KDD2014
ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 24-27, New York, U.S.A.

Reviews For Paper

Track	Research
Paper ID	314
Title	Generating Preview Tables for Entity Graphs

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
How would you rate the novelty of the problem solved in this paper?	A minor variation of some well studied problems
How would you rate the technical ideas and development in this paper?	The technical development is incremental without fundamental contributions
How would you rate the empirical study conducted in this paper?	Acceptable, but there is room for improvement
Repeatability: are the data sets used publicly available (and thus the experiments may be repeated by a third party)?	Yes
How would you rate the quality of presentation?	A very well-written paper, a pleasure to read. No obvious flaws.
Which topic category do you think this paper belongs to?	Other
What is your overall recommendation?	Weak reject. I vote for rejecting it, although would not be upset if it were accepted.
List up to 3 particular strengths of the paper. If none, just say "none".	1. The idea of generating concise tables for entity graph preview is new. 2. The complexity of finding optimal preview is mathematically proven, and several efficient algorithms are proposed. 3. The paper is well-organized. Extensive empirical study is conducted.
List up to 3 particular weaknesses of this	1. The constraints supported by the system, e.g., the max distance between key attributes, seem not so useful and intuitive to users. On one hand, it'd be desired if the system can automatically suggest an optimal value of the distance. On the other hand, a semantically related constraint might be more useful, e.g., generating preview tables, as well as statistics, for a "central" entity type assigned by users, as it would be useful for users who want to use different proportions of the entity graph.

paper. If none, just say "none"..	2. A dynamic programming algorithm and an Apriori-style algorithm for efficient optimal table discovery are proposed, but no direct comparison between these two algorithms is conducted. It should compare which one is better in which situation. 3. The Apriori-style algorithm is as slow as brute-force algorithm in some settings. It seems merely a brute-force algorithm preceded by a graph-pruning step. Is it possible to combine the dynamic programming algorithm with the graph-pruning step to get a more efficient solution?
Detailed comments for the authors; justification of your overall rating	The paper proposes to generate preview tables to provide concise preview of large entity graphs. Important entity/relation types are selected based on defined measures, and organized as flat tables. Users can constrain the size of tables and the distance between the selected entity types. The optimization problem is proven to be NP-hard, and thus two algorithms are developed to speedup. The paper is well-organized, but there are several weak points. First of all, the supported constrains seem not so intuitive for user with different needs. An automatically suggestion of optimal values, as well as other types of constraints, are desired. Secondly, no direct comparison between the proposed dynamic programming alg. and Apriori-style alg. is conducted. It's not clear which one is better in which situation. Finally, flat tables might not be a good form for users to explore the properties of entity graph. E.g., schema graph (pruned if too large) should be a more visual form.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
How would you rate the novelty of the problem solved in this paper?	A minor variation of some well studied problems
How would you rate the technical ideas and development in this paper?	The technical development is incremental without fundamental contributions
How would you rate the empirical study conducted in this paper?	Acceptable, but there is room for improvement
Repeatability: are the data sets used publicly available (and thus the experiments may be repeated by a third party)?	Yes
How would you rate the quality of presentation?	A very well-written paper, a pleasure to read. No obvious flaws.
Which topic category do you think this paper belongs to?	Information extraction
What is your overall recommendation?	Weak reject. I vote for rejecting it, although would not be upset if it were accepted.

List up to 3 particular strengths of the paper. If none, just say "none".	1. This paper studies an interesting problem, that is, generating previews as a compact presentation of large entity graphs. 2. Real datasets are used in the experiment for evaluation. 3. The writing of this paper is of quality and it is pleasant to read the paper.
List up to 3 particular weaknesses of this paper. If none, just say "none"..	1. This paper aims to generate a compact presentation of large entity graphs and thus defines a concept of preview which consists of a set of tables. However, the number of tables as well as the number of attributes contained in a preview must be specified by users. In reality, it is hard to determine proper values for such parameters. In addition, the constraint of the number of non-key attributes is set for a preview instead of the individual tables that constitute the preview. It may result in that one table in the preview contains a lot of attributes while the others only contain the minimum possible number of attributes, for example, the last preview shown in Table 5 and the first preview shown in Table 6. 2. Three types of previews are defined which are concise preview, tight preview and diverse preview. However, from the experiment results shown in Table 6, no matter tight or diverse previews are mined, some attributes such as FILM appear in most of the tables that constitute the preview, which may result in that the mined previews contain highly redundant information. 3. While the large entity graphs are only used for computing the attribute scores, the previews are mined from the much smaller schema graphs. In Section 6.2, the schema graphs used in the efficiency experiments, respectively, contain 6, 23, and 46 nodes, which are too small to well demonstrate the efficiency of the proposed methods. Actually, for a graph containing 6 nodes, generating a preview seems quite unnecessary.
Detailed comments for the authors; justification of your overall rating	Refer to Q10 for detailed comments.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
How would you rate the novelty of the problem solved in this paper?	A well established problem
How would you rate the technical ideas and development in this paper?	Substantial improvement over state-of-the-art methods
How would you rate the empirical study conducted in this paper?	Thorough, including systematic comparison with proper state-of-the-art methods and novel case studies (if applicable)
Repeatability: are the data sets used publicly available (and thus the experiments may be repeated by a third party)?	Yes

How would you rate the quality of presentation?	A very well-written paper, a pleasure to read. No obvious flaws.
Which topic category do you think this paper belongs to?	Graph mining
What is your overall recommendation?	Accept. I vote and argue for acceptance, clearly belongs in the conference
List up to 3 particular strengths of the paper. If none, just say "none".	The paper presents a convincing solution to the problem of summarizing an entity graph; it is the first to include both schema and instance information. The paper comes with a good formal model, includes complexity results for the formal optimization problems, and gives a suite of exact and approximate algorithms. The paper provides a good experimental evaluation that takes both efficiency and effectiveness into account, and the results of this evaluation are convincing.
List up to 3 particular weaknesses of this paper. If none, just say "none"..	The paper focuses on entity graphs only, while an extension to ontologies (including type hierarchies, complex type definitions, etc.) would be very valuable for real applications, especially in the context of Linked Data. It is not fully clear how large a good summary should be (even though this could be determined incrementally by the user), but parameter tuning in general (k vs. n) seems to be difficult. There is no evaluation of the summaries in context of an application.
Detailed comments for the authors; justification of your overall rating	The paper considers the problem of summarizing entity graphs. This is an important problem when it comes to choosing graphs that are relevant to a specific problem out of a large collection of such graphs. Existing solutions have focused on summarizing the underlying schema graph, and there does not seem to be prior work that takes size limitations into account and thus provides an approximate, concise summary. Additionally, the example entities that are provided within a summary make it easy for a non-expert user to understand the summary. Thus, the paper clearly makes an important contribution to this problem. Furthermore, the paper provides a reasonable formal model for the problem of size-limited summarization. It introduces three variants of the model and proves that two of them are NP hard. All three variants are reasonable on their own, but it is not fully clear which should be preferred in which situation. Algorithms are provided for each of the three variants, including both exact and approximate variants (depending on which version of the clique solver is applied). The fastest algorithms demonstrate a good performance, especially as computing the summary is usually an offline operation, so no real-time answer times are required. The experimental evaluation is clear and considers both effectiveness and efficiency. The idea of comparing to the Freebase "gold standard" is good, but it is not clear to me how carefully these types and attributes were chosen. Anyway, the extra AMT-based evaluation is a good indication that the scores are reasonable; it is not clear that a single perfect order exists at all. There is no clear advantage for any of the scores; it would be useful if the paper provided rules when a certain score should be preferred. An interesting extension would be a hybrid score that integrates both elementary scores. The same is true for choosing from the three variants of the optimization problem, and how to set parameters k and n for a given entity graph. The experimental evaluation would be even stronger if the paper examined the quality of summaries in the context of an application. Here, it could be evaluated if a much simpler summary such as a plain list of types without attributes and examples would be less effective when it comes to selecting graphs. This would also allow to compare to other methods that do not generate tables, but summary graphs or trees. This is probably the only true weakness of

	this otherwise very nice paper. While the scenario of entity graphs is important and clearly complex enough, it would be very interesting to examine how things change if moving to more complex schemas such as provided by ontologies (including type hierarchies, complex type definitions, etc.). This would make the results even more valuable in the context of Linked Data.
List of typos, grammatical errors and/or concrete suggestions to improve presentation	"domain-s" -> domains "s-core" -> score