Fact Checking: theory and practice

KDD'18 - Traditional-tutorial proposal

Xin Luna Dong * — Amazon
Christos Faloutsos — SCS CMU
Xian Li — Amazon
Subhabrata Mukherjee — Amazon
Prashant Shiralkar — Amazon

April 24, 2018

Abstract

Was Da Vinci born in Florence? Does patient 'Johnson' really have 300 heart-beats per minute? Checking the accuracy of facts is vital, for question answering, data cleaning, anomaly detection, fraud detection, and more.

Here we present three families of fact-checking approaches, based on the domains to which they apply: (a) *text* documents (b) *graphs* and *knowledge bases* and (c) *relational* databases. The emphasis is on the intuition behind each method, as well as on a practitioner's guide, highlighting the applicability of each method to each setting.

1. Title

Fact Checking: theory and practice

2. Abstract

In recent times, the explosion of information from a variety of sources has made it increasingly important to check the credibility and reliability of the underlying data. Large volumes of data generated from diverse information channels like social media, online news outlets, crowd-sourcing contribute valuable knowledge; however, this comes with additional challenges to ascertain the credibility of user-generated information, resolving conflicts between heterogeneous data sources, identifying outliers and anomalies etc. Given diverse information about an object (e.g., a natural language claim text, an entity, an SPO like triple) from various

* Author names listed in alphabetical order.

sources, how do we establish the credibility of this information, resolve conflicts among noisy bits of information, and identify the true value among several fact candidates? How do we identify high quality and trustworthy sources of information? In order to answer these questions, this tutorial surveys several algorithms focusing on the intuition behind them (as opposed to the mathematical analysis); it highlights their strengths, similarities, and illustrates their applicability to real-world problems. In contrast to prior tutorials on similar topics, we cover a wide breadth of related techniques and methods focusing both on unstructured texts and structured data like relational databases and graphs.

3. Target Audience and prerequisites

Data Scientists and practitioners, with interest in Knowledge Bases, Database quality, Truth Finding and Discovery, Credibility Analysis.

Prerequisites: A B.Sc. in computer science should suffice. The tutorial assumes familiarity with basic linear algebra, calculus, discrete math; as well as with fundamentals of Machine Learning (classification, clustering, matrix factorization).

4. Tutors

In alphabetical order:

- Xin Luna Dong, Amazon, `lunadong@amazon.com`
- Christos Faloutsos, CMU and Amazon, +1-412-576.7932, `christos-sabbatical@cs.cmu.edu`
- Xian Li, Amazon, `xianlee@amazon.com`
- Subhabrata Mukherjee, Amazon, `subhomj@amazon.com`
- Prashant Shiralkar, Amazon, `shiralp@amazon.com`

names and affiliations, plus phone and e-address, here

5. Tutors bio

up to 200 words per tutor

Xin Luna Dong is a Principal Scientist at Amazon, leading the efforts of constructing Amazon Product Knowledge Graph. She was one of the major contributors to the Google Knowledge Vault project, and has led the Knowledge-based Trust project, which is called the "Google Truth Machine" by Washington Post. She has got the VLDB Early Career Research Contribution Award for "advancing the state of the art of knowledge fusion". She co-authored book "Big Data Integration", is the PC co-chair for Sigmod 2018 and WAIM 2015, and is serving in the VLDB advisory committee and the Board of Trustees of the VLDB Endowment.

She has given several tutorials on data integration and knowledge management in top-tier conferences.

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Research Contributions Award in ICDM 2006, and the SIGKDD Innovations Award (2010). He has given over 40 tutorials and over 20 invited distinguished lectures. His research interests include large-scale data mining with emphasis on graphs and time sequences; anomaly detection, tensors, and fractals.

Xian Li is an Applied Scientist at Amazon contributing to the data quality and knowledge fusion in Amazon Product Knowledge Graph. Before joining Amazon, she was a data scientist at LinkedIn working as a major contributor of building the LinkedIn's knowledge base of business entities. She received her Ph.D. from SUNY Binghamton and her research interests include truth finding in structured and unstructured data sources, data quality, and knowledge management.

Subhabrata Mukherjee is a Machine Learning Scientist at Amazon building the Amazon Product Knowledge Graph. He is working on building large-scale machine learning models that extract knowledge from unstructured and semi-structured data. He graduated summa cum laude from Max Planck Institute for Informatics, Germany with a Ph.D. He has previously worked at IBM Research on domain adaptation of question-answering systems, and sentiment analysis. His research interests include probabilistic graphical models, information extraction, and recommender systems.

Prashant Shiralkar is an Applied Scientist in the Product Graph team at Amazon. He currently works on knowledge extraction from semi-structured data. Previously, he received a Ph.D. from Indiana University Bloomington where his dissertation work focused on devising computational approaches for fact checking by mining knowledge graphs. His research interests include machine learning, data mining, information extraction and NLP, and Semantic Web technologies.

6. Corresponding author with her/his email address

Subhabrata Mukherjee: `subhomj@amazon.com`

7 Tutorial outline.

Provide as much detail as possible.

- **[15']** Part 1: Introduction
- **[40']** Part 2: Fact Checking in Text Documents
  - Language-aware Fact Assessment: Approaches using probabilistic graphical models [16, 14, 15], graph algorithms [25], FactChecker [17] etc.

  - Natural Language Claim Checkers: Approaches using background knowledge from the Web [19], probabilistic soft logic [22], and neural networks [26, 20, 4]
- **[40']** Part 3: Fact Checking in Graphs
  - Fact Checking from Knowledge Networks: Knowledge Vault [7], T-verifier [13], FactChecker [5], Knowledge Stream [24], Predicate Path Mining [23], Path Ranking Algorithm [10], DeepPath [28]
  - Anomaly detection in graphs: OddBall [1], Survey on anomaly detection [2]; fraud detection / lockstep behavior (CopyCatch [3], Fraudar [9]).
- **[40']** Part 4: Fact checking in Structured Data
  - Fact checking with iterative models (FACTY, Solomon, etc.) [11, 6, 12]
  - Fact checking with probabilistic graphical models (LCA, LTM, KBT, etc.) [29, 18, 8]
  - Fact checking with combined supervised and unsupervised models: SLiMFast [21]
  - Fact checking by query perturbations: [27]
- **[15']** Conclusions - Future research directions.

8. A list of forums of earlier offerings of this/related tutorials

and their time and locations if the tutorial or a similar/highly related tutorial has been presented by the same author(s) before, and highlight the similarity/difference between those and the one proposed for KDD18 (up to 100 words for each entry)

This tutorial is completely new.

Related tutorials include:

- *Data fusion–Resolving data conflicts for integration* Xin Luna Dong and Felix Naumann. In VLDB, 2009. [*click for PDF*] [*click for Presentation*]
- *Truth Discovery and Crowdsourcing Aggregation: A Unified Perspective* Jing Gao, Qi Li, Bo Zhao, Wei Fan, Jiawei Han, in VLDB, Kohala Coast, HI, August 2015

In our proposed tutorial we consider *broader* types of data (text, graphs, and relational data) and we present more *recent* techniques.

9. References

[1] L. Akoglu, M. McGlohon, and C. Faloutsos. oddball: Spotting anomalies in weighted graphs. In *PAKDD (2)*, volume 6119 of *Lecture Notes in Computer Science*, pages 410–421. Springer, 2010.

[2] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Min. Knowl. Discov.*, 29(3):626–688, 2015.

[3] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *WWW*, pages 119–130. International World Wide Web Conferences Steering Committee / ACM, 2013.

[4] S. Cao, B. Qian, C. Yin, X. Li, J. Wei, Q. Zheng, and I. Davidson. Knowledge guided short-text classification for healthcare applications. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 31–40, Nov 2017.

[5] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. *PLoS ONE*, 10(6):1–13, 2015.

[6] X. L. Dong, L. Berti-Équille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[7] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment*, 8(9):938–949, 2015.

[8] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *PVLDB*, 8(9):938–949, 2015.

[9] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. FRAUDAR: bounding graph fraud in the face of camouflage. In *KDD*, pages 895–904. ACM, 2016.

[10] N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[11] F. Li, X. L. Dong, A. Langen, and Y. Li. Knowledge verification for longtail verticals. *PVLDB*, 10(11):1370–1381, 2017.

[12] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 6(2):97–108, 2012.

[13] X. Li, W. Meng, and C. Yu. T-verifier: Verifying truthfulness of fact statements. *Proceedings - International Conference on Data Engineering*, pages 63–74, 2011.

[14] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 745–754, New York, NY, USA, 2015. ACM.

[15] S. Mukherjee and G. Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*, 2015.

[16] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: Credibility of user statements in health communities. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 65–74, New York, NY, USA, 2014. ACM.

[17] N. Nakashole and T. M. Mitchell. Language-Aware Truth Assessment of Fact Candidates. *Acl*, pages 1009–1019, 2014.

[18] J. Pasternack and D. Roth. Latent credibility analysis. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 1009–1020, 2013.

[19] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. *WWW (Companion Volume)*, pages 1003–1012, 2017.

[20] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2921–2927. Association for Computational Linguistics, 2017.

[21] T. Rekatsinas, M. Joglekar, H. Garcia-Molina, A. G. Parameswaran, and C. Ré. Slimfast: Guaranteed results for data fusion and source reliability. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1399–1414, 2017.

[22] M. Samadi, P. Talukdar, M. Veloso, and M. Blum. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources.

In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 222–228. AAAI Press, 2016.

[23] B. Shi and T. Weninger. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104:123–133, 2016.

[24] P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia. Finding streams in knowledge graphs to support fact checking. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 859–864, 2017.

[25] V. V. Vydiswaran, C. Zhai, and D. Roth. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 974–982, New York, NY, USA, 2011. ACM.

[26] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL (2)*, pages 422–426. Association for Computational Linguistics, 2017.

[27] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Computational fact checking through query perturbations. *ACM Transactions on Database Systems (TODS)*, 42(1):4, 2017.

[28] W. Xiong, T. Hoang, and W. Y. Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.06690*, 2017.

[29] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB*, 5(6):550–561, 2012.

10. Equipment you will bring

Laptop; HDMI and VGA adaptors

11. Equipment you will need

- Projector, with HDMI or VGA input.
- Power sockets

12. Equipment attendees should bring

None

13-16. Hands-on-Tutorial

N/A

17. Slides

Slides will be available at `github`

18 Optional: Video snippet of you teaching

- Xin Luna Dong, *Knowledge Vault and Knowledge-based Trust*, Stanford seminar series, 2015: *click here for video.*
- Christos Faloutsos, *Mining Large Graphs*, Distinguished Lecture Series, UIC, April 2015: *click here for the video* and *here for the foils.*
- Prashant Shiralkar, *RelSifter: Scoring Triples from Type-like Relations*, Intelligent Systems Seminar, Indiana University Bloomington, January 2017: *click here for the video* and *here for the foils*