DandD Instance Generation in Textile Plot Environment

Natsuhiko KUMASAKA
Faculty of Science and Technology
Keio University

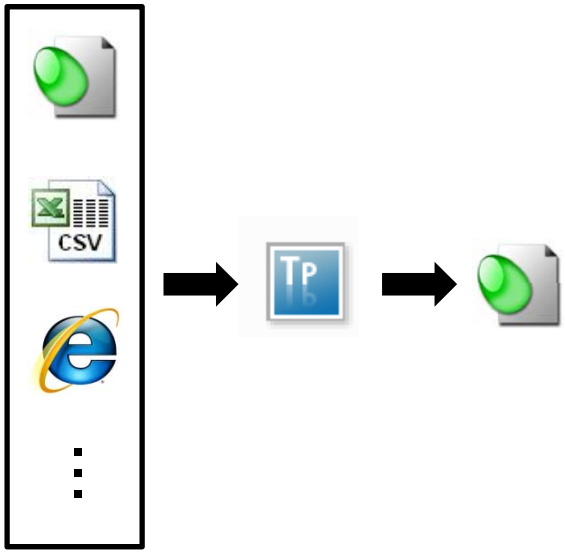
Background

- Variety of data sources
  - Flat files (CSV, TSV, fixed format,…)
  - RDBMS (PostgreSQL, Access, etc…)
  - Spread sheet (Excel,…)
  - Webpage (HTML, CGI,…)
- DandD project
  - DandD instance
    - XML document describing data and its attributes
  - DandD client-server system (Yokouchi and Shibata 2004)
    - Integrate different data sources
    - Employ DandD Instance as an intermediate entity
  - DandDEditor (Yokouchi 2005)
    - Create / Edit DandD instances
    - Restricted in well organised data sources

Textile Plot Environment
(Kumasaka and Shibata 2007)

- Previous features
  - Open DandD instance
  - Modify textile plots
    - Remove/replace columns, cases and observations
    - Add/Remove/Modify attributes
    - Re-organise data structure
  - Save DandD instance
    - Audit description of data analysis
- New feature in 2008
  - Open various data sources
    - Extract tables embedded in data files or databases
    - Separate attributes from tables
    - Divide tables into a set of datavectors
    - Check consistency in datavectors

2007

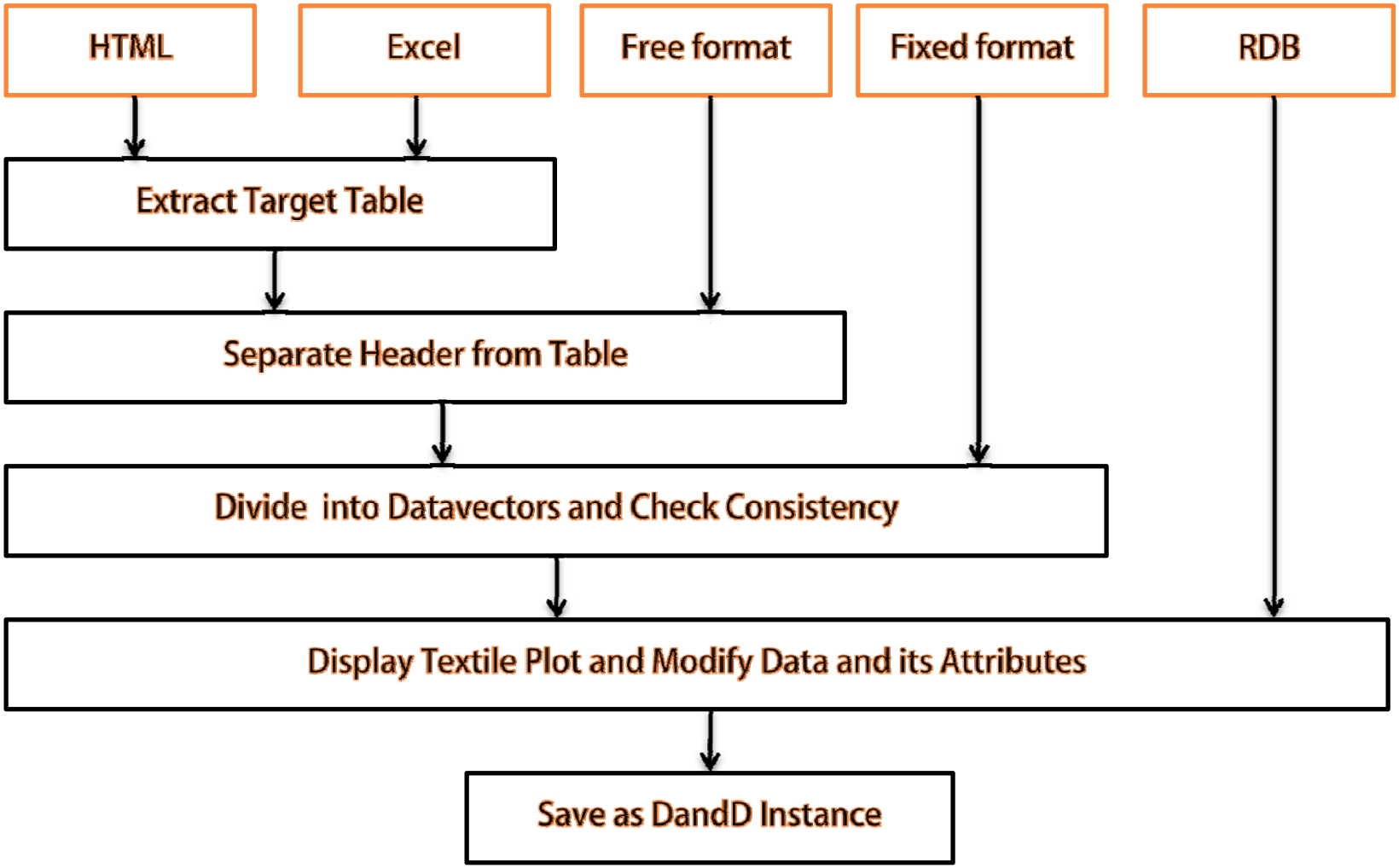


2008

Target Data Sources

- Relational Databases
  - RDBMS
    - PostgreSQL, Access, MySQL
    - Oracle (not tested yet)
- Data File
  - Flat file
    - Fixed format
    - Free format (CSV, TSV)
  - Non-flat file
    - Excel
    - HTML

STRATEGY


HTML
Excel
Free format
Fixed format
RDB
Extract Target Table
Separate Header from Table
Divide into Datavectors and Check Consistency
Display Textile Plot and Modify Data and its Attributes
Save as DandD Instance

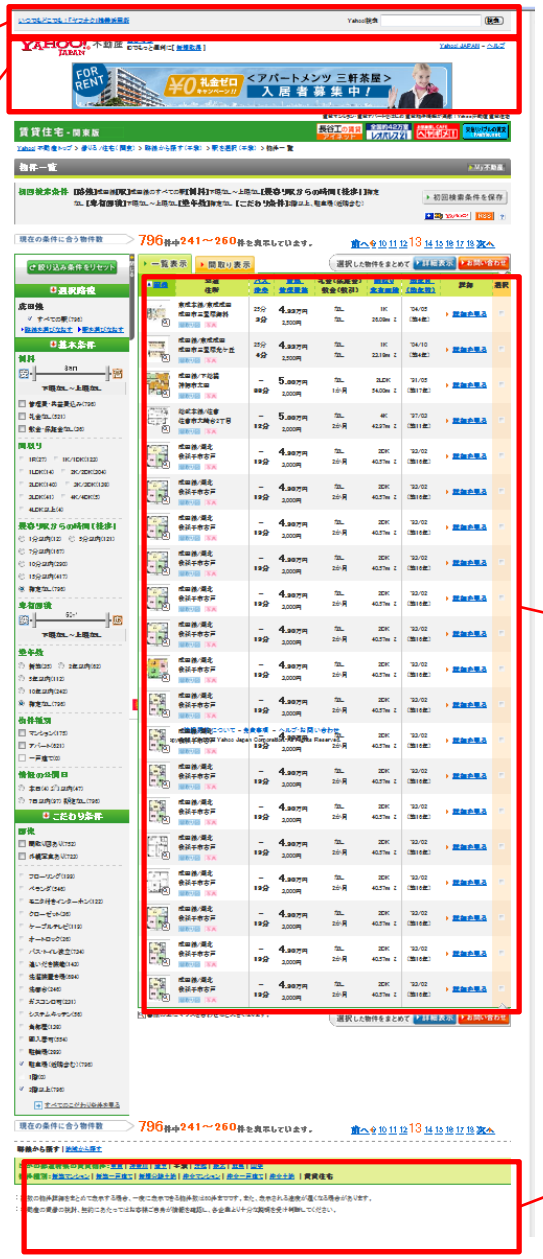
HTML

- Extract target <TABLE>s
  1. Create DOM Tree
     - DOM Level 1 (W3C 2008)
     - CyberNeko HTML Parser 1.9.6.1 (Clark 2008)
  2. Find <TABLE>s
  3. Count the number of <TR> and <TD>
  4. Select <TABLE>s which have a number of cells
- Separate header from table body
  1. Search <TBODY> and <THEAD>
  2. Find disjoint sets of <TR>s
  3. Measure dissimilarity between disjoint <TR>s
  4. Using clustering algorithm
- Merge <TABLE>s according to header information
- Divide into datavectors and check consistency

EXTRACT TARGET <TABLE>S

<TR>: 1, <TD>: 1
Num of cells: 1

<TR>: 1, <TD>: 2
Num of cells: 2



<TR>: 21, <TD>: 9
Num of cells: 189

<TR>: 1, <TD>: 1
Num of cells: 1

http://rent.realestate.yahoo.co.jp/

Separate Header from Body: Find disjoint sets of <TR>s

http://rent.realestate.yahoo.co.jp/

FIND DISJOINT SETS OF <TR>S: NESTED BODY

http://www.athome.co.jp/

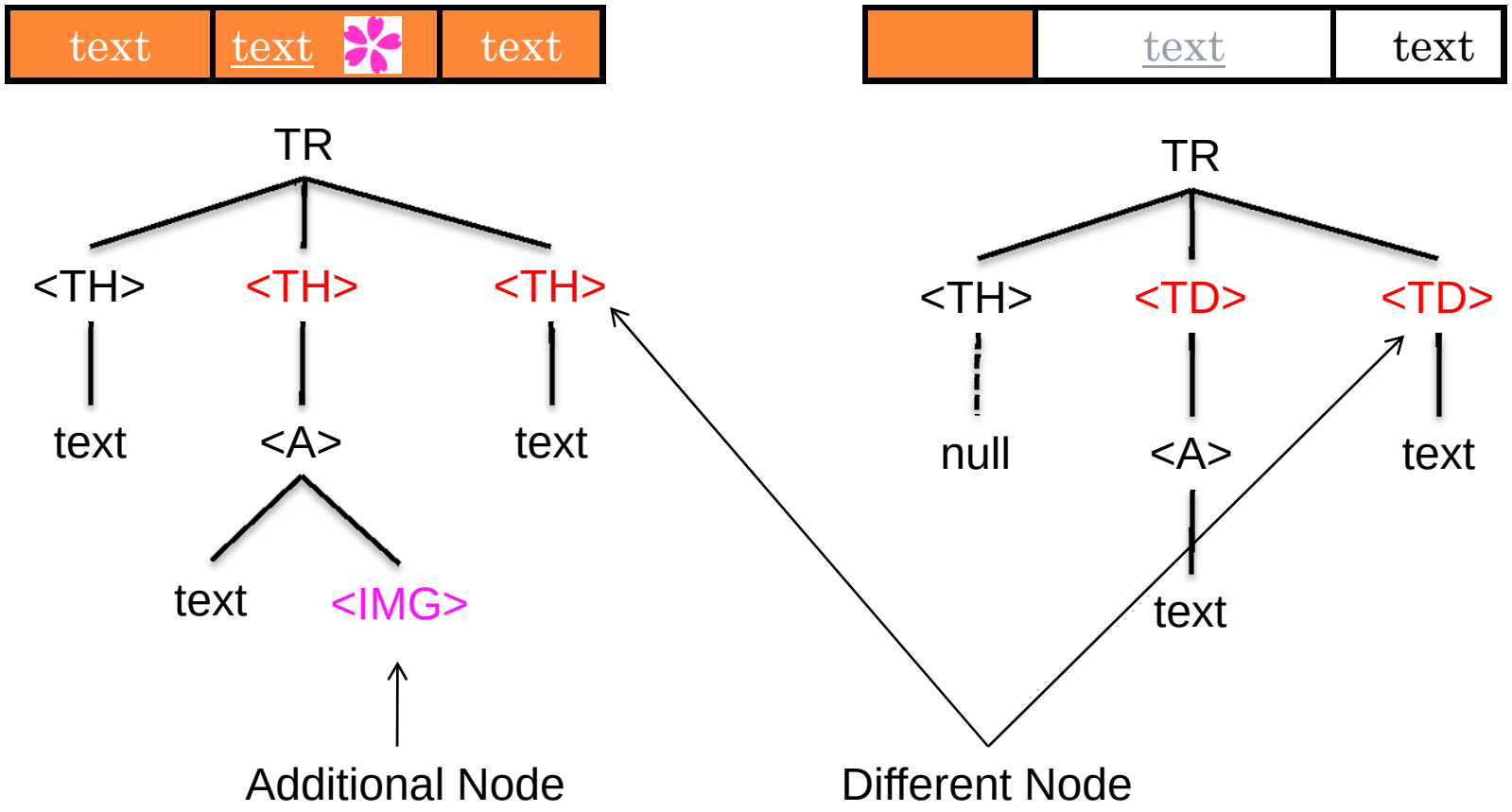
FIND DISJOINT SETS OF <TR>S: NESTED HEADER

毎日の全国データ一覧表（日別値詳細版：2008年03月07日　17時現在）

地点	気圧				気温				蒸気圧	湿度			風向・風速						日照時間	全天日射量	平均雲量	降水量					降雪の深さ合計	最深積雪	天気概況	
	現地	海面	最低海面		平均	最高		最低	平均	平均	最小		平均	最大			最大瞬間					日合計	日最大						0600-1800	1800-翌0600
	平均	平均	値	起時		値	起時	値			値	起時	風速	風速	風向	起時	風速	風向	起時				1時間	起時	10分間	起時				
札幌			1015.6]	15:52		3.9]	13:32	−4.7]	05:44			40]	12:29	6.2]	北北西	14:10	9.0]	北北西	14:16	5.0]	13.2]		0.0/—					3]	76]		
稚内			1013.1]	17:00		3.2]	12:13	−3.4]	06:09			66]	12:30	7.0]	西南西	13:00	11.4]	西	12:33	9.8]	15.4]		0.0/—					3]	35]		
北見枝幸			1013.3]	16:26		3.9]	11:53	−2.8]	06:00			44]	13:23	6.3]	西	14:30	10.6]	西	16:05	5.1]			0.0/—					2]	65]		
旭川			1015.1]	16:07		2.1]	14:59	−6.7]	05:02			56]	14:54	5.9]	西南西	15:20	9.2]	西南西	15:20	3.5]	12.1]		0.5]					4]	61]		
留萌			1015.2]	17:00		2.5]	14:28	−3.6]	06:20			60]	15:05	6.6]	西南西	16:20	9.8]	西	14:48	4.5]	11.6]		5.0]					10]	42]		
羽幌			1014.7]	17:00		2.9]	14:23	−3.8]	04:11			58]	15:01	5.9]	西	15:10	8.9]	西	15:27	5.7]			1.0]					2]	70]		
岩見沢			1015.8]	16:07		3.0]	14:41	−5.6]	03:25			57]	14:36	3.3]	西南西	13:50	5.9]	北西	15:00	6.7]			0.0/—					—	88]		
小樽			1015.6]	16:12		3.7]	12:47	−2.9]	03:32			39]	12:50	4.2]	西南西	17:00	7.7]	西南西	16:32	6.8]			0.0/—					1]	96]		
寿都			1016.5]	15:56		3.5]	12:32	−2.9]	06:21			47]	12:49	3.7]	北西	13:30	9.8]	北西	13:36	5.3]	11.2]		1.0]					2]	37]		
倶知安			1016.1]	16:07		2.8]	12:18	−7.5]	05:27			46]	12:46	6.2]	西北西	15:00	9.8]	西北西	14:44	7.4]			3.5]					8]	171]		
網走			1013.7]	17:00		2.2]	13:35	−4.1]	05:30			41]	14:49	4.2]	北西	13:00	7.6]	西北西	13:07	1.4]	12.0]		0.0/—					6]	62]		
紋別			1013.2]	15:49		3.3]	12:34	−3.1]	03:34			43]	12:20	6.2]	西	16:30	9.7]	西	16:25	3.7]			0.0/—					—	41]		
雄武			1013.3]	16:59		3.8]	14:14	−2.8]	03:06			42]	13:55	6.4]	西南西	14:50	10.2]	西	14:48	3.9]			0.0/—					—	41]		
根室			1013.8]	15:53		1.5]	13:14	−4.4]	05:58			59]	13:15	4.5]	北北西	14:40	6.6]	北北西	14:32	3.9]	13.6]		0.0/—					—	—		
釧路			1014.3]	14:52		1.4]	12:59	−4.6]	06:31			67]	13:21	6.7]	南西	13:30	8.5]	南西	12:49	8.4]			0.0/—					—	—		
帯広			1013.4]	15:39		4.8]	14:01	−6.7]	06:20			37]	13:29	6.8]	西北西	15:10	11.3]	西北西	14:59	10.2]	17.3]		0.5]					2]	16]		
広尾			1014.1]	15:49		1.7]	16:33	−6.5]	06:01			59]	16:33	3.4]	西	06:00	5.0]	東南東	11:36	10.6]			0.0/—					—	47]		
室蘭			1016.2]	16:35		3.3]	14:31	−2.1]	05:36			60]	16:11	5.5]	西	13:40	7.6]	西北西	14:12	9.9]	17.1]		0.0/—					3]	9]		
苫小牧			1016.0]	16:06		2.4]	16:07	−7.3]	06:23			60]	16:17	4.1]	南東	11:20	5.2]	北	13:11	7.3]			0.0/—					1]	8]		
浦河			1015.8]	16:35		2.5]	14:54	−5.2]	05:14			61]	09:12	8.6]	西	13:30	12.1]	西北西	15:18	10.4]	17.2]		0.0/—					—	—		
函館			1016.6]	16:27		4.2]	12:44	−4.6]	04:44			38]	12:21	6.6]	北西	14:20	10.0]	西北西	14:09	7.0]	14.2]		0.5]					3]	18]		
江差			1017.1]	16:05		5.0]	12:27	−0.5]	07:50			43]	12:45	6.6]	西北西	04:00	12.2]	北西	00:52	6.9]			1.5]					1]	1]		
青森			1017.4]	14:54		4.8]	13:12	−2.4]	05:58			54]	15:01	6.4]	西南西	10:30	10.7]	西北西	14:58	4.7]	13.2]		0.5]					2]	43]		
八戸			1017.1]	14:43		4.2]	10:21	−2.3]	06:00			48]	10:24	5.6]	南西	10:00	10.5]	南西	10:59	4.7]			0.0/—					—	1]		
深浦			1018.4]	15:32		4.8]	12:35	−1.0]	08:53			48]	12:00	7.8]	北西	08:50	14.6]	北西	08:43	0.8]			3.0]					5]	22]		
むつ			1017.0]	16:19		2.6]	14:45	−5.6]	01:48			66]	14:43	4.2]	南南西	13:10	7.9]	西	08:06	3.0]			3.5]					8]	11]		
秋田			1018.9]	16:38		3.1]	11:43	−2.4]	02:13			75]	10:54	7.5]	西北西	11:50	14.0]	北北西	11:50	3.2]	10.8]		8.5]					11]	5]		
盛岡			1017.4]	15:00		5.3]	12:29	−5.0]	06:22			37]	12:23	6.4]	西	13:10	10.6]	南西	12:50	6.6]	14.6]		0.5]					—	—		
大船渡			1016.8]	14:52		7.0]	13:41	−2.3]	05:42			32]	13:49	5.6]	北北西	15:30	11.4]	西北西	15:19	7.1]			0.0/—					—	—		
宮古			1017.0]	13:53		6.6]	11:07	−3.8]	06:08			45]	13:13	5.3]	東南東	13:40	7.6]	東南東	13:32	1.8]			0.5]					2]	3]		
仙台			1017.6]	14:29		8.0]	11:32	−0.3]	05:27			32]	11:37	9.2]	西北西	14:10	13.9]	西北西	15:40	7.5]	15.8]		0.0/—					—	—		
石巻			1017.2]	15:21		8.1]	12:09	−0.8]	05:42			31]	14:49	10.1]	西北西	15:00	15.2]	西北西	15:56	9.1]			0.0/—					—	—		

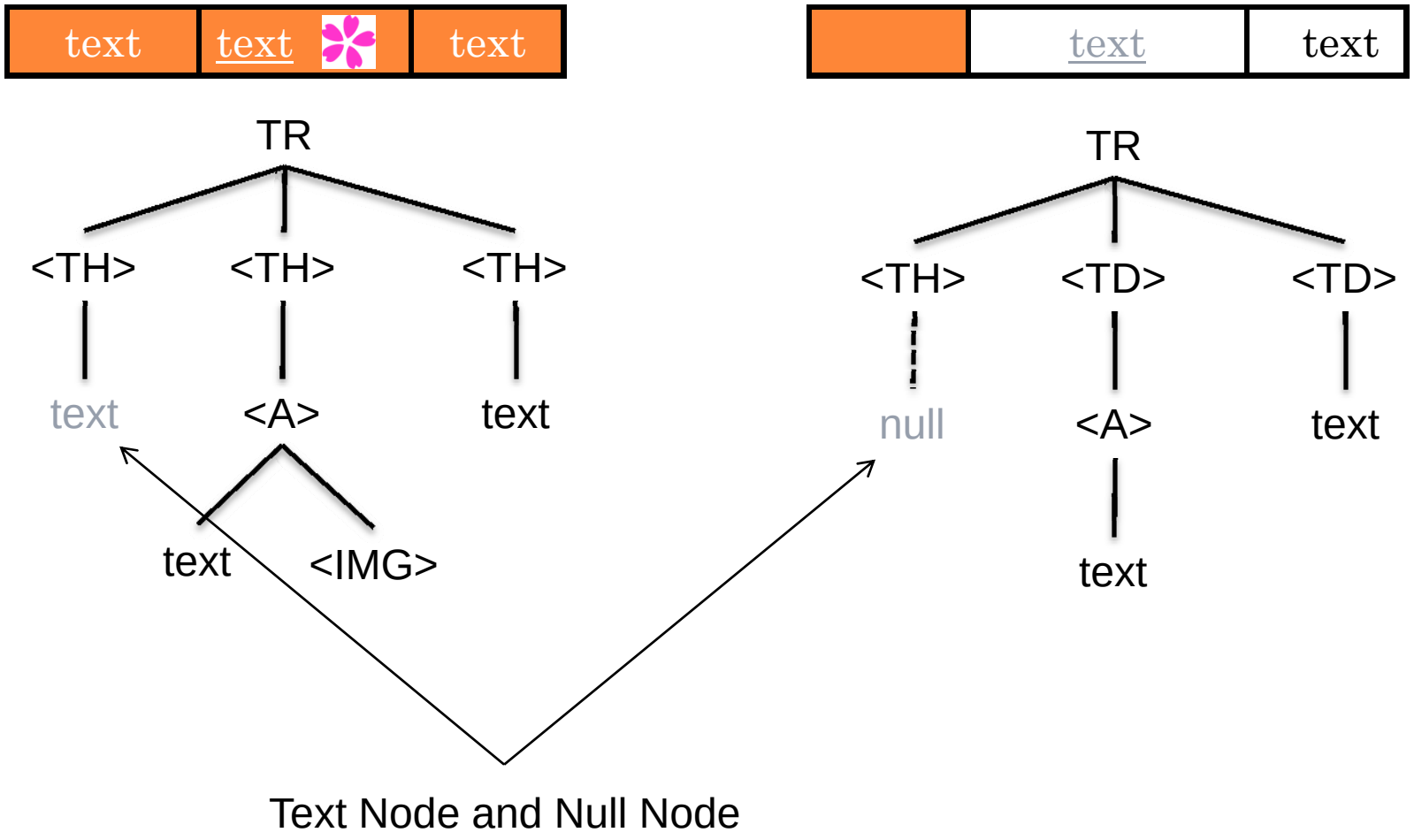
http://www.athome.co.jp/

Dissimilarity Measure between Disjoint <TR>s



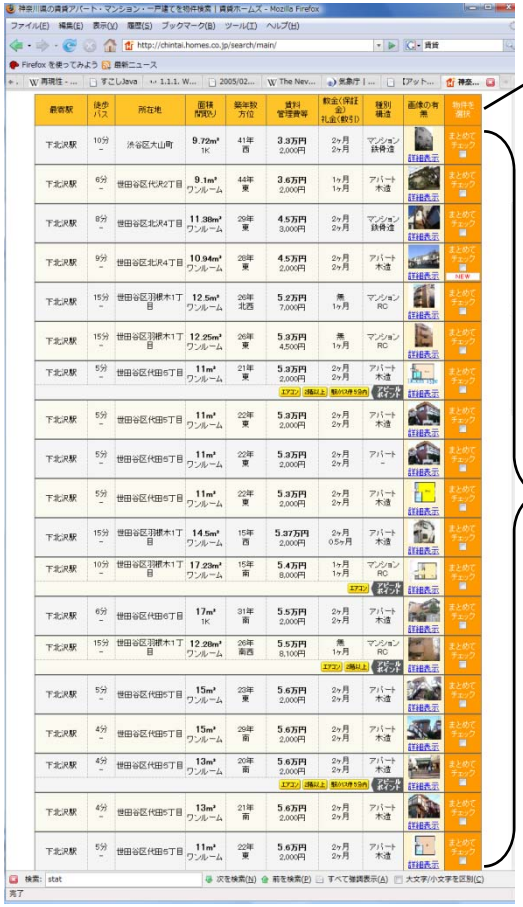
Count the number of differences

Dissimilarity Measure between Disjoint <TR>s

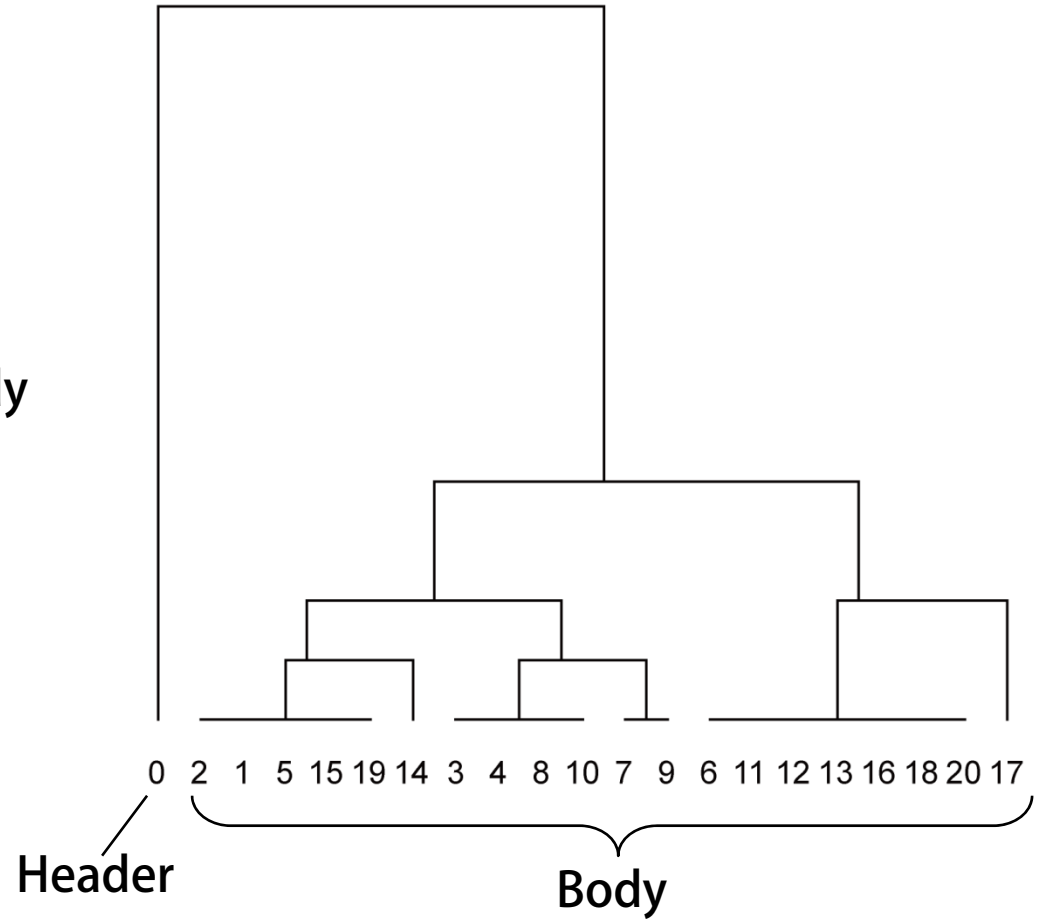


Text Node and Null Node

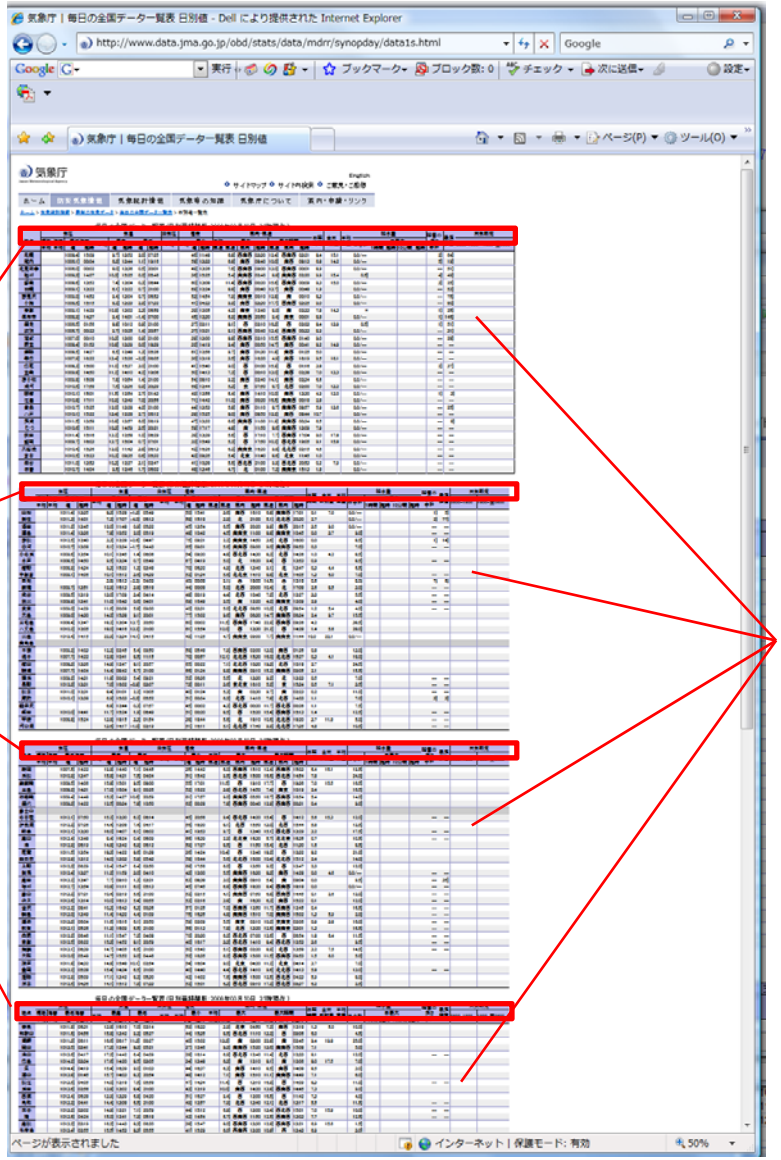
Clustering Algorithm



http://chintai.homes.co.jp/



Merge <TABLE>s according to header information



Same headers

Different bodies

http://www.athome.co.jp/

Check Consistency in Datavector

- Find Missing / Invalid
  - “NA”, “-”, “?”, “”, “.”, space, null
- Estimate Data Type
  - Numeric (1.34, 12,000,…)
    - Check data range
      - Double -> Measurement
      - Integer -> Ordinal Number
      - Non-negative Integer -> Count
      - {0, 1} -> Logical
  - Semi-numeric (2000yen, 5min,…)
    1. Find pattern (e.g. “2,000 yen”, “3,000 yen”)
    2. Separate value and unit (e.g. “2,000yen” -> “2,000”, “yen”)
    3. Go to numeric
  - Character (Hiyoshi, Tsunashima,…)
    - Check repeated values -> Category; otherwise -> ID

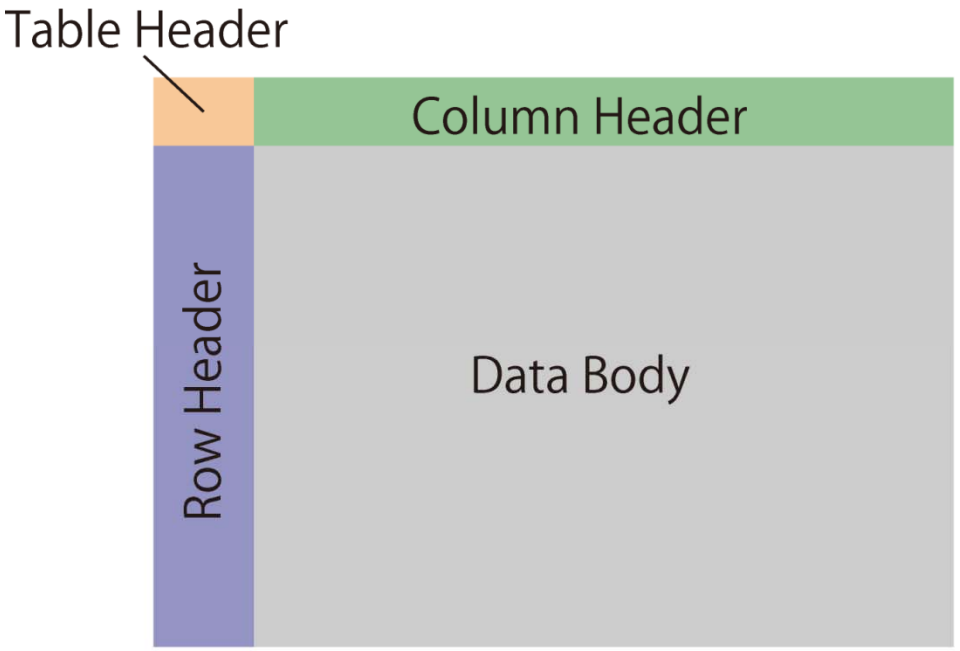
Excel

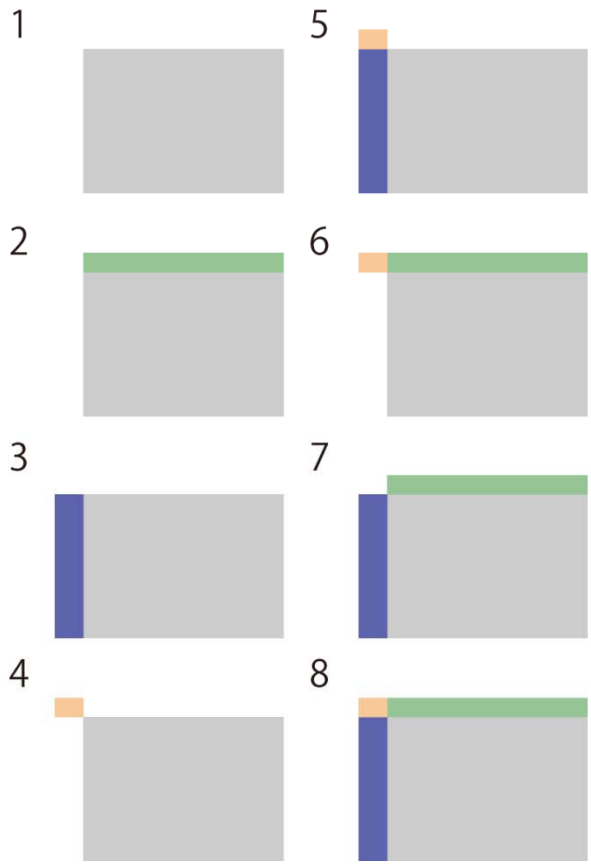
- Extract tables from a sheet
  1. Create Sheet Object
     - JExcelAPI (JExcel API group, 2008)
  2. Find disjoint sets of cells
  3. Determine rectangle hull for each set
- Separate header and body
  1. Check table type
  2. Cell type clustering
- Merge tables
- Divide into datavectors and check consistency

Disjoint cells: LIBOR data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	EUR	1-Dec-99	2-Dec-99	3-Dec-99	6-Dec-99	7-Dec-99	8-Dec-99	9-Dec-99	10-Dec-99	13-Dec-99	14-Dec-99	15-Dec-99	16-Dec-99	17-Dec-99	20-Dec-99	21-Dec-99	22-Dec-99	23-Dec-99	24-Dec-99	27-Dec-99	28-Dec-99	29-Dec-99	30-Dec-99	31-Dec-99
2	1w	3.07375	3.07563	3.07000	3.07375	3.07188	3.06000	3.04750	3.03000	3.00125	3.00000	3.01875	3.05250	3.06125	3.02750	3.00000	3.00375	4.06000	2.97250	4.04375	4.09375	3.11125	3.11750	
3	1m	3.51063	3.51963	3.52125	3.53125	3.54500	3.54000	3.54688	3.54825	3.54438	3.55000	3.55000	3.55375	3.56250	3.57500	3.56500	3.53938	3.53475	3.02700	3.53375	3.53500	3.17875	3.17250	
4	2m	3.45000	3.45000	3.45000	3.46000	3.46000	3.45875	3.45875	3.45700	3.44938	3.45000	3.45000	3.45000	3.45750	3.45750	3.45750	3.44875	3.44938	3.45000	3.44250	3.44625	3.26625	3.26125	
5	3m	3.45000	3.45000	3.45000	3.46000	3.46000	3.45875	3.45875	3.45863	3.45688	3.45563	3.45688	3.45625	3.46375	3.46688	3.46563	3.45750	3.44813	3.45000	3.44750	3.44875	3.34625	3.33875	
6	4m	3.45000	3.45000	3.45575	3.46000	3.46000	3.46000	3.46000	3.46000	3.46000	3.46000	3.46063	3.46875	3.47938	3.48000	3.48000	3.48000	3.47000	3.45000	3.47000	3.47000	3.41375	3.41000	
7	5m	3.46000	3.46000	3.46463	3.46813	3.47000	3.47000	3.47063	3.47375	3.47063	3.47438	3.47875	3.48000	3.50000	3.50000	3.51000	3.51000	3.50375	3.45000	3.50500	3.50625	3.46500	3.46688	
8	6m	3.46938	3.47313	3.48000	3.48113	3.48000	3.48000	3.48313	3.48375	3.48688	3.49000	3.49625	3.50250	3.53938	3.54000	3.55000	3.55000	3.55000	3.45188	3.55000	3.55250	3.51750	3.51750	
9	7m	3.50750	3.51000	3.51563	3.51625	3.51688	3.51875	3.52000	3.52000	3.51938	3.52000	3.53875	3.54750	3.59000	3.59625	3.60063	3.60000	3.60000	3.47700	3.60000	3.60125	3.57750	3.57688	
10	8m	3.55125	3.55588	3.56000	3.56000	3.56000	3.56000	3.56000	3.56000	3.55938	3.56000	3.59000	3.59875	3.64250	3.65125	3.66000	3.66000	3.65125	3.50463	3.65625	3.65500	3.63750	3.63688	
11	9m	3.60250	3.60063	3.60875	3.60713	3.60375	3.60000	3.60000	3.60000	3.60000	3.61000	3.63375	3.65250	3.70375	3.71000	3.72000	3.72000	3.71000	3.54000	3.71875	3.71625	3.69125	3.69813	
12	10m	3.65000	3.65000	3.66000	3.65938	3.65000	3.65000	3.65250	3.65013	3.65000	3.66000	3.68000	3.70000	3.76438	3.77000	3.78000	3.78000	3.77000	3.57838	3.77125	3.77125	3.75500	3.76000	
13	11m	3.70375	3.70875	3.71938	3.71400	3.71000	3.70875	3.71000	3.71013	3.71000	3.72000	3.73750	3.75500	3.82000	3.83000	3.84000	3.83000	3.82250	3.61813	3.82250	3.82375	3.81125	3.81813	
14	12m	3.76375	3.76588	3.78000	3.77825	3.77000	3.77000	3.77438	3.77163	3.77000	3.78000	3.79875	3.81375	3.89000	3.89000	3.89750	3.88500	3.88000	3.66000	3.87875	3.88000	3.87250	3.87813	
15																								
16	USD	1-Dec-99	2-Dec-99	3-Dec-99	6-Dec-99	7-Dec-99	8-Dec-99	9-Dec-99	10-Dec-99	13-Dec-99	14-Dec-99	15-Dec-99	16-Dec-99	17-Dec-99	20-Dec-99	21-Dec-99	22-Dec-99	23-Dec-99	24-Dec-99	27-Dec-99	28-Dec-99	29-Dec-99	30-Dec-99	31-Dec-99
17	1w	5.62172	5.57375	5.56531	5.57406	5.57281	5.57656	5.60750	5.62688	5.63375	5.62438	5.62313	5.60125	5.60125	5.58281	5.61250	5.60281	7.71875	7.75625			5.92344	5.92344	
18	1m	6.47875	6.47875	6.47875	6.46875	6.46750	6.46875	6.46125	6.46250	6.46250	6.46125	6.46000	6.46125	6.46250	6.46125	6.47000	6.47125	6.48125	6.49000			5.82750	5.82750	
19	2m	6.16625	6.16875	6.17125	6.17000	6.17000	6.17000	6.17125	6.17125	6.17125	6.17250	6.18000	6.18000	6.18625	6.18750	6.20000	6.20125	6.20250	6.20250			5.90375	5.90375	
20	3m	6.12125	6.12125	6.12375	6.12000	6.12000	6.12000	6.12000	6.12125	6.12125	6.12000	6.14000	6.14000	6.15250	6.16250	6.17500	6.18000	6.18125	6.18375			6.00375	6.00375	
21	4m	6.09875	6.10125	6.11000	6.11000	6.11000	6.10875	6.11000	6.11125	6.11000	6.11125	6.13000	6.13000	6.14375	6.15375	6.16250	6.17000	6.17250	6.17500			6.05250	6.05250	
22	5m	6.08250	6.08500	6.10000	6.10000	6.10000	6.09875	6.10000	6.10125	6.10000	6.10125	6.12000	6.12125	6.13750	6.14500	6.15375	6.16250	6.16750	6.16750			6.09375	6.09375	
23	6m	6.07125	6.07750	6.09000	6.09000	6.09000	6.08875	6.09000	6.09125	6.09000	6.09125	6.11125	6.11875	6.13250	6.13875	6.15000	6.16000	6.16750	6.16625			6.13000	6.13000	
24	7m	6.10375	6.11000	6.12250	6.12125	6.12125	6.12000	6.12250	6.12125	6.12000	6.12250	6.15000	6.15500	6.18500	6.18875	6.20625	6.21250	6.21750	6.21625			6.19375	6.19375	
25	8m	6.14375	6.14500	6.15750	6.15250	6.15125	6.15000	6.15125	6.15125	6.14875	6.15125	6.18875	6.19375	6.23250	6.23500	6.25625	6.26000	6.26500	6.26375			6.25375	6.25375	
26	9m	6.18000	6.18250	6.19250	6.18125	6.18250	6.18000	6.18375	6.18375	6.18000	6.19000	6.22625	6.23375	6.28125	6.28250	6.30375	6.31000	6.31625	6.31250			6.31000	6.31000	
27	10m	6.22000	6.22375	6.23500	6.22125	6.22250	6.22000	6.22500	6.22375	6.21875	6.23000	6.27125	6.28125	6.33375	6.33750	6.35625	6.36250	6.36750	6.36500			6.37500	6.37500	
28	11m	6.25750	6.26250	6.27625	6.26125	6.26125	6.26000	6.26500	6.26250	6.25750	6.26875	6.31500	6.32875	6.38250	6.38750	6.41000	6.41875	6.42250	6.42250			6.43625	6.43625	
29	12m	6.29250	6.29750	6.31625	6.30000	6.30000	6.30000	6.30500	6.29875	6.29250	6.30875	6.36250	6.37500	6.43000	6.43375	6.46125	6.47375	6.48000	6.48125			6.50000	6.50000	
30																								
31	GBP	1-Dec-99	2-Dec-99	3-Dec-99	6-Dec-99	7-Dec-99	8-Dec-99	9-Dec-99	10-Dec-99	13-Dec-99	14-Dec-99	15-Dec-99	16-Dec-99	17-Dec-99	20-Dec-99	21-Dec-99	22-Dec-99	23-Dec-99	24-Dec-99	27-Dec-99	28-Dec-99	29-Dec-99	30-Dec-99	31-Dec-99
32	1w	5.65094	5.44875	5.37500	5.31094	5.25000	5.37563	5.37688	5.13188	5.15188	5.20203	5.62750	5.28453	5.12625	4.93125	4.90781	4.81063	4.49250	4.23438			4.00000	3.75000	
33	1m	5.90094	5.89141	5.88641	5.87813	5.84391	5.85344	5.84719	5.63641	5.62063	5.62125	5.71984	5.73484	5.72219	5.71000	5.69719	5.67078	5.63250	5.43875			5.41172	5.40813	
34	2m	6.00000	6.00000	6.01125	6.02266	6.00375	6.00125	6.00000	5.87563	5.86859	5.86406	5.87750	5.92016	5.92781	5.92875	5.93125	5.92906	5.93000	5.88813			5.90750	5.90406	
35	3m	6.04906	6.05906	6.08109	6.08625	6.06563	6.06281	6.06344	5.99500	5.98234	5.97969	6.00625	6.04297	6.06094	6.06906	6.06938	6.07438	6.08547	6.06844			6.07719	6.07844	
36	4m	6.08968	6.10547	6.11422	6.11547	6.10641	6.10266	6.10141	6.04781	6.04172	6.04047	6.08781	6.10766	6.12063	6.12250	6.12188	6.13000	6.13875	6.14078			6.15188	6.15109	
37	5m	6.13125	6.14141	6.14656	6.14703	6.14266	6.14406	6.14188	6.09813	6.09344	6.09594	6.15031	6.17172	6.18281	6.18500	6.19141	6.19594	6.20109	6.20469			6.21578	6.22609	
38	6m	6.17656	6.19266	6.19594	6.19344	6.19094	6.19703	6.19719	6.15906	6.15484	6.15734	6.21203	6.24109	6.25000	6.25000	6.26500	6.27750	6.28406	6.28406			6.28563	6.29031	
39	7m	6.25000	6.25125	6.25875	6.25125	6.25125	6.25625	6.25750	6.21359	6.21859	6.22219	6.26891	6.30172	6.32188	6.32188	6.34000	6.35109	6.35422	6.35422			6.35578	6.36422	
40	8m	6.31406	6.31406	6.32438	6.31531	6.31625	6.31875	6.31750	6.28000	6.28266	6.28656	6.33281	6.36938	6.39969	6.40078	6.41078	6.42344	6.42375	6.42375			6.42250	6.42719	
41	9m	6.37813	6.37625	6.37813	6.37813	6.37813	6.38063	6.38063	6.34422	6.33969	6.35281	6.39078	6.43531	6.47328	6.47828	6.48344	6.49625	6.49750	6.49875			6.50000	6.50000	
42	10m	6.43844	6.43594	6.43875	6.43875	6.43875	6.44031	6.44063	6.40234	6.40359	6.41484	6.45047	6.49750	6.53422	6.54063	6.54453	6.55391	6.55906	6.55656			6.56219	6.56094	
43	11m	6.50000	6.50000	6.50000	6.50000	6.50000	6.50000	6.50000	6.46469	6.46469	6.47703	6.51391	6.55938	6.59891	6.60594	6.60984	6.62563	6.62688	6.62375			6.62938	6.63063	
44	12m	6.56281	6.56094	6.56156	6.56156	6.56125	6.56125	6.56094	6.52516	6.52891	6.53781	6.60047	6.62438	6.67016	6.68250	6.68250	6.69281	6.69313	6.69156			6.70578	6.70813	
45																								
46	JPY	1-Dec-99	2-Dec-99	3-Dec-99	6-Dec-99	7-Dec-99	8-Dec-99	9-Dec-99	10-Dec-99	13-Dec-99	14-Dec-99	15-Dec-99	16-Dec-99	17-Dec-99	20-Dec-99	21-Dec-99	22-Dec-99	23-Dec-99	24-Dec-99	27-Dec-99	28-Dec-99	29-Dec-99	30-Dec-99	31-Dec-99
47	1w	0.07125	0.07125	0.07250	0.07250	0.07250	0.07250	0.07250	0.07125	0.07125	0.07375	0.07625	0.09250	0.08875	0.08125	0.08250	1.87500	5.25000	6.06250			0.12750	0.12625	
48	1m	0.62125	0.64000	0.61125	0.61125	0.61750	0.57875	0.49125	0.41250	0.38500	0.36625	0.38625	0.65750	0.68125	0.67125	0.67125	0.87500	1.58125	1.58750			0.14125	0.15250	
49	2m	0.39625	0.39000	0.40250	0.41375	0.40500	0.38000	0.32688	0.29625	0.27000	0.26625	0.27625	0.40750	0.41000	0.40500	0.39313	0.46875	0.65625	0.76875			0.15875	0.16500	
50	3m	0.32000	0.32875	0.33750	0.33125	0.33125	0.31125	0.27750	0.25500	0.23375	0.23125	0.23375	0.32375	0.32750	0.32375	0.32422	0.36125	0.48750	0.56250			0.20000	0.19375	
51	4m	0.28125	0.28375	0.29000	0.28875	0.28625	0.27125	0.25125	0.24750	0.22750	0.22625	0.22750	0.28875	0.29875	0.29375	0.28750	0.31250	0.41375	0.43875			0.20500	0.20875	
52	5m	0.26125	0.26375	0.27250	0.26500	0.26500	0.25375	0.24625	0.24000	0.22250	0.22250	0.22500	0.26625	0.28000	0.27625	0.27500	0.29875	0.38625	0.40000			0.21875	0.21875	
53	6m	0.25500	0.25500	0.26375	0.25750	0.25750	0.24625	0.24250	0.23750	0.22250	0.22125	0.22375	0.26375	0.26875	0.26000	0.26125	0.28375	0.37250	0.38750			0.23750	0.23000	
54	7m	0.25375	0.25250	0.26250	0.25750	0.25625	0.24625	0.24250	0.23625	0.22875	0.22500	0.22625	0.25625	0.26250	0.26000	0.26125	0.28125	0.36625	0.37500			0.24375	0.23625	
55	8m	0.25375	0.25250	0.26250	0.25750	0.25625	0.24750	0.24250	0.23750	0.22875	0.22625	0.22750	0.25500	0.26125	0.26000	0.26125	0.28125	0.36625	0.37500			0.24875	0.24375	
56	9m	0.25375	0.25250	0.26250	0.25750	0.25750	0.24875	0.24250	0.23750	0.23000	0.22750	0.23000	0.25375	0.26000	0.25750	0.26000	0.28125	0.36625	0.37500			0.25750	0.25375	
57	10m	0.25375	0.25375	0.26250	0.25750	0.25750	0.24875	0.24250	0.23750	0.23375	0.23250	0.23125	0.25375	0.26250	0.26000	0.26250	0.28375	0.36625	0.37625			0.26125	0.25875	
58	11m	0.25375	0.25375	0.26250	0.25875	0.25750	0.25000	0.24250	0.23750	0.23625	0.23625	0.23500	0.25375	0.26250	0.26000	0.26250	0.28375	0.36625	0.37375			0.26625	0.26375	
59	12m	0.25375	0.25375	0.26250	0.26000	0.25875	0.25000	0.24250	0.23750	0.23625	0.23625	0.23500	0.25375	0.26250	0.26125	0.26375	0.28375	0.36250	0.37250			0.27125	0.27250	
60																								

Target Table Type





Check Table Type

EUR	1-Dec-99	2-Dec-99	3-Dec-99	6-Dec-99	7-Dec-99	8-Dec-99
1w	3.07375	3.07563	3.07000	3.07375	3.07188	3.06000
1m	3.51063	3.51963	3.52125	3.53125	3.54500	3.54000
2m	3.45000	3.45000	3.45000	3.46000	3.46000	3.45875
3m	3.45000	3.45000	3.45000	3.46000	3.46000	3.45875
4m	3.45000	3.45000	3.45575	3.46000	3.46000	3.46000
5m	3.46000	3.46000	3.46463	3.46813	3.47000	3.47000
6m	3.46938	3.47313	3.48000	3.48113	3.48000	3.48000
7m	3.50750	3.51000	3.51563	3.51625	3.51688	3.51875
8m	3.55125	3.55588	3.56000	3.56000	3.56000	3.56000
9m	3.60250	3.60063	3.60875	3.60713	3.60375	3.60000
10m	3.65000	3.65000	3.66000	3.65938	3.65000	3.65000
11m	3.70375	3.70875	3.71938	3.71400	3.71000	3.70875
12m	3.76375	3.76588	3.78000	3.77825	3.77000	3.77000

Check Table Type

Table header

EUR	1-Dec-99	2-Dec-99	6-Dec-99	7-Dec-99	8-Dec-99
1w	3.07375	3.07563	3.07000	3.07375	3.07188	3.06000
1m	3.51063	3.51963	3.52125	3.53125	3.54500	3.54000
2m	3.45000	3.45000	3.45000	3.46000	3.46000	3.45875
3m	3.45000	3.45000	3.45000	3.46000	3.46000	3.45875
4m	3.45000	3.45000	3.45575	3.46000	3.46000	3.46000
5m	3.46000	3.46000	3.46463	3.46813	3.47000	3.47000
6m	3.46938	3.47313	3.48000	3.48113	3.48000	3.48000
7m	3.50750	3.51000	3.51563	3.51625	3.51688	3.51875
8m	3.55125	3.55588	3.56000	3.56000	3.56000	3.56000
9m	3.60250	3.60063	3.60875	3.60713	3.60375	3.60000
10m	3.65000	3.65000	3.66000	3.65938	3.65000	3.65000
11m	3.70375	3.70875	3.71938	3.71400	3.71000	3.70875
12m	3.76375	3.76588	3.78000	3.77825	3.77000	3.77000

Column Header

Row Header

Table Body

Free Format

- Separate header from body
  - Clustering technique
    - Distance between value types
      - Numeric (1.45, 25000, etc…)
      - Semi-Numeric (4,000yen, 5min, etc...)
      - Character (Hiyoshi, Shibuya, etc…)
    - Minimum distance algorithm
- Divide into datavectors and check consistency

RDB and Fixed Format

- Header and body are separated
- Value type of each column is given by
  - Metadata (VARCHAR, INTEGER, DOUBLE,...)
  - Format text (F3.1, S5,...)

Future Works

- HTML documents generated by CGI and JavaScript
- Excel with Common header
- Free format file with a lot of missing values
- Large RDB

References

Clark, A. (2008) CyberNeko THML Parser Home Page, http://nekohtml.sourceforge.net/.

JExcelApi Group (2008) JExcel API Homepage: http://jexcelapi.sourceforge.net/

Kumasaka, N. and Shibata, R. (2007) Textile Plot Environment, 統計数理特集号「統計データの可視化」, **55** pp. 47-68.

Yokouchi, D. and Shibata, R. (2004) DandD Client-Server System, *Proceedings in COMPSTAT 2004*.

W3C (2008) DOM Homapage: http://www.w3c.org/DOM/