Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier
LIRMM

HYPE

HYPE: Hierarchical Sequential Pattern Mining

MARC PLANTEVIT, ANNE LAURENT, MAGUELONNE TEISSEIRE

LIRMM, UNIVERSITY MONTPELLIER II, FRANCE

DOLAP'06, Arlington, November the 10^{th} 2006

Plan

Plan

OLAP & KDD

User Navigation

- OLAP users are now decision makers.
- Users navigate in the aggregated datacube in order to discover knowledge.
  - ROLL UP, DRILL DOWN, ...

Our Goal:

Providing automatically knowledge thanks to data mining approaches

Sequential Patterns

- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

Sequential Patterns

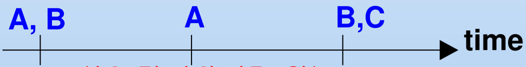
- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

Sequential Patterns

- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

Sequential Patterns

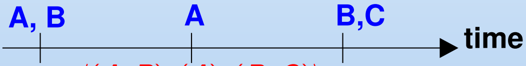
- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .



$$\langle (A, B), (A), (B, C) \rangle$$

Sequential Patterns

- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

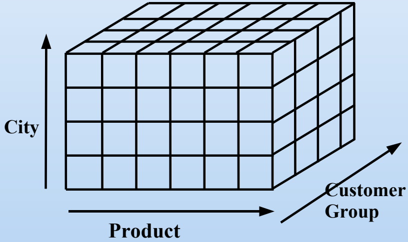


$$\langle (A, B), (A), (B, C) \rangle$$

- ☹: Sequential patterns are quite poor (only one mined dimension)

Data Cube

- Knowledge are mined among one dimension: *product* dimension.
- What about the other ones ?



Combining Several Analysis Dimensions

- Items are not defined on one dimension, they are defined on several dimensions

Combining Several Analysis Dimensions

- Items are not defined on one dimension, they are defined on several dimensions
- Classical item: c

Combining Several Analysis Dimensions

- Items are not defined on one dimension, they are defined on several dimensions
- Classical item: c
- Multidimensional item: $(France, c, 100), (Germany, c, *)$

Combining Several Analysis Dimensions

- Items are not defined on one dimension, they are defined on several dimensions
- Classical item: c
- Multidimensional item: $(France, c, 100), (Germany, c, *)$
- Multidimensional sequence:

$$\langle\{(France, c, 100), (Germany, d, 54)\}\{(*, b, 2)\}\rangle$$

instead of

$$\langle (c, d), b \rangle$$

Taking Hierarchies into account

dilemma Support/#patterns

- Minimal support too high: too few frequent knowledge to be used and to enhance the decision making process.
- Minimal support too low: too much frequent knowledge, unusable for the decision maker.

It is very difficult to choose the right support value for mining relevant knowledge

Taking Hierarchies into account

dilemma Support/#patterns

- Minimal support too high: too few frequent knowledge to be used and to enhance the decision making process.
- Minimal support too low: too much frequent knowledge, unusable for the decision maker.

It is very difficult to choose the right support value for mining relevant knowledge

Taking hierarchies into account to solve this dilemma

- Mining rules on several levels of hierarchy.
- subsumption power.

State-of-the-art

	(1)	(2)	(3)
Multidimensionality	No	No	Yes
Simulation of multi.	No	??	___
Sequential patterns	Yes	No	Yes
Hierarchy in patterns	Several	Single	No

- (1) Agrawal & Srikant (1995): the pioneer approach.
- (2) Han & Fu (2001): an original approach.
- (3) Yu & Chen (2005): Using hierarchies for a smart time representation.

No approach for mining **multidimensional** sequences among **several** levels of hierarchy

Plan

Database & Blocks

BLOCK :

- A database can be partioned into different blocks according to some dimensions

Market	Cust-Grp	Date	Place	Product
Carrefour	Educ.	1	Germany	beer
Carrefour	Educ.	1	Germany	pretzel
Carrefour	Educ.	2	Germany	M2
Carrefour	Educ.	3	Germany	chocolate
Carrefour	Educ.	4	Germany	M1
Carrefour	Employ.	1	France	soda
Carrefour	Employ.	2	France	wine
Carrefour	Employ.	2	France	chocolate
Carrefour	Employ.	3	France	M2
wellmart	retir.	1	UK	whisky
wellmart	retir.	1	UK	pretzel
wellmart	retir.	2	UK	M2
wellmart	Educ.	1	LA	chocolate
wellmart	Educ.	2	LA	M1
wellmart	Educ.	3	NY	whisky
wellmart	Educ.	4	NY	soda

LIRMM
Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

Dimension Set Partition

$$D = \mathscr{D}_{\mathscr{R}} \oplus \mathscr{D}_{\mathscr{A}} \oplus \mathscr{D}_t$$

- D_t: temporal dimensions
- D_A: analysis dimensions
- D_R: reference dimensions

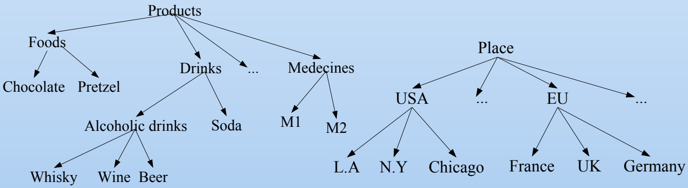
tuple $c = (d_1, \cdots, d_n) = (r, a, t)$ where :

- r: is the restriction of c on $\mathscr{D}_{\mathscr{R}}$
- a: is the restriction of c on $\mathscr{D}_{\mathscr{A}}$
- t: is the restriction of c on $\mathscr{D}_t$

Hierarchies

- Hierarchical relations on each analysis dimensions.
- These relations are materialized in the form of **trees** (Is-a relation).
- Only the leaves can be in the database.

Hierarchies over PLACE and PRODUCT dimensions:



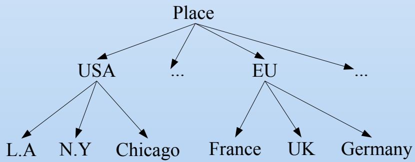
Relation Between Elements

ancestor: $\hat{x}$ is an ancestor of x according to the hierarchy.

descendant: denoted $\check{x}$.

$$E.U = \widehat{France}$$

$$Place = \widehat{Germany}$$



H-generalized(H.G.) Item, Itemset and Sequence

H.G. Multidimensional Item:

A tuple $e = (d_1, \ldots, d_m)$ defined over the set of the analysis dimensions D_A such that $d_i \in \{label(T_i)\}$.

Examples : $(France, Chocolate)$,
$(Germany, Drinks)$

H-generalized(H.G.) Item, Itemset and Sequence

H.G. Multidimensional Item:

A tuple $e = (d_1, \ldots, d_m)$ defined over the set of the analysis dimensions D_A such that $d_i \in \{label(T_i)\}$.

Examples : $(France, Chocolate)$,
$(Germany, Drinks)$

Hierarchical Inclusion

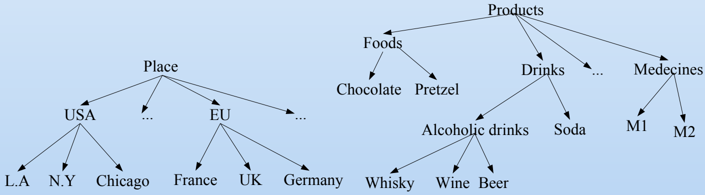
Let $e = (d_1, \ldots, d_m)$ and $e' = (d'_1, \ldots, d'_m)$, then:

- e is **more general** than e' ($e >_h e'$) if $\forall d_i, d_i = \hat{d}'_i$ or $d_i = d'_i$
- e is **more specific** than e' ($e <_h e'$) if $\forall d_i, d_i = \check{d}'_i$ or $d_i = d'_i$
- e and e' are **incomparable** if there is no relation between them ($e \not>_h e'$ et $e' \not>_h e$)

Hierarchical Relations

Example: Relation between items

- $(USA, Drinks) >_h (USA, soda)$.
- $(France, wine) <_h (EU, Alcoholic\ drinks)$.
- $(France, wine)$ and $(USA, soda)$ are incomparable.



H-generalized Itemset and Sequence

H.G. Multidimensional Itemset:

$i = \{e_1, \ldots, e_k\}$ where all items are all incomparable.

$\{(France, wine), (USA, soda)\}$ YES
$\{(France, wine), (EU, Alcohol.D.)\}$ NO:
$(France, wine) <_h (EU, Alcohol.D.)$

H.G. Multidimensional Sequence

$s = \langle i_1, \ldots, i_j \rangle$ is an ordered list of h-generalized itemsets.

$\langle \{(France, wine), (USA, soda)\}\{(Germany, beer)\} \rangle$

Support

Item Supported by a tuple

A tuple b supports an item e if $\Pi_{D_A}(b) <_h e$.

Tuple $(Carrefour, Educ, 1, France, wine)$ supports the item $(EU, Alcohol.D.)$.

Sequence Supported by a Block

A block supports a sequence $\langle i_1, \ldots, i_l \rangle$ if $\forall j = 1 \ldots l, \ \ \exists d_j \in Dom(D_j)$, for each item e from i_j, $\exists t = (r, e, d_j)$ *or* $t = (r, \check{e}, d_j) \in T$ w.r.t $d_1 < d_2 < \ldots < d_l$.

Support of a Sequence

Let :

- D_R: the reference dimensions
- DB: the database partioned into a set of block B_{T,D_R}
- A sequence S

Support of S

$$support(S) = \frac{|\{B \in B_{DB,D_R} \ s.t.\ B \ supports \ S\}|}{|B_{DB,D_R}|}$$

Example

D$_R = \{Market, Cust - Grp\}$, $D_A = \{Place,\ Product\}$ et $D_T = \{Date\}$, $minsupp = 2$
Let us search the support of the sequence
$S = \langle\{(UE, pretzel), (UE, Alcoholic\ Drinks)\}\{(UE, M2)\}\rangle$

Market	Cust-grp	Date	Place	Product
Carrefour	Educ.	1	Germany	beer
Carrefour	Educ.	1	Germany	pretzel
Carrefour	Educ.	2	Germany	M2
Carrefour	Educ.	3	Germany	chocolate
Carrefour	Educ.	4	Germany	M1
Carrefour	Employ.	1	France	soda
Carrefour	Employ.	2	France	wine
Carrefour	Employ.	2	France	chocolate
Carrefour	Employ.	3	France	M2
wellmart	retir.	1	UK	whisky
wellmart	retir.	1	UK	pretzel
wellmart	retir.	2	UK	M2
wellmart	Educ.	1	LA	chocolate
wellmart	Educ.	2	LA	M1
wellmart	Educ.	3	NY	whisky
wellmart	Educ.	4	NY	soda

$\langle\{(UE, pretzel), (UE, Alcohol.D.)\}\{(UE, M2)\}\rangle$

Block 1

Carrefour	Educ.	1	**Germany**	**pretzel**
Carrefour	Educ.	1	**Germany**	**beer**
Carrefour	Educ.	2	**Germany**	**M2**
Carrefour	Educ.	3	Germany	chocolate
Carrefour	Educ.	4	Germany	M1

Block 1 supports S: $support(S) + +$

Block 2

Carrefour	Employ.	1	France	soda
Carrefour	Employ.	2	**France**	**pretzel**
Carrefour	Employ.	2	**France**	**wine**
Carrefour	Employ.	3	**France**	**M2**

Block 2 supports S: $support(S) + +$

$\langle\{(UE, pretzel), (UE, Alcohol.D.)\}\{(UE, M2)\}\rangle$

Block 3

wellmart	retir.	1	**UK**	**pretzel**
wellmart	retir.	1	**UK**	**whisky**
wellmart	retir.	2	**UK**	**M2**

Block 3 supports S: $support(S) + +$

Block 4

wellmart	Educ.	1	LA	chocolate
wellmart	Educ.	2	LA	M1
wellmart	Educ.	3	NY	whisky
wellmart	Educ.	4	NY	soda

Block 4 does not support S

$\langle\{(UE, pretzel), (UE, Alcohol.D.)\}\{(UE, M2)\}\rangle$

Block 3

wellmart	retir.	1	**UK**	**pretzel**
wellmart	retir.	1	**UK**	**whisky**
wellmart	retir.	2	**UK**	**M2**

Block 3 supports S: $support(S) + +$

Block 4

wellmart	Educ.	1	LA	chocolate
wellmart	Educ.	2	LA	M1
wellmart	Educ.	3	NY	whisky
wellmart	Educ.	4	NY	soda

Block 4 does not support S

$support(S) = 3 \geq minsupp$

- s is frequent

Overview

Frequent Item Generation

- to mine all the **maximaly specified items.**
- levelwise generation

Frequent Sequence Generation

- anti-monotonicity property of the support
- Apriori-like method (generate and prune)
- use of prefix tree to store the sequences (PSP)

Support Count:

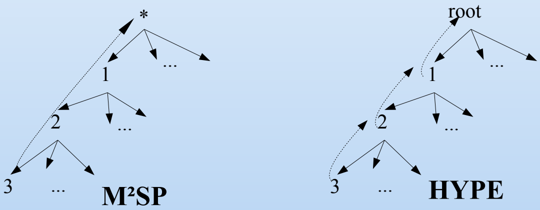
- required: data pre-processing (group by *date*, $D_1, \cdots, D_n$)
- Support Sequence Count: **SupportCount(**$s, DB, \mathscr{D}_{\mathscr{R}}$**)**
  - For each block : **SupportBlock(**s, B**)**
- **anchoring** operation ($\sigma_{condition}(B) \longmapsto C'$ with $B' \subseteq B$)

complexity

- n_B: # tuples in B
- $m = |D_A|$: # analysis dimensions
- P_{max}: maximal depth of the hierarchies
- SupportBlock:$O(P_{max} \times n_B \times m \times \log n_B)$
- SupportCount:$O(l \times P_{max} \times n_{max} \times m \times \log n_{max})$

Why Hype provides a better way to manage star

- (M^2SP): a "binary" management of wild-card value.
- *HYPE*: A better management thanks to the taking hierarchies into account.
- More Accurate knowledge.

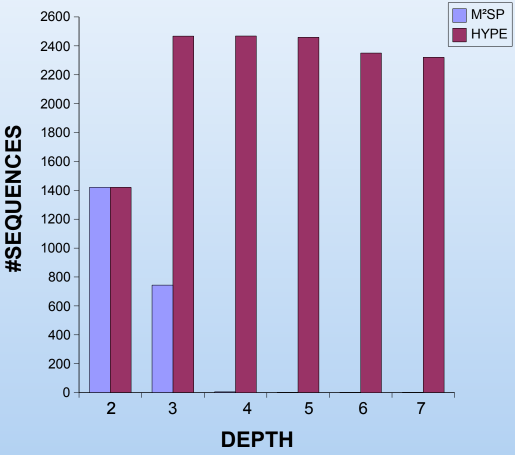


Experimental Protocol

- Synthetic database
- $10,000$ tuples
- $|D_A| = 5$
- Studying # frequent sequences according to:
  - Hierarchy depth (specialization degree)
  - user-defined minimal threshold
- Comparison with M^{2}SP(-α):

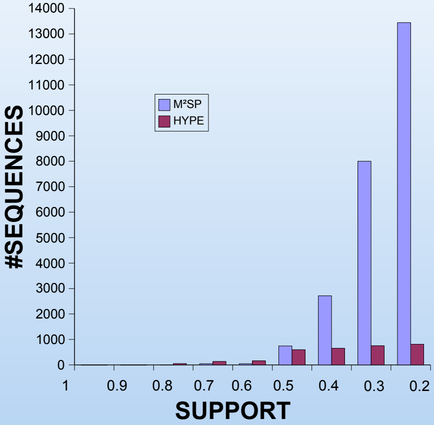
frequent sequences over Hierarchy depth:

- minsup=0.3, average degree = 3



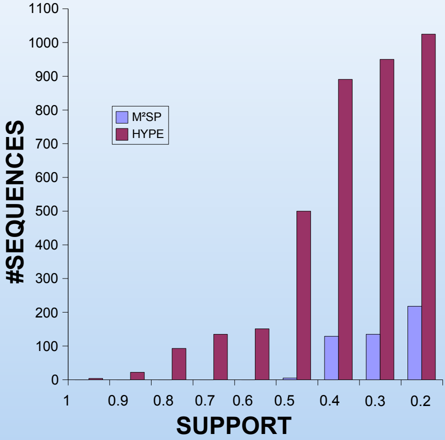
frequent sequences over minimal support:

- Dense Database (lower degree)



frequent sequences over minimal support:

- Sparse Database (higher degree)



Plan

Managing hierarchies in multidimensional sequential pattern mining

- H.G. multidimensional sequential patterns.
- More accurate knowledge.
- An approach more efficient according to "min. support/mined knowledge dilemma" thanks to subsumption ability.

	Agrawal	Jiawei Han	Yu	HYPE
Multidimensionality	No	No	Yes	Yes
Simulation of multi.	No	??	___	Yes
Sequential patterns	Yes	No	Yes	Yes
Hierarchy in patterns	Several	Single	No	Yes

Future Work

Future Work

- Use of *condensed* representation (closed, free patterns).
- Another definition of support to better fit the OLAP framework.
- Modular hierarchy management in order to enhance customized and focused knowledge discovery.

	Agrawal	Jiawei Han	Yu	HYPE
Multidimensionality	No	No	Yes	Yes
Simulation of multi.	No	??	__	Yes
Sequential patterns	Yes	No	Yes	Yes
Hierarchy in patterns	Several	Single	No	Yes
Condensed Representation	No	No	No	No
Counting sup. with aggregates	No	No	No	No
User interaction	No	No	No	Not enough

References

R. Agrawal and R. Srikant.
Mining sequential patterns.
In *Proc. 1995 Int. Conf. Data Engineering (ICDE'95)*, pages 3–14, 1995.

J. Han and Y. Fu.
Mining multiple-level association rules in large databases.
IEEE Trans. Knowl. Data Eng., 11(5):798–804, 1999.

M. Plantevit, Y. W. Choong, A. Laurent, D. Laurent, and M. Teisseire.
M^2SP: Mining sequential patterns among several dimensions.
In *Proc. 2005 PKDD*, pages 205–216, 2005.

C.-C. Yu and Y.-L. Chen.
Mining sequential patterns from multidimensional sequence data.
IEEE Transactions on Knowledge and Data Engineering, 17(1):136–140, 2005.