Meeting notes

- Shu, Ioana
- 20-04-2017

Main points

- Review of the summary (By Ioana) of the paper: SEEDB (Especially some points in terms of word utilization, article structure...)
    - A good example of paper summary by UC-Berkeley (https://people.eecs.berkeley.edu/~fox/summaries/database/)
- Some inspiring ideas

A comparison between SEEDB and our project

Perpective	SEEDB	Our Project
Database	snowflake schema	RDF graph
Q: query	Used to specify a subset of the dataset that an analyst wants to explore In this paper: SPJ queries over snowflake schema	1.Type queries e.g. **x rdf:type T** 2.Ressource having most frequent set of properties[1] 3.Simply, all D
Reference data D_R	May be defined as the entire underlying dataset(D), the complement of D_Q or data selected by any arbitrary query. Analyst can specify.	**If** class-hierarchy (For instance, Pub ←Thesis, Pub←Journals, Pub ←ConferencePapers) and we work on sub-class (For instance, journals, D_T is the set of journals) **Then** D_R is the set of Pub **Else** D_R is resources having most frequent set of properties (maybe also 2nd most frequent...) **OR** simply $D_R = D_Q$
		1. Many potential: **a.** only those present in all D_R resources **b.** all the attributes 2. We may choose only N attributes. But which N dimensions? So, we should decide

a	attributes on which we group-by	which attributes are good dimensions (For instance, 1) if the attribute has few number of distinct values[2]) 2) if it's an integer 3) if it's a string) Problem can thus be reduced to M attributes, but we still have 2^M possibilites, so starting from only one attribute can a solution (if each time we group by only one "good dimension", we have M possibilites)
m, f	measure attributes, on which we apply aggregate function	Measure m: 1. Only those present in all D_R resources 2. all Aggregate function f: 1. If it's a $number \rightarrow$ sum, avg, min, max ... 2. If it's a $String \rightarrow$ COUNT
Deviation	Utility is defined via deviation.Visualizations showing different trends in the query dataset compared to a reference dataset (large deviation) are said to have a high utility	Deviation between (a,m,f) on D_T and D_R **If** no D_R **Then** use **skewness**(3rd moment) or **kurtosis** (4th moment) as measure for interestingness[3]

Remarks

[1] In most of cases, resources having most frequent set of properties may only be a small set of total resources. Consequently, we may need some **support** here: **How can we make sure that we work on a "big" subset of D?** (Should we explicitly tell the users the percentage of total data that we are working on when it comes to a small subset of data? Or should we explictly tell the users the percentage of total data at any time?)

[2] The meaning of "few number of distinct values": image attribute A has 1000 distinct values. DB1 has 1000 resources having A, DB2 has 1000000 resources having A. For DB2, A is a good candidate. Consequently, the threshold of "few distinct" could be $\frac{|D_R|}{K}$ for some value of K.

[3] Reference to chapter 14 of the book Statistical Description of Data page-610

- Problems

What a user can do/say?

- choose D_R from a set of options

- choose at any step between P options

For instance, given a view (a,m,f) = {x type inProceeding group by year count(*)}
How to determine the next P proposals?