(1)

\* The motivation of the problem is still weak. Authors claim that the summary produced by the previous approach is still quite large. But many summarization approaches can be repeatedly executed on a summary until the summary is small enough.

\* One weakness is motivation. This paper says "...assist users in attaining a quick and rough preview of the data". A specific application would be useful.

\* Is there any existing schema summarization method which can address the same problem? At the least, it should be taken into consideration. Otherwise, experiment results without comparison are not strong enough to convince people.

\* Finally, the problem is very similar to the work on relational database summarization. Deeper insights to similarity and difference are important for evaluation and comparison.

(2)

\* The experiments are not enough. Only results on Freebase are tested.

\* In Figure 5, P@10 is about 0.4 in 5 out of the 6 domains.

\* In Table 3, for 5 out of 6 domains, the results show a medium positive correlation between their scoring measures and AMT workers, rather than a strong positive correlation.

(3)

\* the definitions (of scoring definition) are too heuristic. I think the score of the preview should be the weighted sum of different features (for both key attributes and non-key attributes). The different weights for various features could be learned from labeled data, which is more reasonable.

\* The scoring also seems straightforward. The rationality of the proposed scoring mechanism is not sufficiently discussed.

(4) Parameter (k, n) selection is not discussed.

(5) Authors ignore the complexity of the computation of the scoring functions in the complexity analysis of their approach.

(6) Also, the "gold standard" for top attributes shown in the paper may no longer be appropriate since Freebase has changed.

**1. The scores we obtained are -2,+3,-2.**

Must accept (+6) - This is a paper that you estimate to be in the top 25% of the papers typically accepted at WWW. You are prepared to (indirectly) champion it in front of 25 of your senior colleagues (= the senior PC).

Should accept (+3) -- This is a solid paper that you estimate to be within the top 80% of the papers typically accepted at WWW.

Marginal (-2) -- This is not necessarily a bad paper but 4 out of 5 papers at WWW are usually better.

Should reject (-4) -- This is a paper that is under the typical bar of WWW papers. Very few if any of the papers accepted at WWW are at this level.

Must reject (-6) -- This paper is too weak, incomplete, or plain wrong. Given the WWW selectivity, you view it as unacceptable in its present state.

**2. Future submissions:**

KDD 2/21-5/12

CIKM 5/24-7/?

**3. Major issues:**

**3.1 The definition and experiments of scoring measures (criticized by 2 reviewers)**

Previously, we are thinking that by enumerating different scoring measures, we could claim that we are not making contributions over finding any particular good scoring measures but our methods could accommodate other scoring measure definitions for different application scenarios. Actually, this could be true because there might indeed exists no universal good standard for attribute selection. Now looks like this does not work. All reviewers believed that the scoring measure definition is a weakness of this paper. R1 suggests us to learning the measures from labeled data. I'm not sure this is the way we want to go.

*(R1) My main and biggest concern however is related to the scoring definition for measuring the goodness of previews. The intuition is reasonable; however, the definitions are too heuristic. From the experimental results in Section 6.1, the results are not very satisfactory. In Figure 5, P@10 is about 0.4 in 5 out of the 6 domains. In Table 3, for 5 out of 6 domains, the results show a medium positive correlation between their scoring measures and AMT workers, rather than a strong positive correlation. This scoring definition is the key for the success of this algorithm. If this scoring measure cannot measure the goodness of preview tables well, the final results may be not very satisfactory. I think the score of the preview should be the weighted sum of different features (for both key attributes and non-key attributes). The different weights for various features could be learned from labeled data, which is more reasonable.*

*(R2) The rationality of the proposed scoring mechanism is not sufficiently discussed. Authors mainly use the frequency of an attribute in the entity graph as the basic scoring mechanism. It is arguable whether frequency is a good mechanism to summarize a knowledge graph. For example, in many case we hope to get a complete summary instead of a summary focusing on the most frequent entities.*

*Some experimental results seem to be discouraging. For example, in Figure 5, P@K is about 40%. In Table 3, most PCC is less than 0.5. It implies that the user study result has weak correlation with the result of the proposed approach.*

**3.2 Motivation (criticized by 2 reviewers)**

We are planning to show a demo interface to enforce the motivation. This should be finished before next submission. We also need to revise the introduction section.

*(R2) The motivation of the problem is still weak. Authors claim that the summary produced by the previous approach is still quite large. But many summarization approaches can be repeatedly executed on a summary until the summary is small enough.*

*(R3) One weakness is motivation. This paper says "...assist users in attaining a quick and rough preview of the data". A specific application would be useful. In the introduction part, the arguement that summary of the schema graph is inadequate is not convincing. Is there any existing schema summarization method which can address the same problem? At the least, it should be taken into consideration.*

**3.3 Experiments (criticized by 2 reviewers)**

We are planning to have results over other dataset before next submission. But reviewers also want us to add comparison with relational summarization we mentioned in introduction.

*(R2) The experiments are not enough. Only results on Freebase are tested. Freebase is edited by human. It is less convincible if only such a well-edited knowledge base is tested. We expect to see the results on other knowledge bases. Currently, whether the proposed solution can work on other knowledge base is with uncertainty.*

*(R3) Also, the "gold standard" for top attributes shown in the paper may no longer be appropriate since Freebase has changed.*

*(R3) Finally, the problem is very similar to the work on relational database summarization. Although the authors compared them in Section 1, it seems that they are very similar in principle. Deeper insights to similarity and difference are important for evaluation and comparison.*

**3.5 Parameter selection (criticized by 1 reviewer)**

We treat k, n, and d as user input parameters. One reviewer think we should also study these numbers.

*(R2) Parameter selection is not discussed. In this paper, selection of k (the number of key attributes) and n (the number of non-key attributes) are important issues. Unfortunately, how to select them is not known. It is expected to discuss this issue.*

**3.6 Tuple selection (criticized by 1 reviewer)**

We need to add a discussion of tuple selection. Previously it's in future work but removed due to space limitation. R3 thought we need to include all entities into a preview table.

*(R3) Other issues are that the weightings of the entities in the same table could still be different. However, this approach doesn't include this factor in the table generating process. The importance score simply relies on the structure of the schema graph rather than the entire entity graph.*

*In addition, to increase the coverage score, it seems that the proposed methods would prefer selecting the schema graphs that consist of many entities. Thus, even though the number of tables and the number of attributes in a preview table are limited, it may generate tables with many entities, which can still be space costly. It might be more reasonable to add constraints on the number of entities per table as well.*

**4. Solutions**

Currently, I'm working on a demo of generating previews, which might help improving motivation. I'm also planning to add other data experiments but not sure I have enough time to finish this. Including this will definitely help experiment section. We might also need to think about comparing with relational summarization somehow. For 3.1, it might be really hard to address, I'm not we should go with machine learning algorithm for attribute selection, and with current methods, there is no way to make the figures look better.