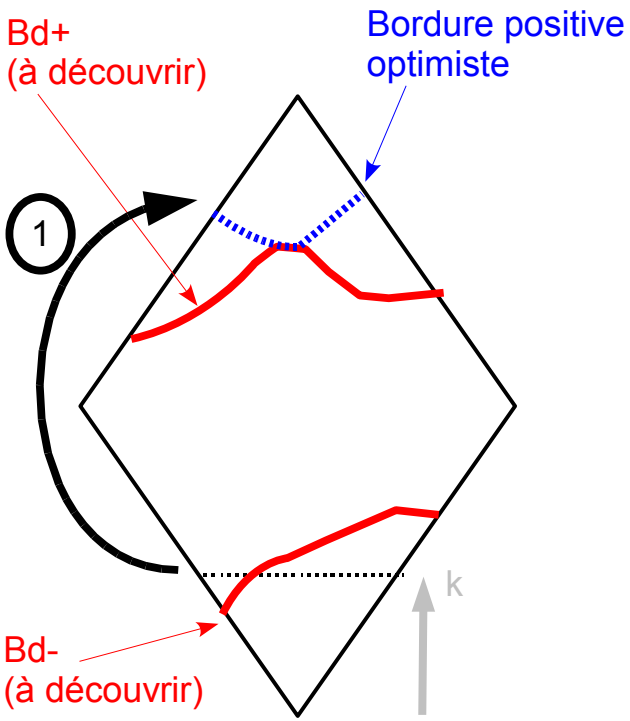
  si proche de 1, dualisation
  sinon continue l'approche par niveaux

ABS : fonctionnement global (1/3)



- Un début de parcours par niveaux
- Niveau k d'arrêt fixé dynamiquement à partir des données
- Découverte d'une partie de la bordure négative (et éventuellement une partie de la bordure positive)

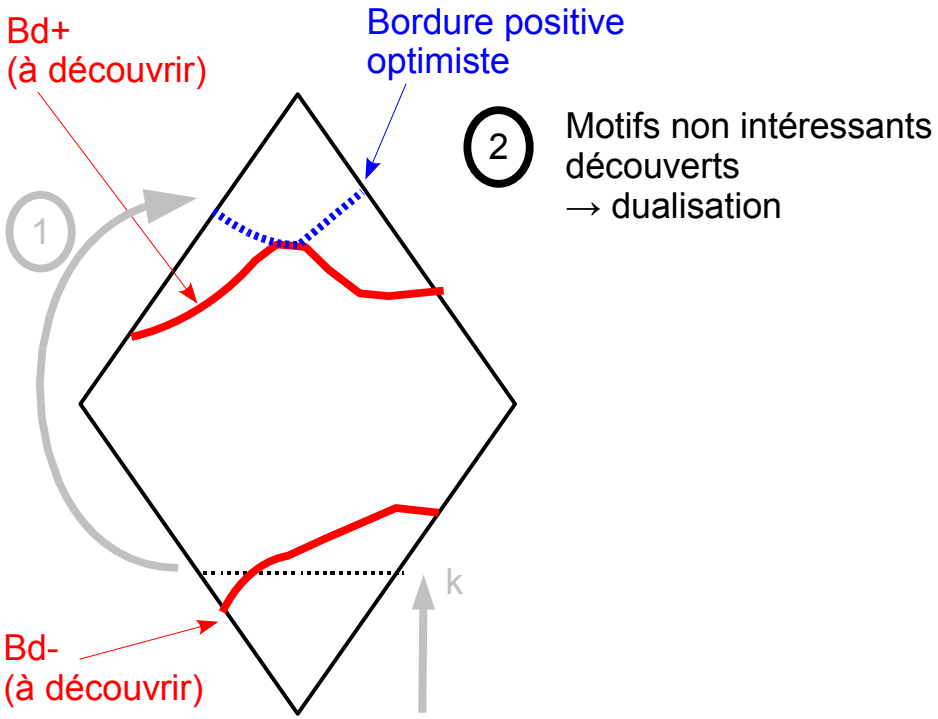
ABS : fonctionnement global (2/3)



Alternance de dualisations entre les deux bordures

1. dualisation de Bd^- en construction vers Bd^+

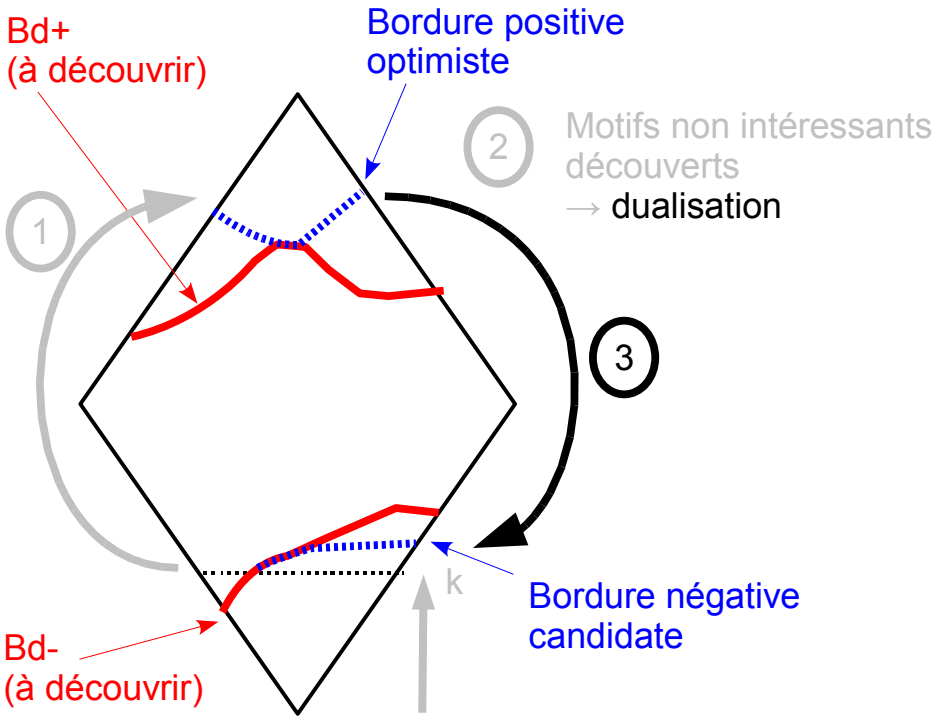
ABS : fonctionnement global (2/3)



Alternance de dualisations entre les deux bordures

1. dualisation de Bd^- en construction vers Bd^+
2. si découverte de motifs non intéressants

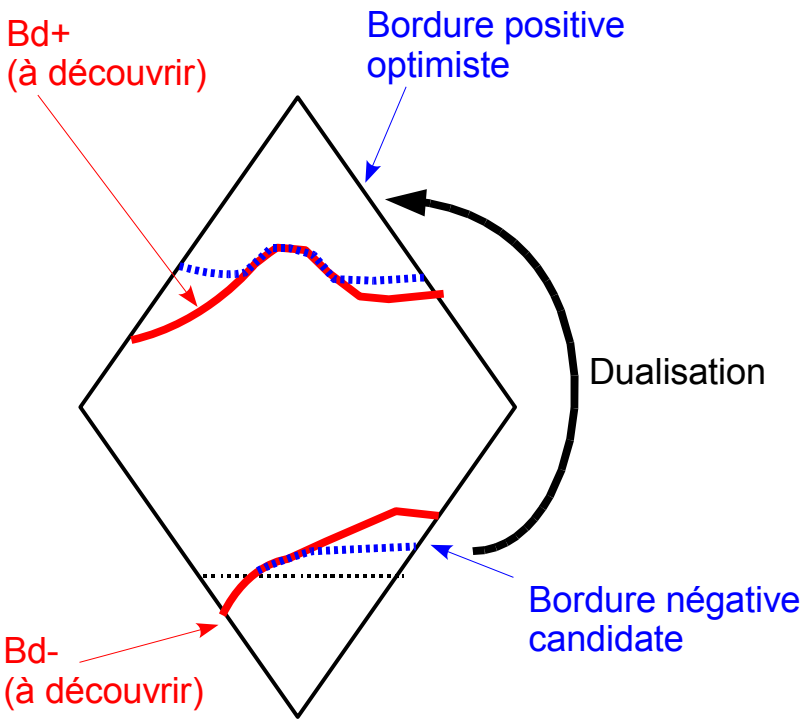
ABS : fonctionnement global (2/3)



Alternance de dualisations entre les deux bordures

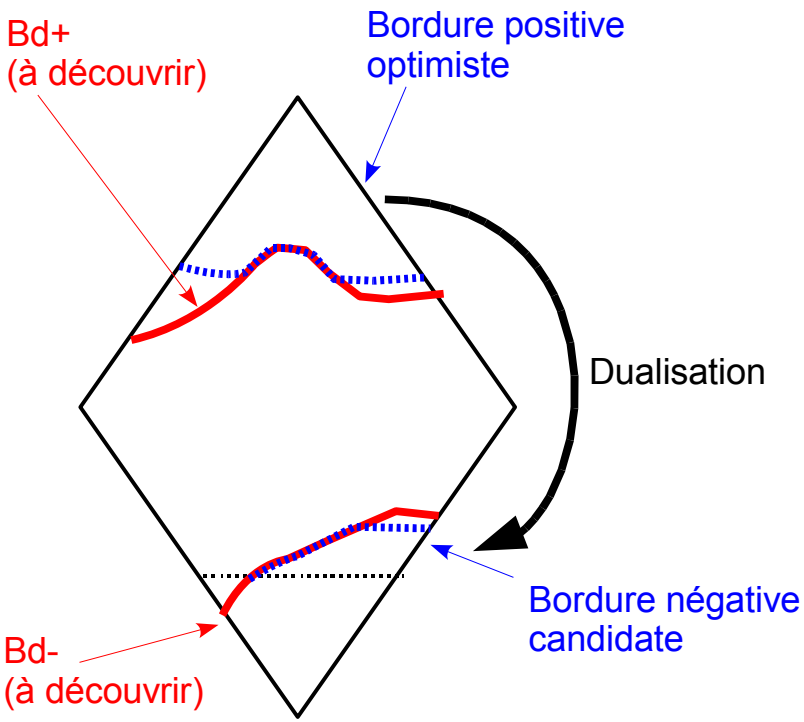
1. dualisation de Bd^- en construction vers Bd^+
2. si découverte de motifs non intéressants
3. ⇨ dualisation de Bd^+ en construction vers Bd^-

ABS : fonctionnement global (3/3)



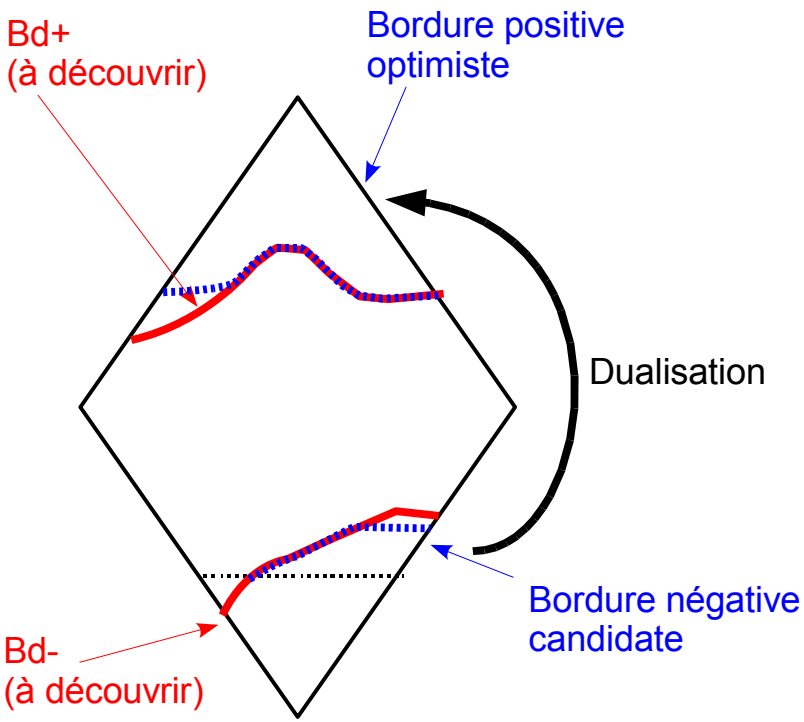
Alternance de dualisations jusqu'à découverte des deux bordures

ABS : fonctionnement global (3/3)



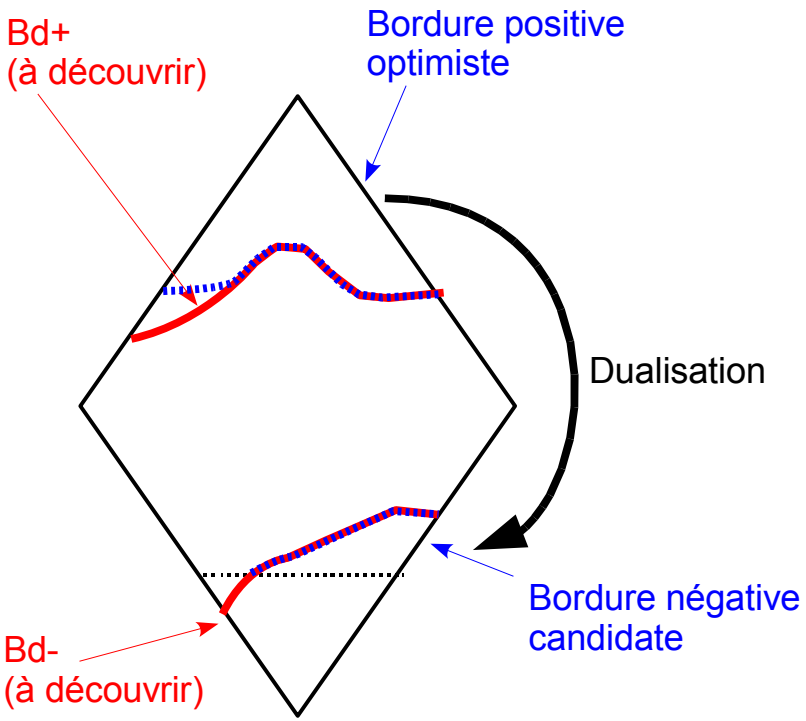
Alternance de dualisations jusqu'à découverte des deux bordures

ABS : fonctionnement global (3/3)



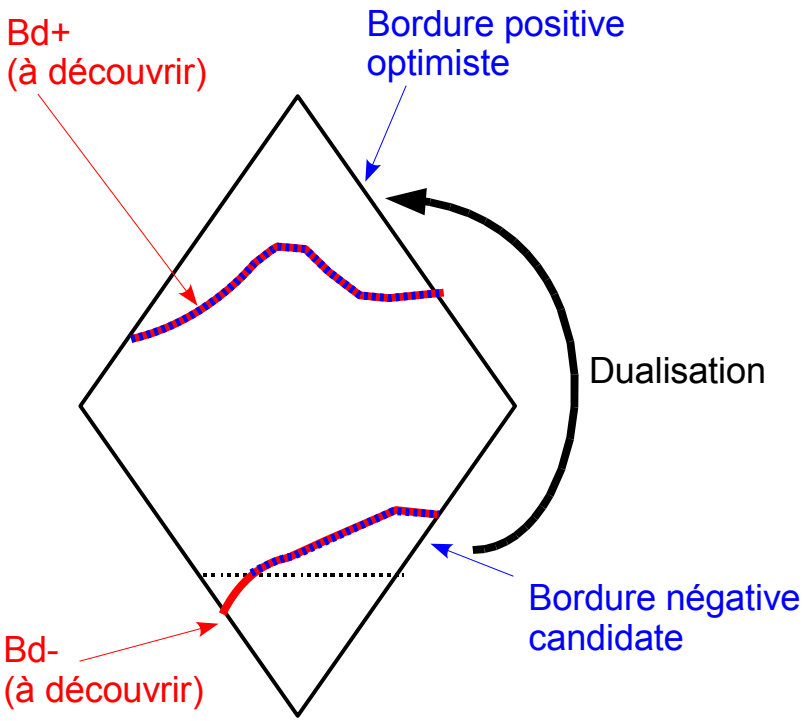
Alternance de dualisations jusqu'à découverte des deux bordures

ABS : fonctionnement global (3/3)



Alternance de dualisations jusqu'à découverte des deux bordures

ABS : fonctionnement global (3/3)



Alternance de dualisations jusqu'à découverte des deux bordures

Rq : nombre de dualisations dépend du niveau du plus "grand" motif de Bd^-

Résultats expérimentaux

Expérimentation pour le problème des ensembles fréquents

- jeux de données, implémentations et bancs d'essais disponibles

⇨Participation au banc d'essais de FIMI 2004 (Flouvat, De Marchi et Petit, FIMI 2004)

- moins bon que les meilleurs algorithmes
  - algorithmes et implémentations fortement optimisés
  - ⇨ difficilement applicables à un autre problème du cadre
- meilleur que les autres algorithmes du cadre théorique (*Apriori* et *IBE*)

⇨Efficace pour des jeux de données de type I et II

Synthèse : Mise en place d'un algorithme adaptatif et générique pour l'extraction de motifs intéressants

Etude de l'influence des données sur les stratégies

- ⇨ nouvelle caractérisation et classification des données basée sur la distribution des bordures

Proposition de l'algorithme *ABS*

- ⇨ adaptatif et générique à tout problème du cadre

⇨Question : Exploitation en pratique du cadre ? outils logiciels génériques ?

Outline

Vers des outils génériques pour l'extraction de motifs

Très difficile d'exploiter en pratique ce cadre théorique

Problèmes :

- implémentations existantes dédiées à un problème donné
- implémentations opaques car très optimisées
- pas de découplage entre données et algorithmes

⇨Adaptation quasiment impossible à un autre problème du cadre

Objectif : Proposer une implémentation générique du cadre

- offrir un bon compromis entre efficacité et rapidité de développement

Contribution : librairie générique de composants C++

Principale contribution allant dans ce sens

Librairie DMTL (*Data Mining Template Library*) Zaki et al. ICFCA 2005

- problèmes d'extraction de motifs fréquents
  - ensembles, séquences, arbres et graphes
- extensible : possibilité de traiter de nouveaux types de motifs

Limite :

- se focalise sur le prédicat "être fréquent" : pas de souplesse sur la définition du problème

Implémentation du cadre théorique de Mannila et Toivonen

➪Proposition d'une librairie C++ structurée d'après le cadre (Flouvat, De Marchi et Petit, BDA 2006)

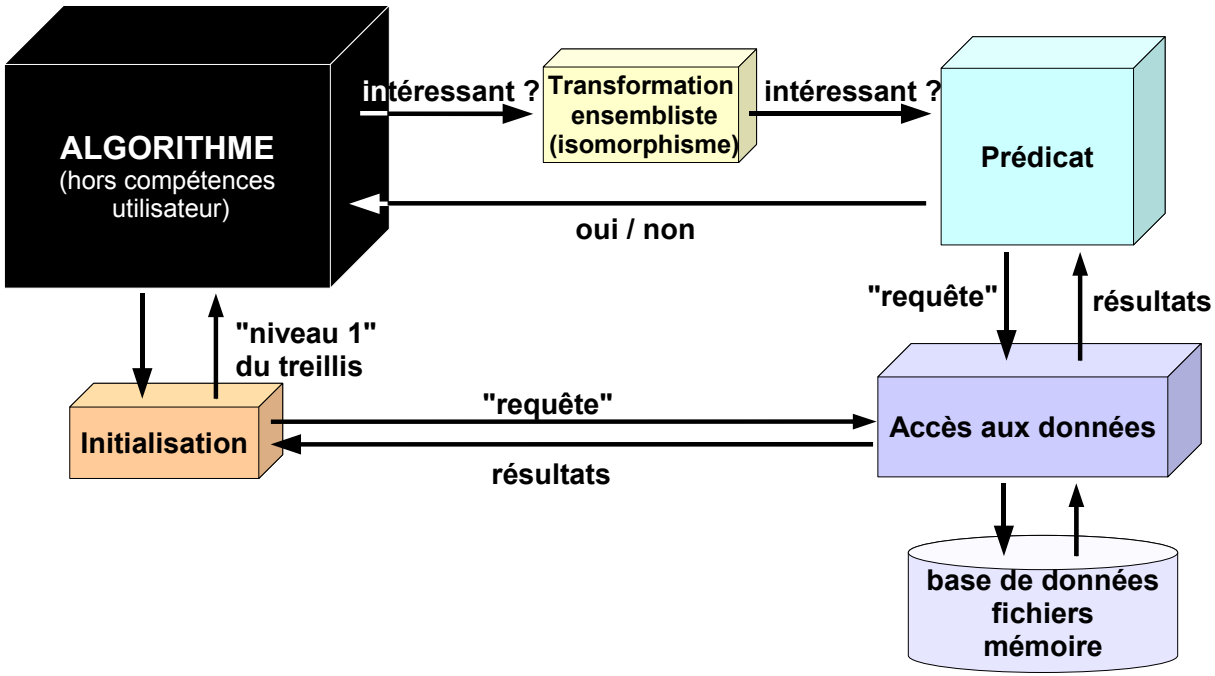


FIG. : Organisation et fonctionnement de la librairie

➪Algorithme transparent pour l'utilisateur

➪Découplage entre données et algorithme

Aspects méthodologiques

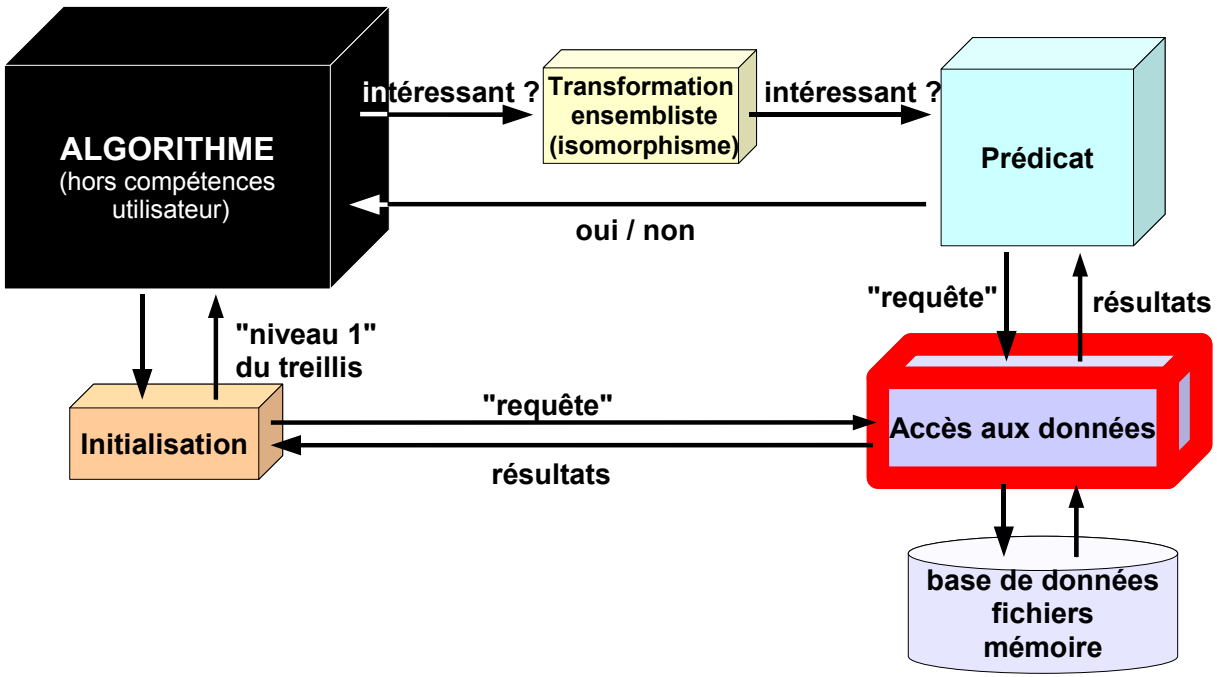
Les questions à se poser avant d'utiliser la librairie

- mon problème rentre-t-il dans le cadre ?
  - quels sont les motifs ? la relation d'ordre ? représentation ensembliste ?
  - quel est le prédicat ? (anti-)monotone ?
  - quelle est la solution ?

- quel algorithme choisir ? cela dépend
  - des caractéristiques des données, des problèmes
  - de la solution

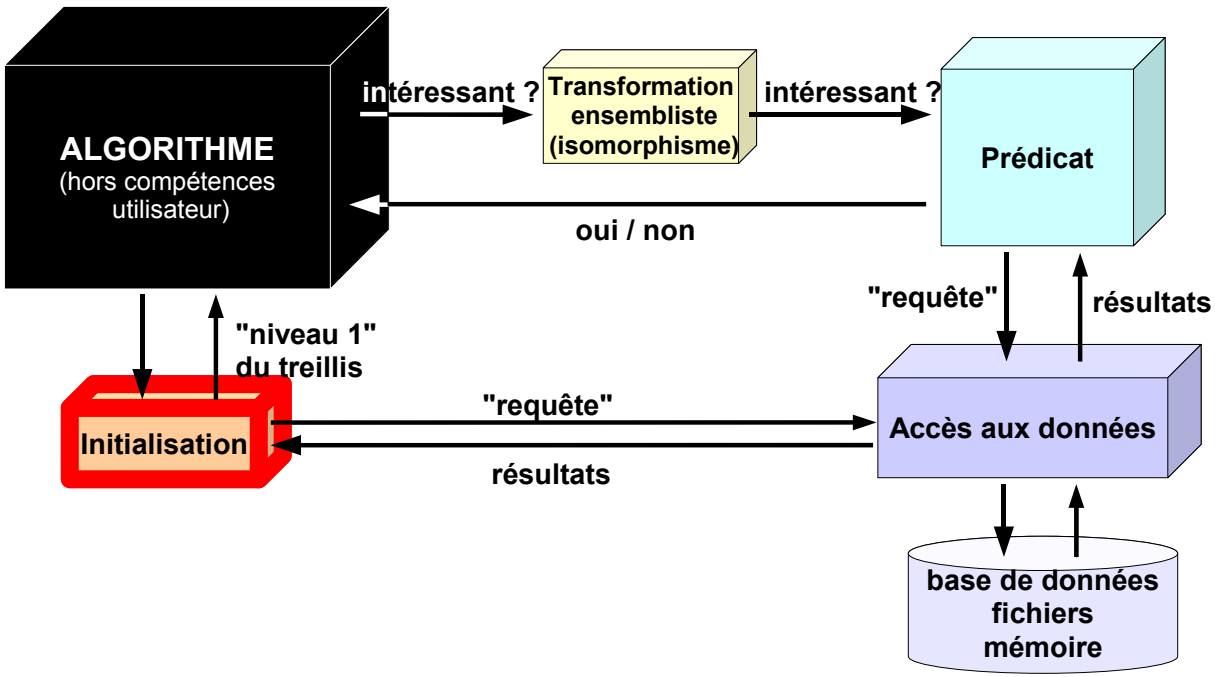
Développer de nouveaux composants - intégrer de nouveaux problèmes

- définir accès aux données
- définir la fonction d'initialisation du treillis
- définir la fonction de transformation ensembliste
- définir le prédicat



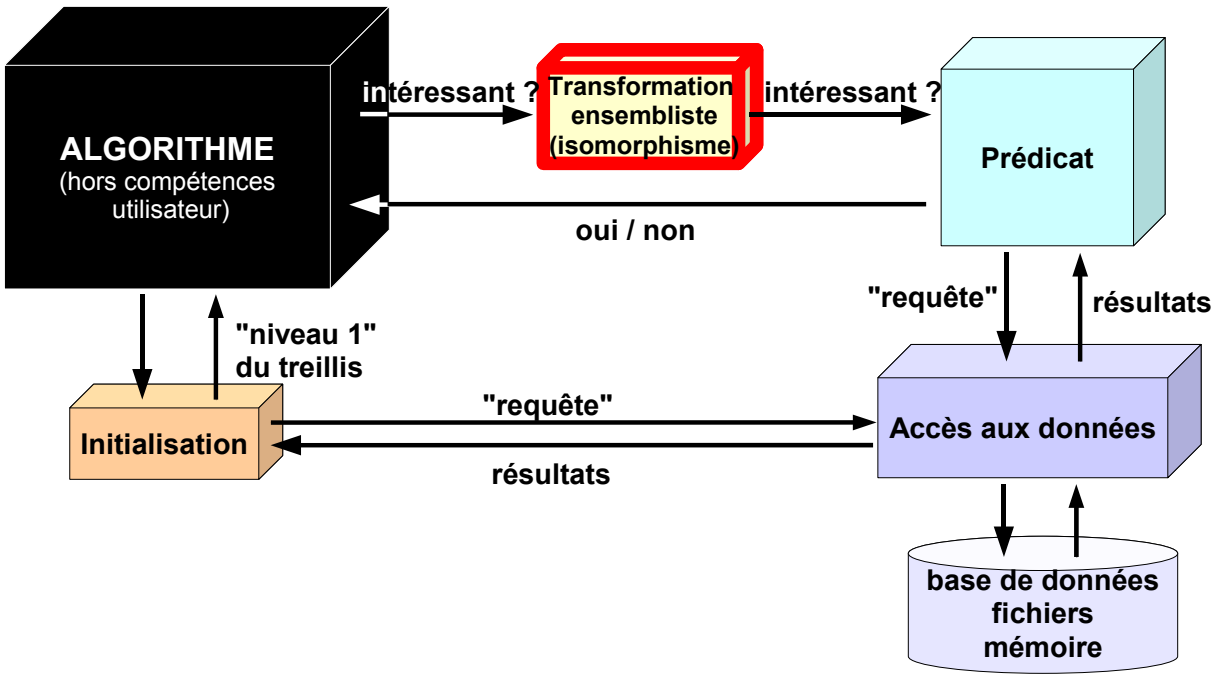
Développer de nouveaux composants - intégrer de nouveaux problèmes

- définir accès aux données
- définir la fonction d'initialisation du treillis
- définir la fonction de transformation ensembliste
- définir le prédicat



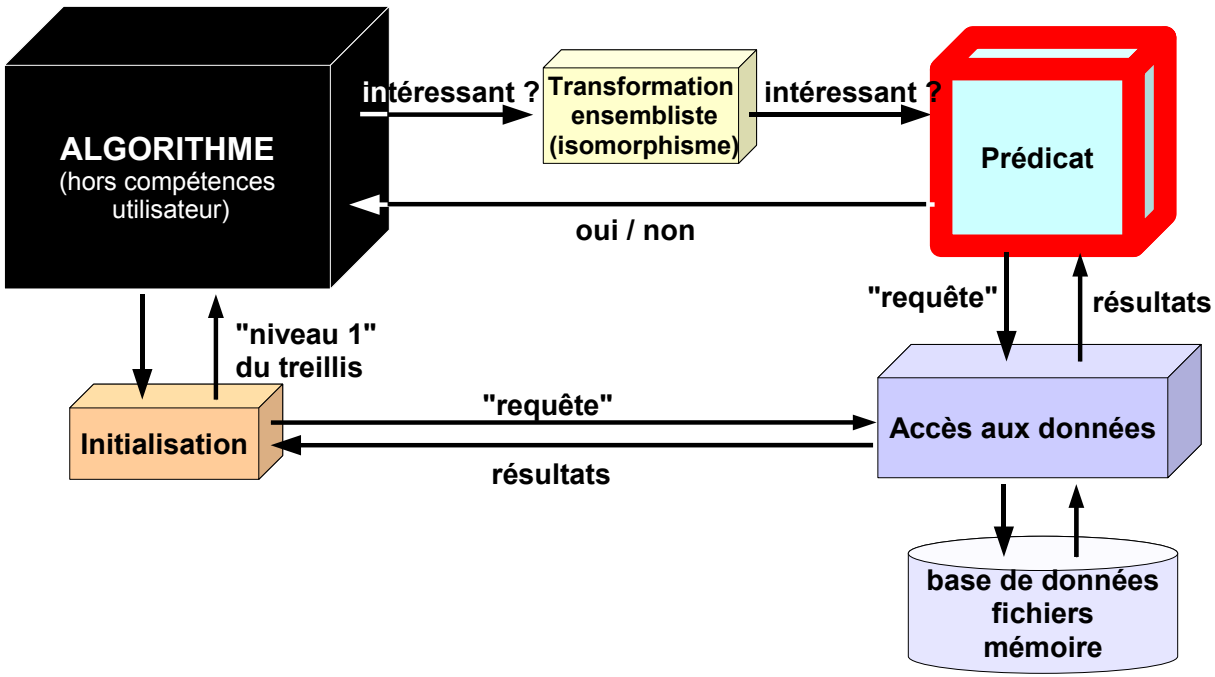
Développer de nouveaux composants - intégrer de nouveaux problèmes

- définir accès aux données
- définir la fonction d'initialisation du treillis
- définir la fonction de transformation ensembliste
- définir le prédicat



Développer de nouveaux composants - intégrer de nouveaux problèmes

- définir accès aux données
- définir la fonction d'initialisation du treillis
- définir la fonction de transformation ensembliste
- définir le prédicat



La librairie : une boîte à outils de composants

Algorithmes : stratégie par niveaux, *ABS*, et deux variantes changeant l'exploration de l'espace de recherche

- solutions : théorie, bordures positive et/ou négative
- contraintes monotones ou anti-monotones
- possibilité d'ajouter des optimisations spécifiques au problème étudié

Sources de données :

- en mémoire dans des structures de données
- dans des fichiers
- dans un SGBD (MySQL)

Problèmes déjà intégrés :

- extraction d'ensembles fréquents et essentiels fréquents
- extraction des clés d'une relation
- extraction des DI satisfaites entre deux relations

Expérimentations

Difficile d'estimer le temps de développement...

- à titre d'indication, développement d'une solution pour le problème des clés d'une relation en quelques heures

Performances ?

Etude du temps d'exécution pour le problème des ensembles fréquents

- comparaison avec deux implémentations optimisées d'*Apriori*, dont la plus performante actuellement

⇨ performances moins bonnes que la meilleure implémentation d'*Apriori*, mais comparables à la deuxième pourtant plus optimisée

Outline

Conclusion

Nouvelle caractérisation et classification des données fondée sur la distribution des bordures

- stable
- en accord avec les performances des algorithmes

Proposition d'un nouvel algorithme, *ABS*, adaptatif dans le cadre théorique de Mannila et Toivonen

- alterne stratégie par niveaux et dualisations
- change dynamiquement de stratégie en fonction des données
- meilleur que les autres algorithmes du cadre

Développement d'une librairie pour l'extraction de motifs intéressants

- implémentation d'un cadre théorique
- simplicité d'utilisation et performances encourageantes

Quelques perspectives

Etudier d'un point de vue théorique les observations faites lors des expérimentations

- stabilité des distributions par rapport au changement de seuil de support
- interactions entre les distributions des deux bordures

Améliorer la stratégie adaptative d'*ABS*

- quelle stratégie intégrer pour être aussi efficace pour les problèmes/données de type III ? en existe-t-il ?

Développer une version déclarative de la librairie évitant à l'utilisateur d'avoir à implémenter

- définition/utilisation d'un langage de requêtes
- mise en place d'un modèle de coûts pour choisir automatiquement l'algorithme le plus adapté

Merci de votre attention