Sharing retrieved information using Linked Credibility Reviews

Ronald Denaux[a], Jose Manuel Gomez-Perez[a]

[a]*Expert System, Madrid, Spain*

Abstract

In recent years, the advent of deep learning for NLP is enabling the accurate retrieval of semantically similar content. Such retrieval, along with check-worthiness and stance detection, is crucial for identifying misinformation and linking it to relevant verified information. While sources of credibility signals (and methods to retrieve them) are abundant on the web, they vary greatly in terms of quality and relevance and can be quite scarce for specific claims. These special requirements for such IR systems suggest the need for good abstractions to represent relevant aspects like the credibility of retrieved information and the confidence of automated systems that retrieved the information. In this paper, we (i) summarise Linked Credibility Reviews, existing work that provides a conceptualisation and exchange format for representing the credibility of retrieved verified information and (ii) discuss the role this conceptualisation can play in information retrieval systems for reducing online misinformation.

Keywords

misinformation detection, shared conceptualisation, information exchange, distributed information retrieval

1. Introduction

The Web and social media have ushered an era of ultra-fast spreading of messages without the need for centralized media like publishing houses, newspapers, radio or TV channels. However, this lack of centralization also entails a lack of editorial and quality control over the messages that spread online. This misinforming capacity of the web and social media is increasingly being exploited and is having a detrimental effect on society as evidenced by problems like political polarization and interference in democratic and policymaking processes.

Information retrieval is bound to play a crucial role in reducing online misinformation as it has the potential to bring back some of the quality control that was lost in the transition from traditional media to the web. However, traditional information retrieval metrics and approaches cannot be directly applied if they are to be used to spot and limit the spread of misinformation. Additional aspects like credibility and harmfulness have to be taken into account when determining relevancy of results and other requirements like explainability, reproducibility, trust and decentralisation have to be taken into account. All of this means that datasets and information retrieval systems need to be aware of the wider online context where they are

CEUR Workshop Proceedings (CEUR-WS.org)

deployed and this requires a well designed conceptual model and vocabulary to publish and share such artifacts on-line.

In this paper, we summarize existing work on such conceptual models and vocabularies produced recently by the semantic web research community. Rather than presenting new research, our intention is to build a bridge between the information retrieval and semantic web communities as we believe such collaboration is necessary to successfully tackle misinformation online.

2. IR Requirements for Misinformation Reduction

As mentioned above, information retrieval systems for online misinformation reduction have several new requirements in addition to standard IR systems. In this section, we enumerate and describe these additional requirements.

1. **New relevancy aspects**: In standard information similarity, content (and metadata) similarity are the main aspects that determine relevancy of retrieved results. In the context of misinformation, aspects like credibility, stance and harmfulness also need to be considered.
2. **Subjectivity**: Aspects like credibility and harmfulness are subjective as they depend on the content being evaluated, the evaluation method and the background information used to evaluate both aspects.
3. **Explainability**: IR systems should be able to explain the reason why retrieved content was deemed relevant, credible or harmful.
4. **Reproducibility**: When possible, the methods used should be reproducible by third parties, especially end users. This allows end users to verify the methods and results even if they are not aligned with their subjective views on, for example, credibility.
5. **Trust**: Since full reproducibility or explainability are not always possible (or desirable), the final credibility sources should be inspectable. This allows end users to build (or remove) trust with those sources.
6. **Decentralizable**: Systems that aim to aid in online misinformation reduction should be designed with web-scale in mind. Similarly, the subjectivity and trust requirements means that a single source of credible information is not possible and should be avoided. All of this is best achieved if such systems are designed to be decentralizable.

All these requirements mean that IR systems to be used for misinformation reduction should be designed to address these needs. In particular, providing a list of document identifiers (and their credibility) as the output of a IR system clearly is insufficient. Recent work by Denaux and Gomez-Perez [1] proposed a conceptual model and vocabulary that can be used, which we summarise next.

3. Credibility Reviews

Linked Credibility Reviews (LCR) [1], is a linked data model for composable and explainable misinformation detection. Its key insight is that calculations of credibility are ultimately subjective

and have to be modeled accordingly. This subjectivity is achieved by modeling steps in the information retrieval and credibility assessment as "Reviews". The approach can be described at a conceptual level and can be implemented by extending the Schema.org vocabulary [2][1].

3.1. Conceptual Model

A *Credibility Review* (CR) is an extension of the generic `Review` data model defined in Schema.org. A Review R can be conceptualised as a tuple (d, r, p) where R:

- reviews a *data item d*, via property `itemReviewed`, this can be any data node (e.g. an article, claim or social media post).

- assigns a numeric or textual *rating r* to (some, often implicit, `reviewAspect` of) d, via property `reviewRating`

- *optionally provides provenance information p*, e.g. via properties `author` and `isBasedOn`.

A Credibility Review (CR) is a subtype of Review, defined as a tuple $\langle d, r, c, p \rangle$, where the CR:

- r must have `reviewAspect credibility` and is recommended to be expressed as a numeric value in range $[-1, 1]$ and is qualified with a *rating confidence c* (in range $[0, 1]$).

- the provenance p is mandatory and must include information about:
  - *credibility signals* (CS) used to derive the credibility rating, which can be either (i) Reviews for data items relevant to d or (ii) *ground credibility signals* (GCS) resources (which are not CRs) in databases curated by a trusted person or organization.
  - the *author* of the review. The author can be a person, organizations or bot. Bots are automated agents that produce CRs.

3.2. Vocabulary to share Credibility Reviews

While `schema.org` provides most of the vocabulary needed to describe CRs, Denaux and Gomez-Perez had to extend it slightly to be able to express the proposed model. The relevant fragment of schema.org and the proposed extensions are depicted in figure 1, focused on CRs for textual web content (although notice that the model is also applicable to other modalities).

Besides the conceptual model and vocabulary, in [1], the authors also propose generic strategies for retrieving relevant credibility signals and propagating both the credibility and confidence values.

4. Acred: example implementation

Acred is an example implementation for the Linked Credibility Review architecture, also presented in [1]. It uses a database of 85K sentences with estimated credibilities. For 45K sentences a known `ClaimReviews` provides a credibility score with fairly high confidence; this

---

[1]See also https://schema.org/
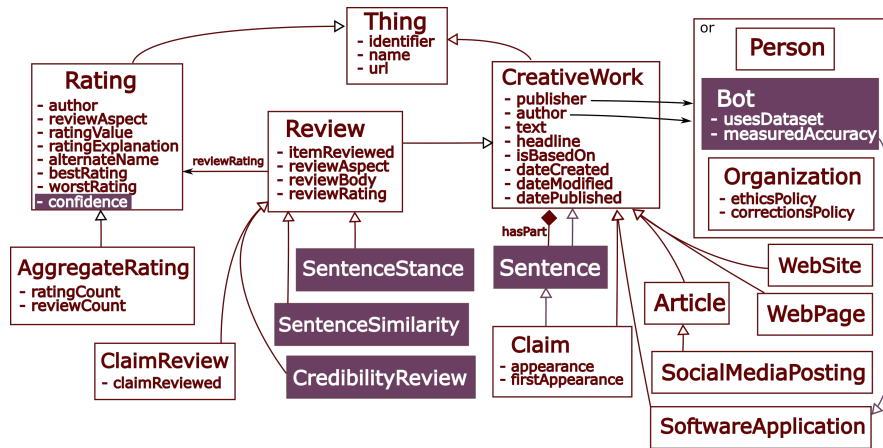
Figure 1: Credibility Review data model, extending schema.org.

was mainly provided by ClaimsKG [3], with trusted credibility labels provided by fact-checkers. The remaining 40K sentences were extracted from a variety of news sites and their credibility was estimated based on the trustworthiness of the publisher as determined by MisinfoMe [4, 5] and ultimately site reputation organizations like NewsGuard and Web Of Trust[2].

For information retrieval, acred uses a sentence encoder trained on STS-B [6] (similar to Sentence BERT [7]) and FAISS [8] to index the 85K sentences as well as a Solr instance to store further metadata about associated `ClaimReviews` or sites where the sentence was published. A second model is used to perform stance detection, which provides polarity which is necessary for correctly propagating credibility values. Heuristic rules are used to combine the credibility confidence and the similarity between the query and database sentences. Given an input document, a third model is used to select query sentences which are factual and check-worthy [9] and further heuristics are used to select the best combination of highest confidence and lowest credibility. See the original paper [1] and corresponding GitHub repository[3] for the full details.

Acred was evaluated on the Clef'18 CheckThat! Factuality Task [10] task and achieved state of the art results with a 0.6475 MAE (compared to the previous score of 0.7050 [11]). Similarly, acred obtained an accuracy of 0.716 on the PolitiFact fragment of FakeNewsNet, which is also state of the art when not using social signals (shares, replies and comments on Twitter to the original article). Both of these results are remarkable when considering that acred does not require a training step since all its steps rely on pre-trained models. Note as well that the representation of credibility score and confidence can be mapped to different labeling schemes: FakeNewsNet only has labels "fake" or "real", while Clef'18 has labels "true", "half-true" and "false" (a third dataset presented in [1] even has 6 different labels including "not verifiable" and "uncertain").

---

[2]https://www.newsguardtech.com/, https://www.mywot.com/. See https://misinfo.me/misinfo/credibility/origins for a full list of such sources.
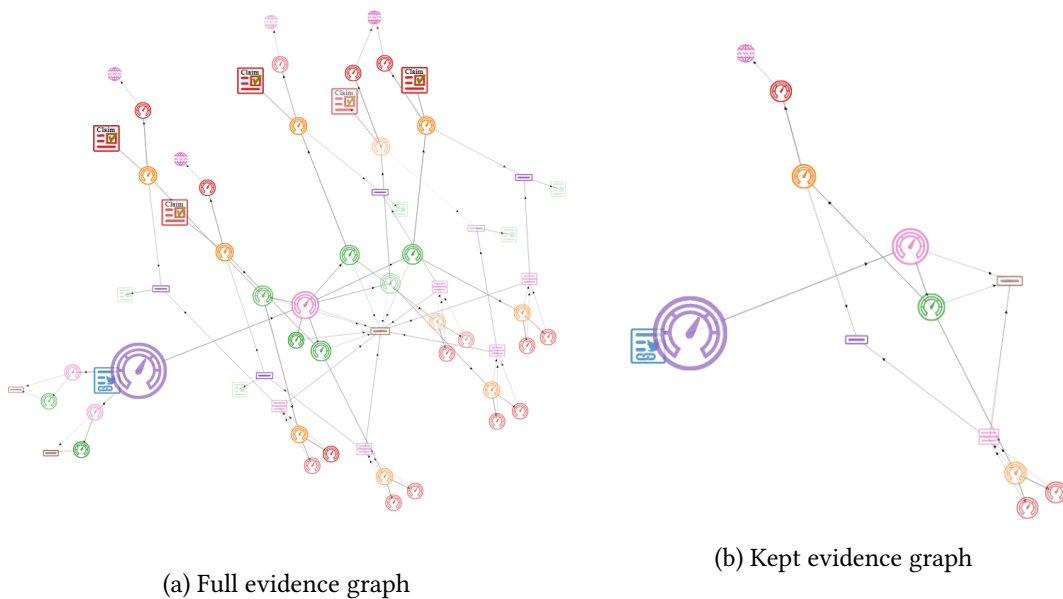
[3]https://github.com/rdenaux/acred

(a) Tweet with label and feedback buttons

(b) Credibility Review with explanation

Figure 2: Example UIs for a (dis)agreement task for a tweet. The user can provide feedback about correct or incorrect labels predicted by acred.

(a) Full evidence graph

(b) Kept evidence graph

Figure 3: Evidence graph for the credibility review and tweet shown in Fig. 1. The big "meter" icon represents the main credibility review, next to the icon for the tweet. All the other nodes form the **evidence** gathered by acred and used to determine the credibility of the tweet.

The output of acred is Credibility Review which can be simplified as a single label (see fig. 2a). However, the provenance and authorship relations can be exploited to render a full (or partial) "evidence graph" (see fig. 3) including intermediate steps (facilitating verifiability and reproducibility) or to generate explanations (see fig. 2b).

5. Conclusion

In this paper, we summarized recent work on a conceptual model and vocabulary for describing and sharing retrieved documents, their credibility and its implication for the credibility of potentially misinforming documents online. We believe that the IR community working on retrieval of content for reducing online misinformation should be aware of this work and can benefit from adopting this conceptual model. As stated in section 2, these IR systems need to deal with a variety of requirements and such conceptual models and vocabularies are a way to ensure these requirements are met. The proposed conceptual model makes it easier to propagate and document results and steps in IR systems, thus facilitating explainability and reproducibility. It also makes explicit where the boundaries are between the subjective credibility and similarity analyses are and the trust in ground credibility sources like fact-checkers and organizations like NewsGuard. The fact that the sample implementation relies on third-party services like ClaimsKG, MisinfoMe, various fact-checkers and site reviewing organizations further shows that the conceptual model facilitates decentralization and handling of subjective reviews. Note finally that acred only uses a small set of "ground credibility signals" (ClaimReviews and website reviews), there are several such sources of signals as identified by the W3C Credibilty Signals working group.[4]. Some of the open issues with acred are discussed in [1, 12].

Although the conceptual model in [1] only provides a model for expressing credibility, a similar approach could be taken for Harmfulness Reviews, which is an orthogonal aspect to credibility (and arguably even more subjective). Finally, it should be mentioned that the semantic web community is also proposing similar vocabularies [13][5]; however these impose certain distinctions which may not be relevant for the IR community.

Acknowledgments

References

[1] R. Denaux, J. M. Gomez-Perez, Linked credibility reviews for explainable misinformation detection, in: J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), The Semantic Web – ISWC 2020, Springer International Publishing, Cham, 2020, pp. 147–163.

[2] R. V. Guha, D. Brickley, S. Macbeth, Schema. org: evolution of structured data on the web, Communications of the ACM 59 (2016) 44–51.

[3] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, K. Todorov, M. Zloch, ClaimsKG: A Knowledge Graph of Fact-Checked Claims. International Semantic Web Conference (2019) 309–324.

---

[4]https://credweb.org/signals-beta/
[5]And more recently the Open Claims model (under review)

[4] M. Mensio, H. Alani, MisinfoMe: Who's Interacting with Misinformation?, in: 18th International Semantic Web Conference: Posters & Demonstrations, 2019.

[5] M. Mensio, H. Alani, News Source Credibility in the Eyes of Different Assessors, in: Conference for Truth and Trust Online, In Press, 2019.

[6] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, in: Proc. of the 10th International Workshop on Semantic Evaluation, 2018, pp. 1–14. `arXiv:1708.00055`.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3973–3983.

[8] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, arXiv preprint arXiv:1702.08734 (2017).

[9] K. Meng, D. Jimenez, F. Arslan, J. D. Devasier, D. Obembe, C. Li, Gradient-Based Adversarial Training on Transformer Networks for Detecting Check-Worthy Factual Claims (2020). `arXiv:2002.07725`.

[10] P. Nakov, A. Barrón-Cedeño, R. Suwaileh, L. M. . Arquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Da, S. Martino, Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2018, pp. 372—-387.

[11] D. Wang, J. G. Simonsen, B. Larsen, C. Lioma, The Copenhagen Team Participation in the Factuality Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 Fact Checking Lab., CLEF (Working Notes) 2125 (2018).

[12] R. Denaux, F. Merenda, J. Manuel, Towards crowdsourcing tasks for accurate misinformation detection, in: Advances in Semantics and Linked Data: Joint Workshop Proceedings from ISWC 2020, SEMIFORM: Semantics for Online Misinformation Detection, Monitoring, and Prediction, volume 2722, CEUR-WS, 2020. URL: http://ceur-ws.org/Vol-2722/semiform2020-paper-2.pdf.

[13] K. Boland, P. Fafalios, A. Tchechmedjiev, Modeling and Contextualizing Claims, in: 2nd International Workshop on Contextualised Knowledge Graphs, 2019.