VLDB 2010
The 36th International Conference on Very Large Data Bases
13 to 17 September 2010

Singapore

Reviews For Paper

Track	Core Database Technology
Paper ID	442
Title	Structured Querying of Annotation-Rich Web Text with Shallow Semantics

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall Rating	Reject
Reject due to technical incorrectness	No
Novelty	Medium
Technical Depth	Low
Presentation	Adequate
Summary of the paper's main contributions and impact (up to one paragraph)	This paper presents an entity search system. The authors identify limitations of existing entity search systems, in the lack of query support for multiple entities and inter-relationships. And then a novel Entity Centric Index is proposed, with the formal query definition and search process methods. Experiments on a Wikipedia dataset demonstrate the efficiency of the new approach.
Three strong points of the paper (please number them S1,S2,S3)	S1. Entity search with relation is an interesting problem. S2. This paper is easy to understand and the authors also provide a demo site. S3. The detailed indexing design and formal definition are presented.
Three weak points of the paper (please number them W1,W2,W3)	W1. This paper is limited in the declarative query interface and simple indexing design. The proposed approach is straightforward and couldn't extend. W2. The relation split and pruning are heuristic. Long join relation predicate sequences or other types of inter-relationships are ignored. W3. Though the authors show various types of test queries, the complexity and characters of relational entity search still look vague to us.
Detailed Comments (please number each point)	Though the entity search with relationship is an interesting problem, I think this paper doesn't provide much insights in this important direction. D1. the relation extraction used in this paper is straightforward and the relation scope is limited in sentence level. Is other possible relation or

	context definition necessary? D2. how to select entities and relations from other rich textual data? flat documents, instead of structured Wikipedia? This is an important question towards relational entity search. D3. the join sequence and simple pruning are similar to query optimization in RDBMS or candidate network selection in keyword search. As this paper only focuses on indexing design, the search space or long join sequences should also be included in this paper. D4. The authors present the pre-join speed-up in Sec 5.2. Is it extensible to the system deployment? D5. The demo site is intuitive and promising. But this paper seems unfinished and leaves many spaces for further improvement.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Overall Rating	Accept
Reject due to technical incorrectness	No
Novelty	High
Technical Depth	Medium
Presentation	Not good
Summary of the paper's main contributions and impact (up to one paragraph)	Proposes an approach that allows structured querying of annotated textual data, allowing some rather interesting queries. This is a neat approach that sidesteps the problems of converting textual data completely to structured data, while providing many of the querying benefits. The example queries are quite well motivated.
Three strong points of the paper (please number them S1,S2,S3)	S1: The query language proposed seems quite powerful. S2: The proposed entity centric indexing approach seems clean and efficient, and experiments demonstrate its superiority over a (somewhat strawman-like) document-centric index.
Three weak points of the paper (please	W1: The performance results only report IO counts. This is NOT the defacto standard, since the costs of sequential and random IO are vastly different. It can be used as a metric only if your technique guarantees the

number them W1,W2,W3)	number of seeks is small and independent of the indexing technique, so the IOs are mostly sequential. If this is so for your indexing techniques, please explain why. In any case, it is essential to present actual execution times on a cold cache (on linux, at least, there are commands to flush the file system cache).
Detailed Comments (please number each point)	D1: In related work, some of the things you say about [30] such as tradeoff of space and time, are also applicable to [11]. Also, it considers efficient indexing of hierarchies of categories, which you do not consider, although clearly you do things that [11] does not address. Please reread [11] and describe its contributions and differences from your work better. D2: The paper has a number of grammatical errors. For example, Theorem 1 mixes up "phrase" and "phase". Many sections such as end of 4.1 talk about "trash evidences", this is not an appropriate term, use "data later found to be irrelevant". In the last para of 4.1, "It is unknown how to (and probably not able to)" badly needs rephrasing. The use of "enlists" should be replaced by "lists" everywhere. Section 5.2 says "hugh to afford", rephrase this. There are others too, which you need to find.
List specific clarifications you seek from the Authors (if you have answered "Yes" to Q. 6) Use this space to respond to author feedback too.	C1: Please provide actual execution time numbers as explained in W1. I trust these should not be hard to obtain, and doable in the limited author feedback time. Without these numbers, I am not confident that the ECR approach is even a good idea.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Overall Rating	Reject
Reject due to technical incorrectness	No
Novelty	Medium
Technical Depth	Low
Presentation	Not good
Summary of the paper's main contributions and impact (up to one paragraph)	The paper presents a declarative SQL-like language called SSQ which enables users to formulate entity-centric queries. Such language is claimed to be more powerful than simple keyword-based search for entity retrieval. To support evaluation of SSQ, an entity-based retrieval algorithm is proposed that uses a new entity centric index (ECI). In

	contrast to existing document-centric retrieval, it orders posting lists by entity id. Some experimental results are presented to demonstrate the merit of the proposed algorithm.
Three strong points of the paper (please number them S1,S2,S3)	S1: The introduction and motivation is well presented. S2: Simplicity of SSQ for querying entities and relationship. S3: A new indexing scheme is proposed to for entity retrieval.
Three weak points of the paper (please number them W1,W2,W3)	W1: The paper lacks of technical depth and have low research quotient. W2: Presentation needs improvement. W3: Lack of experimental omparison with some existing approaches.
Detailed Comments (please number each point)	The paper starts of well. The motivating example in Section 1 is nice and sets the work. However, the meat of the paper needs to be improved from several fronts as follows. W1: The paper lacks technical depth. The main meat of the paper is in Sections 3, 4.2, and 4.3. Although the proposed entity centric index-based retrieval clearly works, there is no indepth analysis of the proposed algorithm. Furthermore, the entity centric indexing scheme is simple and not very novel. While simplicity is always a virtue, a deeper analysis is needed to improve the research quotient of the work. W2: The presentation of the work needs significant improvement. First, there are several typos and careless mistakes in the paper which is annoying. It gives an impression that proper proof reading has not be carried out. Let me highlight some of these mistakes: * The authors have used "phrase" instead of "phase" in several places in the paper. * In Pg 4 3rd para: a1 a3 c2 -> a1 b3 c2, Gats -> Gates * Second last para in Section 4.1: wast -> waste Section 4.1 is verbose. Given that it is not a novel contribution, the entire discussion can be summarized to create space for deeper analysis of the ECR approach. The formal description of Algorithm 3 is basically very similar to Algorithm 2. It can easily be explained in the context of Algo 2 by highlighting on how it can be extended to handle ECR with pruning. Pls note that Appendix should contain materials that are not mandatory to read. The description of the features of the dataset should not be in the Appendix as the reader certainly needs to know the characteristics of the dataset to make sense of the performances discussed subsequently. The quality of SSQ results, presented in Section 5.3, does not add any

value to the work as the ranking technique is not discussed in the paper!

There are several terms used in the paper that are ambiguous. For instance, what are major join and sub-join? What is "pre-joined" posting list - are you referring to materialized views?

W3: The empirical study needs improvement. The entire study is focused on disk I/O comparison. Certainly, this is important. However, I would also like to see study on the pruning power of the proposed algorithm and scalability. Further, why the proposed approach is not empirically compared with [30]?