Experimental study of frequent itemsets datasets

Frédéric Flouvat[1]

Research Report

June 2005

[1] flouvat@isima.fr

Abstract

The discovery of frequent patterns is a famous problem in data mining. While plenty of algorithms have been proposed during the last decade, only a few contributions have tried to understand the influence of datasets on the algorithms behavior. Being able to explain why certain algorithms are likely to perform very well or very poorly on some datasets is still an open question.
In this setting, we describe a thorough experimental study of datasets with respect to frequent itemsets. In addition to frequent itemsets, three other classical representations are considered: frequent closed, frequent free and frequent essential itemsets. For each of them, the collection size for each level (from singletons to the size of the largest sets) is given and moreover, their negative and positive borders are given whenever possible.
From this analysis, we exhibit some properties of datasets and some invariants allowing to better understand and predict the behavior of well known algorithms.
The main perspective of this work is to devise adaptive algorithms with respect to dataset characteristics.

Keywords
Data mining, frequent itemsets, datasets, experimental study, borders

Résumé

Le problème de l'énumération de la bordure positive des motifs fréquents dans une base de données transactionnelle est un problème classique en fouille de données. Alors que beaucoup d'algorithmes ont été proposés pendant la dernière décennie, seulement quelques contributions ont essayé de comprendre l'influence des jeux de données sur le comportement des algorithmes. Pouvoir expliquer pourquoi certains algorithmes sont susceptibles d'être efficaces ou pas sur certains jeux de données est toujours une question en suspens.
Dans ce contexte, nous décrivons une étude expérimentale détaillée des jeux de données utilisés pour les motifs fréquents. En plus des motifs fréquents, trois autres représentations classiques sont considérées: les motifs fermés fréquents, les motifs clés fréquents et essentiels fréquents. Pour chacune d'elles, la taille de la collection pour chaque niveau (des singletons aux plus grands ensembles) est étudiée et, leurs bordures négative et positive sont présentées autant que possible.
A partir de cette analyse, nous mettons en avant certaines caractéristiques des jeux de données et quelques invariants permettant de mieux comprendre et prédire le comportement des algorithmes.
La principale perspective de ce travail est de permettre la conception d'algorithmes adaptatifs par rapport aux données.

Mots-clés
Fouille de données, motifs fréquents, jeux de données, étude expérimentale, bordures

1. Introduction

The discovery of frequent patterns is a famous problem in data mining, introduced in [AGR93] as a first step for mining association rules. While plenty of algorithms have been proposed during the last decade [AGR94, BUR01, GOU01, GRA03], only a few contributions have tried to understand the influence of dataset characteristics on the algorithms behavior [GOE03, GOU01, PAL04].
These studies focus on the number of transactions, average length of transactions, or frequent itemsets distribution, i.e. statistics from frequent itemsets and maximal frequent itemsets are usually given. Nevertheless algorithms could have quite different behaviors for (apparently) similar datasets.
Benchmarks comparing algorithms performances have been done on real and synthetic datasets [BOE03, BAY04] (see FIMI website [GOE04]). As an example, consider the maximal frequent itemset discovery: many algorithms have been proposed [BAY98, BUR01, FLO04] and many datasets are freely available. Even with all these informations, being able to explain why certain algorithms are likely to perform very well or very poorly on some datasets is still an open question.

More generally, studying datasets can provide useful hints for devising adaptive algorithms [FLO04, ORL03], i.e. algorithms which adapt themselves to data characteristics in order to increase their time or memory efficiency. Adaptive behavior of algorithms is not new in the setting of frequent itemsets mining, for example many contributions [BOR03, BUR03] use heuristics to decide when tries-like data structure have to be rebuilt for datasets and/or itemset collection.
The promising results obtained by these algorithms show the interest of applying specific strategies according to dataset features.

Another key point is that some problems have specific invariant characteristics, whatever the datasets studied. Knowledge of datasets and their impact on algorithms could give useful information about the difficulty to solve these problems while giving hints on the more appropriate strategies to cope with these difficulties.

Related works: Classical characteristics of datasets were studied in [GOU01], and more particularly a density criteria. Up to our knowledge no formal definition of density does exist. According to [GOU01], a dataset is *dense* when it produces many long frequent itemsets even for high values of minimum support threshold. The authors studied seven datasets, each of them capturing a fairly large range of typical uses. The result of these experimentations is a classification of datasets in four categories according to the density. The density is estimated by using the characteristics of maximal frequent itemsets.

The main problem of this classification is that it depends on the minimum support threshold. For example, for a given minimum support threshold, a dataset could be of the first category, and for another minimum support threshold of the second category. As a concrete example, this case arises with Pumsb* and a minimum support threshold sets up to 15% and 25% respectively (see experimental study section).
Moreover, there is no clear relationship between the proposed classification and algorithms performances. Even worse, a surprising result was obtained in the last FIMI workshop [BAY04]: algorithms seem to be more efficient on some very dense datasets than on some other sparser datasets.

Based on the works done in [GOU01], [PAL04] proposed a statistical property of transactional datasets to characterize dataset density. Actually, they consider the dataset as a transaction source and measure an entropy signal, i.e. the transactions produced by such a source. Moreover, they show how such a characterization can be used in many fields, from performance prediction, support range determination, sampling, to strategy decisions. As for the previous work, it does not explain algorithms performances anymore. This may be due to the fact that only frequent itemsets are used to calculate the entropy measure.

In [MAN03], the positive border distribution (i.e. the number of elements in each level) is considered as a key parameter to characterize transaction databases. It is proved that any distribution is "feasible", and thus susceptible to be met in practice. Moreover, a constructive theorem is proposed to compute a synthetic transaction database given a positive border distribution as input. Nevertheless, the negative border is never considered and as a result, such synthetic databases do not reflect in general the "complexity" of real-world datasets.

Contribution: In this setting, we describe a thorough experimental study of datasets with respect to frequent itemsets. We study the distribution of frequent itemsets together with the distribution of three concise representations: frequent closed, frequent free and frequent essential itemsets.
For each of them, we study the distribution of their positive and negative borders whenever possible.
From this analysis, we exhibit some properties of datasets and some invariants allowing to better understand and predict the behavior of well known algorithms.
The main perspective of this work is to devise *adaptive algorithms* with respect to dataset/problem characteristics.

Paper organization: In section 2, we introduce some preliminaries. Experimental study of the datasets is given in section 3, including experimental protocol, results and analysis for each dataset. The section 4 presents a synthesis of the main observations done on the datasets. Finally, we conclude and give some perspectives for this work.

2. Preliminaries

Let R be a set of symbols called *items*, and r a database of subsets of R. The elements of r are called transactions. An *itemset* X is a set of some items of r. The support of X is the number of transactions in r that contain all items of X. An itemset is frequent if its support in r exceeds a minimum support threshold, called *minsup*. The goal is given a minimal support threshold and a database, to find all frequent itemsets.
We recall the notion of borders of a set [MAN97]. Let $(I, \preceq)$ be a partially ordered set of elements. A set $S \preceq I$ is *closed downwards* if, for all $X \in S$, all the subsets of X are in S. In these conditions, S can be represented by its *positive border* $Bd^+(S)$ or its *negative border* $Bd^-(S)$ defined by:

$$Bd^+(S) = max_{\subseteq}(X \in S)$$
$$Bd^-(S) = min_{\subseteq}(Y \in I\text{-}S)$$

Notice that if p is an anti-monotone predicate on elements of $(I, \preceq)$, and S is the set of all elements of I satisfying p, then S is closed downwards.
For instance, if FI is the set of all frequent itemsets in a database r, then it is closed downwards and $Bd^+(FI)$ is also called the set of *maximal frequent itemsets* in r.

Usual representation of frequent itemsets

Several concise (or condensed) representations of frequent itemsets have been studied [CAL03,MAN96]. Their goal is twofold: improving efficiency of frequent itemsets mining when possible, and compacting the storage of frequent itemsets for future usages.

Formally, a condensed representation must be equivalent to frequent itemsets: one can retrieve each frequent itemset *together with its frequency* without accessing data [CAL03]. Such a representation is known as *closed sets* [PAS99]. Two other representations are considered in this paper: frequent free itemsets [BAS00, BOU03] and frequent essential itemsets [CAS05]. Notice that these sets are not exactly sufficient to represent frequent sets, since they need a subset of the frequent itemsets border to become condensed representations [CAL03].
We briefly describe these representations in the rest of this section.

Frequent Closed sets: Given an itemset X, *the closure* of X is the set of all items that appear in all transactions where X appears. Formally, given a transaction database r:

$$Cl(X) = \cap\{\, t \in r \mid X \subseteq t \,\}$$

If $Cl(X)=X$ then X is said to be closed.

Frequent free itemsets: An itemset X is said to be free if there is no exact rule of the form $X_1 \rightarrow X_2$ where X_1 and X_2 are distinct subsets of X. Free sets can be efficiently detected through the following property:

$$X \text{ is free} \leftrightarrow \forall x \in X,\ sup(X) < sup(X\text{-}x)$$

Frequent essential itemsets: The notion of essential itemsets has been defined recently in [CAS05]. It is based on the notion of disjunctive rule [BYK01, KRY02]. A *disjunctive rule* is of the form $X \rightarrow A_1 \vee A_2 \vee \ldots \vee A_n$. Such a rule is satisfied if, every transaction that contains X contains at least one of the elements $A_1, \ldots A_n$.
An itemset X is said to be essential if there is no *disjunctive rule* of the form $A_1 \rightarrow A_2 \vee \ldots \vee A_n$, where $(A_i)_{i=1..k}$ are distinct elements in X. As for free sets, they can be efficiently tested exploiting the following property:

$$X \text{ is essential} \leftrightarrow \forall x \in X,\ sup_{dij}(X) > sup_{dij}(X\text{-}x)$$

where $sup_{dij}(X) = |\{t \in r \mid t \cap X \neq \varnothing\}|$

In the following, we study the distributions of these three important subfamilies of frequent itemsets. Moreover, the three predicates "being a frequent itemset", "being a frequent free itemset" and "being a frequent essential set" are anti-monotone w.r.t. inclusion.

Other concise representations based on the notion of disjunctive rules have been defined, the reader is referred to the general framework proposed in [CAL03] for more details. To end up, we believe that our choice of concise representations covers a fairly large range of typical cases.

3. Experimental study

Experimental protocol

For the frequent itemset problem, fourteen datasets are commonly used for benchmarks [GOE04], most of them being real-life datasets. Two datasets are synthetic ones, generated by the generator from the IBM Almaden Quest research group [IBM]. All experiments have been done on these datasets.
Each dataset has been studied for many representative minimum support thresholds, from very high to very low supports. For each one, frequent itemsets, frequent closed, frequent free and frequent essential itemsets, have been collected. We have studied their distribution, i.e. the number of elements in each level (from singletons to the size of the largest itemsets). Moreover, we have studied borders distributions of frequent, frequent free and frequent essential itemsets.
To perform these tests, we used algorithms available in the FIMI website [GOE04]. The discovery of frequent itemsets and frequent closed itemsets has been done using *FPClose* and *FP-growth** algorithms from [GRA03]. *ABS* algorithm [FLO04] has been used to find the other characteristics.
From FIMI, algorithms execution times on these datasets have been studied. For each dataset, we compared algorithms performances with the characteristics of the dataset. We choose to study maximal frequent itemsets mining algorithms to focus on the exploration strategy of the search space.

To the best of our knowledge, this work is the first one to address the understanding of datasets for frequent itemsets by using the negative border as a first class citizen.

Notations

Some notations used in the sequel are reported in the following Table.

FI	**Frequent Itemsets**
FCI	**Frequent Closed Itemsets**
FFI	**Frequent Free Itemsets**
FEI	**Frequent Essential Itemsets**

ACCIDENTS

Data description : This data set of traffic accidents is obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the period 1991-2000. The traffic accident data contains a rich source of information on the different circumstances in which the accidents have occurred: details about the accident (type of collision, injuries, …), traffic conditions (maximum speed, priority regulation, …), environmental conditions (weather, light conditions, time of the accident, …), road conditions (road surface, obstacles, …), human conditions (alcohol, …) and geographical conditions (location, physical characteristics,…).

Characteristics :

Number of items : 468
Number of transactions : 340 183
Average size of transactions : 33.8

Experimental results

Minsup 50 % (170092)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	24	24	24	24	444		444		444	
2	202	202	202	202	74		74		74	6
3	815	815	815	718	210	2	210	2	307	67
4	1808	1808	1808	1075	177	5	177	5	362	222
5	2352	2352	2352	548	60	12	60	12	330	365
6	1826	1826	1826	45	9	41	9	41	33	45
7	819	819	819		1	75	1	75		
8	193	193	193			63		63		
9	18	18	18			18		18		
Total	8057	8057	8057	2612	975	216	975	216	1550	705

Minsup 40 % (136074)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	30	30	30	30	438		438		438	
2	294	294	294	294	141		141		141	5
3	1452	1452	1452	1319	264		264		397	73
4	4229	4229	4229	2957	553	4	553	4	884	300
5	7700	7700	7700	2807	391	17	391	17	1258	1194
6	8896	8896	8895	620	127	65	127	65	341	526
7	6434	6434	6434	16	31	141	31	141	12	16
8	2784	2784	2784		6	263	6	263		
9	646	646	646		1	219	1	219		
10	63	63	63			52		52		
11	1	1	1			1		1		
Total	32529	32529	32528	8043	1952	762	1952	762	3471	2114

Minsup 30 % (102055)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	32	32	32	32	436		436		436	
2	406	406	406	406	90		90		90	5
3	2549	2548	2549	2393	564	1	564	1	720	103
4	9242	9238	9241	6996	1035	10	1036	10	1661	517
5	21536	21530	21532	10523	1230	18	1230	18	3152	2255
6	34054	34050	34048	6051	1136	77	1136	77	2497	2868
7	37093	37092	37089	1010	453	259	453	260	431	853
8	27387	27387	27386	25	109	582	109	581	7	25
9	13070	13070	13070		35	723	35	723		
10	3664	3664	3664		8	712	8	712		
11	495	495	495			330		330		
12	17	17	17			17		17		
Total	149545	149529	149529	27436	5096	2729	5097	2729	8994	6626

Minsup 25 % (85046)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	38	38	38	38	430		430		430	
2	473	473	473	473	230		230		230	9
3	3285	3284	3285	3103	507		507		689	80
4	13650	13639	13649	10831	1730	7	1731	7	2558	665
5	36242	36207	36231	19423	2376	29	2376	29	5305	3298
6	65069	65021	65034	14729	2143	120	2143	120	4882	5219
7	82081	82050	82033	4774	1358	390	1358	392	1657	2689
8	73790	73781	73759	460	477	838	478	843	147	401
9	46591	46590	46582	7	146	1326	146	1321		7
10	19652	19652	19651		22	1255	22	1254		
11	5005	5005	5005		4	1069	4	1069		
12	626	626	626		2	392	2	392		
13	23	23	23			23		23		
Total	346525	346389	346389	53838	9425	5449	9427	5450	15898	12368

Minsup 20 % (68037)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	48	47	48	48	420		420		420	
2	623	616	622	622	505		506		506	7
3	4395	4371	4388	4141	754	2	754	2	1001	135
4	19899	19822	19875	16167	1963	12	1966	12	3037	822
5	60804	60547	60727	36350	4218	36	4235	36	8295	3909
6	128017	127452	127760	36665	5133	139	5139	144	10915	10564
7	189432	188722	188866	16647	3499	508	3500	522	4433	6617
8	200883	200377	200173	3730	1758	1263	1760	1291	1003	2252
9	154794	154563	154289	275	692	2374	692	2390	46	235
10	86621	86529	86390	4	211	3097	211	3077		4
11	34215	34190	34123		31	2484	31	2549		
12	8827	8827	8802		3	1329	3	1304		
13	1259	1259	1259			586		586		
14	66	66	66			66		66		
Total	889883	887388	887388	114649	19187	11896	19217	11979	29656	24545

Minsup 15 % (51028)

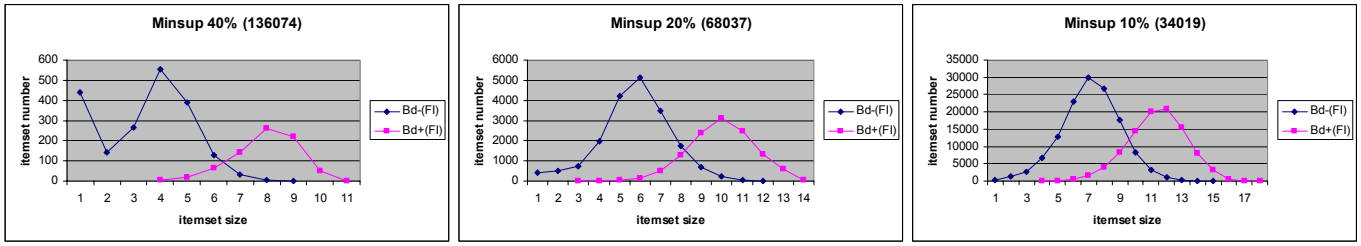
Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	60	58	60	60	408		408		408	
2	863	833	861	861	907		909		909	9
3	6534	6352	6504	6150	1101	2	1104	2	1458	175
4	31681	31011	31499	25917	3264	11	3281	11	4919	1257
5	106497	104547	105827	67140	6556	45	6605	52	12334	5148
6	256680	251840	254728	91995	10446	246	10469	261	21910	18302
7	451130	441885	446275	59889	10553	780	10580	858	13954	17956
8	583185	570757	573898	20342	6900	2013	6906	2164	4903	9666
9	555601	544409	543116	2937	3088	4195	3093	4430	510	1723
10	388871	382326	377702	213	1213	6251	1215	6492	19	154
11	198621	196107	192144	6	336	7197	340	7229	1	6
12	73446	72734	70956		52	5941	52	5788		
13	19391	19244	18680		14	2660	14	2736		
14	3464	3456	3317		3	722	3	670		
15	346	346	338			197		189		
16	11	11	11			11		11		
Total	2676381	2625916	2625916	275510	44841	30271	44979	30893	61325	54396

Minsup 10 % (34019)

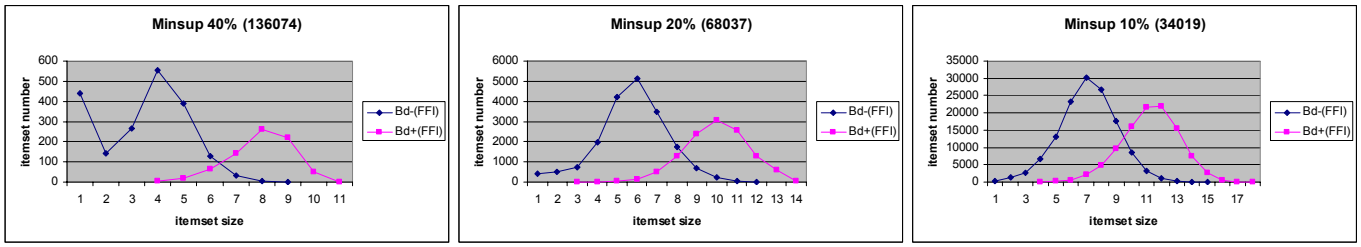
Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	75	70	75	75	393		393		393	
2	1325	1209	1320	1320	1450		1455		1455	11
3	11458	10436	11342	10860	2658		2682		3164	223
4	60237	55220	59211	49641	6663	18	6756	25	9481	2094
5	219471	202766	214399	141399	12804	117	12993	160	22745	9447
6	590662	548839	573724	243542	23067	461	23165	605	44314	31660
7	1203151	1120764	1160829	227332	30018	1637	30151	2074	43854	50370
8	1869839	1742148	1786742	116891	26641	3921	26722	4777	21476	35904
9	2222695	2069402	2094131	30856	17632	8240	17726	9588	5022	14084
10	2019620	1879454	1865541	3774	8347	14470	8422	16093	356	2441
11	1396106	1300378	1255846	203	3294	20051	3338	21602	26	138
12	726195	678569	631344	6	1063	20728	1077	21848		6
13	278605	261923	232242		210	15560	211	15403		
14	76211	72278	60340		38	8043	38	7384		
15	14095	13492	10421		4	3137	4	2655		
16	1668	1603	1106			579		418		
17	130	126	68			37		33		
18	6	6	2			6		2		
Total	10691549	9958683	9958683	825899	134282	97005	135133	102667	152286	146378

Some graphical representations

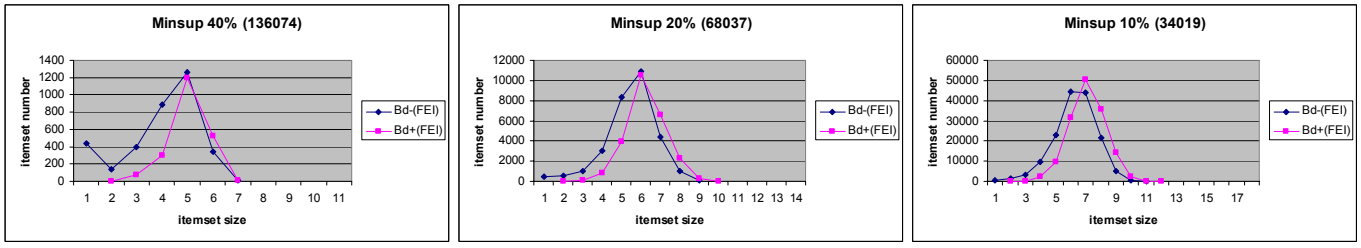
Borders of frequent itemsets



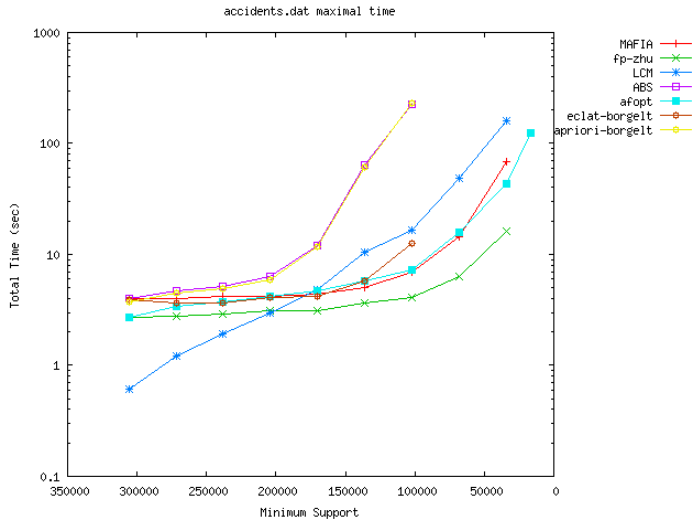
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. The borders distributions are very close, i.e. the mean of the negative border curve is only few levels before the mean of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the borders distribution of free itemsets is very similar to the borders distribution of frequent itemsets, since the frequent free itemsets are nearly equal to frequent itemsets.
- Frequent essential itemsets: the two borders are very close. As expected the number of frequent essentials is much smaller than the number of frequent free sets.

Performances of algorithms for MFI mining

For minimum support thresholds with few frequent itemsets (until a minsup of 40% to 50%), algorithms do not have difficulties. But when the number of frequent itemsets becomes more important, the algorithms performances grow exponentially.

Remarks:

- Note that there is a small number of exact association rules for this dataset ($|FCI| \approx |FI|$)
- The number of frequent essential itemsets is much smaller than the number of frequent closed itemsets by a factor of 5 to 10.

BMS-WebView-1

Data description : BMS-WebView-1 representing a real life dataset comes from a small dot-com company called Gazelle.com, a legwear and legcare retailer, which no longer exists. It contains several months of clickstream data from an ecommerce web site. Each transaction in these datasets is a web session consisting of all the product detail pages viewed in that session. That is, each product detail view is an item. The goal for both of these datasets is to find associations between products viewed by visitors in a single visit to the web site. A portion of these data was used in the KDD-Cup 2000 competition.

Characteristics :

Number of items : 497
Number of transactions : 59 602
Average size of transactions : 2.5
Maximal size of transactions : 267

Experimental results

Minsup 0.1% (60)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	343	343	343	343	154	131	154	131	154	131
2	1573	1573	1573	1573	57080	780	57080	780	57080	780
3	1492	1481	1492	1492	8301	767	8301	789	8301	767
4	522	516	514	522	1014	340	1022	345	1014	340
5	59	59	54	59	79	47	83	42	79	47
6	2	2	2	2	1	2	1	2	1	2
Total	3991	3974	3978	3991	66629	2067	66641	2089	66629	2067

Minsup 0.08% (50)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	350	350	350	350	147	109	147	109	147	109
2	2319	2319	2319	2319	58756	1055	58756	1055	58756	1055
3	3049	3019	3049	3049	15805	1254	15805	1293	15805	1254
4	1790	1668	1761	1790	2903	892	2932	958	2903	892
5	508	368	383	508	361	200	452	264	361	200
6	141	66	28	141	18	25	28	28	18	25
7	31	18		31		8				8
8	3	3		3		3				3
Total	8191	7811	7890	8191	77990	3546	78120	3707	77990	3546

Minsup 0.07% (40)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	362	362	362	362	135	87	135	87	135	87
2	3593	3572	3593	3593	61748	1240	61748	1249	61748	1240
3	9315	8401	9291	9315	30550	2257	30574	2751	30550	2257
4	12861	9279	11660	12861	13274	2842	14399	4447	13274	2842
5	10353	5367	5716	10353	4439	2090	6469	3203	4439	2090
6	6089	1920	855	6089	689	1017	857	734	689	1017
7	3303	422	23	3303	38	232	40	23	38	232
8	1718	101		1718	1	42			1	42
9	707	39		707		8				8
10	203	17		203		8				8
11	36	6		36		1				1
12	3	3		3		3				3
Total	48543	29489	31500	48543	110874	9827	114222	12494	110874	9827

Minsup 0.06% (35)

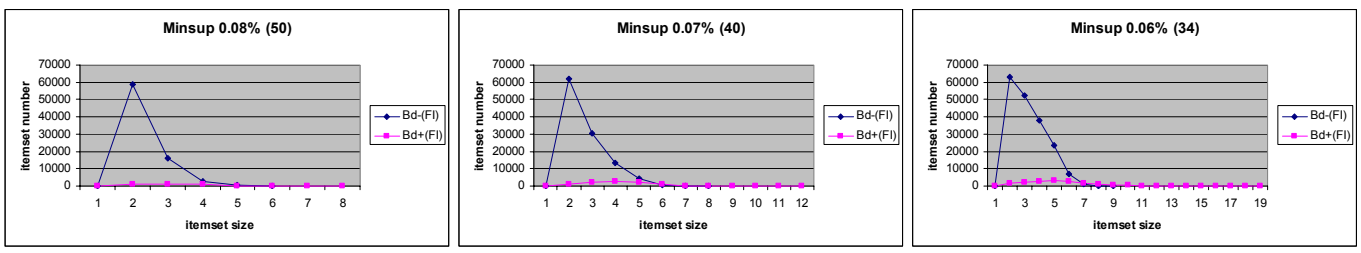
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	369	369	369	369	128	66	128	66	128	66
2	4916	4680	4916	4916	62980	1408	62980	1520	62980	1408
3	20735	13373	20186	20735	45641	2355	46190	4933	45641	2355
4	62019	19437	44165	62019	32049	3080	46966	14806	32049	3080
5	129240	16814	35946	129240	18709	3172	40152	16187	18709	3172
6	198760	10617	11573	198760	5496	2160	8277	6511	5496	2160
7	233225	5709	1500	233225	977	1443	1050	1230	977	1443
8	214111	2744	41	214111	79	845	79	41	79	845
9	156992	1336		156992		540				540
10	92577	620		92577		320				320
11	43479	299		43479		171				171
12	15883	135		15883		79				79
13	4350	70		4350		41				41
14	842	39		842		29				29
15	103	12		103		10				10
16	6	6		6		6				6
Total	1177607	76260	118696	1177607	166059	15725	205822	45294	166059	15725

Minsup 0.06% (34)

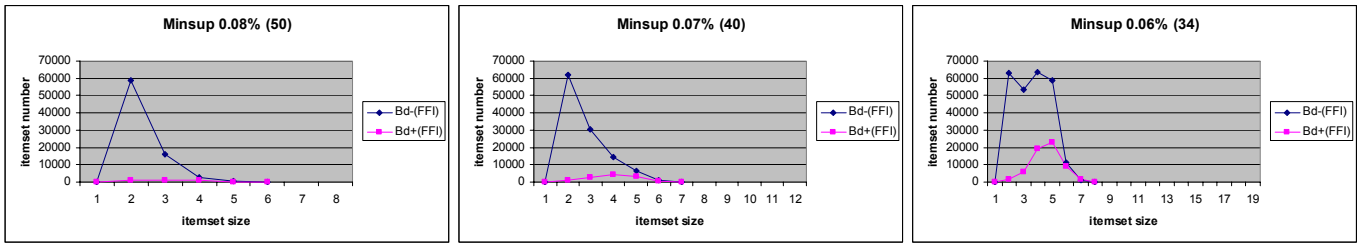
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	370	370	370	370	127	64	127	64	127	64
2	5351	4959	5351	5351	62914	1410	62914	1583	62914	1410
3	25360	14360	24234	25360	52171	2340	53297	5996	52171	2340
4	90678	21454	58864	90678	37681	2932	63372	19420	37681	2932
5	235394	19323	50988	235394	23744	2956	58554	22969	23744	2956
6	465587	12866	16822	465587	7207	2419	11351	8888	7207	2419
7	719116	7165	2475	719116	1058	1507	1177	1772	1058	1507
8	881196	3625	120	881196	84	895	84	120	84	895
9	869471	1834		869471	2	503			2	503
10	698300	949		698300		332				332
11	458479	481		458479		182				182
12	245450	266		245450		131				131
13	106179	155		106179		85				85
14	36533	84		36533		47				47
15	9756	50		9756		39				39
16	1947	23		1947		17				17
17	273	12		273		12				12
18	24	5		24		5				5
19	1	1		1		1				1
Total	4849465	87982	159224	4849465	184988	15877	250876	60812	184988	15877

Some graphical representations

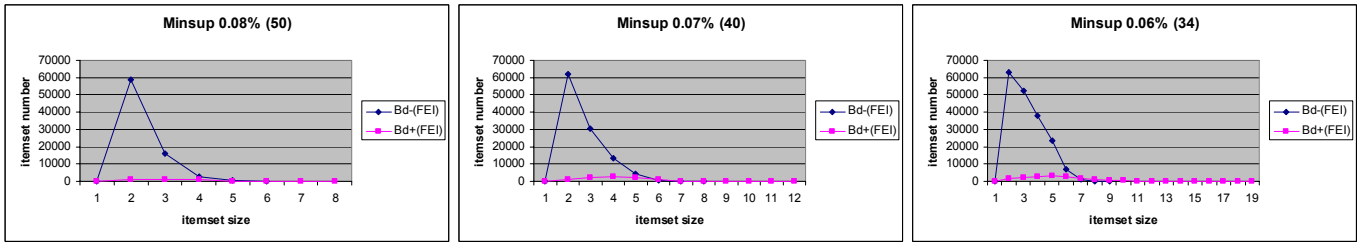
Borders of frequent itemsets



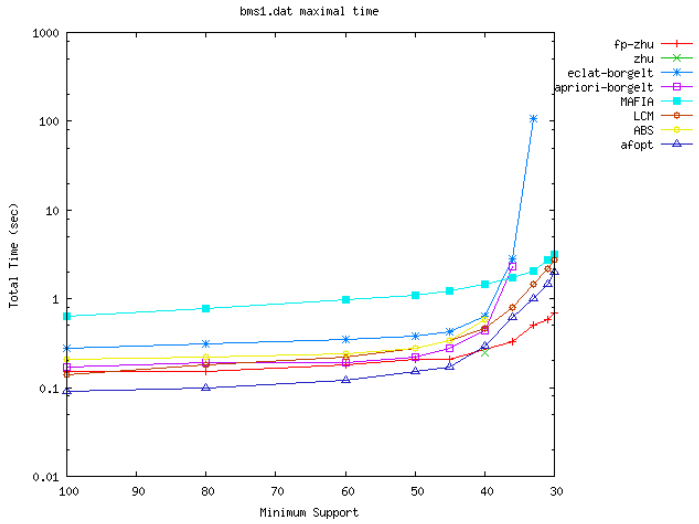
Borders of frequent free itemsets



Borders of essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. Even if the positive border appears to have some longer itemsets than the negative border, most itemsets in the two borders have relatively the same size. Consequently, the borders distributions are very close, i.e. the mean of the negative border curve is only few levels before the mean of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the two borders are very close, since the frequent free itemsets are relatively small.
- Frequent essential itemsets: the borders distribution of essential itemsets is very similar to the borders distribution of frequent itemsets, since the frequent essential itemsets are equal to frequent itemsets for all the minimum support thresholds tested.

Performances of algorithms for MFI mining

For minimum support thresholds with few frequent itemsets (until a minsup of 0.07%), algorithms do not have difficulties. When the number of frequent itemsets becomes more important, the algorithms performances grow, but still good since most of the frequent itemsets are relatively small.

Remarks:

- We have to study this dataset for very low support thresholds to have a significant number of frequent itemsets.
- Note that there is an important number of exact association rules for this dataset (see the number of frequent free and closed itemsets), and some rules are relatively long (see for example the results for a minsup of 34).
- The number of frequent essential itemsets is much more important than the number of frequent closed itemsets.

BMS-WebView-2

Data description : BMS-WebView-2 contains the same data type than BMS-WebView-1 for the same ecommerce web site. Recall that these datasets represent a real life dataset, and come from a small dot-com company called Gazelle.com, a legwear and legcare retailer, which no longer exists. It contains several months of clickstream data from an ecommerce web site. On the other hand, the objective is different of BMS-WebView-1. The goal of this data file is to determine starting from a whole of visited pages, if the user will visit another page of the web site or if it will leave it.

Characteristics :

Number of items : 3341
Number of transactions : 77 512
Average size of transactions : 5.6
Maximal size of transactions : 161

Experimental results

Minsup 0.1% (80)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	1113	1113	1113	1113	2227	737	2227	737	2227	737
2	1307	1307	1307	1307	617521	650	617521	650	617521	650
3	1872	1872	1872	1872	2523	393	2523	393	2523	393
4	3284	3273	3284	3284	1336	457	1336	459	1336	457
5	4514	4411	4503	4514	869	365	880	374	869	365
6	4744	4432	4636	4744	533	440	588	513	533	440
7	3213	2907	2887	3213	413	287	504	384	413	287
8	1338	1241	1057	1338	113	196	121	211	113	196
9	271	264	195	271	31	110	31	82	31	110
10	19	19	13	19	1	19	1	13	1	19
Total	21675	20839	20867	21675	625567	3654	625732	3816	625567	3654

Minsup 0.08% (60)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	1409	1409	1409	1409	1931	858	1931	858	1931	858
2	2132	2132	2132	2132	989804	1105	989804	1105	989804	1105
3	2692	2690	2692	2692	6062	642	6062	643	6062	642
4	4970	4938	4968	4970	1584	559	1586	562	1584	559
5	7867	7446	7833	7867	1504	649	1536	713	1504	649
6	9414	8006	8938	9414	1019	551	1308	724	1019	551
7	8734	6469	7120	8734	478	475	759	690	478	475
8	5920	3953	3614	5920	273	247	309	382	273	247
9	2697	1843	1078	2697	81	156	82	239	81	156
10	707	561	155	707	11	113	11	76	11	113
11	83	81	8	83		71		8		71
12	1	1		1		1				1
Total	46626	39529	39947	46626	1002747	5427	1003388	6000	1002747	5427

Minsup 0.06% (50)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	1597	1597	1597	1597	1743	887	1743	887	1743	887
2	2837	2837	2837	2837	1271569	1473	1271569	1473	1271569	1473
3	3725	3722	3725	3725	9157	1080	9157	1080	9157	1080
4	6249	6177	6246	6249	2777	767	2780	783	2777	767
5	10250	9572	10170	10250	1606	777	1678	864	1606	777
6	13085	10609	12280	13085	1424	750	1905	1043	1424	750
7	13132	8638	10040	13132	708	581	1162	859	708	581
8	10394	5535	5354	10394	248	291	309	530	248	291
9	6189	2899	1758	6189	52	140	55	261	52	140
10	2550	1243	303	2550	17	62	17	140	17	62
11	634	412	18	634	2	68	2	18	2	68
12	70	70		70		70				70
Total	70712	53311	54328	70712	1289303	6946	1290377	7938	1289303	6946

Minsup 0.05% (40)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	1822	1822	1822	1822	1518	937	1518	937	1518	937
2	3963	3963	3963	3963	1654968	1994	1654968	1994	1654968	1994
3	5505	5495	5505	5505	15593	1614	15593	1615	15593	1614
4	9252	9004	9241	9252	4334	1384	4345	1432	4334	1384
5	14554	13177	14276	14554	3117	1141	3347	1427	3117	1141
6	18814	14901	17192	18814	1767	1099	2557	1467	1767	1099
7	19038	12193	14078	19038	1279	991	1914	1422	1279	991
8	15693	7451	7259	15693	375	590	464	790	375	590
9	10979	3694	2339	10979	55	225	60	352	55	225
10	6145	1657	412	6145	9	29	9	141	9	29
11	2479	701	29	2479		9		29		9
12	638	262		638		23				23
13	85	67		85		46				46
14	3	3		3		3				3
Total	108970	74390	76116	108970	1683015	10085	1684775	11606	1683015	10085

Minsup 0.04% (30)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	2122	2122	2122	2122	1218	956	1218	956	1218	956
2	6052	6052	6052	6052	2244329	2962	2244329	2962	2244329	2962
3	9022	8971	9022	9022	30794	2564	30794	2585	30794	2564
4	15591	14616	15536	15591	7879	2320	7934	2545	7879	2320
5	26068	21321	24932	26068	5348	2096	6233	2835	5348	2096
6	35714	24428	29887	35714	3898	1794	5743	3350	3898	1794
7	36968	21272	23793	36968	2465	1748	3454	2937	2465	1748
8	28884	14053	12243	28884	1218	1375	1360	1685	1218	1375
9	17873	7060	3680	17873	357	920	373	932	357	920
10	9407	2626	545	9407	53	508	53	224	53	508
11	4372	816	34	4372	1	58	1	34	1	58
12	1657	292		1657		6				6
13	449	106		449		1				1
14	76	34		76		1				1
15	6	6		6		6				6
Total	194261	123775	127846	194261	2297560	17315	2301492	21045	2297560	17315

Minsup 0.03% (20)

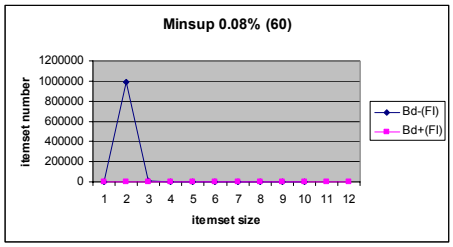
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	2434	2434	2434	2434	906	814	906	814	906	814
2	10672	10667	10672	10672	2950289	5023	2950289	5029	2950289	5023
3	17998	17658	17993	17998	70646	5226	70651	5345	70646	5226
4	31717	27474	31302	31716	17482	4196	17889	5161	17483	4195
5	57845	37760	51930	57845	10144	3900	14278	6671	10143	3900
6	93065	41944	62901	93065	8083	3107	13881	7618	8083	3107
7	122110	37988	50375	122110	4619	2434	6812	6083	4619	2434
8	123813	28319	26342	123813	2004	1735	2268	3747	2004	1735
9	93791	18229	8589	93791	621	1348	651	1687	621	1348
10	52125	10232	1517	52125	134	1152	134	627	134	1152
11	20985	4693	112	20985	9	839	9	89	9	839
12	6059	1563	2	6059		476		2		476
13	1249	362		1249		119				119
14	182	71		182		33				33
15	18	10		18		2				2
16	1	1		1		1				1
Total	634064	239405	264169	634063	3064937	30405	3077768	42873	3064937	30404

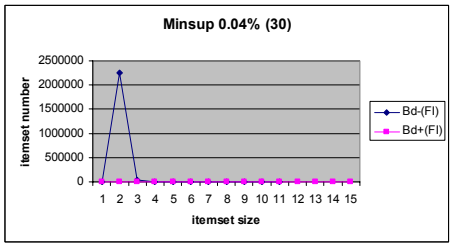
Minsup 0.01% (10)

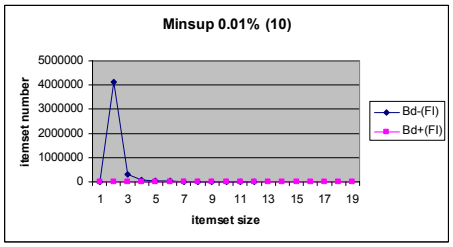
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	2885	2885	2885	2885	455	555	455	555	455	555
2	28003	27782	28003	28003	4132167	12434	4132167	12611	4132167	12435
3	55120	50176	54876	55118	308992	13688	309236	15798	308994	13690
4	111136	74623	104169	111134	71022	12431	77666	22799	71023	12430
5	206928	87454	147548	206918	41402	10496	73230	30390	41405	10505
6	338172	85728	141656	338138	19332	8669	40613	22288	19337	8678
7	493584	71528	101953	493565	5386	6178	9791	12213	5385	6179
8	638317	52408	51382	638314	2144	3848	2565	7785	2144	3845
9	712247	34857	15669	712247	1126	2247	1161	3248	1126	2247
10	672851	21359	2610	672851	174	1362	174	1279	174	1362
11	532202	12273	175	532202	7	791	7	142	7	791
12	348943	6730	3	348943	2	849	2	3	2	849
13	186708	3250		186708		429				429
14	79561	1530		79561		143				143
15	26132	679		26132		50				50
16	6348	279		6348		28				28
17	1076	108		1076		50				50
18	115	25		115		8				8
19	6	6		6		6				6
Total	4440334	533680	650929	4440264	4582209	74262	4647067	129111	4582219	74280

Some graphical representations

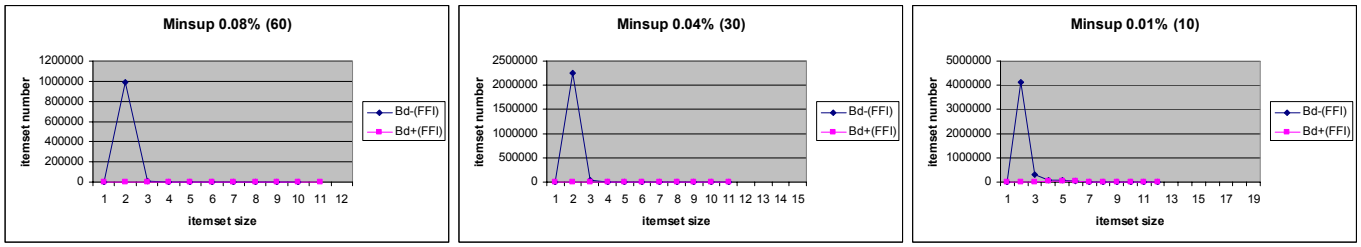
Borders of frequent itemsets



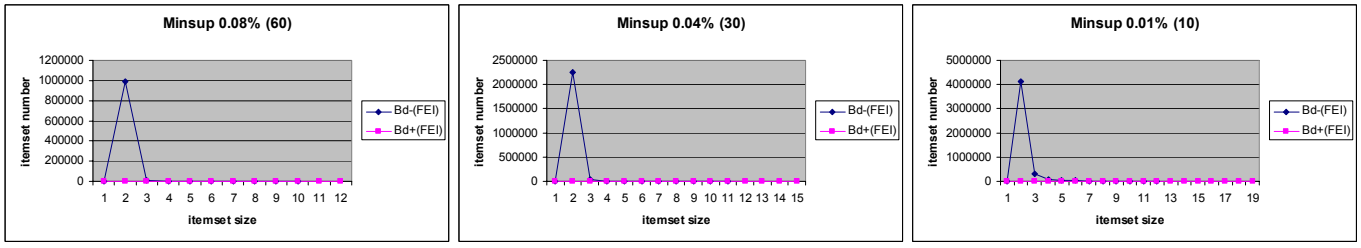




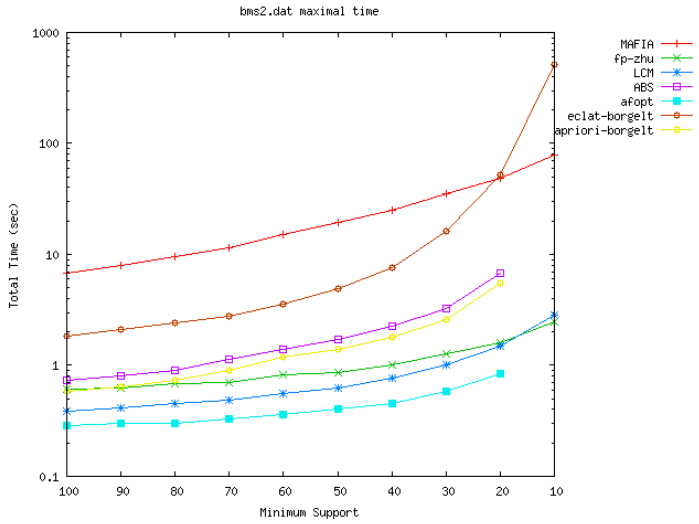
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. Even if the positive border appears to have some longer itemsets than the negative border, most itemsets in the two borders have relatively the same size. More precisely approximately 80% of the positive border shares the same levels than the negative border. Consequently, the borders distributions are very close.

Borders of other concise representations

- Frequent free itemsets: the two borders are very close (see result tables).
- Frequent essential itemsets: the borders distribution of essential itemsets is very similar to the borders distribution of frequent itemsets, since the frequent essential itemsets are approximately equal to frequent itemsets.

Performances of algorithms for MFI mining

For minimum support thresholds with few frequent itemsets (until a minsup of 0.08%), algorithms do not have difficulties. When the number of frequent itemsets becomes more important, the algorithms performances grow, but still good since most of the frequent itemsets are relatively small.

Remarks:

- We have to study this dataset for very low support thresholds to have a significant number of frequent itemsets.
- The number of frequent essential itemsets (approximately equal to the number of frequent itemsets) is much more important than the number of frequent closed itemsets.

BMSPOS

Data description : The BMS-POS dataset contains several years worth of point-ofsale data from a large electronics retailer. Since this retailer has so many different products, product categories are used as items. The transaction in this dataset is a customer's purchase transaction consisting of all the product categories purchased at one time. The goal for this dataset is to find associations between product categories purchased by customers in a single visit to the retailer.

Characteristics :

Number of items : 1658
Number of transactions : 515 597
Average size of transactions : 7.5
Maximal size of transactions : 164

Experimental results

Minsup 0.29% (1500)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	278	278	278	278	1379	15	1379	15	1379	15
2	2126	2126	2126	2126	36377	259	36377	259	36377	259
3	4307	4307	4307	4307	11938	970	11938	970	11938	970
4	3973	3973	3973	3973	3503	1400	3503	1400	3503	1400
5	1737	1737	1737	1737	900	952	900	952	900	952
6	298	298	298	298	88	261	88	261	88	261
7	7	7	7	7		7		7		7
Total	12726	12726	12726	12726	54185	3864	54185	3864	54185	3864

Minsup 0.19% (1000)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	335	335	335	335	1322	26	1322	26	1322	26
2	3436	3436	3436	3436	52509	326	52509	326	52509	326
3	8488	8488	8488	8488	25925	1582	25925	1582	25925	1582
4	9701	9701	9701	9701	9134	2762	9134	2762	9134	2762
5	5753	5753	5753	5753	3052	2459	3052	2459	3052	2459
6	1617	1617	1617	1617	526	1023	526	1023	526	1023
7	159	159	159	159	23	151	23	151	23	151
8	1	1	1	1		1		1		1
Total	29490	29490	29490	29490	92491	8330	92491	8330	92491	8330

Minsup 0.15% (750)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	370	370	370	370	1287	22	1287	22	1287	22
2	4710	4710	4710	4710	63555	384	63555	384	63555	384
3	13500	13500	13500	13500	43609	2184	43609	2184	43609	2184
4	17743	17742	17743	17743	17593	4501	17593	4502	17593	4501
5	12576	12576	12575	12576	6505	4540	6506	4539	6505	4540
6	4704	4704	4704	4704	1653	2519	1653	2519	1653	2519
7	763	762	763	763	154	582	154	586	154	582
8	31	31	30	31	4	31	5	30	4	31
Total	54397	54395	54395	54397	134360	14763	134362	14766	134360	14763

Minsup 0.1% (500)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	443	443	443	443	1214	23	1214	23	1214	23
2	7090	7090	7090	7090	90813	507	90813	507	90813	507
3	25122	25122	25122	25122	87396	3101	87396	3101	87396	3101
4	40105	40092	40105	40105	42640	8304	42640	8313	42640	8304
5	35261	35213	35248	35261	18253	10214	18266	10278	18253	10214
6	17933	17894	17885	17933	6126	7279	6169	7303	6126	7279
7	4848	4831	4809	4848	1129	2749	1157	2768	1129	2749
8	548	547	531	548	70	456	83	446	70	456
9	12	12	11	12		12		11		12
Total	131362	131244	131244	131362	247641	32645	247738	32750	247641	32645

Minsup 0.06% (300)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	547	547	547	547	1110	21	1110	21	1110	21
2	11415	11415	11415	11415	137916	711	137916	711	137916	711
3	51671	51622	51671	51671	200388	4935	200388	4966	200388	4935
4	105207	104584	105158	105207	119786	16230	119835	16653	119786	16230
5	118999	117101	118376	118999	63695	26720	64290	28044	63695	26720
6	81561	79705	79663	81561	26899	24646	28411	25498	26899	24646
7	34015	33194	32159	34015	8133	13666	9044	13375	8133	13666
8	7878	7687	7057	7878	1251	4347	1523	4194	1251	4347
9	820	814	629	820	71	666	90	516	71	666
10	20	20	14	20		20		14		20
Total	412133	406689	406689	412133	559249	91962	562607	93992	559249	91962

Minsup 0.04% (200)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	637	637	637	637	1020	38	1020	38	1020	38
2	16240	16236	16240	16240	186326	839	186326	841	186326	839
3	88672	88108	88668	88672	365958	7002	365962	7335	365958	7002
4	216201	210358	215637	216201	262460	26891	263021	30153	262460	26891
5	295222	278951	289379	295222	163509	53354	168863	61128	163509	53354
6	249244	231475	232972	249244	82050	60211	92467	65180	82050	60211
7	134543	124794	116775	134543	30960	42138	36140	40576	30960	42138
8	45349	42547	35600	45349	7243	18256	8300	15345	7243	18256
9	8631	8273	5829	8631	947	4904	1003	3771	947	4904
10	711	706	353	711	44	619	44	303	44	619
11	10	10	5	10		10		5		10
Total	1055460	1002095	1002095	1055460	1100517	214262	1123146	224675	1100517	214262

Minsup 0.03% (150)

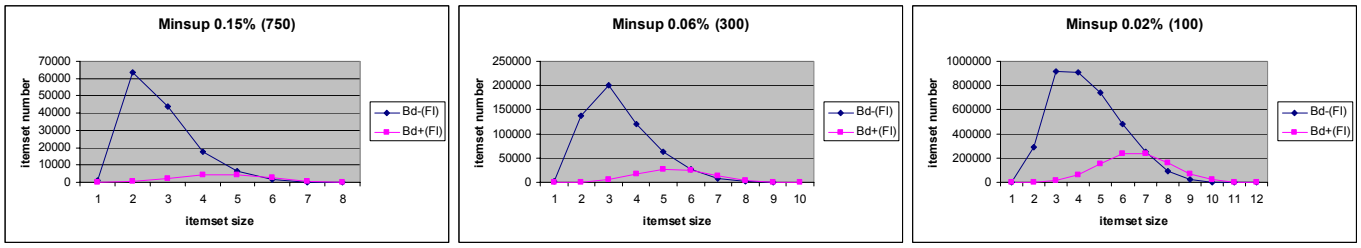
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	707	707	707	707	950	30	950	30	950	30
2	20631	20601	20631	20631	228940	1023	228940	1044	228940	1025
3	128097	125868	128067	128096	550027	8771	550057	9978	550028	8770
4	352956	332522	350727	352956	444800	37832	447014	47530	444800	37832
5	547832	491582	527397	547832	310085	84134	327896	105791	310085	84134
6	531482	464215	475224	531482	174182	108788	203324	123864	174182	108788
7	337195	295531	269931	337195	75344	88573	88009	84233	75344	88573
8	139497	124970	97838	139497	22411	46892	24467	36564	22411	46892
9	35476	32735	20953	35476	3894	15549	3991	10447	3894	15549
10	4863	4676	2122	4863	316	3142	317	1623	316	3142
11	246	245	59	246	6	234	6	59	6	234
12	1	1		1		1				1
Total	2098983	1893653	1893656	2098982	1810955	394969	1874971	421163	1810956	394970

Minsup 0.02% (100)

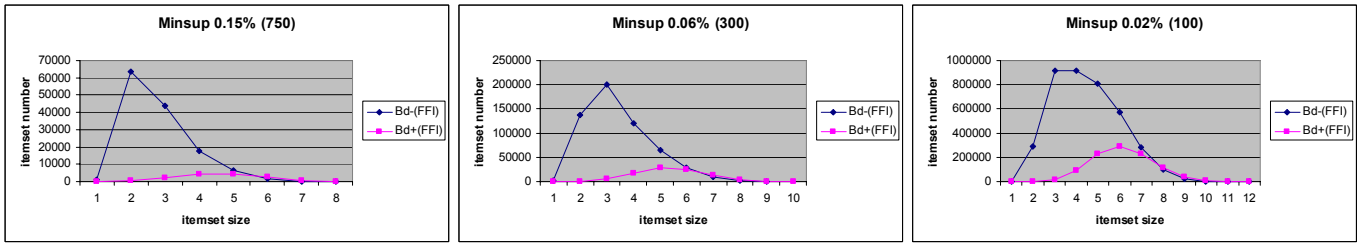
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	796	796	796	796	861	29	861	29	861	29
2	28153	27992	28153	28153	288257	1242	288257	1344	288257	1242
3	210086	199546	209925	210085	913291	11843	913452	16793	913292	11847
4	684693	595908	674147	684691	909223	59212	919619	93712	909224	59210
5	1264760	1017238	1175888	1264760	741232	153082	811667	225389	741232	153082
6	1481982	1152132	1234299	1481982	479882	236621	570276	292444	479882	236621
7	1157805	912680	827903	1157805	248911	237160	279815	228276	248911	237160
8	612864	502199	367960	612864	93313	157291	97112	116758	93313	157291
9	216491	186346	105958	216491	23555	70342	23701	40512	23555	70342
10	47846	43430	17720	47846	3457	20538	3459	9427	3457	20538
11	5725	5497	1310	5725	236	3787	236	1117	236	3787
12	246	246	18	246	2	246	2	18	2	246
Total	5711447	4644010	4644077	5711444	3702220	951393	3908457	1025819	3702222	951395

Some graphical representations

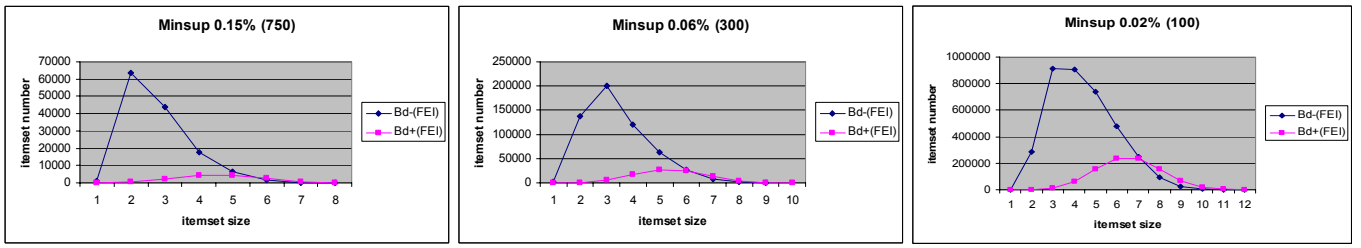
Borders of frequent itemsets



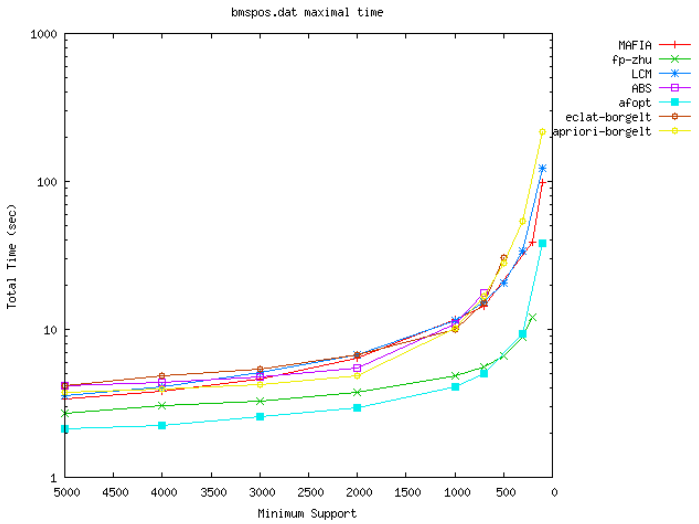
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. The borders distributions are very close, since most itemsets in the two borders are relatively small.

Borders of other concise representations

The borders distributions of frequent free and essential itemsets are very similar to the borders distribution of frequent itemsets, since the frequent free and essential itemsets are approximately equal to frequent itemsets.

Performances of algorithms for MFI mining

For minimum support thresholds with few frequent itemsets (until a minsup of ≈0.19%, i.e 1000), algorithms do not have difficulties. When the number of frequent itemsets becomes more important, the algorithms performances grow exponentially.

Remarks:

- We have to study this dataset for very low support thresholds to have a significant number of frequent itemsets.
- There is few exact association rules (|FCI|≈|FI|).
- The number of frequent itemsets, frequent closed and essential itemsets are approximately equal.

CHESS

Data description : The Chess dataset is derived from the steps of Chess games. The format for the transactions in this database is a sequence of 37 attribute values. Each transaction is a board-descriptions for this chess endgame. The first 36 attributes describe the board. The last (37th) attribute is the classification: "win" or "no win". There are 0 missing values.

Characteristics :

Number of items : 75
Number of transactions : 3 196
Average size of transactions : 37
Minimal size of transactions : 37
Maximal size of transactions : 37

Experimental results

Minsup 70% (2238)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	24	19	24	24	51		51		51	
2	238	153	228	228	38		48	1	48	83
3	1237	674	1104	303	160	2	166	12	967	218
4	3857	1884	3100	36	316	10	316	30	211	36
5	7891	3629	5424		323	35	323	86		
6	11125	5003	6159		248	77	248	176		
7	11113	5159	4652		124	147	124	249		
8	7916	4015	2335		49	197	49	198		
9	3895	2290	733		13	172	13	108		
10	1216	872	125			134		48		
11	204	181	8			115		8		
12	14	12				1				
13	1	1				1				
Total	48731	23892	23892	591	1322	891	1338	916	1277	337

Minsup 60% (1918)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	34	25	34	34	41		41		41	
2	389	234	371	371	172		190	2	190	124
3	2325	1174	2057	697	295	4	305	19	1665	364
4	8831	3795	6979	209	584	23	584	77	772	194
5	23155	8714	15605	3	1070	67	1070	201	13	3
6	43106	14985	23556		1228	150	1228	423		
7	57479	19806	23937		809	329	809	722		
8	55062	20157	16273		356	437	356	834		
9	37876	15645	7276		125	603	125	689		
10	18607	8983	2014		47	841	47	398		
11	6419	3676	299		8	533	8	126		
12	1466	1025	17			239		17		
13	187	165				89				
14	8	8				8				
Total	254944	98392	98418	1314	4735	3323	4763	3508	2681	685

Minsup 50% (1598)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	37	26	37	37	38		38		38	
2	531	285	508	508	135		158	1	158	127
3	4000	1716	3509	1476	503	4	533	21	2566	611
4	18565	6621	14421	742	1198	21	1209	114	1981	578
5	58172	17982	38330	46	2408	93	2413	493	64	46
6	129952	36451	69727		3468	262	3468	1123		
7	214297	57191	89110		3347	652	3347	1904		
8	266635	70698	79951		2427	1220	2427	2729		
9	252853	69314	49858		1158	1753	1158	2641		
10	181970	53949	20874		449	2150	449	2039		
11	97307	33137	5444		81	2041	81	920		
12	37232	15599	788		13	1729	13	287		
13	9667	5236	46			1078		46		
14	1574	1119				380				
15	136	122				76				
16	4	4				4				
Total	1272932	369450	372603	2809	15225	11463	15294	12318	4807	1362

Minsup 40% (1279)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	40	26	40	40	35		35		35	
2	673	313	641	641	107		139	1	139	132
3	6111	2215	5304	2632	785	1	823	17	3495	753
4	34560	10142	26371	2329	2205	23	2229	171	4640	1530
5	131925	32461	85203	319	5362	68	5384	856	392	244
6	358757	77483	187840	16	9328	413	9328	2576	2	16
7	721503	143555	291475		11370	1138	11370	5119		
8	1101076	211687	323496		10185	2456	10185	7926		
9	1296912	251459	257911		6697	4539	6697	9166		
10	1189378	241313	145980		3127	6041	3127	7696		
11	849824	187351	56571		1112	7096	1112	5132		
12	469982	116884	13962		239	6499	239	2375		
13	198718	57538	1939		30	4887	30	702		
14	63101	21493	109		3	3247	3	109		
15	14604	5892				1138				
16	2309	1185				438				
17	219	150				56				
18	10	10				10				
Total	6439702	1361157	1396842	5977	50585	38050	50701	41846	8703	2675

Minsup 30% (959)

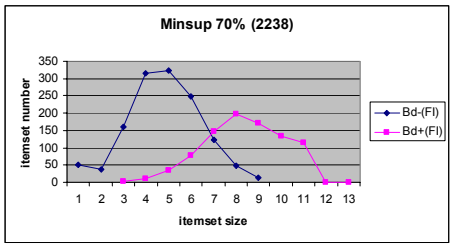
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	50	27	50	50	25		25	1	25	1
2	896	338	828	828	329		397	2	397	149
3	9049	2568	7628	4240	928	1	988	34	4376	853
4	59589	13221	44096	6283	3371	9	3440	268	8519	3178
5	273069	49002	170161	1635	10118	89	10178	1343	1764	1186
6	907800	137564	456826	116	21405	439	21416	4876	36	109
7	2255159	303661	875938	1	33711	1369	33720	11963		1
8	4276852	540861	1216501		39910	3686	39910	22521		
9	6291848	787143	1231162		33890	8200	33890	31137		
10	7263312	940504	903996		21894	14804	21894	32243		
11	6626801	923310	474618		10160	21183	10160	25491		
12	4790827	740773	172688		3507	24638	3507	15326		
13	2738089	481499	41186		791	23766	791	6403		
14	1227702	250715	5787		114	18088	114	1951		
15	425896	102977	360		8	10934	8	314		
16	111726	32875	3			5085		3		
17	21328	7908				1734				
18	2757	1370				496				
19	206	145				97				
20	6	6				6				
Total	37282962	5316467	5601828	13153	180161	134624	180438	153876	15117	5477

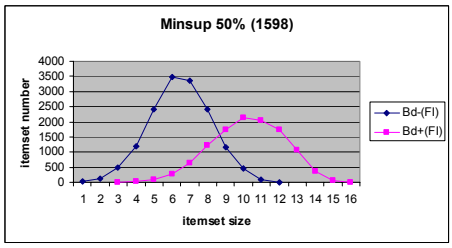
Minsup 25% (799)

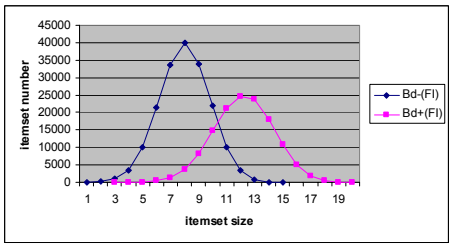
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	51	27	51	51	24		24		24	
2	1018	343	948	948	257		327		327	145
3	11251	2697	9349	5484	1050		1139	37	5004	927
4	79977	14564	57245	10013	4706	18	4831	303	11511	4260
5	396274	57404	236347	3638	13514	87	13622	1788	3602	2603
6	1438219	172895	690297	303	29974	402	30014	6673	112	240
7	3945550	411795	1461149	11	53992	1482	54021	17925	2	11
8	8349392	797211	2268598		73782	4192	73789	36928		
9	13814044	1277083	2596739		73659	10115	73661	55819		
10	18036226	1706042	2186617		55024	19993	55024	64080		
11	18703479	1900452	1339086		29910	31678	29910	57392		
12	15465778	1753885	582874		12397	42521	12397	38488		
13	10211888	1327284	173615		3588	48174	3588	18701		
14	5375535	814454	33186		641	43242	641	6760		
15	2242491	399746	3565		73	30561	73	1548		
16	731679	154329	156		3	17048	3	156		
17	182128	46575				7415				
18	33128	10841				2453				
19	4105	1882				606				
20	309	204				93				
21	11	11				11				
Total	99022533	10849724	11639822	20448	352594	260091	353064	306598	20582	8186

Some graphical representations

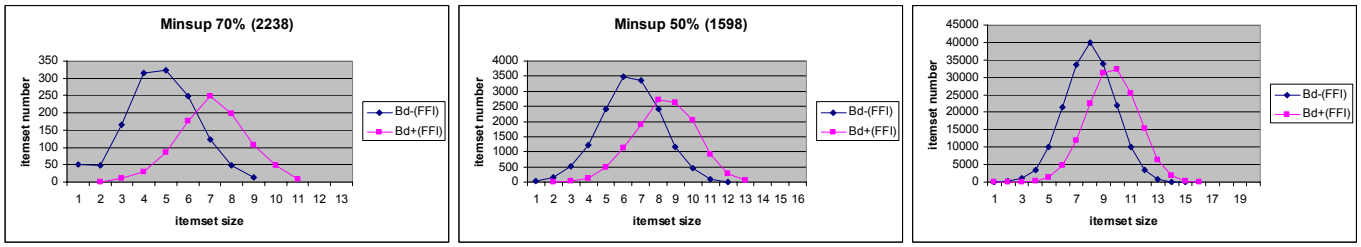
Borders of frequent itemsets



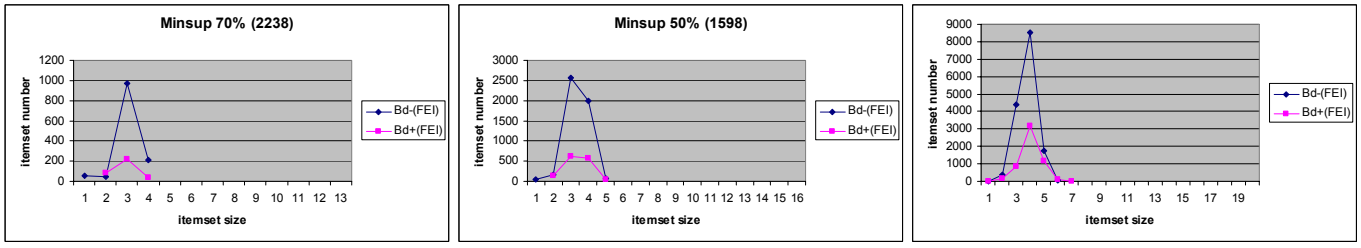




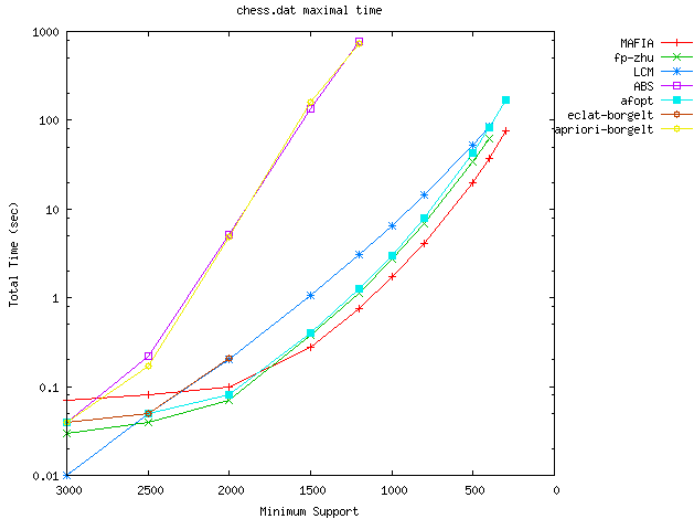
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. The borders distributions are very close, i.e. the mean of the negative border curve is only few levels before the mean of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the borders distributions of free itemsets are very close
- Frequent essential itemsets: the two borders are very close, since essential itemsets are very small. As expected the number of frequent essentials is much smaller than the number of frequent free sets.

Performances of algorithms for MFI mining

Algorithms performances are growing exponentially with the diminution of the minimum support threshold, since the number of frequent itemsets is quickly very important and frequent itemsets are relatively long. Note that Chess dataset have a relatively small number of transactions (3196).

Remarks:

- The number of frequent essential itemsets is much smaller than the number of frequent closed itemsets by a factor of 80 to more than 4000!

CONNECT

Data description : This database contains all legal 8-ply positions in the game of connect-4 in which neither player has won yet, and in which the next move is not forced.

Information about attributes (x = player x has taken, o = player o has taken, b = blank) :

The board is numbered like:

```
6 . . . . . . .
5 . . . . . . .
4 . . . . . . .
3 . . . . . . .
2 . . . . . . .
1 . . . . . . .
  a b c d e f g
```

1. a1: {x,o,b}
2. a2: {x,o,b}
3. a3: {x,o,b}
4. a4: {x,o,b}
5. a5: {x,o,b}
6. a6: {x,o,b}
7. b1: {x,o,b}
8. b2: {x,o,b}
9. b3: {x,o,b}
10. b4: {x,o,b}
11. b5: {x,o,b}
12. b6: {x,o,b}
13. c1: {x,o,b}
14. c2: {x,o,b}
15. c3: {x,o,b}
16. c4: {x,o,b}
17. c5: {x,o,b}
18. c6: {x,o,b}
19. d1: {x,o,b}
20. d2: {x,o,b}
21. d3: {x,o,b}
22. d4: {x,o,b}
23. d5: {x,o,b}
24. d6: {x,o,b}
25. e1: {x,o,b}
26. e2: {x,o,b}
27. e3: {x,o,b}
28. e4: {x,o,b}
29. e5: {x,o,b}
30. e6: {x,o,b}
31. f1: {x,o,b}
32. f2: {x,o,b}
33. f3: {x,o,b}
34. f4: {x,o,b}
35. f5: {x,o,b}
36. f6: {x,o,b}
37. g1: {x,o,b}
38. g2: {x,o,b}
39. g3: {x,o,b}
40. g4: {x,o,b}
41. g5: {x,o,b}
42. g6: {x,o,b}
43. Class: {win,loss,draw}

Missing Attribute Values: None
Class Distribution: 44473 win(65.83%), 16635 loss(24.62%), 6449 draw(9.55%)

Characteristics :

Number of items : 129
Number of transactions : 67 557
Average size of transactions : 43
Minimal size of transactions : 43
Maximal size of transactions : 43

Experimental results

Minsup 80% (54046)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	28	7	28	28	101		101		101	
2	319	28	276	276	59		102	1	102	276
3	2091	84	1345		124		124	11		
4	8962	209	3544		214		214	57		
5	26867	447	5101		264		264	186		
6	58541	839	3732		164		164	224		
7	94729	1395	1081		50	5	50	1081		
8	115018	2023				15				
9	104912	2531				41				
10	71316	2700				47				
11	35443	2333				118				
12	12434	1549				142				
13	2894	733				177				
14	397	205				104				
15	24	24				24				
Total	533975	15107	15107	304	976	673	1019	1560	203	276

Minsup 70% (47290)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	31	7	31	31	98		98		98	
2	420	28	366	366	45		99		99	300
3	3333	84	2145	93	158		158	9	2210	93
4	17586	210	6855		335		335	68		
5	65994	458	12002		470		470	238		
6	183361	893	10685		449		449	410		
7	387087	1574	3791		119	3	119	3791		
8	631321	2508				6				
9	803144	3656				3				
10	799637	4808				17				
11	621376	5576				52				
12	373307	5593				120				
13	170310	4746				175				
14	57250	3272				270				
15	13482	1718				248				
16	2033	621				227				
17	163	119				95				
18	4	4				4				
Total	4129839	35875	35875	490	1674	1220	1728	4516	2407	393

Minsup 60% (40535)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	36	8	36	36	93		93		93	
2	539	35	468	468	91		162	2	162	323
3	4737	112	3055	318	249		249	22	2986	318
4	28098	292	11001		565		565	83		
5	120688	651	21938		847	1	847	336		
6	391168	1265	22495		804	0	804	663		
7	982450	2182	9285		288	12	288	8910		
8	1945230	3344	65		3	34	3	65		
9	3068880	4637				59				
10	3880435	6060				79				
11	3939476	7550				68				
12	3203650	8830				55				
13	2072446	9380				96				
14	1053018	8774				207				
15	411697	7062				282				
16	119911	4689				332				
17	24693	2424				389				
18	3281	875				346				
19	233	168				138				
20	5	5				5				
Total	21250671	68343	68343	822	2940	2103	3011	10081	3241	641

Minsup 44.41% (30000)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	39	9	39	39	90		90		90	
2	685	42	605	605	56		136		136	308
3	7161	140	4738	1009	308		308	10	4037	1009
4	50575	377	20441		1017		1017	102		
5	259130	862	49566		1807	1	1807	465		
6	1006748	1710	64209		2077	3	2077	1305		
7	3054413	3011	37372		1089	18	1089	19434		
8	7387411	4863	4979		197	27	197	4979		
9	14454860	7422				36				
10	23119996	10730				28				
11	30433763	14627				79				
12	33089149	18595				83				
13	29732019	21694				148				
14	22020945	22993				322				
15	13360251	22091				415				
16	6569484	19095				561				
17	2575195	14641				734				
18	784886	9831				760				
19	178991	5626				634				
20	28680	2607				548				
21	2873	848				399				
22	135	135				135				
Total	188117389	181949	181949	1653	6641	4931	6721	26295	4263	1317

Minsup 40% (27023)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	41	9	41	41	88		88		88	
2	744	42	654	654	76		166		166	306
3	8039	140	5310	1253	328		328	14	4385	1253
4	58775	378	23963		1090		1090	108		
5	312955	881	61346		2186		2186	550		
6	1269427	1827	84644		2741		2741	1569		
7	4039060	3418	53995		1565	5	1565	22276		
8	10285215	5812	9419		286	12	286	9419		
9	21260694	9080				27				
10	36037804	13116				46				
11	50437121	17577				84				
12	58525921	21959				139				
13	56392099	25736				201				
14	45068783	28256				264				
15	29755746	28730				435				
16	16108418	26567				544				
17	7065781	21975				752				
18	2466702	15927				923				
19	667254	9933				967				
20	134225	5202				808				
21	18775	2139				585				
22	1613	605				358				
23	63	63				63				
Total	339915255	239372	239372	1948	8360	6213	8450	33936	4639	1559

Minsup 30% (20268)

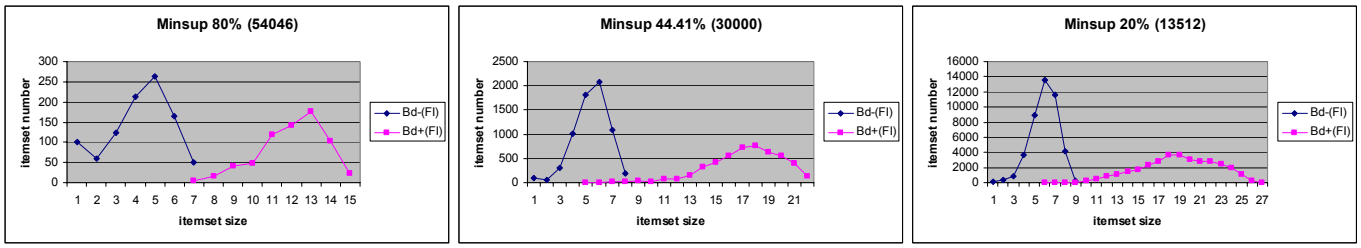
Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	46	12	46	46	83		83		83	
2	894	60	794	794	141		241	1	241	307
3	10455	206	7060	2111	456		456	11	5405	1942
4	83195	569	35299	93	1514	2	1514	133	4253	93
5	483881	1331	100954		3611	7	3611	769		
6	2152746	2713	159568		4872	16	4872	2194		
7	7552212	4976	123709		3356	15	3356	29390		
8	21342261	8420	32926		834	19	834	32926		
9	49330046	13228				34				
10	94277282	19453				70				
11	150106047	26822				92				
12	200071615	34626				191				
13	223798390	41921				258				
14	210163452	47715				368				
15	165379261	51067				477				
16	108590456	51065				779				
17	59080847	47165				904				
18	26356931	39783				1119				
19	9496507	30125				1419				
20	2703835	20147				1509				
21	589067	11508				1428				
22	93420	5300				1246				
23	9912	1769				773				
24	596	362				300				
25	13	13				13				
Total	1331673367	460356	460356	3044	14867	11039	14967	65424	9982	2342

Minsup 20% (13512)

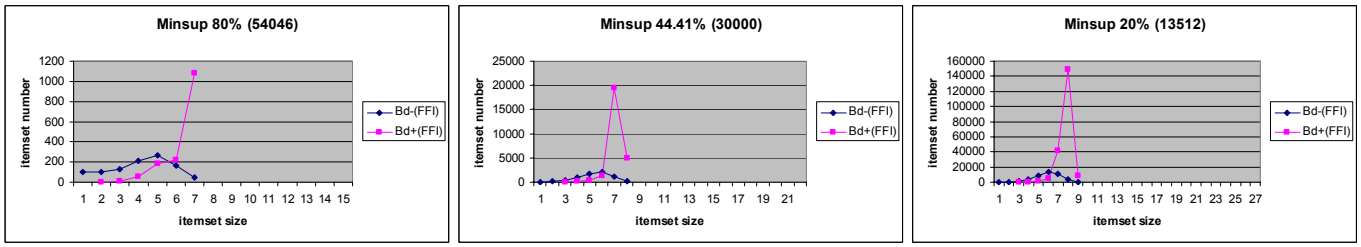
Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	54	20	54	54	129		75		75	
2	1150	120	1050	1050	75		381		381	460
3	13839	455	9916	2750	281		767	9	7933	2270
4	111214	1330	51464	342	767		2437	196	6012	342
5	654164	3242	153019		2437		5684	1076		
6	2969816	6830	255634		5684	1	7845	3507		
7	10742508	12628	216716		7845	30	5947	34386		
8	31587192	20696	69540		5947	58	1737	63863		
9	76522654	30356	1090		1737	156	39	1090		
10	154202043	40276			39	310				
11	260252819	49288				537				
12	369621033	56847				698				
13	442977574	63342				903				
14	448392113	68711				940				
15	382938885	72373				967				
16	275061309	72926				1061				
17	165254295	69296				1162				
18	82345043	61123				1384				
19	33620441	49423				1774				
20	11054628	36005				1955				
21	2855391	23110				1916				
22	558341	12624				1898				
23	77940	5476				1445				
24	7000	1718				950				
25	322	265				247				
26	3	3				3				
Total	2751821771	758483	758483	4196	24941	18395	24912	104127	14401	3072

Some graphical representations

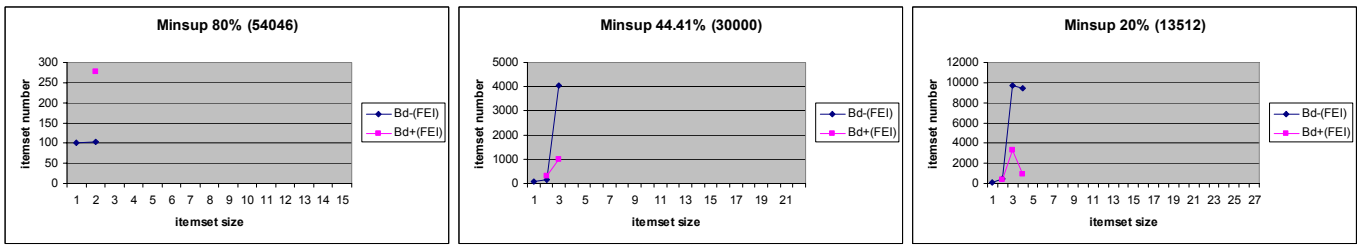
Borders of frequent itemsets



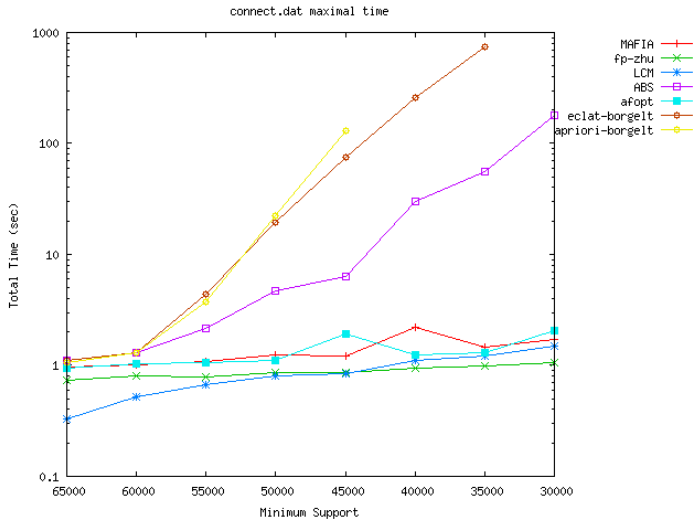
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. A large distance between the borders does exist, i.e. the mean of the negative border curve is many levels before the mean of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the borders distributions of free itemsets are very close..
- Frequent essential itemsets: the two borders are very close, since essential itemsets are very small. As expected the number of frequent essentials is much smaller than the number of frequent free sets.

Performances of algorithms for MFI mining

Most algorithms are efficient and stable, even for relatively low supports with a huge number of frequent itemsets. Note that Chess is growing exponentially with the diminution of the minimum support threshold, whereas Connect remains stable. However, Connect has more and longer transactions, and more items than Chess.

Remarks:

- There are many long exact association rules.
- The number of frequent essential itemsets is much smaller than the number of frequent closed itemsets by a factor of 1750 to more than 650 000!

KOSARAK

Data description : This dataset contains data corresponding to click carried out on a Hungarian website of news.

Characteristics : Number of items : 41 270
Number of transactions : 990 002
Average size of transactions : 8.1

Experimental results

Minsup 0.25% (2476)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	405	405	405	405	40865	28	40865	28	40865	28
2	1599	1596	1599	1599	80211	112	80211	112	80211	112
3	2621	2597	2618	2621	1990	279	1993	309	1990	279
4	2245	2222	2221	2245	343	408	364	410	343	408
5	1179	1167	1156	1179	48	234	65	256	48	234
6	488	488	476	488	18	103	19	91	18	103
7	209	209	209	209	3	8	3	8	3	8
8	74	74	74	74		3	0	4	0	4
9	12	12	12	12		4	2	12	2	12
Total	8832	8770	8770	8832	123478	1179	123522	1230	123480	1188

Minsup 0.2% (1981)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	568	568	568	568	40702	43	40702	43	40702	43
2	2469	2464	2469	2469	158559	180	158559	180	158559	180
3	4622	4496	4617	4622	3981	410	3986	482	3981	410
4	5642	5266	5516	5642	673	611	790	674	673	611
5	6231	5505	5855	6231	491	493	539	575	491	493
6	6840	5892	6114	6840	492	435	497	445	492	435
7	6255	5471	5307	6255	342	298	342	278	342	298
8	4289	3842	3505	4289	136	183	136	159	136	183
9	1971	1815	1524	1971	141	247	141	214	141	247
10	501	471	345	501	49	102	49	104	49	102
11	71	69	41	71	2	16	2	6	2	16
12	5	5	3	5		5		3		5
Total	39464	35864	35864	39464	205568	3023	205743	3163	205568	3023

Minsup 0.15% (1486)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	851	851	851	851	40419	135	40419	135	40419	135
2	4066	4053	4066	4066	357609	202	357609	204	357609	202
3	8820	8503	8807	8820	9637	652	9650	882	9637	652
4	11686	10773	11369	11686	1717	1191	2013	1368	1717	1191
5	14152	12007	13237	14152	343	762	597	929	343	762
6	20665	15671	18497	20665	92	508	193	480	92	508
7	31266	21485	26065	31266	60	152	89	162	60	152
8	40606	25912	30248	40606	125	55	130	131	125	55
9	42174	25566	26848	42174	264	99	264	227	264	99
10	34108	20224	17807	34108	231	161	231	476	231	161
11	21054	12655	8905	21054	141	210	141	197	141	210
12	9834	6243	3535	9834	7	149	7	16	7	149
13	3477	2403	1070	3477		37		7		37
14	912	700	229	912		10		4		10
15	167	143	31	167		2		0		2
16	19	18	2	19		2		2		2
17	1	1		1		1				1
Total	243858	167208	171567	243858	410645	4328	411343	5220	410645	4328

Minsup 0.1% (991)

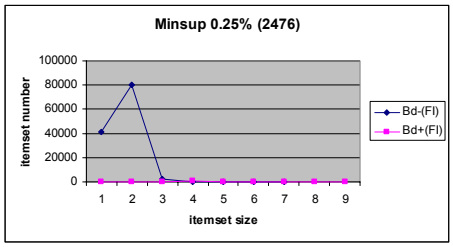
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	1259	1259	1259	1259	40011	82	40011	82	40011	82
2	8632	8548	8632	8632	783279	459	783279	476	783279	459
3	23055	21723	22971	23055	41759	1452	41843	2373	41759	1452
4	36902	33622	35568	36902	8420	3286	9601	4225	8420	3286
5	50110	44176	46806	50110	3063	2951	4284	3510	3063	2951
6	71373	57785	65312	71373	2056	2213	2291	2321	2056	2213
7	100355	73374	86190	100355	1528	1565	1612	1603	1528	1565
8	123614	82978	95126	123614	1702	1213	1726	1685	1702	1213
9	125110	78714	82209	125110	1675	1324	1675	1731	1675	1324
10	101803	61038	54054	101803	1032	1017	1032	1277	1032	1017
11	65890	38253	26521	65890	423	618	423	689	423	618
12	33298	18898	9535	33298	152	469	152	425	152	469
13	12966	7039	2520	12966	22	281	22	208	22	281
14	3929	1992	483	3929	9	79	9	25	9	79
15	926	453	63	926		12	0	17	0	25
16	159	83	3	159		6	1	3	1	6
17	18	12		18		0				0
18	1	1		1		1				1
Total	759400	529948	537252	759400	885131	17028	887961	20650	885132	17041

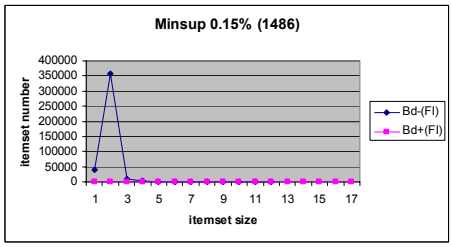
Minsup 0.09% (892)

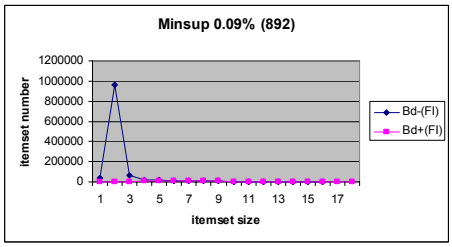
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	1392	1392	1392	1392	39878	87	39878	87	39878	87
2	10679	10530	10679	10679	957457	549	957457	577	957457	549
3	32531	30495	32382	32531	61703	1763	61852	3016	61703	1763
4	65838	59773	63790	65838	17580	4849	19320	6448	17580	4849
5	117593	103446	111443	117593	13935	5935	16229	7424	13935	5935
6	194375	160263	179597	194375	13011	7303	13996	8458	13011	7303
7	270946	205697	234866	270946	12139	7618	12619	9067	12139	7618
8	305307	212839	237712	305307	9403	7165	9469	8934	9403	7165
9	279964	179636	187567	279964	5470	5749	5470	7079	5470	5749
10	216929	129195	119702	216929	1974	3207	1974	3386	1974	3207
11	146022	81809	62991	146022	432	1351	432	1161	432	1351
12	85019	45490	26729	85019	94	533	94	503	94	533
13	41748	21777	8784	41748	24	179	24	274	24	179
14	16710	8754	2138	16710	9	79	9	71	9	79
15	5225	2865	361	5225	4	10	4	45	4	10
16	1187	737	37	1187	1	11	1	4	1	11
17	167	131	2	167		17		2		17
18	10	10		10		10				10
Total	1791642	1254839	1280172	1791642	1133114	46415	1138828	56536	1133114	46415

Some graphical representations

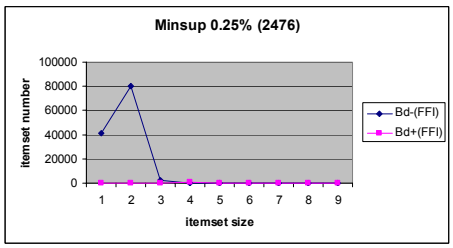
Borders of frequent itemsets

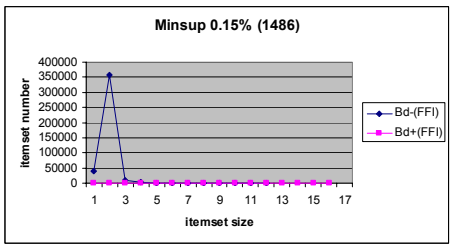


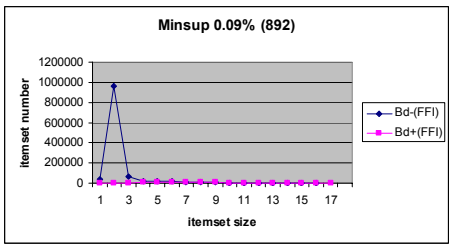




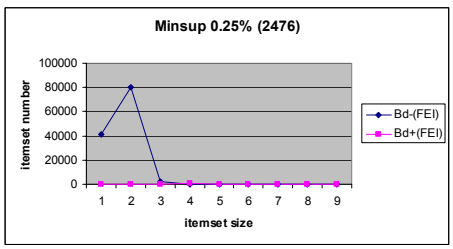
Borders of frequent free itemsets

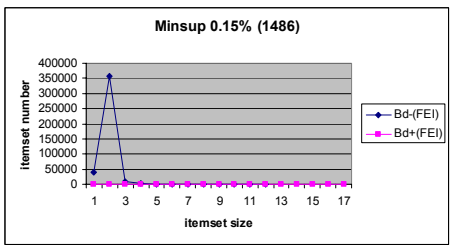


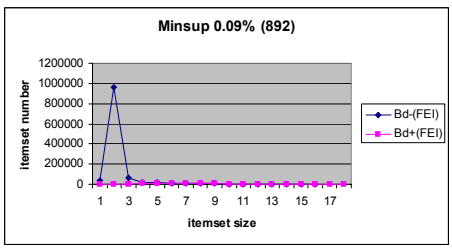




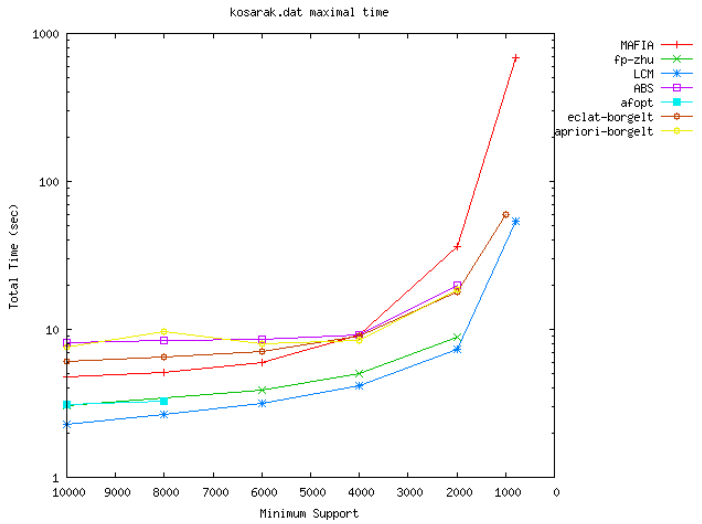
Borders of frequent essential itemsets







Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border, but only few levels lower. The borders distributions are very close, most itemsets of the positive and negative borders have approximately the same size.

Borders of other concise representations

- Frequent free itemsets: the borders distribution of free itemsets is very similar to the borders distribution of frequent itemsets, i.e. the two borders are very close.
- Frequent essential itemsets: the borders distribution of essential itemsets is very similar to the borders distribution of frequent itemsets, since the frequent essential itemsets are equal to frequent itemsets for all minimum support thresholds tested.

Performances of algorithms for MFI mining

Algorithms have relatively good performances w.r.t. the number of transactions (990 002) for minimum support thresholds with few frequent itemsets (until a minsup of ≈0.2%, i.e 2000). But when the number of frequent itemsets becomes more important, the algorithms performances grow exponentially.

Remarks:

- We have to study this dataset for very low support thresholds to have a significant number of frequent itemsets.
- The number of frequent essential itemsets is more important than the number of frequent closed itemsets.

MUSHROOM

Data description : This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy.

Class Distribution: edible: 4208 (51.8%)
poisonous: 3916 (48.2%)

Characteristics :

Number of items : 119
Number of transactions : 8 124
Average size of transactions : 23

Experimental results

Minsup 20% (1625)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	43	1	43	43	76		76	5	76	5
2	376	21	272	272	527		631	28	631	39
3	1472	99	605	461	292	6	427	129	571	160
4	3559	169	529	255	89	8	203	150	126	77
5	6267	234	231	67	27	15	47	88	14	55
6	8802	290	53	2	4	48	4	20	1	2
7	10151	231	6			51		6		
8	9488	97				25				
9	7010	33				3				
10	4004	12				0				
11	1729	4				0				
12	546	4				0				
13	119	0				0				
14	16	1				1				
15	1	1				1				
Total	53583	1197	1739	1100	1015	158	1388	426	1419	338

Minsup 10% (813)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	56	1	56	56	63		63	3	63	3
2	763	23	608	608	777	1	932	78	932	41
3	4593	140	2026	1931	1284	2	1791	335	1886	374
4	16150	358	2742	2097	640	7	1157	787	806	465
5	38800	612	1601	1084	274	26	349	512	111	380
6	69835	958	515	203	45	50	45	204	28	178
7	98846	1139	67	4	6	130	6	67	1	4
8	111786	908				180				
9	100660	434				81				
10	71342	187				37				
11	39171	65				9				
12	16292	20				3				
13	4956	17				0				
14	1039	1				0				
15	134	14				13				
16	8	8				8				
Total	574431	4885	7615	5983	3089	547	4343	1986	3827	1445

Minsup 5% (407)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	73	1	73	73	46		46	8	46	8
2	1329	23	1033	1033	1299		1595	140	1595	50
3	10618	156	4544	4944	2992		4348	989	3948	472
4	48226	488	7502	8335	2988	9	4287	1703	3099	1265
5	144928	939	5778	6200	1208	21	1475	2044	545	1386
6	315873	1711	1904	2133	427	47	434	815	96	1025
7	527176	2548	303	241	54	154	54	238	11	201
8	692740	2798	9	6	7	249	7	9		6
9	723735	2176				398				
10	600196	1112				269				
11	391578	480				110				
12	197889	170				45				
13	75624	64				11				
14	21041	27				1				
15	4000	25				11				
16	461	101				93				
17	24	24				24				
Total	3755511	12843	21146	22965	9021	1442	12246	5946	9340	4413

Minsup 1% (82)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	96	1	96	96	23		23	6	23	6
2	2426	23	1860	1860	2134		2700	210	2700	90
3	28416	161	12157	14426	5559		9404	2311	7135	927
4	192619	558	30709	48486	11885	2	16833	7780	12956	2800
5	854957	1258	34242	77118	13089	6	14440	10203	8604	5768
6	2684895	2494	18787	61290	5712	6	5786	6905	1451	9672
7	6272420	4556	4927	23282	1484	22	1484	2413	99	7978
8	11259560	7006	577	3714	168	103	168	429	6	2391
9	15841445	8757	22	201	11	377	11	22		191
10	17649729	8808		1		663				1
11	15612711	7011				1037				
12	10918055	4309				909				
13	5967341	2106				621				
14	2498192	807				301				
15	775059	418				160				
16	168614	373				97				
17	23236	1676				1179				
18	1610	1298				1265				
19	20	20				20				
Total	90751401	51640	103377	230474	40065	6768	50849	30279	32974	29824

Minsup 0.5% (41)

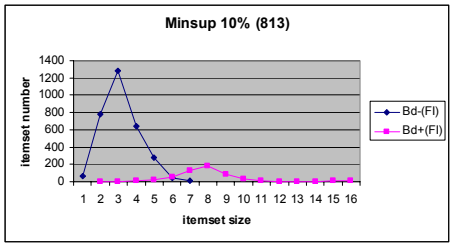
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	104	1	104	104	15		15	9	15	9
2	2791	23	2097	2097	2565		3259	173	3259	73
3	35797	161	15500	18576	5605		10451	2636	7375	894
4	270331	558	43785	74457	15040		21605	10211	16468	3629
5	1352352	1282	55897	146953	20652	3	22799	16402	17298	9296
6	4820636	2591	34690	152028	12755	4	12940	12126	4693	13932
7	12835974	4755	10783	85175	3716	19	3725	4757	345	16460
8	26330031	7495	1570	23618	504	34	504	890	22	9939
9	42428091	9970	99	2690	24	141	24	89		1979
10	54309549	11259	1	91		408		1		91
11	55468859	10579				809				
12	45138700	8049				1130				
13	29055199	4972				928				
14	14588853	2520				728				
15	5584895	1012				340				
16	1571490	742				242				
17	305741	1876				200				
18	36680	6367				5069				
19	2054	1982				1978				
20	4	4				4				
Total	294138131	76198	164526	505789	60876	12037	75322	47294	49475	56302

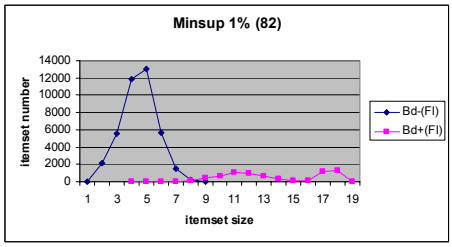
Minsup 0.1% (9)

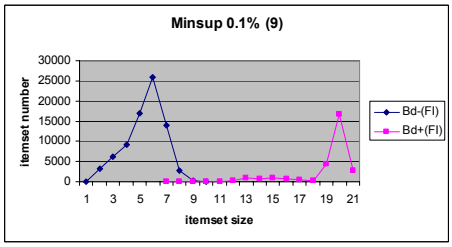
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	115	1	115	115	4		4	3	4	3
2	3324	23	2458	2458	3231		4097	144	4097	88
3	45710	161	18757	22892	6104		12271	1197	8136	643
4	379371	558	64459	110070	9141		18139	7068	11060	1845
5	2144872	1288	106970	288977	16977		20639	22469	15973	9448
6	8869210	2636	86095	429198	25857		26279	23734	20807	22730
7	28045239	4897	36194	385285	14056	5	14089	13240	7645	28346
8	69772431	7736	7618	215265	2816	5	2816	3503	223	27658
9	139053182	10457	744	68959	310	11	310	549	10	17110
10	224411720	12433	22	11084	1	27	1	22		6062
11	294884601	13322		659		66				636
12	315849140	12882		2		165				2
13	274961242	10618				821				
14	193182216	7374				713				
15	108239073	4985				893				
16	47483392	2674				655				
17	15862473	2758				433				
18	3865217	7586				222				
19	639863	19720				4332				
20	63024	23120				16688				
21	2676	2676				2676				
Total	1727758091	147905	323432	1534964	78497	27712	98645	71929	67955	114571

Some graphical representations

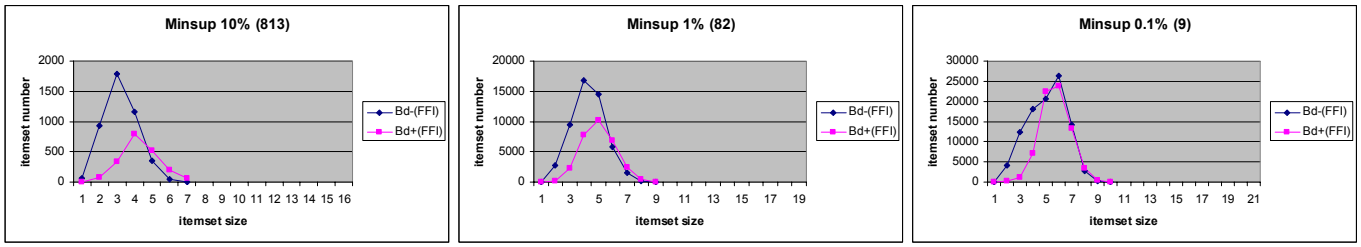
Borders of frequent itemsets



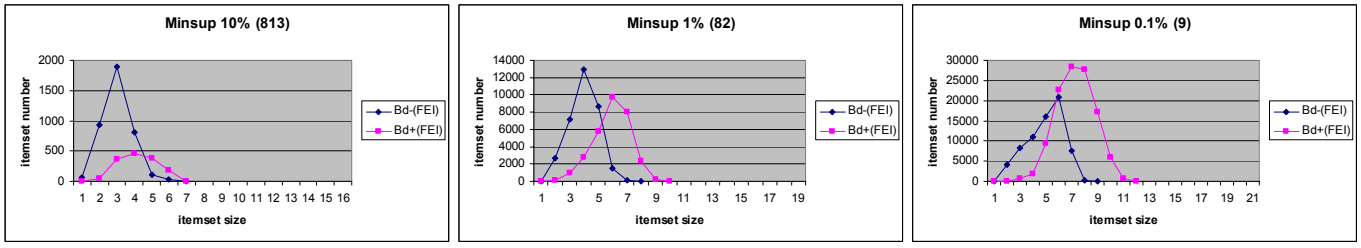




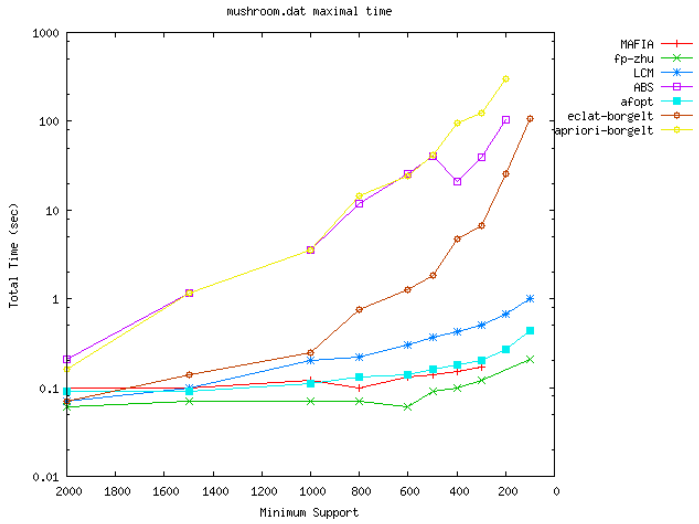
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of frequent free and essential itemsets. For the borders of frequent itemsets, there is always the same kind of distribution with an increase after the discovery of the negative border, a decrease until high levels, and another increase in the higher levels. But the intensity of the different increase changes between high values of support and low values. For high values of minimum support threshold the first peak is the most important, whereas for the low supports this is always the second peak which is the most important. Note that for all minimum support thresholds, the negative border of frequent itemsets is always "lower" than its corresponding positive border.

Borders of frequent itemsets

For high values of minimum support thresholds, even if the distance between the two borders is not very important, most of the negative border is "lower" than its corresponding positive border and there is some long itemsets in the positive border.

For low values of support (from minsup 1%), a large distance between the borders does exist, i.e. the peak of the negative border curve is many levels before the most important peak of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the borders distributions of free itemsets are very close.
- Frequent essential itemsets: the two borders are very close. As expected the number of frequent essentials is much smaller than the number of frequent free sets.

Performances of algorithms for MFI mining

For high values of minimum support threshold (1%<) , algorithms are efficient since the number of frequent itemsets is not very important, and most itemsets are not very long.

For relatively low supports with a huge number of frequent itemsets, most algorithms still efficient and stable. Note for example that for a minimum support threshold of 0.1%, the search space is composed of more than 1.7 billion of frequent itemsets!

Remarks:

- There are many long exact association rules.
- The number of frequent essential itemsets is much smaller than the number of frequent closed for high values of minimum support threshold. But for low supports, the number of frequent closed itemsets is smaller (for example for minsup 0.1%, there is more than 10 times more essential itemsets than closed).

PUMSB

Data description : Pumsb contains census data from PUMS (Public Use Microdata Samples). Each transaction represents the answers to a census questionnaire, including the age, tax-filing status, marital status, income, sex, veteran status, and location of residence of the respondent.

Characteristics :

Number of items : 2113
Number of transactions : 49 046
Average size of transactions : 74
Minimal size of transactions : 74
Maximal size of transactions : 74

Experimental results

Minsup 90% (44142)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	20	14	20	20	2093		2093		2093	
2	140	88	137	137	50		53		53	9
3	459	269	429	350	104	15	104	39	183	288
4	786	412	685	22	228	45	228	61	594	22
5	746	380	563		149	82	149	152		
6	364	224	181		56	72	56	104		
7	85	71	15		10	38	10	15		
8	7	7				7				
Total	2607	1465	2030	529	2690	259	2693	371	2923	319

Minsup 80% (39237)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	25	16	25	25	2088		2088		2088	
2	276	122	256	256	24		44		44	5
3	1760	577	1483	1376	82	1	82	24	189	482
4	6999	1796	5420	1381	566	25	566	115	3680	1151
5	18215	3960	12898	68	1733	98	1733	371	541	68
6	31532	6414	19177		2873	252	2873	1424		
7	36382	7670	17092		2666	526	2666	2176		
8	27758	6636	8748		1402	784	1402	2235		
9	13786	4109	2400		397	705	397	934		
10	4380	1639	326		41	506	41	231		
11	895	328	10		4	228	4	10		
12	131	22				17				
13	16	5				2				
14	1	1				1				
Total	142156	33295	67835	3106	11876	3145	11896	7520	6542	1706

Minsup 70% (34333)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	34	23	34	34	2079		2079		2079	
2	441	213	418	418	120	1	143	5	143	38
3	2951	967	2479	2198	417	21	417	34	698	604
4	13462	2959	9667	3402	621	59	621	198	5238	2232
5	47691	7274	28711	521	1189	130	1189	468	2839	479
6	131945	14852	66311	8	2704	218	2704	852		8
7	279661	25702	115691		5629	375	5629	2159		
8	451931	37297	149992		7399	607	7399	4748		
9	557216	44220	141740		8872	1299	8872	7308		
10	523996	42473	93526		6653	1929	6653	8410		
11	374743	32633	39624		3170	2426	3170	6687		
12	202388	19651	9276		926	2061	926	3648		
13	81619	9033	902		110	1450	110	806		
14	24203	3027	8		16	762	16	8		
15	5156	716				306				
16	755	129				73				
17	69	24				17				
18	3	3				3				
Total	2698264	241196	658379	6581	39905	11737	39928	35331	10997	3361

Minsup 60% (29428)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	39	27	39	39	2074		2074		2074	
2	610	283	586	586	131	4	155	8	155	38
3	5701	1845	5051	4314	364	3	364	9	1101	833
4	34080	7911	27065	9432	2215	57	2215	240	12097	5089
5	138306	23811	93947	2778	8950	258	8950	1929	8726	2366
6	400457	53149	219070	105	20199	977	20199	7098	144	92
7	877898	93070	367929	2	27397	2082	27397	13866		2
8	1548429	133172	480813		24488	3772	24488	18997		
9	2305272	159369	518640		16727	5095	16727	17783		
10	2958670	162940	468753		11480	5774	11480	14919		
11	3257918	145444	347102		8537	5414	8537	12004		
12	3026825	115724	202284		5311	4701	5311	9229		
13	2331941	82445	88411		2317	3342	2317	6550		
14	1468616	51596	27409		769	2354	769	3110		
15	746766	27218	5406		194	1702	194	1440		
16	302887	11491	520		26	965	26	280		
17	96676	3731	16			559		16		
18	23850	988				210				
19	4417	299				78				
20	582	93				33				
21	49	19				6				
22	2	2				2				
Total	19529991	1074627	2853041	17256	131179	37388	131203	107478	24297	8420

Minsup 55% (26976)

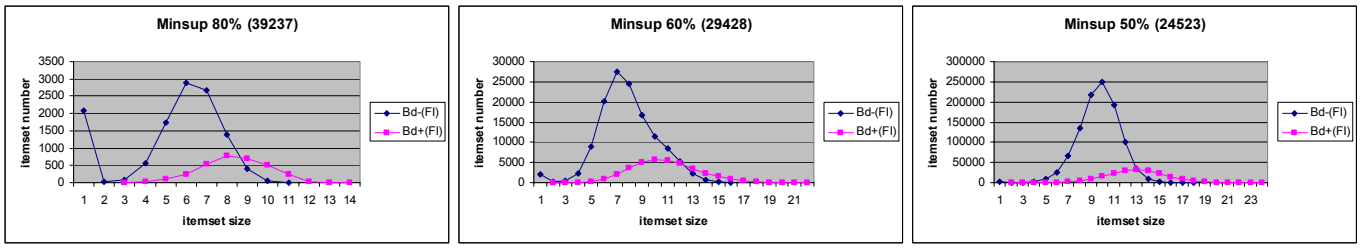
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	44	29	44	44	2069		2069		2069	
2	740	336	704	704	206		242	5	242	44
3	7118	2210	6166	5281	505	12	505	48	1390	979
4	46231	9927	35773	13569	1724	55	1724	267	14399	6332
5	215975	33250	146745	5736	7882	201	7882	1413	14318	3909
6	740946	85469	428126	553	26330	807	26330	6941	358	474
7	1911581	173118	895935	13	60295	2244	60295	23278		13
8	3794609	283036	1358940		94038	4971	94038	53212		
9	5937864	381996	1532733		95046	9164	95046	77583		
10	7529137	436629	1350654		66815	12704	66815	75339		
11	7958468	428394	966909		34416	15008	34416	52335		
12	7187499	358649	557411		13531	15036	13531	31015		
13	5631981	253988	251227		4403	12922	4403	16522		
14	3836178	151955	88300		1339	9401	1339	6753		
15	2245758	77504	24289		448	5373	448	2220		
16	1107408	34183	4708		112	2524	112	904		
17	448895	13030	496		8	1077	8	240		
18	145532	4254	16			445		16		
19	36487	1262				160				
20	6747	401				73				
21	853	139				23				
22	64	34				19				
23	2	2				2				
Total	48790117	2729795	7649176	25900	409167	92221	409203	348091	32776	11751

Minsup 50% (24523)

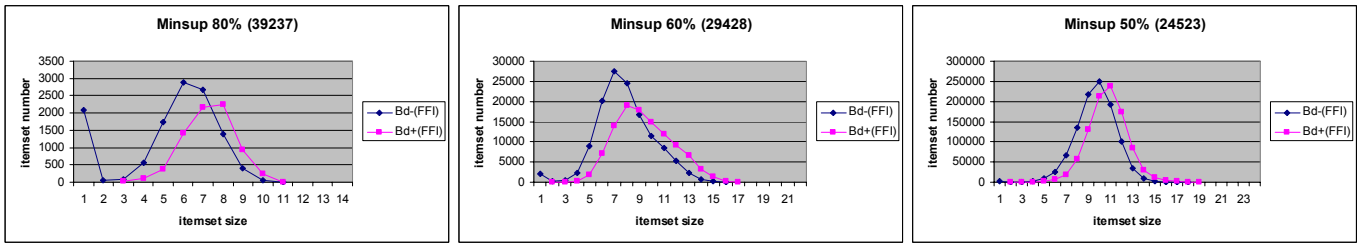
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	52	31	52	52	2061		2061		2061	
2	961	393	916	916	365	2	410	11	410	54
3	9803	2795	8448	7144	803	7	803	54	2107	1143
4	66142	13021	49516	19183	3313	83	3313	524	19380	8694
5	324271	43958	205498	9829	9390	393	9390	3046	20859	6011
6	1226917	117882	659212	1436	24771	1041	24771	6821	864	1012
7	3645422	261084	1645118	82	66052	2169	66052	19034	3	82
8	8553772	482935	3117738		136126	4531	136126	56301		
9	15948349	750674	4413430		217738	8909	217738	130252		
10	23813841	987830	4638061		248938	15303	248938	211849		
11	28745889	1111738	3649668		193450	22838	193450	238000		
12	28352901	1078882	2223156		101712	28734	101712	174082		
13	23122188	903558	1113975		35274	31172	35274	85467		
14	15784211	648317	474518		9354	28639	9354	28723		
15	9114023	394498	160070		2736	22367	2736	11309		
16	4471507	201019	37345		494	14597	494	4429		
17	1854995	84124	5278		90	8024	90	1411		
18	640966	28189	396		4	3492	4	116		
19	180076	7605	16			1196		16		
20	39802	1957				311				
21	6618	580				74				
22	775	165				45				
23	57	27				10				
24	2	2				2				
Total	165903540	7121264	22402411	38642	1052671	193939	1052716	971445	45684	16996

Some graphical representations

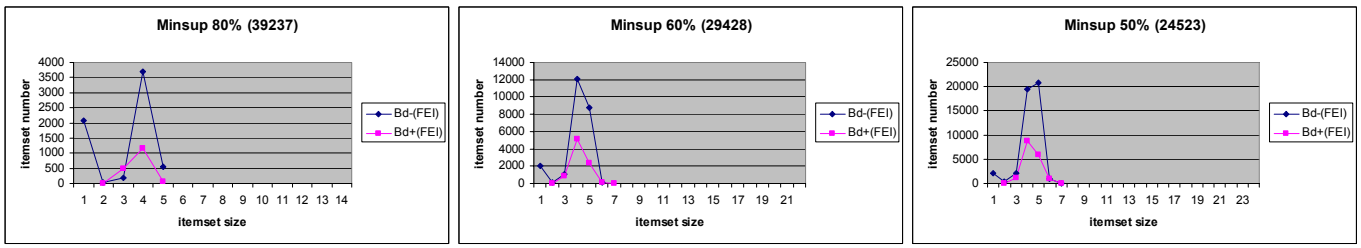
Borders of frequent itemsets



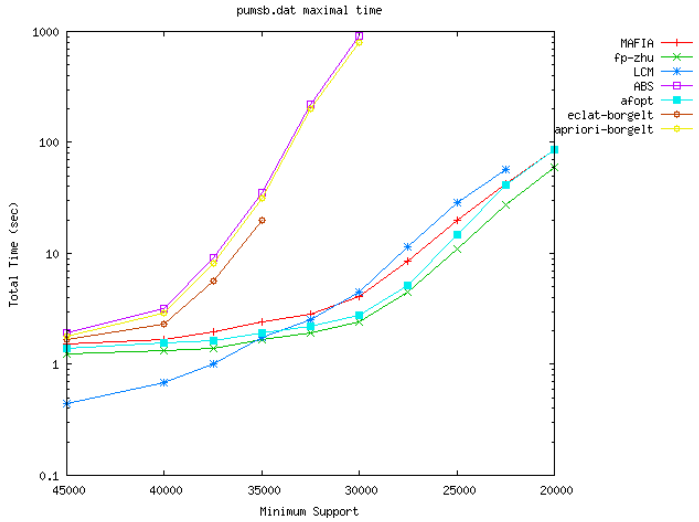
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. The borders distributions are very close, i.e. the mean of the negative border curve is only few levels before the mean of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the borders distributions of free itemsets are very close.
- Frequent essential itemsets: the two borders are very close, since essential itemsets are very small. As expected the number of frequent essentials is much smaller than the number of frequent free sets.

Performances of algorithms for MFI mining

Algorithms performances are growing exponentially with the diminution of the minimum support threshold. For a 50% minimum support threshold with 165 903 540 frequent itemsets on Pumsb, most algorithms execution time is more than 10 sec. If we compare with Connect dataset with a 44.41% minimum support threshold and 188 117 389 frequent itemsets, their execution time is less than 1sec. This difference is even more important for lower supports. Finally note that Pumsb has fewer transactions than Connect.

Remarks:

- The number of frequent essential itemsets is much smaller than the number of frequent closed itemsets by a factor of 45 to more than 4000!

PUMSB*

Data description : Pumsb* contains census data from PUMS (Public Use Microdata Samples). Each transaction represents the answers to a census questionnaire, including the age, tax-filing status, marital status, income, sex, veteran status, and location of residence of the respondent. In Pumsb* all items with 80% or more support in the original PUMS data set are deleted.

Characteristics :

Number of items : 2088
Number of transactions : 49 046
Average size of transactions : 50.5

Experimental results

Minsup 40% (19619)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	46	24	46	46	2042	2	2042	4	2042	4
2	282	109	251	251	753	8	784	12	784	12
3	974	230	638	611	155	7	211	72	238	88
4	2337	355	1001	656	47	7	50	19	229	128
5	4088	443	982	303	40	11	44	65	20	107
6	5468	465	519	70	15	3	15	62		16
7	5650	392	138	9		10		59		9
8	4451	269	12			9		12		
9	2605	161				10				
10	1090	90				3				
11	307	48				4				
12	52	20				4				
13	4	4				4				
Total	27354	2610	3587	1946	3052	82	3146	305	3313	364

Minsup 30% (14714)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	60	31	60	60	2028		2028		2028	
2	708	217	640	640	1062	5	1130	42	1130	45
3	3284	646	2158	2071	1115	23	1226	142	1313	164
4	10527	1237	4611	3333	281	26	332	229	977	433
5	25485	1881	6948	2564	130	27	139	142	272	450
6	47971	2328	6810	1130	99	26	111	167	33	218
7	71473	2411	4068	255	19	19	19	288		110
8	84545	2146	1410	22	1	31	1	235		22
9	79021	1719	254		1	43	1	148		
10	57804	1300	12		4	21	4	12		
11	32612	930				30				
12	13876	635				14				
13	4296	403				17				
14	911	201				13				
15	118	62				22				
16	7	7				7				
Total	432698	16154	26971	10075	4740	324	4991	1405	5753	1442

Minsup 25% (12262)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	68	34	68	68	2020	1	2020	1	2020	1
2	1028	279	943	943	1250	5	1335	39	1335	51
3	6106	1024	4256	4072	2551	28	2743	89	2927	233
4	23331	2222	10283	8031	829	39	940	606	1969	1011
5	65625	3684	16969	7263	432	51	461	637	1179	1227
6	142594	5084	20043	3538	280	49	304	454	135	753
7	245939	5874	15544	977	171	53	171	524	4	267
8	341388	5840	7551	143	33	55	33	731		86
9	383371	5134	2193	7		85		269		7
10	348404	4133	351			103		180		
11	255380	3196	18			45		18		
12	149810	2418				55				
13	69329	1758				46				
14	24710	1131				47				
15	6524	616				42				
16	1196	257				26				
17	135	65				20				
18	7	7				7				
Total	2064945	42756	78219	25042	7566	757	8007	3548	9569	3636

Minsup 20% (9810)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	86	39	86	86	2002	2	2002	2	2002	2
2	1720	356	1448	1448	1935	8	2207	29	2207	46
3	14966	1540	8518	8203	4850	22	5288	240	5603	446
4	87048	4022	25625	21600	2812	43	3236	1006	4880	1703
5	405470	7523	49384	29470	1356	74	1633	1256	3530	3093
6	1606900	11242	64802	23378	980	85	1162	1566	799	2353
7	5536081	14112	56261	11942	449	125	481	2472	40	1259
8	16668011	15454	32237	3741	197	138	197	1939		722
9	43863703	15252	11861	597	49	139	49	1538		324
10	100909177	13692	2580	33	8	207	8	621		33
11	203220250	11392	292			174		162		
12	359124401	9065	12			136		12		
13	558404826	6927				146				
14	765931653	4984				129				
15	928655589	3225				121				
16	996607692	1792				101				
17	947176327	849				77				
18	796952610	358				36				
19	592911099	149				11				
20	389163476	70				7				
21	224599831	38				1				
22	113452560	21				0				
23	49850385	16				0				
24	18899595	15				0				
25	6117111	18				0				
26	1666575	17				0				
27	374968	13				0				
28	67832	9				1				
29	9484	5				0				
30	962	2				1				
31	63	2				0				
32	2	2				2				
Total	7122280453	122201	253106	100498	14638	1786	16263	10843	19061	9981

Minsup 15% (7357)

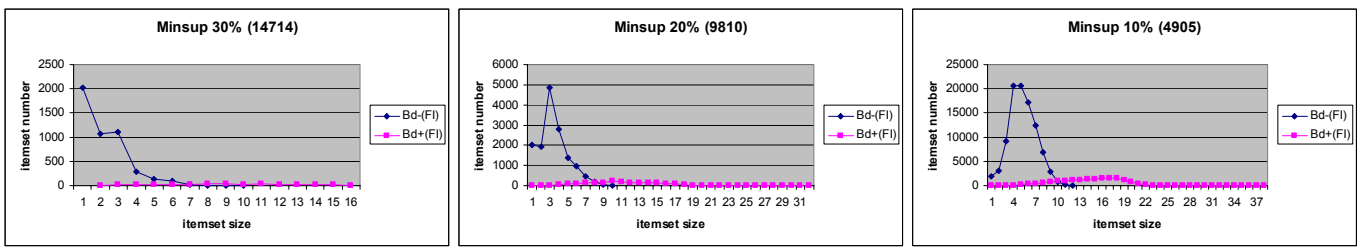
Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	94	44	94	94	1994	3	1994	3	1994	3
2	2136	433	1856	1856	2235	7	2515	18	2515	41
3	22493	2140	14410	13934	6171	24	6692	175	7168	325
4	149351	6439	56391	48690	8231	79	9174	1291	11679	2738
5	756351	13898	133892	85886	5560	122	6485	3307	10602	7420
6	3193889	23643	211934	86930	3833	166	4577	4819	4412	8697
7	11759397	33295	225851	58339	2278	200	2644	7231	401	5017
8	38417845	40675	162071	27298	903	265	939	7657	16	2172
9	111634968	44609	76725	7863	261	285	261	5776		1775
10	287749550	44510	22664	1101	85	416	85	3234		600
11	656378008	40941	3681	54	2	442	2	1463		54
12	1324350858	35334	236			401		236		
13	2365983076	29334				350				
14	3749466756	23544				405				
15	5281232615	17861				393				
16	6623118339	12260				420				
17	7404612322	7282				320				
18	7385054339	3588				283				
19	6571093524	1426				160				
20	5212906344	503				57				
21	3682070205	198				25				
22	2310704778	108				3				
23	1284476700	69				0				
24	629921760	45				0				
25	271113141	37				0				
26	101717877	34				0				
27	32982454	35				0				
28	9140626	35				0				
29	2133763	25				0				
30	411431	18				0				
31	63768	11				0				
32	7633	6				1				
33	662	6				0				
34	37	5				2				
35	1	1				1				
Total	5,5353E+10	382392	909805	332045	31553	4830	35368	35210	38787	28842

Minsup 10% (4905)

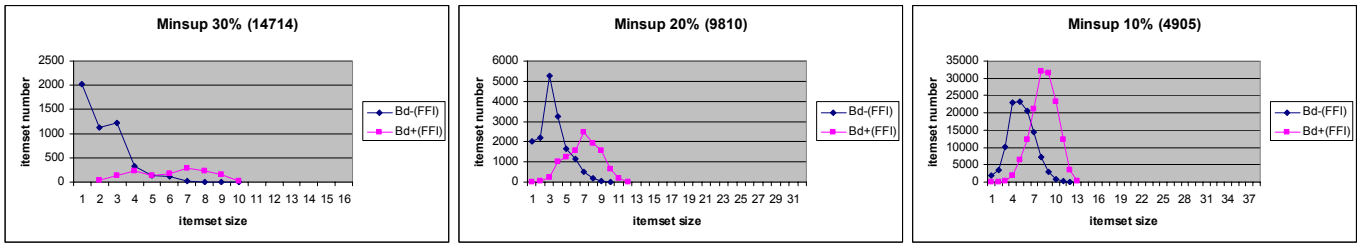
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	110	52	110	110	1978	1	1978	1	1978	1
2	2919	614	2593	2593	3076	18	3402	53	3402	64
3	36712	3254	24829	24267	9249	46	10068	274	10630	417
4	295204	10931	125442	113123	20682	93	22870	1977	25988	4273
5	1762618	26994	381234	269754	20559	192	23274	6541	34676	17992
6	8523093	51952	762253	360280	17209	307	20657	12164	23719	32072
7	35036385	82236	1027736	310449	12428	389	14384	20975	5479	24486
8	125693594	112116	934743	196324	6840	579	7201	32042	431	13025
9	398331199	137346	567278	89727	2941	721	2941	31553	19	6953
10	1119899325	154650	221843	26563	797	915	797	23202		3898
11	2796709999	161075	52146	4231	196	1025	196	12292		2011
12	6208751794	156472	6443	244	42	1227	42	3371		244
13	12270020576	143668	297			1191		297		
14	21628215724	126771				1282				
15	34080650791	107538				1321				
16	48114407028	86300				1516				
17	60979098457	63494				1539				
18	69488373790	41770				1489				
19	71276990523	24158				1102				
20	65849035735	12150				752				
21	54794082345	5240				412				
22	41046527064	1979				192				
23	27649632135	717				73				
24	16718542392	312				32				
25	9051293502	190				1				
26	4372882046	157				0				
27	1877074675	143				0				
28	711941504	135				1				
29	236920663	122				0				
30	68561150	101				0				
31	17056758	74				3				
32	3594002	51				1				
33	628735	38				2				
34	88832	31				6				
35	9735	20				5				
36	776	9				1				
37	40	4				2				
38	1	1				1				
Total	5,50931E+11	1512865	4106947	1397665	95997	16437	107810	144742	106322	105436

Some graphical representations

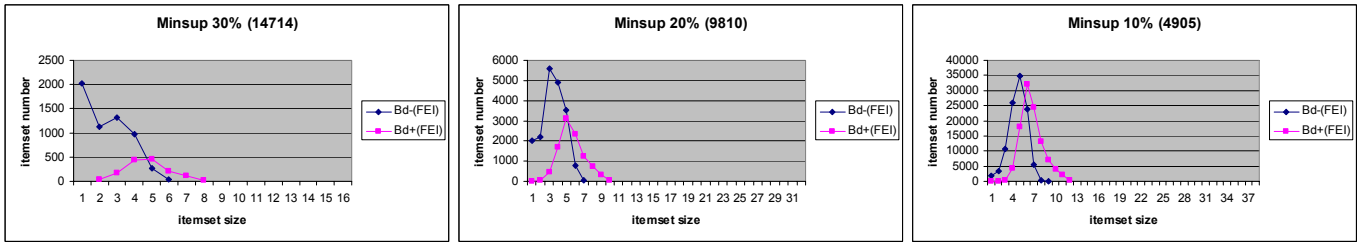
Borders of frequent itemsets



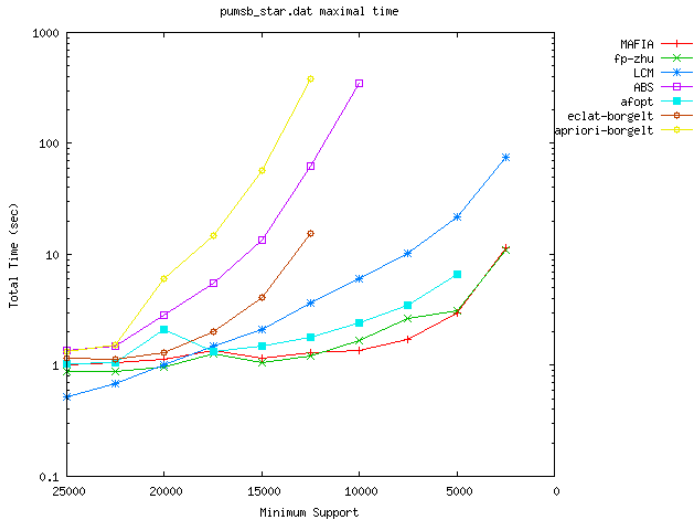
Borders of frequent free itemsets



Borders of frequent essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. A distance between the borders does exist, i.e. the mean of the negative border curve is many levels before the mean of the positive border curve. But it is less visible since the positive border is more spread out, i.e. the peak of the positive border curve is distributed on many levels. Nevertheless, most itemsets of the positive border are much longer than those of the negative border.

Borders of other concise representations

- Frequent free itemsets: the borders distributions of free itemsets are very close, since free itemsets are relatively small.
- Frequent essential itemsets: the two borders are very close too.

Performances of algorithms for MFI mining

Algorithms are efficient and stable, even for relatively low supports with a huge number of frequent itemsets. For example, for a 15% minimum support threshold with more than 55 billion of frequent itemsets, algorithms execution time is near 1 sec. Note that Pumsb* and Pumsb are very similar w.r.t. the transactions and number of items, but algorithms performances on these datasets are very different.

Remarks:

- This dataset have many long exact association rules.

RETAIL

Data description : The dataset contains information about the market basket of clients in a Belgian supermarket. Data were collected over three non-consecutive periods. This results in approximately 5 months of data. Each record in the data set contains information about the date of purchase, the receipt number, the article number, the number of items purchased, the article price in Belgian Francs and the customer number. Although most of the products are identified by a unique barcode, some article numbers in the dataset represent a group of products rather than an individual product item. In total, 5,133 customers have purchased at least one product in the supermarket during the data collection period.

Some data statistics :

1. The average number of distinct items (i.e. different products) purchased per shopping visit equals 13 and most customers buy between 7 and 11 items per shopping visit.
2. Most customers have visited the store from 4 to 24 times over the entire period (24 weeks), the average number of visits to the store equals 25, which corresponds to about once per week.
3. Most of the visits to the store take place on Thursday, Friday and Saturday.
4. 89% the items in the assortment are slow moving, i.e. sold on average less than once per day.

Characteristics :

Number of items : 16 470
Number of transactions : 88 162
Average size of transactions : 10.3

Experimental results

Minsup 0.05% (45)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	3926	3925	3926	3926	12544	1291	12544	1292	12544	1292
2	8198	8154	8197	8197	7696577	2982	7696578	3007	7696578	2981
3	5411	5352	5367	5410	10835	2701	10879	2716	10836	2705
4	1534	1511	1475	1529	866	1012	899	1019	868	1016
5	170	169	147	168	85	154	89	136	85	152
6	3	3	2	3	2	3	2	2	2	3
Total	19242	19114	19114	19233	7720909	8143	7720991	8172	7720913	8149

Minsup 0.01% (9)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	8919	8902	8919	8919	7551	1491	7551	1504	7551	1504
2	67992	62825	67975	67975	39701829	21026	39701846	23673	39701846	21054
3	83417	70792	76993	83382	828985	26119	835404	33020	829015	26121
4	47180	35732	31888	47141	24605	19527	31026	20009	24613	19558
5	17095	8803	5024	17060	2569	5845	3532	3819	2568	5851
6	7727	1429	420	7718	168	881	220	250	168	872
7	4617	393	45	4617	19	207	24	37	19	207
8	2489	122	1	2489	1	56	1	1	1	56
9	1045	49		1045		26				26
10	309	19		309		15				15
11	57	6		57		2				2
12	5	5		5		5				5
Total	240852	189077	191265	240717	40565727	75200	40579604	82313	40565781	75271

Minsup 0.008% (7)

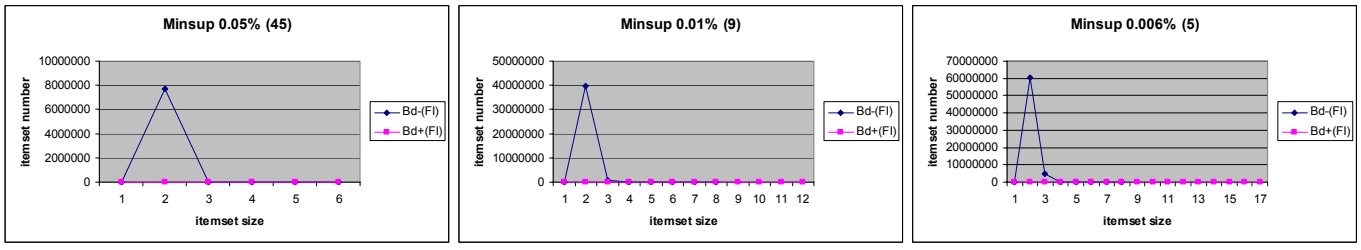
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	9788	9736	9788	9788	6682	1529	6682	1564	6682	1564
2	100851	89286	100795	100795	47796727	30674	47796783	37073	47796783	30707
3	136656	106187	121483	136574	1735940	39297	1751088	55283	1735997	39305
4	91173	58590	51629	91088	43996	30934	60499	32213	43996	30985
5	46056	17097	8681	45976	4759	10788	7081	6465	4757	10813
6	30309	3595	782	30275	398	2111	513	520	402	2094
7	24478	1091	75	24475	44	545	52	60	44	542
8	18730	487	2	18730	2	217	2	2	2	217
9	12321	228		12321		136				136
10	6579	84		6579		52				52
11	2698	28		2698		16				16
12	801	9		801		1				1
13	160	12		160		6				6
14	19	4		19		4				4
15	1	1		1		1				1
Total	480620	286435	293235	480280	49588548	116311	49622700	133180	49588663	116443

Minsup 0.006% (5)

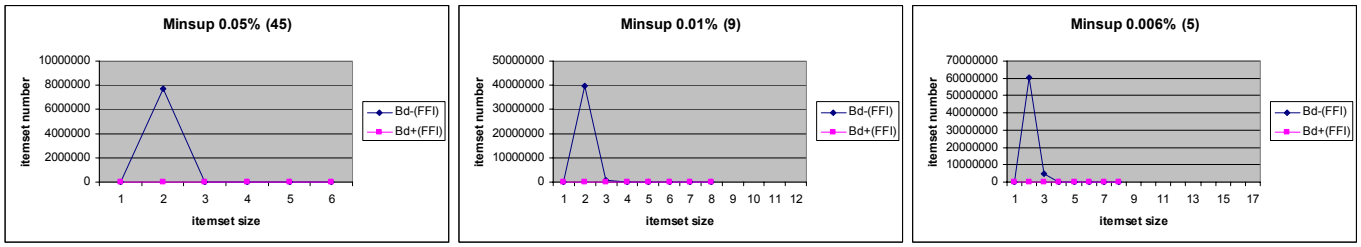
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	10988	10853	10988	10988	5482	1566	5482	1655	5482	1655
2	175951	142039	175803	175803	60186627	50023	60186775	70294	60186775	50082
3	278283	182726	229900	278057	4626656	70913	4674918	116474	4626761	70944
4	235071	111666	96872	234727	103193	57719	159251	61453	103207	57812
5	175873	38737	16972	175461	10560	22688	16735	12166	10551	22752
6	154014	10701	1640	153694	895	6119	1159	1087	900	6113
7	139913	3847	157	139730	52	1943	63	89	53	1943
8	119659	1867	10	119585	4	995	4	10	4	993
9	93337	928		93319		497				499
10	63750	456		63748		261				259
11	36264	195		36264		126				126
12	16389	64		16389		41				41
13	5629	29		5629		16				16
14	1397	15		1397		2				2
15	233	11		233		7				7
16	23	7		23		6				6
17	1	1		1		1				1
Total	1506775	504142	532342	1505048	64933469	212923	65044387	263228	64933733	213251

Some graphical representations

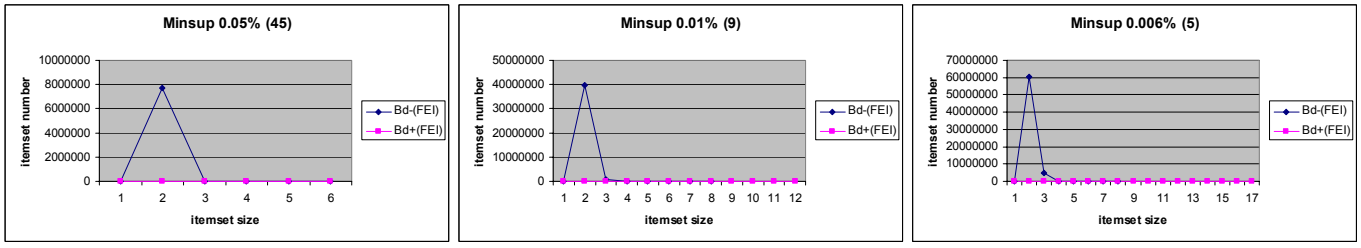
Borders of frequent itemsets



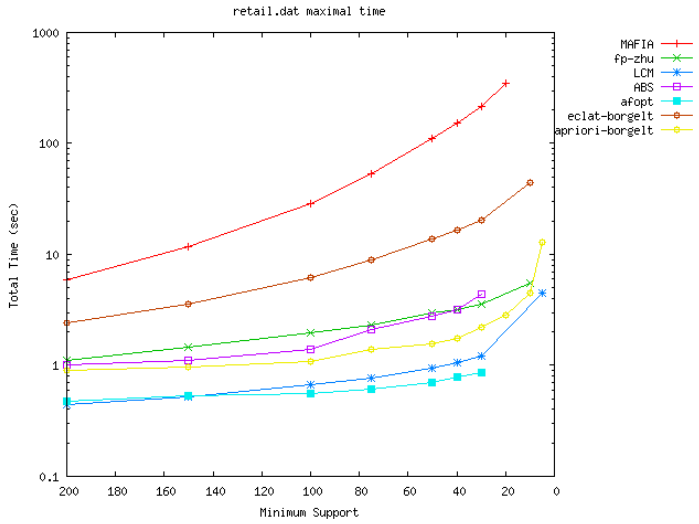
Borders of frequent free itemsets



Borders of essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. Even if the positive border appears to have some longer itemsets than the negative border, most itemsets in the two borders have relatively the same size. Consequently, the borders distributions are very close, i.e. the mean of the negative border curve is only one level before the mean of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the two borders are very close. Actually the two borders are distributed on the same levels.
- Frequent essential itemsets: the borders distribution of essential itemsets is very similar to the borders distribution of frequent itemsets, since the frequent essential itemsets are approximately equal to frequent itemsets for all the minimum support thresholds tested.

Performances of algorithms for MFI mining

Algorithms are efficient for this dataset, since most of the frequent itemsets are small.

Remarks:

- We have to study this dataset for very low support thresholds to have a significant number of frequent itemsets.
- The number of frequent essential itemsets is much more important than the number of frequent closed itemsets.

T10I4D100K

Data description : This dataset is a synthetic dataset generated using the generator from the IBM Almaden Quest research group [IBM], based on the method presented in [AGR94]. To summarize, the goal of this generation is to create transactions similar to those obtained in a supermarket environment. By applying certain distribution laws, the datasets tend to model the real world compared to given characteristics. The data are generated in order to correspond, on average, to the input characteristics while respecting a certain distribution and the existence of exceptions.

The different parameters for the generation are the following ones :

1. Number of transactions (D)
2. Average size of transactions (T)
3. Average size of long potentially maximal itemsets (I)
4. Number of long potentially maximal itemsets (L)
5. Number of items (N)

Characteristics :

Number of items : 870
Number of transactions : 100 000
Average size of transactions : 10
Average size of maximal itemsets : 4

Experimental results

Minsup 0.15% (151)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	766	766	766	766	104	59	104	59	104	59
2	5508	5505	5508	5508	287487	1535	287487	1535	287487	1551
3	4794	4662	4791	4699	7912	296	7915	359	8007	357
4	3839	3667	3702	3634	88	239	216	275	120	254
5	2371	2260	2205	2213	54	111	59	126	55	122
6	1160	1122	1063	1088	13	73	15	89	14	71
7	417	414	386	399	10	53	10	38	10	51
8	110	110	108	108		11		9		9
9	21	21	21	21		1		1		1
10	2	2	2	2		2		2		2
Total	18988	18529	18552	18438	295668	2380	295806	2493	295797	2477

Minsup 0.1% (100)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	797	796	797	797	73	22	73	22	73	22
2	8831	8813	8830	8830	308375	2888	308376	2892	308376	2914
3	7130	6905	7113	6978	35947	459	35959	579	36094	544
4	5499	5251	5274	5181	169	309	357	360	216	330
5	3188	3039	2957	2950	75	222	80	230	72	224
6	1429	1365	1294	1323	9	85	11	91	10	84
7	503	485	448	476	2	51	2	46	2	48
8	130	127	116	127	1	13	1	17	1	10
9	23	23	21	23		3		1		3
10	2	2	2	2		2		2		2
Total	27532	26806	26852	26687	344651	4054	344859	4240	344844	4181

Minsup 0.011% (11)

Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	867	846	867	867	3		3	3	3	3
2	102446	101297	102405	102405	272965	55140	273006	55661	273006	55210
3	121005	93601	118982	120529	9432170	42062	9434064	70305	9432517	42319
4	62298	28812	30081	60974	49985	8385	78732	17338	50062	8989
5	30792	13241	6410	29263	209	4317	2144	1209	214	4431
6	12901	5022	2248	12126	11	1635	13	257	12	1587
7	5341	1615	690	5118	1	422	1	86	1	387
8	2284	482	160	2246		118		25		107
9	927	112	25	924		22		5		19
10	316	19	2	316		8		2		8
11	80	3		80		2				2
12	13	2		13		0				0
13	1	1		1		1				1
Total	339271	245053	261870	334862	9755344	112112	9787963	144891	9755815	113063

Minsup 0.006% (6)

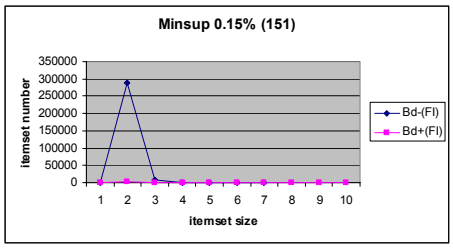
Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	869	846	869	869	1		1	3	1	3
2	157542	150874	157490	157490	219604	49018	219656	52622	219656	49061
3	484367	301935	469889	483836	19651957	130736	19666256	314629	19652309	132058
4	329545	83032	97404	323827	522472	31736	730073	79162	522617	35931
5	189478	40079	8078	179590	708	19077	10588	2557	731	20314
6	83469	16797	2319	77454	8	9072	11	308	8	9188
7	29526	5712	693	27554		3443		89		3146
8	8748	1750	160	8405		1097		25		970
9	2254	423	25	2228		286		5		260
10	509	76	2	509		48		2		48
11	95	18		95		17				17
12	13	2		13		0				0
13	1	1		1		1				1
Total	1286416	601545	736929	1261871	20394750	244531	20626585	449402	20395322	250997

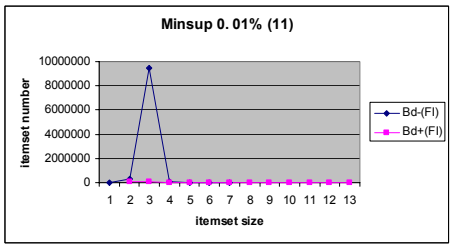
Minsup 0.003% (3)

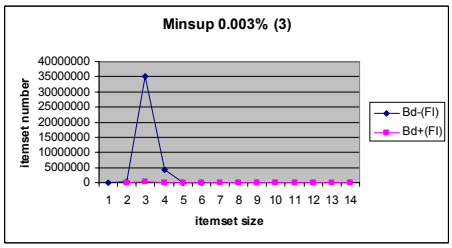
Itemset size	Fl	FCl	FFl	FEl	Bd-(Fl)	Bd+(Fl)	Bd-(FFl)	Bd+(FFl)	Bd-(FEl)	Bd+(FEl)
1	869	846	869	869	1		1	3	1	3
2	220988	191763	220936	220936	156158	22634	156210	39171	156210	22675
3	1650009	780242	1572848	1648999	35173474	303090	35250002	1027965	35173851	308240
4	1696318	251468	449641	1669404	4224179	114790	5308171	417049	4224556	128245
5	1279623	122802	12046	1224429	6478	62444	55850	6297	6562	68424
6	756202	61827	2364	710710	16	36241	20	353	32	38282
7	361441	26897	693	338067		18239		89	2	17767
8	142925	10243	160	133954		7478		25		6844
9	46576	3249	25	43363		2600		5		2362
10	12127	901	2	10940		743		2		727
11	2404	329		2009		304				294
12	340	46		242		41				47
13	30	17		15		16				15
14	1	1				1				
Total	6169853	1450631	2259584	6003937	39560306	568621	40770254	1490959	39561214	593925

Some graphical representations

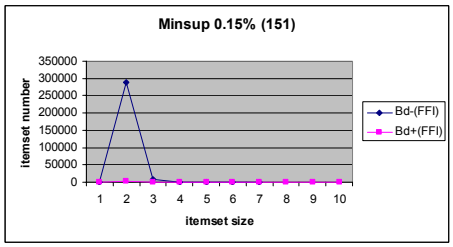
Borders of frequent itemsets

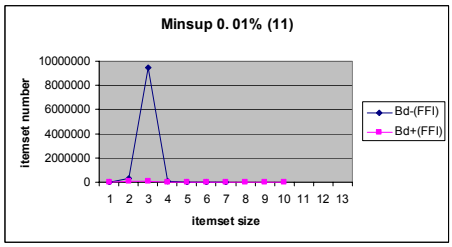


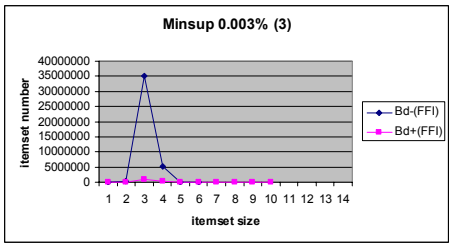




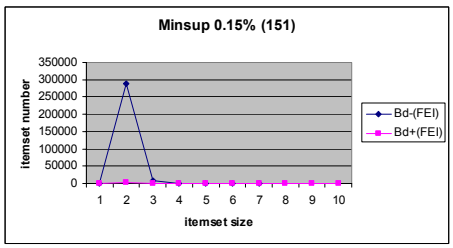
Borders of frequent free itemsets

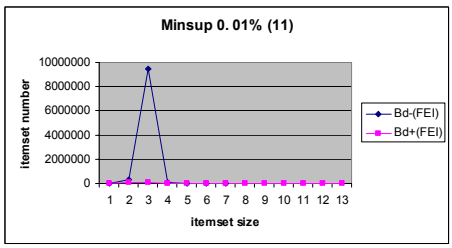


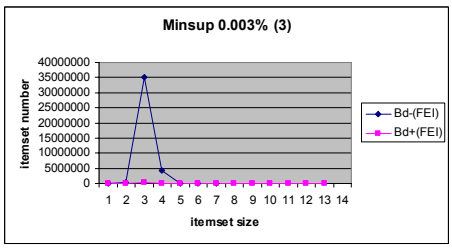




Borders of essential itemsets







Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. Even if the positive border appears to have some longer itemsets than the negative border, most itemsets in the two borders have relatively the same size. Consequently, the borders distributions are very close, i.e. the mean of the negative border curve is at the same level or one level before the mean of the positive border curve.

Borders of other concise representations

- Frequent free itemsets: the two borders are very close.
- Frequent essential itemsets: the borders distribution of essential itemsets is very similar to the borders distribution of frequent itemsets, since the frequent essential itemsets are approximately equal to frequent itemsets for all the minimum support thresholds tested.

Performances of algorithms for MFI mining

Algorithms are efficient for this dataset, since most of the frequent itemsets are small.

Remarks:

- We have to study this dataset for very low support thresholds to have a significant number of frequent itemsets.
- The number of frequent essential itemsets is much more important than the number of frequent closed itemsets.

T40I10D100K

Data description : This dataset is a synthetic dataset generated using the generator from the IBM Almaden Quest research group, based on the method presented in [AGR 94]. See T10I4D100K for more details.

Characteristics :

Number of items : 942
Number of transactions : 100 000
Average size of transactions : 40
Average size of maximal itemsets : 10

Experimental results

Minsup 1.5% (1500)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	682	682	682	682	260	296	260	296	260	296
2	4408	4408	4408	4408	227813	4227	227813	4227	227813	4227
3	259	259	259	259	41938	55	41938	55	41938	55
4	366	366	366	366	6	1	6	1	6	1
5	483	483	483	483	0	34	0	34	0	34
6	340	340	340	340	129	333	129	333	129	333
7	1	1	1	1	143	1	143	1	143	1
Total	6539	6539	6539	6539	270289	4947	270289	4947	270289	4947

Minsup 1% (1001)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	755	755	755	755	187	177	187	177	187	177
2	12921	12921	12921	12921	271714	10856	271714	10856	271714	10856
3	4359	4359	4359	4359	233972	1446	233972	1446	233972	1446
4	6819	6819	6819	6819	1308	669	1308	669	1308	669
5	12535	12535	12535	12535	1117	308	1117	308	1117	308
6	15214	15214	15214	15214	6492	3179	6492	3179	6492	3179
7	7588	7588	7588	7588	4885	4550	4885	4550	4885	4550
8	2011	2011	2011	2011	1	19	1	19	1	19
9	1000	1000	1000	1000	2	10	2	10	2	10
10	363	363	363	363	1	0	1	0	1	0
11	91	91	91	91		1		1		1
12	14	14	14	14		1		1		1
13	1	1	1	1		1		1		1
Total	63671	63671	63671	63671	519679	21217	519679	21217	519679	21217

Minsup 0.8% (801)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	783	783	783	783	159	117	159	117	159	117
2	20201	20201	20201	20201	285952	16091	285952	16091	285952	16091
3	8820	8820	8820	8794	500030	2608	500030	2608	500056	2626
4	14321	14321	14321	14111	1798	802	1798	802	1835	808
5	26871	26871	26871	26377	711	139	711	139	711	140
6	44585	44585	44585	43793	470	854	470	854	470	854
7	63020	63020	63020	62096	358	172	358	172	358	172
8	76401	76401	76401	75609	144	439	144	439	144	439
9	77445	77445	77445	76950	93	8	93	8	93	8
10	65156	65156	65156	64936	0	87	0	87	0	87
11	44769	44769	44769	44703	43	2009	43	2009	43	2009
12	22710	22710	22710	22698	2160	3810	2160	3810	2160	3798
13	8607	8607	8607	8607	867	25	867	25	866	25
14	3061	3061	3061	3061		1		1		1
15	816	816	816	816		0		0		0
16	153	153	153	153		0		0		0
17	18	18	18	18		0		0		0
18	1	1	1	1		1		1		1
Total	477738	477738	477738	473707	792785	27163	792785	27163	792847	27176

Minsup 0.5% (501)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	838	838	838	838	104	59	104	59	104	59
2	42914	42914	42914	42914	307789	29437	307789	29437	307789	29439
3	39230	39135	39230	39171	1820324	12029	1820324	12032	1820383	12058
4	61712	60586	61611	61345	15475	2936	15576	3174	15560	2965
5	106758	104207	105524	105966	5289	2819	6045	3260	5244	2809
6	155170	152496	152643	154046	3673	3099	3910	3064	3669	3100
7	193527	191514	190939	192335	1517	1850	1517	1870	1517	1850
8	204597	203520	202659	203658	4105	9077	4105	9115	4104	9073
9	176957	176574	175915	176410	6504	2929	6504	2963	6504	2933
10	135706	135625	135332	135475	792	871	792	815	792	871
11	88234	88221	88153	88167	219	493	219	490	219	492
12	47237	47236	47224	47225	138	166	138	166	138	167
13	20470	20470	20469	20469	20	9	20	8	20	8
14	6970	6970	6970	6970		5		5		5
15	1787	1787	1787	1787		3		3		3
16	324	324	324	324		1		1		1
17	37	37	37	37		1		1		1
18	2	2	2	2		2		2		2
Total	1282470	1272456	1272571	1277139	2165949	65786	2167043	66465	2166043	65836

Minsup 0.25% (251)

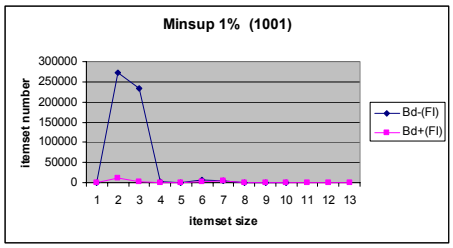
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	894	894	894	894	48	22	48	22	48	22
2	92740	92735	92740	92740	306431	52105	306431	52105	306431	52121
3	201433	200534	201428	201077	7289380	102316	7289385	102336	7289736	102395
4	232006	216547	230320	230305	462305	10224	463950	11478	462524	10297
5	433029	363949	407077	429217	12760	6395	26454	11650	12779	6410
6	697758	540925	596901	691909	9389	5693	19389	10683	9386	5684
7	968981	706871	758090	962259	6462	4897	7422	7490	6367	4783
8	1162086	805357	834404	1156117	6824	7139	6831	10039	6625	6936
9	1189035	790232	782516	1184442	10382	9274	10382	10461	10328	9215
10	1020342	648799	611785	1017101	16111	22759	16111	19884	16108	22765
11	726588	433572	387854	724617	12195	14551	12195	12576	12195	14551
12	445877	251818	211756	444913	1660	2771	1660	1700	1660	2771
13	231984	126096	98096	231630	986	2187	986	1548	986	2185
14	97303	50676	34697	97213	1305	2426	1305	2404	1305	2427
15	31958	15885	8539	31944	493	983	493	989	493	985
16	8419	4240	1609	8418	32	40	32	47	32	39
17	1719	948	235	1719	3	12	3	10	3	12
18	250	160	22	250		3		3		3
19	23	18	1	23		3		1		3
20	1	1		1		1				1
Total	7542426	5250257	5258964	7506789	8136766	243801	8163077	255426	8137006	243605

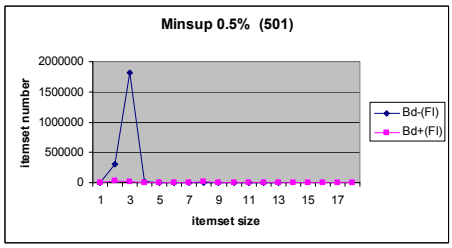
Minsup 0.2% (201)

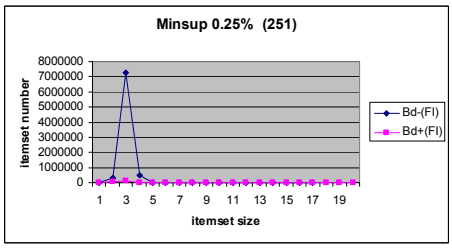
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	906	905	906	906	36	11	36	11	36	11
2	112247	112229	112246	112246	297718	59681	297719	59681	297719	59704
3	325594	323841	325574	325088	10337908	187790	10337916	187816	10338402	187884
4	322701	293125	319601	320442	1210147	14552	1213098	16853	1210330	14601
5	614970	485203	567291	609985	12521	6925	37669	17082	12504	6898
6	1019720	724397	836334	1011938	9557	7089	27839	14076	9512	7038
7	1460519	957226	1073870	1451089	5757	7616	7424	8842	5748	7619
8	1812495	1103142	1193987	1803365	6981	8678	6991	8945	6976	8682
9	1941405	1106490	1134120	1934244	6709	5315	6709	6531	6688	5317
10	1784233	967077	917486	1779637	4362	4131	4362	6175	4350	4114
11	1382495	725070	620157	1380087	6338	5075	6338	9063	6321	5004
12	881436	455717	339071	880413	7018	7992	7018	9700	7003	7959
13	452656	233687	146603	452300	3960	8069	3960	4502	3960	8065
14	187949	96679	50793	187859	936	4091	936	1191	936	4092
15	63715	33060	14082	63701	77	570	77	135	77	572
16	17335	9541	2999	17334	2	29	2	37	2	28
17	3570	2178	451	3570	1	4	1	19	1	4
18	515	363	43	515		5		5		5
19	44	39	2	44		24		2		24
20	1	1		1		1				1
Total	12384506	7629970	7655616	12334764	11910028	327648	11958095	350666	11910565	327622

Some graphical representations

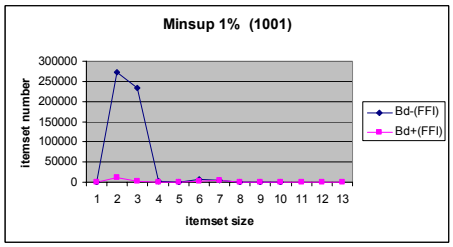
Borders of frequent itemsets

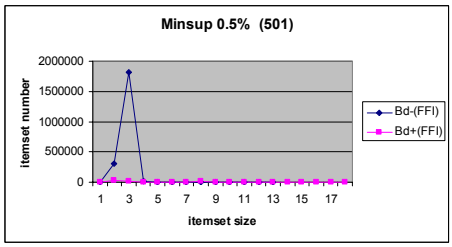


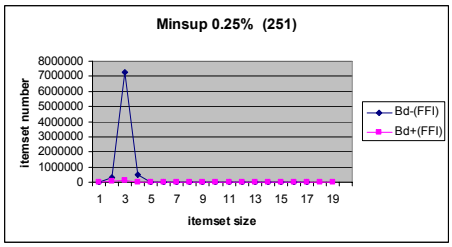




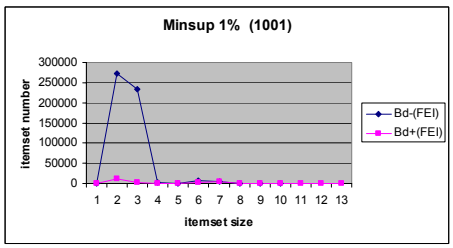
Borders of frequent free itemsets

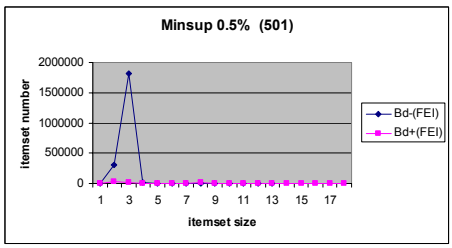


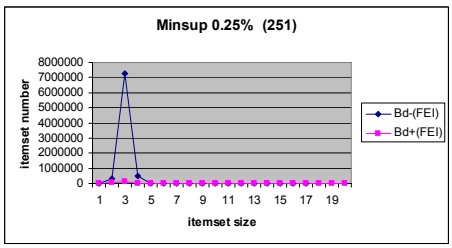




Borders of essential itemsets







Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. The two borders have specific distributions, since the curves have two peaks (see result tables). The borders distributions are very close, i.e. the two peaks of the negative border curve are at the same level or one level before the two peaks of the positive border curve.

Borders of other concise representations

The borders distributions of frequent free and essential itemsets have the same distribution than the borders of frequent itemsets.

Performances of algorithms for MFI mining

Algorithms are efficient for this dataset, since most of the frequent itemsets are relatively small.

Remarks:

- We have to study this dataset for very low support thresholds to have a significant number of frequent itemsets.
- For many minimum support thresholds, there is few exact association rules ($|FFI| \approx |FI|$).
- The number of frequent essential itemsets is equal or more important than the number of frequent closed itemsets.

WEBDOCS

Data description : This dataset was built from a spidered collection of web html documents. The whole collection contains about 1.7 millions documents, mainly written in English, and its size is about 5GB. The transactional dataset was built from the web collection in the following way. All the web documents were preliminarily filtered by removing html tags and the most common words (stopwords), and by applying a stemming algorithm. Then from each document a distinct transaction containing the set of all the distinct terms (items) appearing within the document itself, was generated. The resulting dataset has a size of about 1,48GB.

Characteristics :

Number of items : 5 267 656
Number of transactions : 1 692 082
Maximal size of transactions : 71 472

Experimental results

Minsup 17.73% (300000)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	83	83	83	83	5267573	14	5267573	14	5267573	14
2	400	400	400	400	3003	19	3003	19	3003	19
3	896	896	896	896	505	30	505	30	505	30
4	1061	1061	1061	1061	158	76	158	76	158	76
5	715	715	715	715	29	98	29	98	29	98
6	266	266	266	266	4	89	4	89	4	89
7	45	45	45	45	1	30	1	30	1	30
8	2	2	2	2		2		2		2
Total	3468	3468	3468	3468	5271273	358	5271273	358	5271273	358

Minsup 14.77% (250000)

Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	122	122	122	122	5267534	12	5267534	12	5267534	12
2	768	768	768	768	6613	21	6613	21	6613	21
3	2154	2154	2154	2154	1504	68	1504	68	1504	68
4	3299	3299	3299	3299	577	144	577	144	577	144
5	2995	2995	2995	2995	199	238	199	238	199	238
6	1632	1632	1632	1632	33	279	33	279	33	279
7	513	513	513	513	5	153	5	153	5	153
8	78	78	78	78		54		54		54
9	3	3	3	3		3		3		3
Total	11564	11564	11564	11564	5276465	972	5276465	972	5276465	972

Minsup 11.82% (200000)

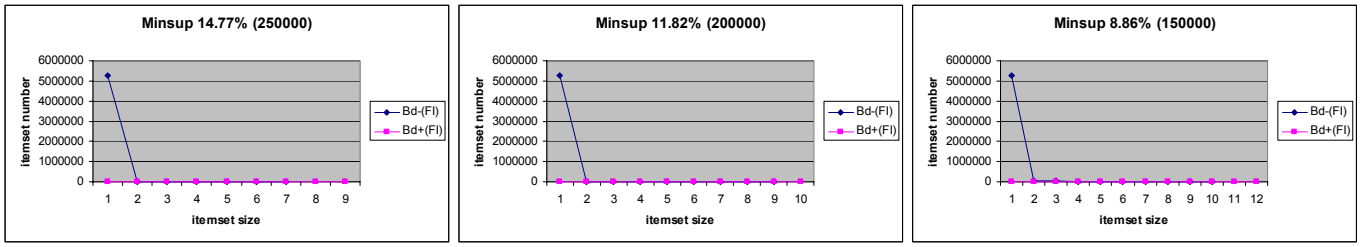
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	195	195	195	195	5267461	14	5267461	14	5267461	14
2	1596	1596	1596	1596	17319	44	17319	44	17319	44
3	5933	5933	5933	5933	5144	126	5144	126	5144	126
4	12393	12393	12393	12393	2810	289	2810	289	2810	289
5	15978	15978	15978	15978	1261	566	1261	566	1261	566
6	13075	13075	13075	13075	430	970	430	970	430	970
7	6724	6724	6724	6724	83	859	83	859	83	859
8	2052	2052	2052	2052	5	568	5	568	5	568
9	328	328	328	328	1	153	1	153	1	153
10	22	22	22	22		22		22		22
Total	58296	58296	58296	58296	5294514	3611	5294514	3611	5294514	3611

Minsup 8.86% (150000)

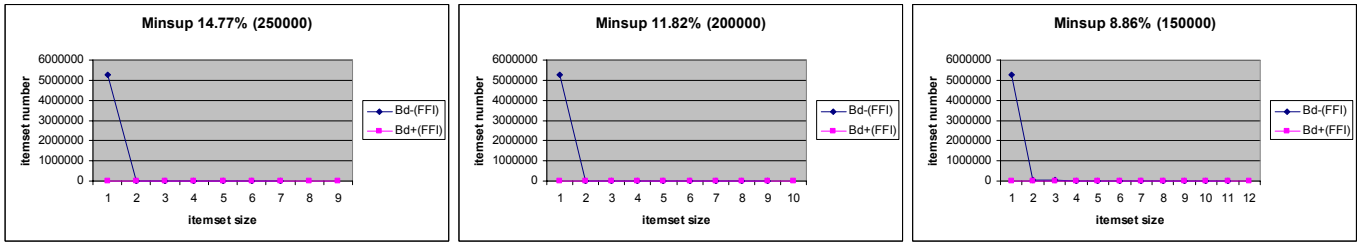
Itemset size	FI	FCI	FFI	FEI	Bd-(FI)	Bd+(FI)	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	313	313	313	313	5267343	11	5267343	11	5267343	11
2	3929	3929	3929	3929	44899	85	44899	85	44899	85
3	20486	20486	20486	20486	24539	297	24539	297	24539	297
4	61646	61642	61646	61646	18861	677	18861	677	18861	677
5	118628	118468	118624	118628	13016	1740	13020	1740	13016	1740
6	151945	151125	151785	151945	7344	3702	7482	3746	7344	3702
7	131494	129922	130674	131494	2754	5657	3071	5847	2754	5657
8	76153	74740	74581	76153	703	6011	947	6413	703	6011
9	28392	27746	26979	28392	83	4294	159	4640	83	4294
10	6276	6135	5630	6276	4	1948	19	1995	4	1948
11	690	680	549	690		451	1	411		451
12	23	23	13	23		23		13		23
Total	599975	595209	595209	599975	5379546	24896	5380341	25875	5379546	24896

Some graphical representations

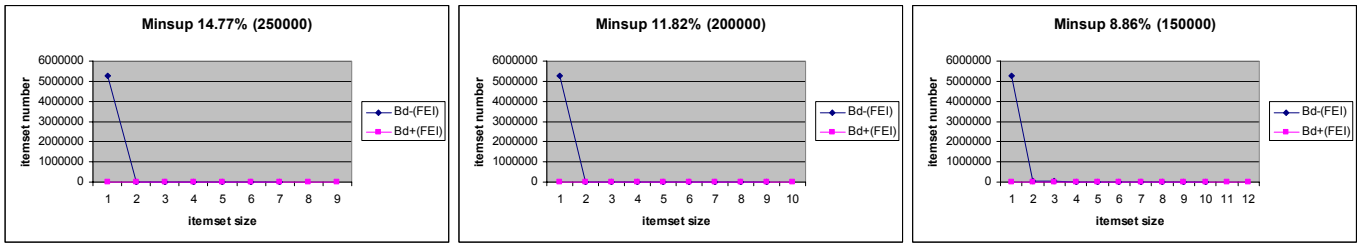
Borders of frequent itemsets



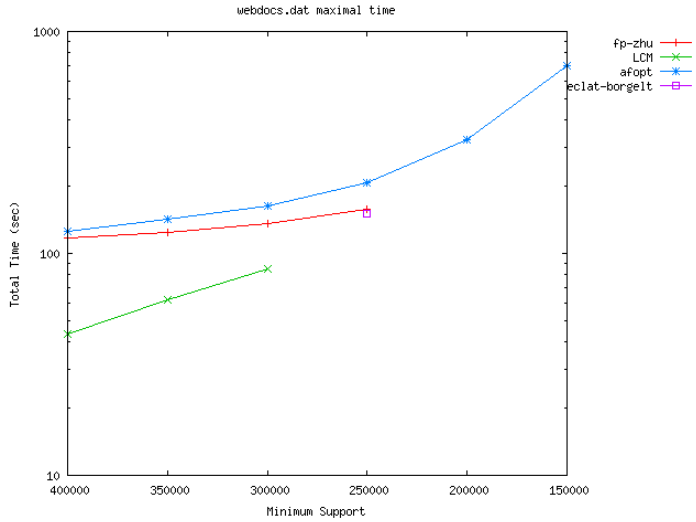
Borders of frequent free itemsets



Borders of essential itemsets



Performances of algorithms for maximal frequent itemsets mining



Analysis

Distributions stability

The relative position of the different collections being studied w.r.t. to the others is relatively stable whenever minimum support threshold value is changing. As shown by the graphical representations, it is more particularly the case for the borders distributions of the different collections. In other words, this observation suggests, a kind of invariant global structure for frequent itemsets distribution.

Borders of frequent itemsets

The negative border is always "lower" than its corresponding positive border. The borders distributions are very close, i.e. the mean of the negative border curve is only one level before the mean of the positive border curve.

Borders of other concise representations

The borders distributions of frequent free and essential itemsets are identical to the borders distribution of frequent itemsets, since the number of frequent itemsets, frequent free and essential itemsets are equal.

Performances of algorithms for MFI mining

Algorithms have difficulties on this dataset even for minimum support thresholds with few frequent itemsets. This should be mainly due to the huge number of transactions (1 692 082).

Remarks:

- There is no exact association rules for the tested minimum support thresholds (|FFI| = |FI|)

4. Synthesis

A surprising observation is that the relative position of the different collections and borders distributions is relatively stable whenever the minimum support threshold studied. In other words, this observation suggests, for each dataset, a kind of invariant global structure.

We observe "bell curve" distributions in almost all datasets for all the collections, which is not necessarily the case in theory. For example, every positive border distribution of frequent itemsets is feasible in theory [MAN03], other properties should exist to explain these distributions. Moreover, the negative and the positive borders seem to follow the same behavior even if the negative border is always "lower" than its corresponding positive border. From [MAN97], the negative border can have elements just one level after the positive border. This case never occurs in our experiments for frequent itemsets.

From FIMI results, we study algorithms execution times for all the datasets. We have presented the performances for the discovery of maximal frequent itemsets to focus on the exploration strategy of the search space.
For example, algorithms execution times on Chess increase exponentially with the diminution of the minimum support threshold, whereas on Connect they remain stable. Moreover, note that Connect has more and longer transactions, and more items than Chess. Consequently, Chess should be easier than Connect. The same kind of behavior can be noticed for datasets such as Pumsb and Pumsb*. These two datasets are very similar w.r.t. the transactions and number of items, but their borders distribution is very different. Algorithms for Pumsb* are still very effective for very low support, whereas for Pumsb, algorithms do not perform very well for relatively high minimum support.
Therefore, we deduce that pruning strategies are more efficient on datasets having a large distance between the positive and negative borders. A possible explanation could be obtained by looking at algorithms pruning strategies since most of them take advantage of unfrequent itemsets to find maximal frequent itemsets and prune the search space.

Moreover, we denote globally two different behaviours (to simplify) of the "distance" between the two borders. For datasets such as Accidents, Chess and Pumsb, the borders distributions of frequent itemsets are very close, i.e. the mean of the negative border curve is only few levels before the mean of the positive border curve.
For datasets Connect, Pumsb* and Mushroom, a large distance between the borders does exist.
For the datasets BMS-WebView-1, BMS-WebView-2, BMSPOS, Kosarak, T10I4D100K and T40I10D100K, the borders distributions of frequent itemsets are very close too. But the difference from the previous ones is that there is few frequent itemsets until very low values of minimum support threshold (>1%). Moreover their borders are principally composed of relatively small itemsets.
Principally based on the “distance” between positive and negative borders distributions of frequent itemsets, different types of datasets have been identified. This leads us to devise a new classification for datasets w.r.t. borders distribution. This classification is made of the three following types:

- Datasets where borders distributions are very close.
- Datasets where there is a large distance between the two borders distributions.
- Datasets where the two distributions are very close, but they are concentrated in very low levels.

For frequent free an essential itemsets, the two borders are very close for all tested datasets. This observation suggests a kind of invariant property for the predicates being frequent free and being frequent essential itemsets.

5. Conclusion

In this paper, we have thoroughly studied datasets for the problems related to frequent itemset mining. We have shown that the distributions of all the collections are stable w.r.t. minimum support threshold evolution. Moreover, we have shown that the distribution of the negative and positive borders have an important impact on datasets classification and algorithms performances. This work is a first step towards a better understanding of the behavior of algorithms with respect to the search space to be discovered.

This work has two main perspectives. The former is to find out theoretical foundation of "bell curves" and stability obtained for the different collections in most of our experiments. The latter is the design of adaptive algorithms, i.e. changing dynamically their strategy is also a challenging perspective.

Bibliography

[AGR93] AGRAWAL R., IMIELINSKI T., SWAMI A. N., “Mining Association Rules between Sets of Items in Large Databases”, BUNEMAN P., JAJODIA S., Eds., *SIGMOD conference, Washington, D.C.*, ACM Press, 1993, p. 207-216.

[AGR94] AGRAWAL R, SRIKANT R, “Fast Algorithms for Mining Association Rules in Large Databases”, In *Proc. VLDB ’94*, pp. 487–499, 1994.

[BAS00] BASTIDE Y., TAOUIL R., PASQUIER N.,STUMME G., and LAKHAL L., “Mining frequent patterns with counting inference”, In *SIGKDD Explorations 2(2)*, pages 66–75, 2000.

[BAY98] BAYARDO R. J., “Efficiently Mining Long Patterns from Databases”, HAAS L. M., TIWARY A., Eds., *ACM SIGMOD Conference, Seattle, USA*, 1998, p. 85-93.

[BAY04] BAYARDO R., GOETHALS B., and ZAKI M. J., FIMI’04 workshop on frequent itemset mining implementations, 2004.

[BOR03] BORGELT C., “Efficient Implementations of Apriori and Eclat”, *FIMI’03 Workshop on Frequent Itemset Mining Implementations*, November 2003.

[BOU03] BOULICAUT J.-F., BYKOWSKI A., and RIGOTTI C., “Free-sets: A condensed representation of boolean data for the approximation of frequency queries”, *Data Mining and Knowledge Discovery*, 7(1):5–22, 2003.

[BUR01] BURDICK D., CALIMLIM M., GEHRKE J., “MAFIA : A Maximal Frequent Itemset Algorithm for Transactional Databases”, *ICDE’01, Heidelberg, Germany*, IEEE CS, 2001, p. 443-452.

[BYK01] BYKOWSKI A. and RIGOTTI C., “A condensed representation to find frequent patterns”, In *PODS’01, Santa Barbara, California, USA*. ACM, 2001.

[CAL03] CALDERS T. and GOETHALS B., “Minimal k-free representations of frequent sets”, In *PKDD conference*, Lecture Notes in Computer Science. Springer, 2003.

[CAS05] CASALI A., CICCHETTI R., and LAKHAL L., “Essential patterns: A perfect cover of frequent patterns”, In *DaWaK conference, Copenhagen, Denmark*, Lecture Notes in Computer Science, 2005.

[FLO04] FLOUVAT F., DE MARCHI F., PETIT J.M., “ABS : Adaptive Borders Search of frequent itemsets”, *FIMI’04 Workshop on Frequent Itemset Mining Implementations*, November 2004.

[GOE03] GOETHALS B., ZAKI M., “Advances in Frequent Itemset Mining Implementations : Introduction to FIMI03”, rapport, 2003, http://sunsite.informatik.rwth-aachen.de/Publications/CEURWS//Vol-90/intro.pdf.

[GOE04] GOETHALS B., Frequent Itemset Mining Implementations Repository, http://fimi.cs.helsinki.fi/ , 2004.

[GOU01] GOUDA K., ZAKI M. J., “Efficiently Mining Maximal Frequent Itemsets”, CERCONE N., LIN T. Y., WU X., Eds., *ICDM’01, San Jose, USA*, IEEE Computer Society, 2001.

[GRA03] GRAHNE G. and ZHU J. “Efficiently using prefix-trees in mining frequent itemsets”, In *FIMI’03 Workshop on Frequent Itemset Mining Implementations*, November 2003.

[IBM] Synthetic Data Generation Code for Associations and Sequential Patterns.Intelligent Information Systems, IBM Almaden Research Center. http://www.almaden.ibm.com/software/quest/Resources/index.shtml.

[KRY02] KRYSZKIEWICZ M. and GAJEK M., “Concise representation of frequent patterns based on generalized disjunction-free generators”, In *PAKDD’02, Taipei, Taiwan*, 2002.

[MAN96] MANNILA H., TOIVONEN H., “Multiple Uses of Frequent Sets and Condensed Representations (Extended Abstract)”, *KDD 1996*, 1996, p. 189-194.

[MAN97] MANNILA H., TOIVONEN H., “Levelwise Search and Borders of Theories in Knowledge Discovery”, *Data Mining and Knowledge Discovery*, vol. 1, n° 1, 1997, p. 241-258, Kluwer.

[MAN03] MANIATTY W. A., RAMESH G, and ZAKI M. J., “Feasible itemset distributions in data mining: Theory and application”. In *SIGMOD conference, San Diego, USA*, June 2003.

[PAL04] PALMERINI P., ORLANDO S., and PEREGO R. , “Statistical properties of transactional databases”. In *ACM symposium on Applied computing*, pages 515–519. ACM Press, 2004.

[PAS99] PASQUIER N., BASTIDE Y., TAOUIL R., and LAKHAL L., “Efficient mining of association rules using closed itemset lattices”, *Information Systems*, 24(1):25–46, 1999.

[ORL03] ORLANDO S., LUCCHESE C., PALMERINI P., PEREGO R., and SILVESTRI F., “kdci: a multi-strategy algorithm for mining frequent sets”. In *Workshop on frequent Itemset Mining Implementations FIMI’03*, November 2003.