[paper title]

***Abstract*—[ABSTRACT]**
***Index Terms*—[Items]**

I. Introduction

Modern Knowledge Base Management Systems allow the ingestion of various data representations, such as relational, graph and full-text data [1]. Current KBMS usually ingest reliable data sources, such as encyclopaedic data [1], business data [2] and medical journals and clinical data [3]. Despite such KBMS are able to reconcile different representation towards an uniform one [4], no technique is currently exploited for detecting *contradicting* facts: in particular both data collected from on-line social network [5] or even medical diagnoses [6] may contain contradictions.

Example 1: One of the simplest ways to find contradictions is to check whether there are facts that are both affirmed and denied at the same time. E.g., fact "Casu Marzu is not a good cheese" is a rebuttal of the fact "Casu Marzu is a exquisite cheese". Alternatively, we can focus on factual representations that do not allow the contemporaneity of two alternative hypotheses, thus violating a functional dependency. E.g., fact "Yesterday Alice flew to Berlin" contradicts "Yesterday Alice took a trip to Kearney" because Alice cannot be found in two different places at the same time.

The inherent inconsistency of spurious data sources leads to the generation of conflicting hypotheses in response to a query. The inability of detecting such inconsistencies prohibits to weight how many KBMS facts either support or discard a given hypothesis, thus preventing from correctly ranking the generated hypotheses. As a result, this paper provides a formal definition of information inconsistencies, which is later on exploited to define such ranking metric; given a type of facts t associated with a functional dependency f, we say that facts c and $\tilde{c}$ of type t are contradictory either if $\tilde{c}$ is the negation of c or the coexistence of c and $\tilde{c}$ in t violates the functional dependency f. In order to meet our goal, we represent both events and relationships as multidimensional facts of different types t via common-sense knowledge integration [8]; to each type t we associate a functional dependency by exploiting the ontologies' ability to describe a domain knowledge of interest [7].

Example 2: Provide a semi-formal description of the previous example, and how former relationships and facts can be represented

To improve efficiency and accuracy, we provide some link and rule mining techniques that [TODO]. To summarize, we make the following contributions:

- Section description here.

II. Related Work

Please note that this section has to be edited accordingly to the type of paper we intend to write.

A. Natural language query answering

- Natural language querying in relational databases: [2], [9]
- Natural language querying in graph databases: [10]

B. Rule Mining

Miguel, here we should quote the paper on temporal rule mining.

C. Approximated graph pattern matching

- Graph matching over dimensions: [11]
- Approximated graph matching: [12], [13]

D. Knowledge Bases

- KB with precise information: [1]
- KB with uncertainty: [4], [14]

III. Preliminaries

Definition 1 (Knowledge Base): A **knowledge base** KB is a graph containing ground atoms.

Definition 2 (Query): A **query** q is a graph containing unbounded variables $?x$ to which a common-sense type τ is associated. U_q is a finite set of all the unbounded atoms in q.

Definition 3 (Answer): An **answer** A to a query q over a knowledge base KB (namely, $q(KB) = A$) is a set of distinct finite functions $\alpha : U_q \to KB$, named **candidate alternative**.

We denote $\alpha \oplus q$ as the operation replacing each unbounded atom $?x$ in the query q with the candidate $\alpha(?x)$.

Definition 4 (Hypothesis): An **hypothesis** h supporting a candidate alternative α for query q is a subgraph of KB matching with q ($h \subseteq KB \wedge \text{sim}(h, \alpha \oplus q)$).

Definition 5 (Support): For each hypothesis h generated from a knowledge base KB, the support is a pair $s(h) = \langle C, \tilde{C} \rangle$, where C is a set of ground KB atoms validating h and $\tilde{C}$ is a set of facts discarding the hypothesis.

As a consequence of the support definition, we want enforce that the hypotheses scoring function σ must enjoy the following properties:

- if $s(h) = \langle C, \emptyset \rangle$ with $C \neq \emptyset$, then h is "global truth" in KB and then $\sigma(h) = 1$.
- if $s(h) = \langle \emptyset, \tilde{C} \rangle$ with $C \neq \emptyset$, then h is "global falsehood" in KB and then $\sigma(h) = 0$.
- $0 < \sigma(h) < 1$ otherwise.

[More definitions compliant with the learning steps] In the learning phase, we can approximate each query q with a set of probable hypotheses that may be returned by the KB

Less equal (<=) is $\leq$: http://web.ift.uib.no/Teori/KURS/WRK/TeX/symALL.html

IV. Architecture

Fragment 1: Our architecture is based upon the Closed World Assumption: each hypothesis that cannot be directly generated or inferred by our internal data representation should be considered inadmissible.

Fragment 2: The system architecture which is used to query the data is defined by the following steps:

1) Given a knowledge base KB, perform the query q returning an answer $A = \{ \alpha_1, \ldots, \alpha_n \}$
2) For each candidate alternative α_i, return a set of hypotheses $\{ h_i^1, \ldots h_i^m \}$
3) For each hypothesis h_i^j, evaluate its support $s(h_i^j)$ and rank it through σ.

V. Algorithm

- Rename this section with the algorithm name.
- Are there more than just one algorithm to describe?
- Shall we describe the main pipeline instead, and use other papers as back references?

VI. Experimental Evaluation

Fragment 3: In order to evaluate our architecture, we use the LDC dataset [TODO: which?], in which each query q is described as a set of hypotheses with its support set. Therefore, we must define a similarity score between support sets for each hypothesis in order to evaluate the compliance of our KB with respect to the returned hypotheses.

VII. Conclusion

Acknowledgment

[DARPA Project ?]

References

[1] I. J. of Research and Development, *This is Watson*, J. W. Murdock, Ed. IBM Co., May/July 2012, vol. 56(3/4).

[2] D. Saha, A. Floratou, K. Sankaranarayanan, U. F. Minhas, A. R. Mittal, and F. Özcan, "Athena: An ontology-driven system for natural language querying over relational data stores," *Proc. VLDB Endow.*, vol. 9, no. 12, pp. 1209–1220, Aug. 2016.

[3] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34 – 49, 2018.

[4] F. Niu, C. Zhang, C. Ré, and J. Shavlik, "Elementary: Large-scale knowledge-base construction via machine learning and statistical inference," *Int. J. Semant. Web Inf. Syst.*, vol. 8, no. 3, pp. 42–73, Jul. 2012.

[5] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018. [Online]. Available: http://science.sciencemag.org/content/359/6380/1094

[6] D. Costa and M. A. Martins, *Measuring Inconsistency in Information*. College Publications, 2018, ch. Inconsistency Measures in Hybrid Logics, pp. 169–194.

[7] V. Fortineau, A. Cornire, T. Paviot, and S. Lamouri, "Modeling domain knowledge using inference ontologies: an application to business rules management," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 573 – 578, 2015, 15th IFAC Symposium onInformation Control Problems inManufacturing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S240589631500381X

[8] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," *CoRR*, vol. abs/1612.03975, 2016.

[9] F. Li and H. V. Jagadish, "Understanding natural language queries over relational databases," *SIGMOD Rec.*, vol. 45, no. 1, pp. 6–13, Jun. 2016.

[10] S. Hu, L. Zou, J. X. Yu, H. Wang, and D. Zhao, "Answering natural language questions by subgraph matching over knowledge graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 5, pp. 824–837, 2018.

[11] A. Petermann, G. Micale, G. Bergami, A. Pulvirenti, and E. Rahm, "Mining and ranking of generalized multi-dimensional frequent subgraphs," in *Twelfth International Conference on Digital Information Management, ICDIM 2017, Fukuoka, Japan, September 12-14, 2017*, 2017, pp. 236–245.

[12] R. De Virgilio, A. Maccioni, and R. Torlone, "Approximate querying of rdf graphs via path alignment," *Distributed and Parallel Databases*, vol. 33, no. 4, pp. 555–581, Dec 2015.

[13] J. Aligon, E. Gallinucci, M. Golfarelli, P. Marcel, and S. Rizzi, "A collaborative filtering approach for recommending olap sessions," *Decision Support Systems*, vol. 69, pp. 20 – 30, 2015.

[14] Y. Chen and D. Z. Wang, "Knowledge expansion over probabilistic knowledge bases," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: ACM, 2014, pp. 649–660.