Полуавтоматическое извлечение данных из таблиц

Никита Астраханцев, Денис Турдаков, Наталья Васильева

Институт системного программирования РАН, HP Labs

RCDL, 2013

Пример

ACT	PRD	COR
2577	2722	2101
Prec	**Rec**	**F-Meas**
0.7719	0.8153	0.7930

Table 2: Binary relation classification results for the maximum entropy classifier. ACT: actual number of related pairs, PRD: predicted number of related pairs and COR: correctly identified related pairs.

System	Prec	Rec	F-Meas
NE	0.4588	0.6995	0.5541
MC	0.5812	0.7315	0.6480
PC	0.6303	0.7726	0.6942

Table 3: Full relation classification results. For a relation to be classified correctly, all the entities in the relation must be correctly identified.

First we observe that the maximal clique method combined with maximum entropy (system **MC**) reduces the relative error rate over naively enumerating and classifying all instances (system **NE**) by 21%. This result is very positive. The system based

System	2-ary	3-ary	4-ary
NE	760:1097:600	283:619:192	175:141:60
MC	760:1025:601	283:412:206	175:95:84
PC	760:870:590	283:429:223	175:194:128

Table 4: Breakdown of true positive relations by type that were returned by each system. Each cell contains three numbers, Actual:Predicted:Correct, which represents for each arity the actual, predicted and correct number of relations for each system.

$(\textit{point mutation}, \textit{codon 12}, \perp, \perp)$ is a 2-ary relation. Clearly the probabilistic clique method is much more likely to find larger non-binary relations, verifying the motivation that there are some low probability edges that can still contribute to larger cliques.

6 Conclusions and Future Work

We presented a method for complex relation extraction, the core of which was to factorize complex relations into sets of binary relations, learn to identify binary relations and then reconstruct the complex relations by finding maximal cliques in graphs that

Использование таблиц

Поиск информации

Пример

Таблица VIII. СИНУСЫ

A	0′	6′	12′	18′	24′	30′	36′	42′	48′	54′	60′		1′	2′	3′
70°	0,9397	9403	9409	9415	9421	9426	9432	9438	9444	9449	0,9455	19°	1	2	3
71°	9455	9461	9466	9472	9478	9483	9489	9494	9500	9505	9511	18°	1	2	3
72°	9511	9516	9521	9527	9532	9537	9542	9548	9553	9558	9563	17°	1	2	3
73°	9563	9568	9573	9578	9583	9588	9593	9598	9603	9608	9613	16°	1	2	2
74°	9613	9617	9622	9627	9632	9636	9641	9646	9650	9655	0,9659	15°	1	2	2
75°	0,9659	9664	9668	9673	9677	9681	9686	9690	9694	9699	9703	14°	1	1	2
76°	9703	9707	9711	9715	9720	9724	9728	9732	9736	9740	9744	13°	1	1	2
77°	9744	9748	9751	9755	9759	9763	9767	9770	9774	9778	9781	12°	1	1	2
78°	9781	9785	9789	9792	9796	9799	9803	9806	9810	9813	9816	11°	1	1	2
79°	9816	9820	9823	9826	9829	9833	9836	9839	9842	9845	0,9848	10°	1	1	2
80°	0,9848	9851	9854	9857	9860	9863	9866	9869	9871	9874	9877	9°	0	1	1
81°	9877	9880	9882	9885	9888	9890	9893	9895	9898	9900	9903	8°	0	1	1
82°	9903	9905	9907	9910	9912	9914	9917	9919	9921	9923	9925	7°	0	1	1
83°	9925	9928	9930	9932	9934	9936	9938	9940	9942	9943	9945	6°	0	1	1
84°	9945	9947	9949	9951	9952	9954	9956	9957	9959	9960	9962	5°	0	1	1
85°	0,9962	9963	9965	9966	9968	9969	9971	9972	9973	9974	9976	4°	0	0	1
86°	9976	9977	9978	9979	9980	9981	9982	9983	9984	9985	9986	3°	0	0	0
87°	9986	9987	9988	9989	9990	9990	9991	9992	9993	9993	9994	2°	0	0	0
88°	9994	9995	9995	9996	9996	9997	9997	9997	9998	9998	0,9998	1°	0	0	0
89°	9998	9999	9999	9999	9999	0000	0000	0000	0000	0000	1,0000	0°	0	0	0
90°	1,0000														
	60′	54′	48′	42′	36′	30′	24′	18′	12′	6′	0′	A	1′	2′	3′

Пример

1. Revenue other

	2007 Actual	2007 Supplementary estimates	2007 Main estimates	2006 Actual
		$000		
Revenue from other government departments	5,498	6,715	6,522	4,369
Revenue from third parties	1,994	2,184	2,184	1,899
Total	7,492	8,899	8,706	6,268

Revenue from third parties consists of revenue derived from the sale of Statistics New Zealand's outputs to third parties other than the Crown or other government departments.

2. Operating costs

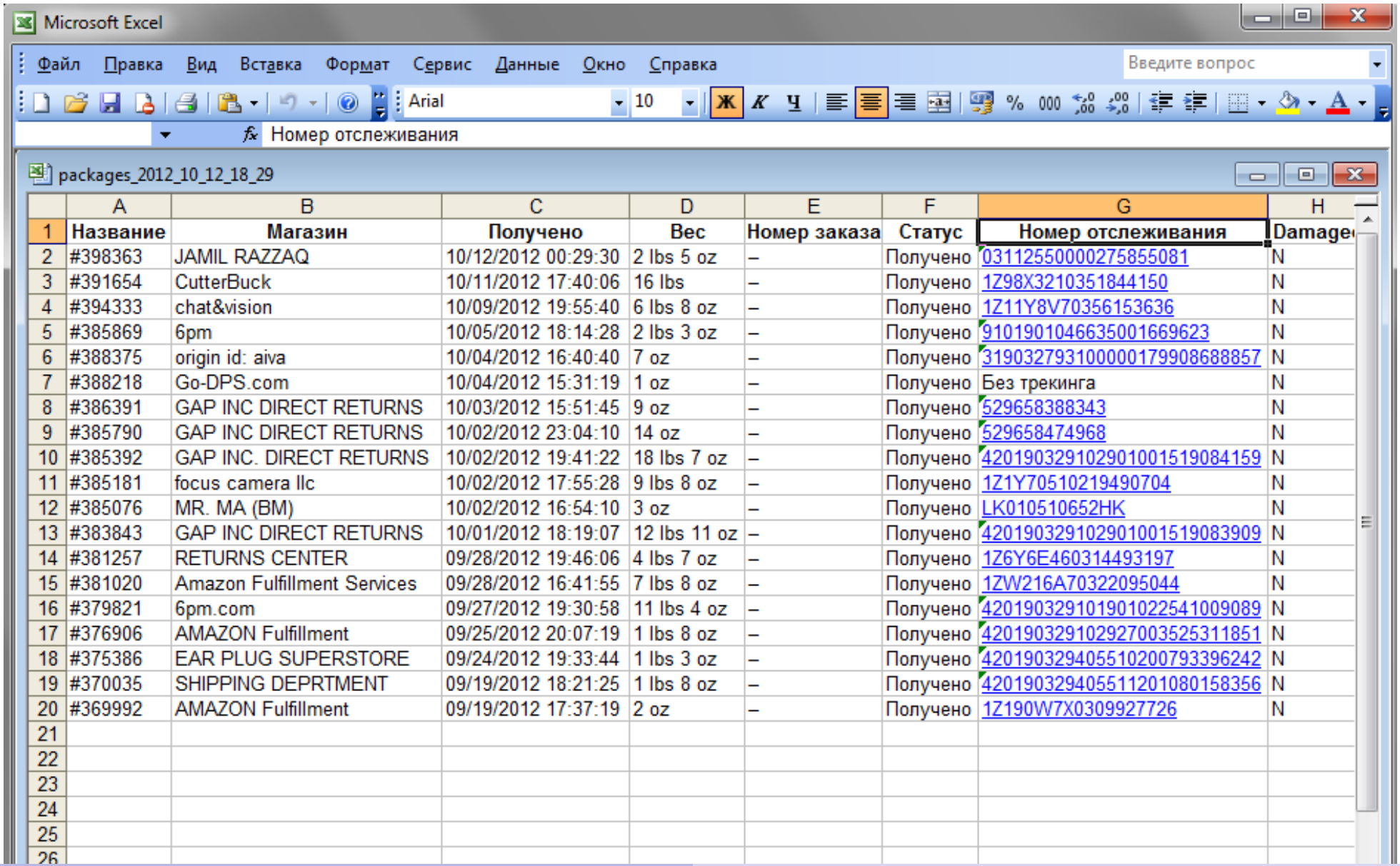
	2007 Actual	2007 Supplementary estimates	2007 Main estimates	2006 Actual
		$000		
Audit fees to auditors	62	65	45	59
Consultancy	1,211	800	800	5,741
Asset write-offs	1,075	0	0	2

Извлечение информации из таблиц

Специальные программы

Извлечение информации из таблиц

Электронные таблицы



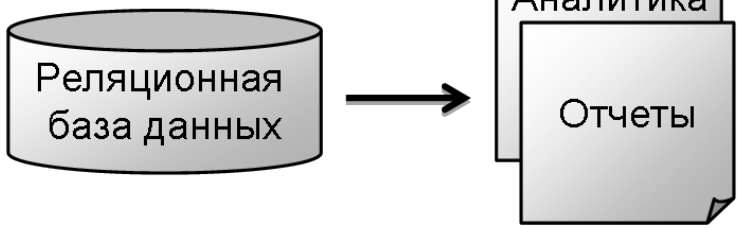
	A	B	C	D	E	F	G	H
1	**Название**	**Магазин**	**Получено**	**Вес**	**Номер заказа**	**Статус**	**Номер отслеживания**	**Damage**
2	#398363	JAMIL RAZZAQ	10/12/2012 00:29:30	2 lbs 5 oz	–	Получено	03112550000275855081	N
3	#391654	CutterBuck	10/11/2012 17:40:06	16 lbs	–	Получено	1Z98X3210351844150	N
4	#394333	chat&vision	10/09/2012 19:55:40	6 lbs 8 oz	–	Получено	1Z11Y8V70356153636	N
5	#385869	6pm	10/05/2012 18:14:28	2 lbs 3 oz	–	Получено	9101901046635001669623	N
6	#388375	origin id: aiva	10/04/2012 16:40:40	7 oz	–	Получено	31903279310000179908688857	N
7	#388218	Go-DPS.com	10/04/2012 15:31:19	1 oz	–	Получено	Без трекинга	N
8	#386391	GAP INC DIRECT RETURNS	10/03/2012 15:51:45	9 oz	–	Получено	529658388343	N
9	#385790	GAP INC DIRECT RETURNS	10/02/2012 23:04:10	14 oz	–	Получено	529658474968	N
10	#385392	GAP INC. DIRECT RETURNS	10/02/2012 19:41:22	18 lbs 7 oz	–	Получено	420190329102901001519084159	N
11	#385181	focus camera llc	10/02/2012 17:55:28	9 lbs 8 oz	–	Получено	1Z1Y70510219490704	N
12	#385076	MR. MA (BM)	10/02/2012 16:54:10	3 oz	–	Получено	LK010510652HK	N
13	#383843	GAP INC DIRECT RETURNS	10/01/2012 18:19:07	12 lbs 11 oz	–	Получено	420190329102901001519083909	N
14	#381257	RETURNS CENTER	09/28/2012 19:46:06	4 lbs 7 oz	–	Получено	1Z6Y6E460314493197	N
15	#381020	Amazon Fulfillment Services	09/28/2012 16:41:55	7 lbs 8 oz	–	Получено	1ZW216A70322095044	N
16	#379821	6pm.com	09/27/2012 19:30:58	11 lbs 4 oz	–	Получено	420190329101901022541009089	N
17	#376906	AMAZON Fulfillment	09/25/2012 20:07:19	1 lbs 8 oz	–	Получено	420190329102927003525311851	N
18	#375386	EAR PLUG SUPERSTORE	09/24/2012 19:33:44	1 lbs 3 oz	–	Получено	420190329405510200793396242	N
19	#370035	SHIPPING DEPRTMENT	09/19/2012 18:21:25	1 lbs 8 oz	–	Получено	420190329405511201080158356	N
20	#369992	AMAZON Fulfillment	09/19/2012 17:37:19	2 oz	–	Получено	1Z190W7X0309927726	N
21								
22								
23								
24								
25								
26								

Извлечение информации из таблиц

Ручное копирование-вставка

Dealer	Model	Price, ₽	Kickback, %	Quantity
Minor-ford	Ford Focus	449000	10	116
	Ford Mondeo	699000	15	365
Smallion	Ford Focus II	165000	70	3
Ralf	Ford Focus	Free	100	0

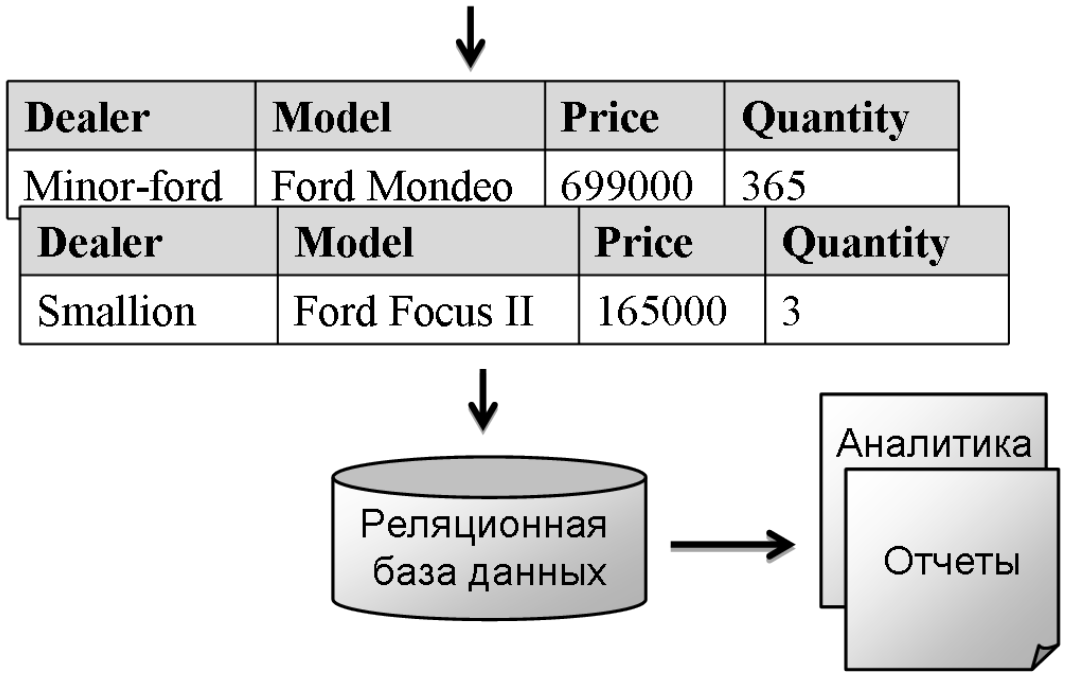
Dealer	Model	Price	Quantity
Minor-ford	Ford Mondeo	699000	365



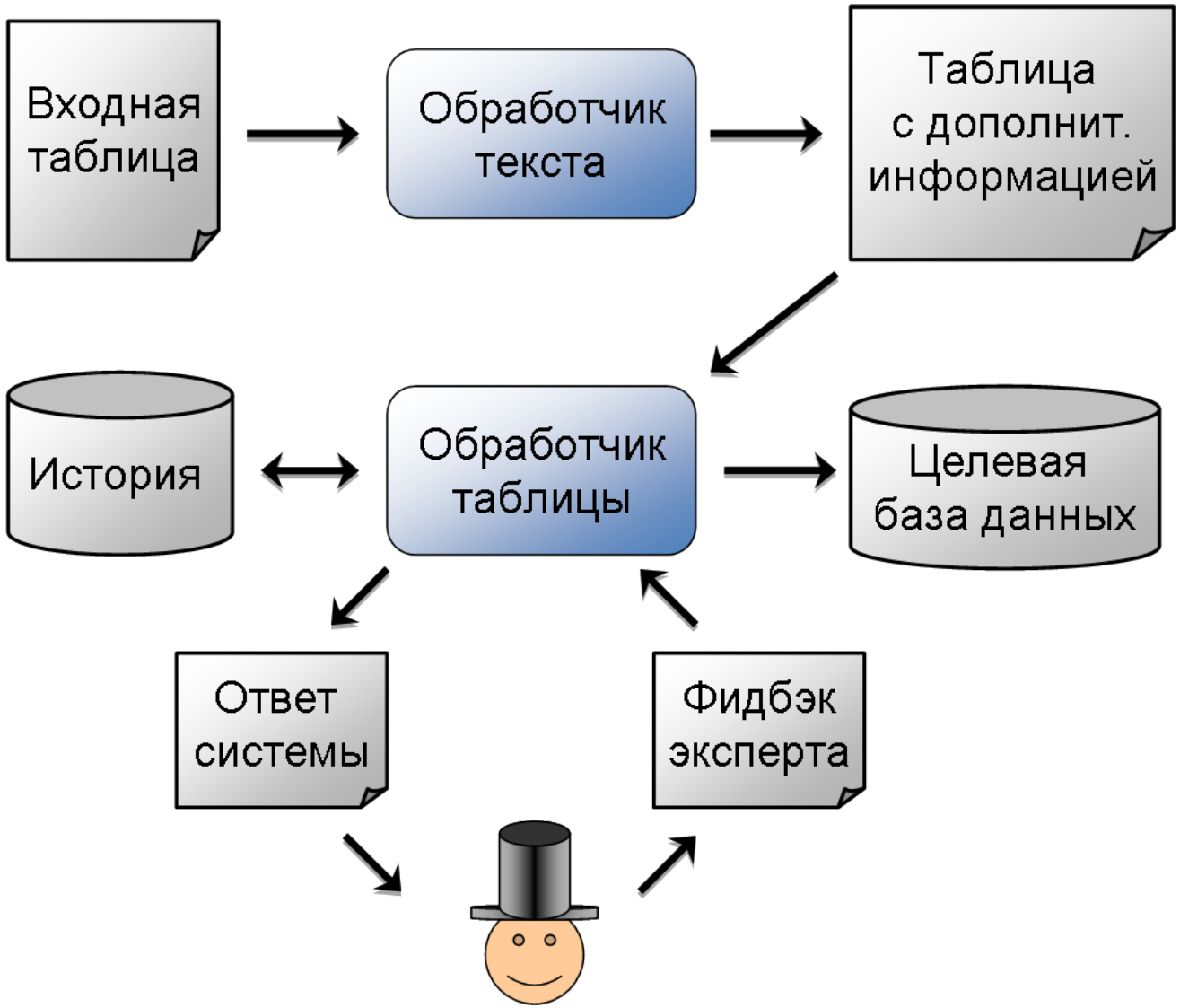
Извлечение информации из таблиц

Ручное копирование-вставка в динамике

Dealer	Model	Price, ₽	Kickback, %	Quantity
Minor-ford	Ford Focus	449000	10	116
	Ford Mondeo	699000	15	365
Smallion	Ford Focus II	165000	70	3
Ralf	Ford Focus	Free	100	0



Общая схема работы



Фидбэк

Пример таблицы

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

Пример фидбэка

Relation name	Attribute name	Cell position	Value
Report	Company	A4	HP
Report	Operating	B4	10
Report	Financial Year	B2	2009

Сдвиг

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

- Сдвиг - пара фидбэков
- Сдвиг ("шаблон сдвига") - набор смещений ячеек, например:
  - 1 ячейка: вниз на 1
  - 2 ячейка: вниз на 1
  - 3 ячейка: на месте

Подход на основе сдвигов

Этапы

1. Регулярный этап
2. Этап с одним фидбэком
3. Этап без фидбэков

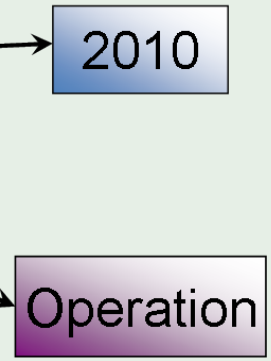
Регулярный этап

Проблемы

1. Извлечение значения

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78



Извлечение значения

- Алгоритм: взять первый подходящий паттерн извлечения значений из ячейки на основе регулярных выражений
- Виды паттернов:
  1. взять подстроку (указанную группу)
  2. заменить все подстроки на указанную

Пример поддерживаемых паттернов

Паттерн	Описание	Ячейка	Фидбэк
.*:::0	Взять значение как есть	HP	HP
(.*)\s(.*):::1:::2	Разбить по пробелу и попробовать каждую часть	Financial year	year
(\d+):::1	Взять только цифры	146%	146
thousand:::000::: ReplaceAll	Заменить все слова *thousand* на 3 нуля	45 thousand	45000

Регулярный этап

Проблемы

1. Извлечение значения
2. Выбор сдвига

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

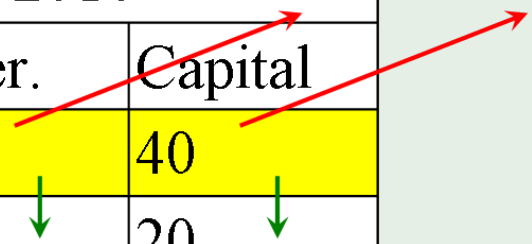
Регулярный этап

Проблемы

1. Извлечение значения
2. Выбор сдвига

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78



Выбор сдвига

Алгоритм

1. Оценить каждый сдвиг линейной комбинацией признаков:
   1. Средняя длина сдвига
   2. Согласованность смещений ячеек
   3. Средняя мера сходства текстовых значений ячеек
   4. Согласованность именованных типов ячеек
2. Использовать сдвиг с наибольшей оценкой

Средняя длина сдвига

Определение

$$AvgLength = 1 - \frac{1}{n}\left(\sum_{i=1}^{n} \frac{offset_i}{maxOffset}\right)$$

где $offset_i$ - длина смещения i-ой ячейки

$maxOffset$ - заданная константа (в экспериментах: 6)

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

$$AvgLength = 1 - \frac{1}{3}\left(\frac{1}{6} + \frac{1}{6} + 0\right) = \frac{8}{9}$$

Согласованность смещений ячеек

Определение

$$OffsetConsistency = \frac{k}{n}$$

где k - число самых частых смещений ячеек

n - всего число смещений ячеек в сдвиге

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

$$OffsetConsistency = \frac{2}{3}$$

Средняя мера сходства текстовых значений ячеек

Определение

$$AvgTextSim = \frac{1}{n} * (\sum_{i=1}^{n} sim(text_{1i}, text_{2i}))$$

где $text_{1i}$ - текстовое содержимое i-ой ячейки первого фидбека

$text_{2i}$ - текстовое содержимое i-ой ячейки второго фидбека

sim - строковая метрика (на основе расстояния Левенштейна)

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

$$AvgTextSim = \frac{1}{3} * (sim(HP, Oracle) + sim(10, 50) + sim(FY2009, FY2009)) = \frac{1}{3} * (0 + 0.5 + 1) = 0.5$$

Средняя мера сходства текстовых значений ячеек

Изменения в строковой метрике

- Все цифры считаются одинаковыми символами

Пример

Sim('10','50')=0.5 $\rightarrow$ ModSim('10','50')=1.0
Sim('10','500')=0.333 $\rightarrow$ ModSim('10','50')=0.667

- Близость с пустой ячейкой равна 0.5

Пример

Sim('10','')=0.0 $\rightarrow$ ModSim('10','')=0.5

- Близость длинных строк (больше 3 слов) равна 0.8

Пример

Sim('Truth is out there','Lie is always here')=0.333 $\rightarrow$
ModSim('Truth is out there','Lie is always here')=0.8

Согласованность именованных типов ячеек

Определение

$AvgTextSim = \frac{1}{n} * (\sum_{i=1}^{n} Consistency(type_{1i}, type_{2i}))$, где

$type_{ji}$ - тип именованной сущности в i-ой ячейки j-го фидбека

$$Consistency = \begin{cases} 1, & \text{если } type_{1i} = type_{2i} \neq undefined; \\ 0.5, & \text{если } type_{1i} = type_{2i} = undefined; \\ 0, & \text{если } type_{1i} \neq type_{2i}. \end{cases}$$

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

$TypeConsistency = \frac{1}{3}(Cons(ORG, ORG) + Cons(NUM, NUM) + Cons(Undef, Undef)) = \frac{1}{3}(1 + 1 + 0.5) = \frac{5}{6}$

Выбор сдвига

Коэффициенты линейной комбинации

1. Средняя длина сдвига (0)
2. Согласованность смещений ячеек (0.4)
3. Средняя мера сходства текстовых значений ячеек (0.2)
4. Согласованность именованных типов ячеек (0.4)

Этап с одним фидбэком

Способы построения сдвига

1. Адаптация сдвига из ранее обработанных таблиц
2. Конструирование нового сдвига

Адаптация сдвига из ранее обработанных таблиц

Алгоритм

Для каждого сохраненного сдвига:

1. Проверить сдвиг на применимость

Пример

Шаблон сдвига:

- 1 ячейка: вниз на 1
- 2 ячейка: вниз на 1
- 3 ячейка: вниз на 1

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

Адаптация сдвига из ранее обработанных таблиц

Алгоритм

Для каждого сохраненного сдвига:

1. Проверить сдвиг на применимость

Пример

Шаблон сдвига:

- 1 ячейка: вниз на 1
- 2 ячейка: вниз на 1
- 3 ячейка: на месте

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

Адаптация сдвига из ранее обработанных таблиц

Алгоритм

Для каждого сохраненного сдвига:

1. Проверить сдвиг на применимость
2. Применить шаблон сдвига к фидбэку

Фидбэк от пользователя

Relation name	Attribute name	Cell position	Value
Report	Company	A4	HP
Report	Operating	B4	10
Report	Financial Year	B2	2009

"Фидбэк"после применения шаблона сдвига

Relation name	Attribute name	Cell position	Value
Report	Company	A5	Samsung
Report	Operating	B5	30
Report	Financial Year	B2	2009

Адаптация сдвига из ранее обработанных таблиц

Алгоритм

Для каждого сохраненного сдвига:

1. Проверить сдвиг на применимость
2. Применить шаблон сдвига к фидбэку
3. Оценить сдвиг - линейная комбинация тех же признаков с другими коэффициентами:
   - Средняя длина сдвига (0.2)
   - Согласованность смещений ячеек (0)
   - Средняя мера сходства текстовых значений ячеек (0.6)
   - Согласованность именованных типов ячеек (0.2)

Использовать лучший сдвиг, если его оценка больше заданного порога (0.5)

Конструирование сдвига

1. Для каждой ячейки: перебрать возможные сдвиги вниз (не более 5) и вправо (не более 3)
2. Оценить каждый полученный сдвиг и выбрать лучший

Пример

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

Этап без фидбэков

Поиск наложимых фидбэков

Сохраненный фидбэк

Relation name	Attribute name	Cell position	Col span	Row span	Value
Universe	Question	**A1**	1	1	What
Universe	Answer	**B1**	1	1	42

Текущая таблица

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

Этап без фидбэков

Поиск наложимых фидбэков

Сохраненный фидбэк

Relation name	Attribute name	Cell position	Col span	Row span	Value
Universe	Question	**A2**	1	1	Who
Universe	Answer	**B2**	1	1	He

Текущая таблица

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

Этап без фидбэков

Поиск наложимых фидбэков

Сохраненный фидбэк

Relation name	Attribute name	Cell position	Col span	Row span	Value
Universe	Question	**A4**	1	1	Where
Universe	Answer	**B4**	1	1	Here

Текущая таблица

	A	B	C	D	E
1	Secret budget				
2		FY 2009		FY 2010	
3		Oper.	Capital	Oper.	Capital
4	HP	10	20	30	40
5	Oracle	50	60	10	20
6	Samsung	12	34	56	78

Этап без фидбэков

Построение фальшивого сдвига

Пример фальшивого сдвига

Relation name	Attribute name	Cell position	Value
Universe	Question	A4	**Where**
Universe	Answer	B4	**Here**

↓ ↓

Relation name	Attribute name	Cell position	Value
Universe	Question	A4	**HP**
Universe	Answer	B4	**10**

Текущая таблица

	A	B	C
1	Secret budget		
2		FY 2009	
3		Oper.	Capital
4	HP	10	20
5	Oracle	50	60
6	Samsung	12	34

Этап без фидбэков

Алгоритм

1. Найти наложимые фидбэки из ранее обработанных таблиц
2. Создать фальшивый сдвиг от сохраненного фидбэка к полученному от наложения сохраненного
3. Оценить по тому же алгоритму с другими коэффициентами:
   - Средняя длина сдвига (0.2)
   - Согласованность смещений ячеек (0)
   - Средняя мера сходства текстовых значений ячеек (0.4)
   - Согласованность именованных типов ячеек (0.4)

Экспериментальные результаты

Описание тестирования

Описание тестового набора

- 30 таблиц (финансовые отчеты) в XML-файлах
- 150 фидбэков, подготовленных вручную

Метрики качества

- Accuracy - доля полностью правильных ответов системы
- Precision - отношение числа правильных отображений ячеек к общему числу отображений ячеек, возвращенных системой
- Recall - отношение числа правильных отображений ячеек к общему числу правильных отображений ячеек

Экспериментальные результаты

Регулярный этап

Accuracy	Precision	Recall
74%	84%	80%

Этап с одним фидбэком

Сохраненных фидбэков	Accuracy	Precision	Recall
0	43%	91%	90%
5	48%	87%	86%
все	86%	87%	86%

Этап без фидбэков

Сохраненных фидбэков	Accuracy	Precision	Recall
0	4%	11%	7%
5	4%	8%	6%
все	16%	28%	25%

Направления дальнейшей работы

- Набор данных для обучения и тестирования
- Более интеллектуальное извлечение данных из ячейки

Пример одной ячейки прайс-листа

HP "Pavilion dm4-2102er"QJ453EA (Core i5 2430M-2.40GHz, 6144MB, HD6470M, WebCam)

- Использование и изменение информации об отношении и атрибуте из фидбэка

Спасибо за внимание!

Precision and Recall

Definition

- **Precision** is the fraction of extracted objects that are correct.

$$Precision = \frac{|\{correct\ \ objects\} \cap \{extracted\ \ objects\}|}{|\{extracted\ \ objects\}|}$$

- **Recall** is the fraction of correct objects that are extracted.

$$Recall = \frac{|\{correct\ \ objects\} \cap \{extracted\ \ objects\}|}{|\{correct\ \ objects\}|}$$