University of Windsor
thinking forward

EMERGING NON-TRADITIONAL DATABASE SYSTEMS:

DATA WAREHOUSING AND MINING (60-539)

SEMINAR REPORT ON TOPIC 1

TOPIC: Finding high quality content in social Media

Article published by: Agichtein,Castillo,Donato,Gionis,Mishne

Cited by 712 other papers

Report Submitted By:

[redacted]

[redacted]

Finding high Quality Content in Social Media

4 1/2
6

INTRODUCTION: In the 1990s onwards, the majority of web users were consumers of content, created by a relatively small amount of publishers. In the early 2000s, user-generated content has become increasingly popular on the web as more users participate in content creation, but not only in consumption of content .Social media contains a rich variety of information sources related to different kind of field content. There is also a wide array of non content information such as links between the items and ratings of content quality by the community members on web. Popular user generated content domains include web forums, photo and video sharing communities, as well as social networking platforms such as Facebook and MySpace. Community-driven question/answering portals are those portals which containing the user-generated content and gaining an attention of people in recent years in wide range. The portals provide an alternative channel for obtaining information on the web: rather than browsing results of search engines in which users answer questions posed by other users. The paper is based on Data Mining.

THE PROBLEM WHICH THE RESEARCHERS/AUTHORS ADDRESSED: The authors state that finding the high quality content in social media is the main task and demand of time. To do it automatically there is a need to find some methods, which can find the quality of information by the feedback of community members available on web. To test the method proposed by author they used Yahoo! Answers.

BACKGROUND AND RELATED WORK: The author focused on Yahoo! Answers portal in this paper. It is basically a system where people ask questions on many topics, some user's give answers some give rating to best answers by voting etc and creates a social network on web. The central element of the Yahoo! Answers portal is questions.

Related work: In the previous work the author refer to the early methods by which the accuracy and quality of content was measured as by using methods like: Link analysis in social media, Propagating reputation, Question-answering portals and forums, Expert finding, Text analysis for content quality, Implicit feedback for ranking. Among all of them the author expanded the work on question/answering portals and forums and implicit feedback for ranking mainly [1] [2] [3].

CONTRIBUTION AND NEW PROPOSED MODEL/APPROCAH AND ANALYSIS

Content Quality Analysis in Social Media: The authors propose a model based on content analysis of social media and the main approach used by authors is **exploiting the features** of social media that are correlated with quality.

Discuss this algm.

-1/2 exceeding page limit of 10. You have 13 or 14 pages.

***Intrinsic content quality*:** The intrinsic quality metrics focus on the text. As a baseline, here using textual features only with all word n-grams up to length 5 that appear in the collection more than 3 times used as features. Also checks Grammar as using several linguistically-oriented features, like part-of-speech (POS) tags.

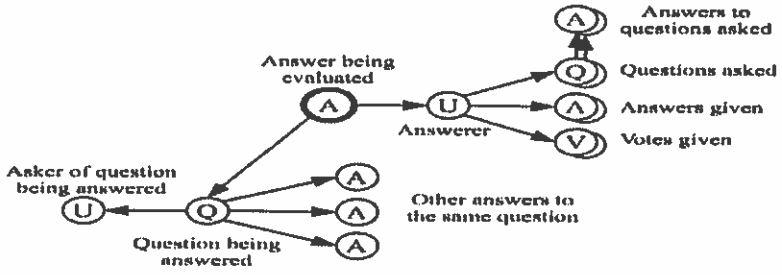
User relationships: Quality information can be obtained from the relationships between users and items. For example, it could apply to QA system, the main challenge is that the dataset, viewed as a graph : nodes of multiple types. *(e.g., questions, answers, users)* edges represent a set of interaction among the nodes having different semantics. *(e.g., "answers", "gives best answer", "votes for", "gives a star to").*

Usage statistics: For example, all items within a popular category such as celebrity topics may receive orders of magnitude more clicks than, for instance, science topics. So it should be normalized by the item category by using per category average, standard deviation and question age.

Modeling Content Quality in Community Question/Answering: The relationships are explicitly modelled in the relational features described as given below:

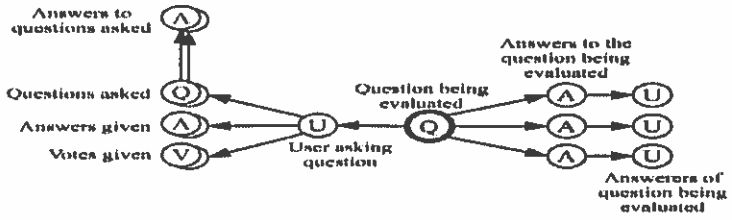
Application-Specific User Relationships: The main forms of interaction among the users are asking a question, answering a question, selecting best answer, voting on an answer. The relationships between questions, users asking and answering questions, & answers can be captured by a tripartite. Since a user is not allowed to answer his/her own questions, there are no triangles in the graph .

Answer features are the types of the data related to a particular answer form a tree, in which the type "Answer" is the root. Two subtrees starting from the answer being evaluated and the types of features on the question subtree are :



Firgure:1 Answer features(Agichetin et al.(2008),page 187)

Question features are the types of the data related to a particular question form a tree, in which the type "Question" is the root. Two subtrees starting from the answer being evaluated and the types of features on the question subtree are :



are you defining terms or what?

Firgure:2 Types of Question features(Agichetin et al.[2008],page 188)

Implicit user-user relations.After the user-question-answer graph, author also consider the user-user graph. This is the graph $G = (V,E)$ in which the set of vertices V is composed of the set of users and the set $E = Ea \cup Eb \cup Ev \cup Es \cup E+ \cup E-$ represents the relationships between users and Ea=answers, Eb =best answers, Ev=votes, $E+$ "thumbs up", $E-$ "thumbs down".

Example to find quality and used algorithmic steps: The input dataset will be the sources of information like Intrinsic content quality, usage statistics, community rating and the new features added by authors that are question features, answer features, user-user relationships. For all of them we can make the graphs and from graphs we can compute adjacency matrix and by apply the HITS, PageRank algorithm in which we can normalize the data at each iteration and can get the pageRank scores. By combining all of these given methods we find relevance and from relevance we find quality. We can take one graph as an example and can compute the scores by applying PageRank algorithm. The number of out-going links is an important parameter. We use the notation "out-degree of a node" to stand for the number of out-going links contained in a node/question/answer/user. In example, there are four nodes. A contains a link to B, a link to C, and a link to D, so B contains one single link to D, node C points to A and D, and node D points to nodes A and C. They are represented by the following graph. We have L(A) = 3, L(B) = 1 and L(C) = L(D) = 2 where L is the number of links outgoing. Note that the sum of the entries in each column is equal to 1. In general, a matrix is said to be column-stochastic if the entries are non-negative and the sum of the entries in each column is equal to 1.The matrix A is by design a column-stochastic matrix, provided that each node contains at least one outgoing link.

what graphs & how are they constructed?

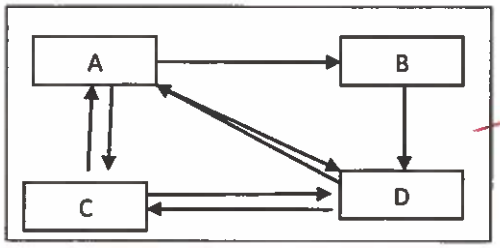
how computed?

how? with example

We don't know what this represents. What are nodes? (web contents or what?)

how is matrix constructed from the graph?

3



What is this & where does it come from?

Let N be the total number of pages. We create an $N \times N$ matrix $\mathbf{A}$ by defining the (i,j)-entry as

$$a_{i,j} = \begin{cases} \frac{1}{L(j)} & \text{if there is a link from } j \text{ to } i, \\ 0 & \text{otherwise.} \end{cases}$$

In Example 1, the matrix $\mathbf{A}$ is the 4×4 matrix

$$\begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1 & 1/2 & 0 \end{bmatrix}.$$

what are the column & row headings?

The simplified Pagerank algorithm is: Initialize x to an Nx1 column vector with non-negative components, and then repeatedly replace x by the product Ax until it converges. We call the vector x the pagerank vector. Usually, we initialize it to a column vector whose components are equal to each other. We can imagine a bored surfer who clicks the links in a random manner. If there are k links in the page, he/she simply picks one of the links randomly and goes the selected question/answer. After a sufficiently long time, the N components of the pagerank vector are directly proportional to the number of times this surfer visits the N web question/answers or links on per category. For example, we let the components of the vector x be xA, xB, xC and xD. Initialize x to be the all-one column vector, i.e.

$$\mathbf{x} = \begin{bmatrix} x_A \\ x_B \\ x_C \\ x_D \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

The evolution of the pagerank vector is shown in the following table.

iteration	x_A	x_B	x_C	x_D
0	1	1	1	1
1	1	0.3333	0.8333	1.8333
2	1.3333	0.3333	1.25	1.0833
3	1.667	.4444	.9861	1.4028
4	1.1944	0.3889	1.0903	1.3264
5	1.2083	0.3981	1.0613	1.3322
6	1.1968	0.4028	1.0689	1.3316
7	1.2002	0.3989	1.0647	1.3361

-1/2 poor tech content & poor clarity

We observe that the algorithm converges quickly in this example. Within 10 iterations, we can see that node D has the highest rank. In fact, D has 3 incoming links, while the others

4

have either 1 or 2 incoming links. It conforms to the rationale of the pagerank algorithm that a node with larger incoming links has higher importance/quality.

EXPERIMENTAL SETTING AND RESULTS: The dataset consists of 6,665 questions and 8,366 question-answer pairs .In the evaluation dataset there is a positive correlation between question quality and answer quality. Question and answer quality is not independent.

In this way by using propagation based metrices we can compute Pagerank score , HITS hub score, HITS authority score from all of the above grpahs and then finally we have source of information gathered from text analysis,clicks and community and relation between the all elements. By getting all these data from matrixs in the form of scores we can now ready to find precision and recall data ie.measure of relevant data from the reterived question/answers and how many of them are relevent and how many of them are not relevant and in this way on the basis of same we can find for answers also we can find the high quality data.

$$\text{Precision} = \frac{tp}{tp + fp_{\text{and}}} \qquad \text{Recall} = \frac{tp}{tp + fn}$$ where *tp*=relevant data reterived,*fp*=reterived but not relevant, *fn*=relevant data but not retrived

Methods	precision	recall	AUC
N-grams(N)	65%	48%	0.52
N+Text analysis	76%	65%	0.65
N+clicks	68%	57%	0.58
N+relations ALL	74% 79%	65% 77%	0.66 0.76

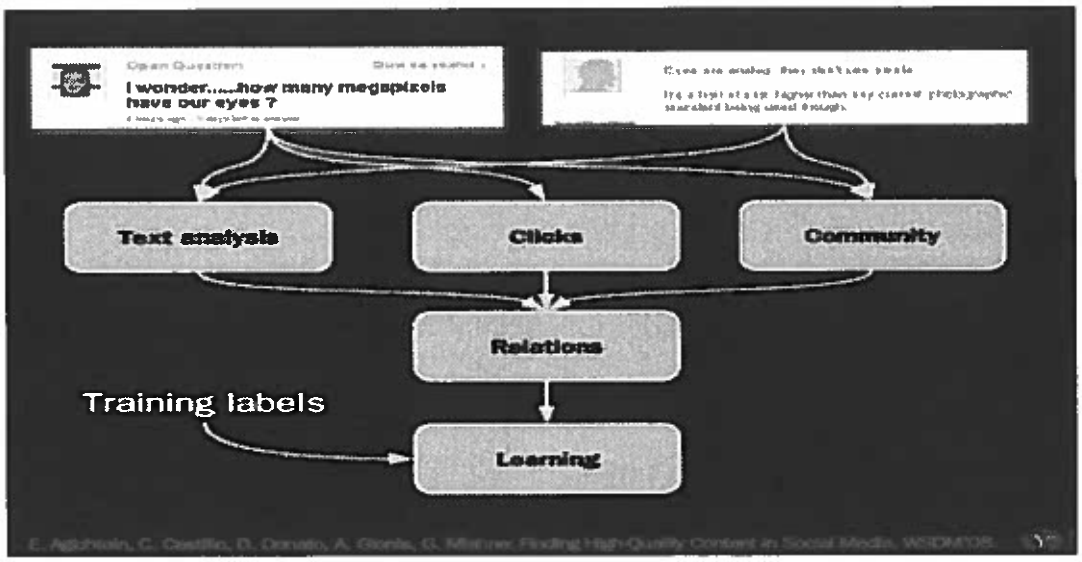
Table2: High quality questions and same we can find for answers also

	Question Quality		
Answer quality	HIigh	medium	low
high	41%	15%	8%
medium	53%	76%	74%
low	6%	9%	18%
Total	100%	100%	100%

Table3: high quality answers and questions

Good answers are much more likely to be written in response to good questions, and bad questions are the ones that attract more bad answers.

5



http://www.mathcs.emory.edu/~eugene/papers/wsdm2008quality.pdf

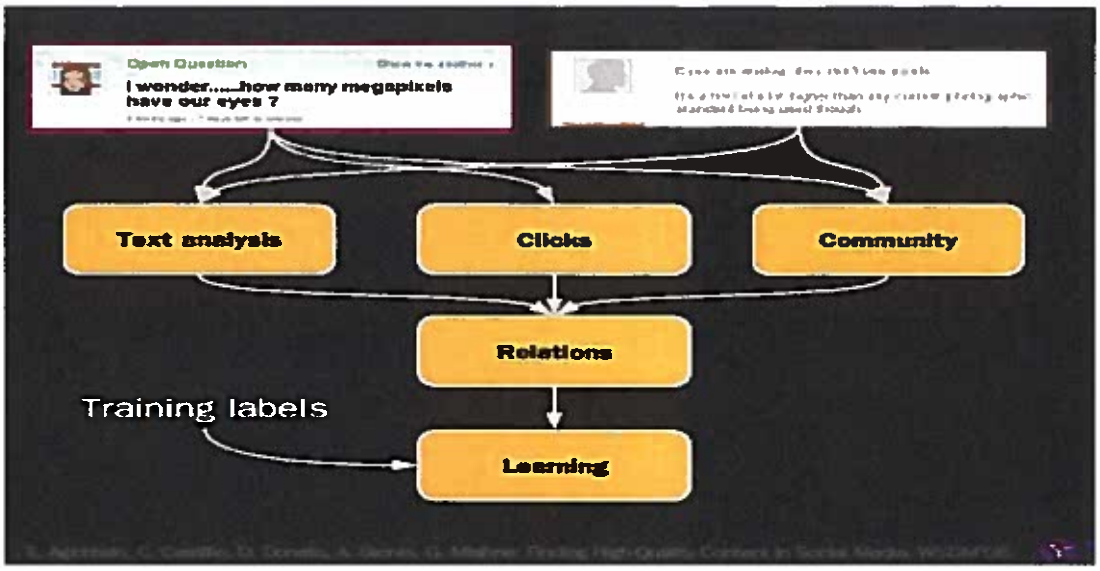
Claims made by the authors with respect to the contribution that they have made: In this paper, authors investigate methods for exploiting such community feedback to automatically identify high quality content. Developing a comprehensive graph-based model of contributor relationships and combined it with content- and usage-based features. By using content of question/answering portal of Yahoo authors show that proposed method is able to separate high-quality data from the rest data just like very close to that of humans.

REFRENCES:

[1] E. Agichtein, E. Brill, S. T. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR*, pages 3–10, 2006.

[2] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of CIKM*, pages 528–531, New Orleans, LA, USA, 2003.

[3] J. Jeon, B. W. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 228–235, New York, NY, USA, 2006. ACM Press.



http://www.mathcs.emory.edu/~eugene/papers/wsdm2008quality.pdf

Claims made by the authors with respect to the contribution that they have made: In this paper, authors investigate methods for exploiting such community feedback to automatically identify high quality content. Developing a comprehensive graph-based model of contributor relationships and combined it with content- and usage-based features. By using content of question/answering portal of Yahoo authors show that proposed method is able to separate high-quality data from the rest data just like very close to that of humans.

REFRENCES:

[1] E. Agichtein, E. Brill, S. T. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR*, pages 3–10, 2006.

[2] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of CIKM*, pages 528–531, New Orleans, LA, USA, 2003.

[3] J. Jeon, B. W. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 228–235, New York, NY, USA, 2006. ACM Press.

7

University of Windsor
thinking forward

EMERGING NON-TRADITIONAL DATABASE SYSTEMS

DATA WAREHOUSING AND MINING (60-539)

SEMINAR REPORT ON TOPIC 2

Title: Descriptive and prescriptive data cleaning

Authors: Anup chalamalla, Ihab F.Ilyas, Mourad Ouzzani,Paolo P

Proceeding : ACM SIGMOD

Year of publication : 2014

Report Submitted By:

INTRODUCTION: In this paper, the authors state that Data cleaning techniques usually depends on some quality rules to identify violating tuples, and then fix these violations using some repair algorithms. Mostly, the rules, which are related to the business logic, can only be defined on some target report generated by transformations over multiple data sources. This creates a situation where the violations detected in the report are decoupled in space and time from the actual source of errors. The main motive of the paper is to explain errors when they are found at a later stage and different space from where they are actually born. The system takes quality rules over the output of a transformation and the errors seen on the output.

THE PROBLEM WHICH THE RESEARCHERS/AUTHORS ADDRESSED: The authors state that in a situation where the violations detected in the report are decoupled in space and time from the actual source of errors are there so in this case applying the repair on the report would need to be repeated whenever the changes has been made in data sources. Hence, if repairing the report is possible and affordable, this would be of little help towards identifying and analyzing the actual sources of errors for future prevention of violations at the target.

CONTRIBUTION AND NEW PROPOSED MODEL/APPROCAH: In this paper, authors propose a system to address this decoupling. The system takes quality rules defined over the output of a transformation and computes explanations of the errors seen on the output. This is performed both at the target level to describe these errors and at the source level to prescribe actions to solve them. . Given a relation R, corresponding VIOLATION Table V(R), and an ERROR Function for V(R), solution for descriptive and prescriptive data cleaning Problem is an EXPLANATION ε_{opt} s.t. ε_{opt} = argmin (|(cover(ε)= E(R))).They present scalable techniques to detect, propagate, and explain errors. They also study the effectiveness and efficiency of our techniques using the TPC-H Benchmark for different scenarios and classes of quality rules.

Example 1: Consider the report T about shops for an international franchise. The HR department enforces a set of policies in the franchise workforce. In the same shop the average salary of the manager (GRD=2) should be greater than the average salary of the staff (GRD=1).

Should lose more marks for page limit

8

T	Shop	Size	Grd	AvgSal	#Emps	Region
t_a	NY1	46 ft^2	2	99 $	1	US
t_b	NY1	46 ft^2	1	100 $	3	US
t_c	NY2	62 ft^2	2	96 $	2	US
t_d	NY2	62 ft^2	1	90 $	2	US
t_e	LA1	35 ft^2	2	105 $	2	US
t_f	LND	38 ft^2	1	65 £	2	EU

Emps	EId	Name	Dept	Sal	Grd	SId	JoinYr
t_1	e4	John	S	91	1	NY1	2012
t_2	e5	Anne	D	99	2	NY1	2012
t_3	e7	Mark	S	93	1	NY1	2012
t_4	e8	Claire	S	116	1	NY1	2012
t_5	e11	Ian	R	89	1	NY2	2012
t_6	e13	Laure	R	94	2	NY2	2012
t_7	e14	Mary	E	91	1	NY2	2012
t_8	e18	Bill	D	98	2	NY2	2012
t_9	e14	Mike	R	94	2	LA1	2011
t_{10}	e18	Claire	E	116	2	LA1	2011

Shops	SId	City	State	Size	Started
t_{11}	NY1	New York	NY	46	2011
t_{12}	NY2	New York	NY	62	2012
t_{13}	LA1	Los Angeles	CA	35	2011

Figure 1: A report T on data sources Emps & Shops.

clarity
tech content
-1/2

Solution of the example: The system takes quality rules as input so, the language in which the constraints are expressed plays an important role. Mostly, constraints can be expressed in any arbitrary code. But, expressing them in Denial constraints makes it uncomplicated. Actual cause of error: t4.grd = 1 instead of 2. In the example, the explanation [T.Region = USΛT.Shop = NY 1] is more specific, if we believe that tb.Grd is the erroneous cell. The process of explaining data errors is two-fold: identifying a set of potential erroneous tuples (cells); and finding concise descriptions that summarize these errors and that can be consumed by users or by other analytics layers.

They focus on the query over source relations Emps and Shops for the US region (Figure 1). Query: *SELECT SId as Shop, Size, Grd, AVG(Sal) as AvgSal, COUNT(EId) as #Emps, 'US' as Region FROM US.Emps JOIN US.Shops ON SId GROUP BY SId, Size, Grd*. To trace back the tuples that contributed to the problems in the target. Tuples ta − td are in violation in T and their lineage is {t1 − t8} and {t11 − t12} over Tables Emps and Shops. By removing these tuples from any of the sources, the violation is removed. Two possible explanations of the problems are therefore [Emps.JoinY r = 2012] On Table Emps, and [Shops.State = NY] on Table Shops.

Example 2: Rules for the running example corresponding to the following denial constraints

$c_1 :\neg (t_\alpha.\text{Shop} = t_\beta.\text{Shop} \wedge t_\alpha.\text{Avgsal} > t_\beta.\text{Avgsal} \wedge t_\alpha.\text{Grd} < t_\beta.\text{Grd})$

9

c2 : $\neg$ (t_α.Size > t_β.Size $\wedge$ t_α.#Emps < t_β.#Emps)

Denial constraint states that all the predicates can not be true at the same time otherwise, there is a violation. Hence, violations in target (T) with respect to a set of constraints (ε) over a transformation on a finite set of data sources is recorded by a function DETECT as a Violation table V(T). A repair function identifies the minimal number of cells to be updated satisfying constraints. Without repair function it's difficult to explain violations and their lineage.

Example 3: Given the rules in the running example a repair function would compute the following updates

repair(c1): (t_α.Shop= t_β.Shop)V(t_α.Avgsal $\leq$ t_β.Avgsal)V (t_α.Grd $\geq$ t_β.Grd)

repair(c2) : (t_α.Size $\leq$ t_β.Size) V (t_α.#Emps $\geq$ t_β.#Emps)

Target to source:

The lineage of target attributes is spread across multiple relations over source. Propagating errors from target to source can be accomplished by tracing the lineage of violations.

Example 4: The lineage of cells t_a,t_b over source relation Emps is t_1-t_4 and Shops is t_{11}.

A novel weight based algorithm is used to calculate the values of tuples corresponding to errors because each tuple in the lineage contributed the same to the error.

The *contribution score*(cs_v) and *removal score*(rs_v) are calculated to weight the tuples violated.A *Contribution score* $cs_v(c)$ is assigned to cell c procedurally based on its value and the SQL operator applied on c. These are computed in top down fashion over a operator tree with each leaf a source tuple and problematic cells annotated with CSV's. A *Removal score* $rs_v(t)$ is 1 if removing a tuple t removes a target violation v.

Algorithm 1: ComputeCSV

A *Contribution score* $cs_v(c)$ is assigned to cell c procedurally based on its value and the SQL operator applied on c. These are computed in top down fashion over a operator tree with each leaf a source tuple and problematic cells annotated with CSV's.

A *Removal score* $rs_v(t)$ is 1 if removing a tuple t removes a target violation v. The algorithm takes input a target relation T with corresponding violation table V(T) over source relations S. It computes CSV's of problematic cells by defining a state as a triplet $(I^l, O^l, INP(O^l))$.The

root is initialized to (T,O^T,inp(^{T})) where O^t is the top operator in the tree that computed T. For all relations in INP(O^T) compute scores of each problematic cell c for each violation v based on procedure 1. For each intermediate relation apply procedure 2 to accumulate the scores and derive CS_v of each problematic cell while ignoring non problematic cells. Thus the scores are computed until the stack is empty and every node is visited.

Most likely error tuples are discovered by using a distance based greedy algorithm. At each step add a tuple with next highest score until gain attains local maxima

$$Gain(H_v) = \sum_{s \in H_v} c_v(s) - \sum_{1 \le j \le |H_v|} \sum_{j < k \le |H_v|} D(s_j, s_k)$$

Where $D(s_j, s_k) = |c_v(s_j) - c_v(s_k)|$

I_1^1	Sid [CSV]	Size [CSV]	Grd [CSV]	Sal [CSV]	Eid [CSV]
t_1^1	NY1 [$\frac{1}{3}$,'']	46 ft^2 ['',$\frac{1}{4}$]	1 [$\frac{1}{4}$,$\frac{1}{4}$]	91 [$\frac{91}{300}$,'']	e4 ['',$\frac{1}{4}$]
t_2^1	NY1 [1,'']	46 ft^2	2 [1,'']	99 [0,'']	e5
t_3^1	NY1 [$\frac{1}{3}$,]	46 ft^2 ['',$\frac{1}{3}$]	1 [$\frac{1}{3}$,$\frac{1}{3}$]	93 [$\frac{93}{300}$,'']	e7 ['',$\frac{1}{3}$]
t_4^1	NY1 [$\frac{1}{3}$,'']	46 ft^2 ['',$\frac{1}{3}$]	1 [$\frac{1}{3}$,$\frac{1}{3}$]	116 [$\frac{116}{300}$,'']	e8 ['',$\frac{1}{3}$]
t_5^1	NY2	62 ft^2 ['',$\frac{1}{3}$]	1 ['',$\frac{1}{3}$]	89	e11 ['',$\frac{1}{3}$]
t_6^1	NY2	62 ft^2	2	94	e13
t_7^1	NY2	62 ft^2 ['',$\frac{1}{2}$]	1 ['',$\frac{1}{2}$]	91	e14 ['',$\frac{1}{2}$]
t_8^1	NY2	62 ft^2	2	98	e18
t_9^1	LA1	35 ft^2	2	94 $	e19
t_{10}^1	LA1	35 ft^2	2	116 $	e20

Figure 4: Procedure 1 Applied on Intermediate Source I_1^1.

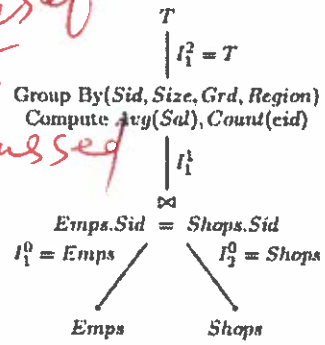


Figure 5: Query Tree.

Emps	Eid [CSV]	Sal [CSV]	Grd [CSV]	Sid[CSV]		[RSV]
t_1	e4 ['',$\frac{1}{4}$]	91 [$\frac{91}{300}$,'']	1 [$\frac{1}{4}$,$\frac{1}{4}$]	NY1 [$\frac{1}{3}$,'']		[0,1]
t_2	e5	99 [0,'']	2 [1,'']	NY1 [1,'']		[1,'']
t_3	e7 ['',$\frac{1}{3}$]	93 [$\frac{93}{300}$,'']	1 [$\frac{1}{3}$,$\frac{1}{3}$]	NY1 [$\frac{1}{3}$,'']		[0,1]
t_4	e8 ['',$\frac{1}{3}$]	116 [$\frac{116}{300}$,'']	1 [$\frac{1}{3}$,$\frac{1}{3}$]	NY1 [$\frac{1}{3}$,'']		[1,1]
t_5	e11 ['',$\frac{1}{3}$]	89	1 ['',$\frac{1}{3}$]	NY2		[',0]
t_6	e13	94	2	NY2		
t_7	e14 ['',$\frac{1}{2}$]	91	1 ['',$\frac{1}{2}$]	NY2		[',0]
t_8	e18	98	2	NY2		
t_9	e14	94	2	LA1		
t_{10}	e18	116	2	LA1		

Figure 6: Procedures 1 and 2 Applied on Emps.

Shops	Sid [CSV]	Size [CSV]		[RSV]
t_{12}	NY1 [2,'']	46 ['',1]		[1,1]
t_{13}	NY2	62 ['',1]		[',1]
t_{14}	LA1	35		

Figure 7: Procedures 1 and 2 Applied on Shops.

Once CSVs are computed for cells, we can compute them for tuples by summing up the cell scores along the same violation while ignoring non contributing cells.

LIKELY ERRORS DISCOVERY: They present two approaches to solve this problem in the first approach; they design an outlier function to separate high and low scoring tuples for each violation. In the second approach, show a reduction from the facility location problem and apply a polynomial time log n-approximation algorithm to compute the likely source errors 1. Distance Based Local Error Separation and 2. Global Error Separation

EXPLANATIONS DISCOVERY: Explanation discovery is performed in two steps basically as given below:

Generating candidate queries for a source S with d dimensions such that the query covers at least one tuple in E(R) for each attribute A^l of R. Data structure p stores all the problematic cells. Then the algorithm expands in a depth first manner by performing a conjunction with attributes $A^l \cdots A^d$

Algorithm 2: The second algorithm is Greedy PDC in which at each step algorithm adds to ε the query that maximizes marginal cover and minimizes weight. So, marginal cover mcover(q) is the number of tuples from R That are in Q and not already in ε and Weight W(q) is calculated by adding Errors not covered by q like bcover(q) = E(R) ∩ cover(q), similarly for bcover(Eopt). Therefore, to clean tuples covered by q they set the parameter λ to be the error rate, as it reflects the proportion between errors and clean tuples.

Metrics. Authors introduce two metrics to test DBRx. For each metric, they show how to compute precision (P) and recall (R).

Error Discovery Quality: Evaluates the quality of the likely errors discovery and the scoring. By comparing the errors computed by Error over the lineage versus the changes introduced in the errors induction step (B)

$$P_{Err} = \frac{E(T) \cap B}{E(T)} \quad R_{Err} = \frac{E(T) \cap B}{B}$$

Explanation Quality: Evaluates the quality of the discovered explanations. To measure the quality of an explanation computed by DBRx by testing the tuples overlap with the ground explanations

$$P_{\mathcal{E}} = \frac{cover(\mathcal{E}_{opt}) \cap cover(\mathcal{E}_g)}{cover(\mathcal{E}_{opt})} \quad R_{\mathcal{E}} = \frac{cover(\mathcal{E}_{opt}) \cap cover(\mathcal{E}_g)}{cover(\mathcal{E}_g)}$$

Advantages: Explanation about errors will be useful for error detection and prevention for users in future.

CONCLUSIONS: To make these explanations easy to consume for users, they formulated a problem that minimise their size while guaranteeing coverage of the violations. The main intuitions focus was (i) violations at the target level can be expressed as evidence of

12

problems over the sources and (ii) summarising such evidence leads to meaningful explanations of the problems.

FUTURE WORK: Authors state that in future they plan to extend this work by considering multi-level transformations, such as ETL processes, where at each step rules for data cleaning can be enforced and Transformations can expressed in black box (JAVA) which is not possible in existing system.