December 2023

Automatic Validation of User-contributed Content Using Learned Similarity Function

Hari Bhaskar S

Nikhil Shirish Ketkar

Automatic Validation of User-contributed Content Using Learned Similarity Function

ABSTRACT

The quality of user-contributed content at online platforms can vary. Manual validation of such content for quality and relevance is difficult and not scalable. No current techniques utilize known high quality content to rank user-contributed content. This disclosure describes techniques for scalable automatic validation of user-contributed content (UCC) provided to an online platform. A similarity function is learned based on merchant media and known high-quality user submitted media. The similarity function is used to score new UCC media and determine whether the new UCC media is relevant and of sufficient quality to include on the platform. The techniques improve the experience of using online platforms by ensuring that UCC is relevant to the entity with reference to which it is contributed and is of good quality. Further, the techniques can generate feedback on improving UCC contributions.

KEYWORDS

- User-contributed content (UCC)
- User-generated content (UGC)
- Content validation
- Content scoring
- Similarity function
- Merchant media
- Digital map
- Review website
- Media similarity

BACKGROUND

Many online platforms such as review websites, digital maps, e-commerce platforms, etc. enable users to submit media contributions, e.g., photographs, audio, video, etc. These are hosted by the platform and provided to other users. For example, users may contribute photographs of a hotel room or a restaurant to a digital map, a review website, or other online service.

Validation of user-contributed content (UCC)/user-generated content (UGC) as related to a specific place, location, or business presents a challenge. The quality of user-contributed content may sometimes be poor, e.g., photographs or videos taken in poor lighting, with inappropriate framing/ angle, etc. and/or not pertinent to the intended place. Establishing factual accuracy and ground truth from UCC is difficult in the absence of confirmation from an authoritative source such as the actual business owner or merchant. Also, in current online platforms, merchant contributions (provided by business/ entity owners) and user contributions are evaluated separately.

When the volumes of such content, scaling such validation processes is difficult. Currently, there are no automated techniques to validate UCC against high-quality, verified content. Furthermore, when users upload media that is not relevant for the identified entity, e.g., photographs in a restaurant review that are in fact not from the restaurant, no mechanisms are available to identify such media based on a factual, benchmarked standard for appropriate action to be taken, e.g., remove the media from being associated with the entity.

DESCRIPTION

This disclosure describes techniques for scalable automatic validation of user-generated content (audio, video, or other media) provided to an online platform such as a review website, a digital map, an e-commerce platform, etc. The techniques improve the experience of using online

platforms that include UCC by ensuring that UCC is relevant to the entity with reference to which it is contributed and is of good quality. Further, the techniques can generate feedback on improving UCC contributions. The UCC contribution section of the online platform can provide such feedback to the contributors, enabling contributors to raise submission quality and providing transparency for the content validation process.

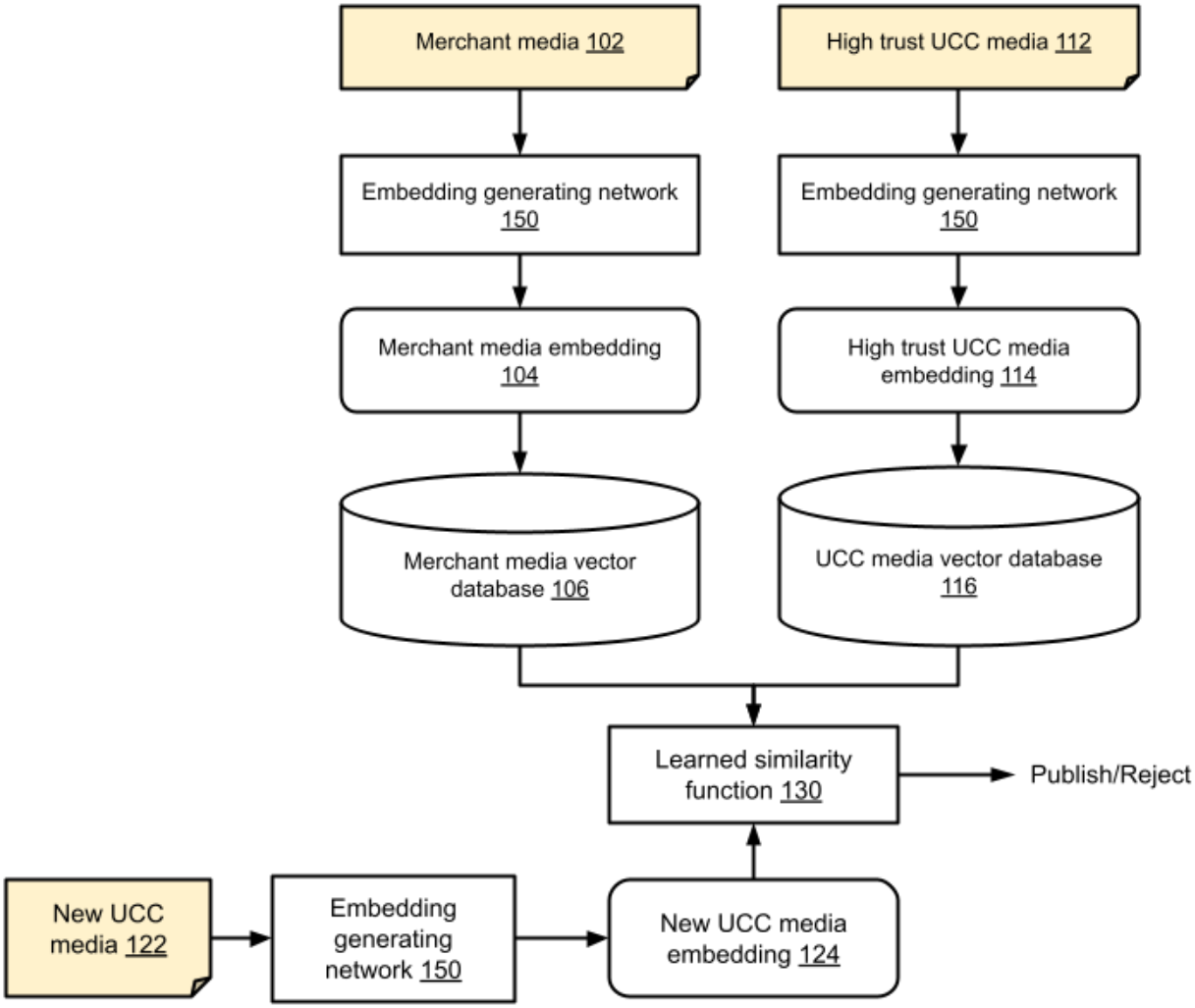

Fig. 1: Learning a similarity function and using it to score UCC media

Fig. 1 illustrates a process to learn a similarity function and use it to score UCC media.

Merchant media (102) and high trust UCC media (112) (UCC media known to be of high quality) are both provided to an embedding generating network (150) that generates corresponding embeddings (vector representations in a multi-dimensional space) - merchant media embedding (104) and high trust UCC media embedding (114). The embeddings represent the attributes associated with high quality media. The embeddings are stored in respective databases - merchant media vector database (106) and UCC media vector database (116).

A similarity function (130) is learned from the embeddings. A large dataset of merchant media and high trust UCC media can be utilized, with appropriate permissions, for this purpose to ensure that the learned similarity function can reliably score arbitrary UCC media that are received. For example, techniques such as K Nearest Neighbors (KNN) or other suitable clustering/scoring techniques can be used to analyze the content similarity in the high quality content.

The learned similarity function can then be utilized to score new UCC media (122) by first using the embedding generating network (150) to generate new UCC media embedding (124). The new UCC media embedding can be scored by the learned similarity function to determine whether to publish the new UCC media at the online platform or to reject the media with appropriate feedback to the contributing user. The feedback may include reasons for rejection and/or suggestions to improve the UCC media. An opportunity to correct/resubmit the UCC media may be provided to the contributing user.

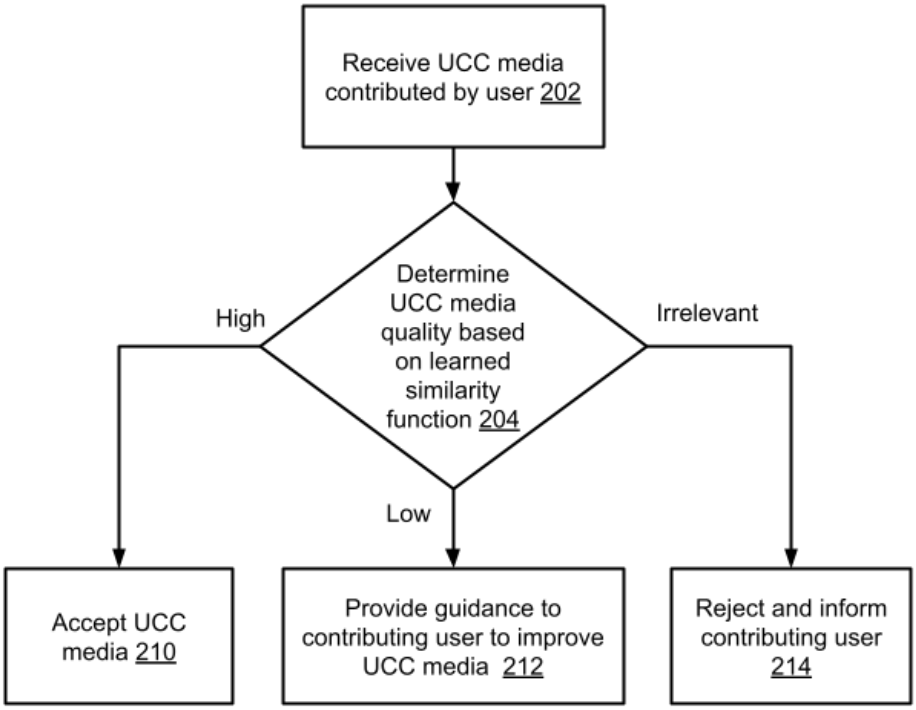

Fig. 2: Content validation workflow for UCC media contributions

Fig. 2 illustrates a content validation workflow for UCC media contributions. UCC media contributed by a contributing user are received (202). The UCC media are evaluated (204), e.g. using the learned similarity function, to determine the quality of the UCC media. If the UCC media is determined to be of high quality, the media is accepted (210) for publication at the online platform. If the UCC media is of low quality, guidance is provided (212) to the contributing user to improve the UCC media. If the UCC media is irrelevant to the entity or factually incorrect, it is rejected (214) and the contributing user is informed of the reason for rejection.

Example of use

Consider a hotel review website. A contributing user provides photographs of a hotel room. If the photograph does not match the hotel's rooms (e.g., as known from merchant media),

the photographs are rejected. If the photograph matches the hotel's rooms but is determined to be of low quality, feedback is provided to the user, e.g., "not accepted due to poor lighting, upload a different photograph." If the photograph is determined to be of high quality, the hotel review website can incorporate the UCC content, e.g., as part of a review contributed by the user.

The described techniques can enhance any online platform that hosts UCC media. The techniques can reduce or eliminate low quality UCC and improve the user experience. The techniques also improve the user experience of contributing UCC content by providing feedback and guidance regarding rejected content and improve transparency, thus providing user education that can improve the quality of UCC. The techniques can also provide merchants (or entities) with insight about the content uploaded regarding their business and guide their promotional strategies.

CONCLUSION

This disclosure describes techniques for scalable automatic validation of user-contributed content (UCC) provided to an online platform. A similarity function is learned based on merchant media and known high-quality user submitted media. The similarity function is used to score new UCC media and determine whether the new UCC media is relevant and of sufficient quality to include on the platform. The techniques improve the experience of using online platforms by ensuring that UCC is relevant to the entity with reference to which it is contributed and is of good quality. Further, the techniques can generate feedback on improving UCC contributions.

REFERENCES

1. “Image embeddings. Image similarity and building… | by Romain Beaumont | Medium” available online at https://rom1504.medium.com/image-embeddings-ed1b194d113e, accessed November 21, 2023.