Article

Not peer-reviewed version

When AI Meets Topology—Why Collaborative Filtering and Contextual Word Embeddings Fail

Hao Wang *

Posted Date: 14 October 2025

doi: 10.20944/preprints202510.1025.v1

Keywords: collaborative filtering; vector field; Poincare-Hopf Index Theorem; contextual word embedding; recommender system

 doi:10.20944/preprints202510.1025.v1

Article

When AI Meets Topology—Why Collaborative Filtering and Contextual Word Embeddings Fail

Hao Wang

CEO & Founder, Ratidar Technologies LLC; Haow85@live.com

Abstract

Collaborative Filtering algorithm requires computation of similarities. By exploring the similarity structure using vector field-based topology, we prove that the algorithm is invalid. Similarly, we can apply an analogous approach to the Contextual Word Embedding algorithms. By reducing the similarity structures into transverse fields which can be decomposed into tangential vectors and normal vectors of the underlying topological structures, we utilize the Poincare-Hopf Index Theorem to prove that the shape of contextual word embedding vectors is not a manifold.

Keywords: collaborative filtering; vector field; Poincare-Hopf Index Theorem; contextual word embedding; recommender system

1. Introduction

The study of the similarity structures of data mining and NLP algorithms is relatively new. However, the impact of such research is tremendous - not only could we analyze the efficiency of algorithms better [1], but also we were able to justify or overthrow existing milestone algorithms [2].

The first published recommender system is collaborative filtering [3]. Although there have been recent efforts [2,4] which try to overthrow the algorithm, such efforts are essentially theoretically flawed themselves. Overthrowing milestone algorithms turns out to be as difficult as establishing them. Most researchers and engineers seem not to be concerned with the theoretical foundation of collaborative filtering.

The basic idea of collaborative filtering is to recommend new items to customers based on other people whose preference is similar to them. This means we need to compute the similarity structures of the input data. As we claimed at the beginning of this paper, good understanding of this similarity structures paves way for theoretical proof or disproof of the algorithm, which will be detailed in the following sections of this paper.

Since 2012, when Word2vec was introduced, contextual word embedding algorithms have become one of the research focuses of the NLP community. The major breakthrough of contextual word embeddings compared with prehistoric word representations such as arrays of TF-IDF's is that for the first time in human history, words can be represented as dense float vectors which accept arithmetic operations.

In spite of wide application of contextual word embeddings, some researchers have started to work on disproof of such algorithms [5,6]. While some research [5] is empirical, others [6] are theoretic (although wrong because of a mistake similar to [4]).

In this paper, we are inspired by other researchers' work [4,6] to disprove collaborative filtering and examine the shape of the contextual word embeddings. Unlike [4,6], which are theoretically flawed due to similar reasons, our proof is theoretically solid.

2. Related Work

Collaborative filtering algorithms [3,7] are the first recommender systems. Although they are out-of-dated nowadays, they serve as the first algorithms to be studied by researchers and engineers

 doi:10.20944/preprints202510.1025.v1

in the field. In recent years, deep learning-based technologies such as DeepFM [8] and Wide & Deep [9] have become the de facto standard recommendation engine approach in the industry.

Contextual word embedding is the corner stone of the natural language processing field. Famous approaches such as Word2vec [10], GloVe [11] and BERT [12] emerged as the milestone in the field one after another. One can safely claim that without contextual word embeddings, natural language processing wouldn't be a successful area in today's IT industry.

Very few researchers have examined the theoretical foundation of collaborative filtering or contextual word embeddings. One notable exception is Wang [2,4,6], who studied the topic using vector field based topology. However, his analysis is theoretically flawed themselves.

3. Domain of User Preference Vectors of Collaborative Filtering

To examine the validity of collaborative filtering, we pick the User-based Collaborative Filtering algorithm and study the 2D projection of the user preference vectors. We define 1-sim(i,j) as the distance between user i and user j whose similarity value between their preference vectors is sim(i,j).

Lemma 1. *The 2D Projection of the domain of user preference vectors is a 2D circle of radius 0.5.*

Proof: By using collaborative filtering algorithm we assume that the domain of similarity values between 2 user preference vectors are [0,1]. Since the distance between user i and j is 1-sim(i,j), the domain of the distances is [0,1]. The largest distance between 2 points is 1, so the users' preferences are distributed within a circle of radius 0.5. The 2D vector projection point can not lie inside the circle, because otherwise the point whose distance is 1.0 away from the point would fall out of the circle. Therefore we have proved that all the user preferences fall on the circle - not inside, not outside of the circle, just on the circle. Since the domain of the distances is the full interval [0,1], there is a 1-to-1 correspondence between the user preference vectors and the 2D circle of radius 0.5.

4. Construction of Vector Field

For an arbitrary user j, we construct a vector field on the 2D projection circle of the user preference vectors: We pick another vector k and compute sim(j, k) and get the value sim(j,k)=C. We then assign each point i on the circle the following vector (sim(i,k)-C, sim(i,k)-C). The vector field is denoted as V.

Notice that we can decompose a vector in the vector field V into a linear combination of the vector tangential to the 2D circle and the normal vector perpendicular to the tangential vector. Since the vectors in V is parallel to the line y=x, the tangential vector is 0 is equivalent to the original vector in V equals 0 except at 2 points of the circle where the tangential vector is perpendicular to the line y=x. We only consider the majority of situations when these 2 points are not zero points. We use the phrase most of the time to represent the situations when these 2 points are not zero points. Therefore, to count the number of zero points in V is equivalent to count the number of zero's in the decomposed tangential vector field V'.

5. Disproof of Collaborative Filtering

In the vector field V', since the domain of points is a manifold, namely a 2D circle. By the Poincare-Hopf Index Theorem, we have the index of its tangential vector field is a constant number, i.e., the Euler Characteristic. We can fixate each zero point of the tangential vector field to a specific type of singular point, so as to make the index of the vector field equivalent to the number of zero points. Since the zero points in the tangential vector field are equivalent to zero points (when sim(i,k) = C = sim(j,k))in V, we can safely draw the conclusion: The number of similarity scores of a specific value C is the same across all real values C in V **most of the time**. This implies the following lemma:

 doi:10.20944/preprints202510.1025.v1

Lemma 2.*The number of similarity scores in the computation of user similarity in user based collaborative filtering that equal a specific value C is a constant for all real value C most of the time.*

Lemma 2 disproves the user-based collaborative filtering algorithm. In real world application, we often notice that the number of similarity scores equal to a specific value C in the computation of user similarity exhibit power-law effect for different C's. This is an empirical evidence that disproves the validity of collaborative filtering algorithm if Lemma 2 holds.

6. Contextual Word Embeddings

Analogous to the way we examine the user-based collaborative filtering algorithm, we first reduce all the word embeddings in the language corpus into 2D space. Then we choose an arbitrary word j, and another word k, and set their similarity to be C = sim(j,k). Then at each 2D projection point i, we assign a vector (sim(i,j)-C, sim(i,j)-C). Then by analysis analogous to Section 4 and 5, we have :

Lemma 3. *If the domain of word embeddings is a manifold, then the number of similarity scores in the computation of vector similarity for word embeddings that equal a specific value C is a constant for all real value C most of the time.*

This would mean that the number of vector pairs whose similarity is 0.5 is the same as the number of vector pairs whose similarity is 0.1, and so on. This is simply untrue for human language. Therefore, the domain of the contextual word embeddings is not a manifold. There might be holes in the domain, so we can not safely apply arithmetic operations such as addition or subtraction arbitrarily for word embeddings.

7. Conclusion

Collaborative filtering and contextual word embedding algorithms are well established technologies that have deep impact on the world. However, few researchers have questioned the validity of the theoretical foundation of them. In this paper, we use vector-field based topological approaches to disprove the collaborative filtering algorithms and claim that the domain of word embeddings is not a mainfold.

In future work, we would like to apply topological analysis to other algorithms to examine the validity of them. We would also like to study the shape of the domain of word embedding vectors.

References

1. H. Wang, Z. Wang and W. Zhang, Quantitative analysis of Matthew effect and sparsity problem of recommender systems, ICCCBDA 2018
2. H. Wang, Collaborative Filtering Is a Lie or Not? It Depends on the Shape of Your Data Domain, CECNet 2023
3. D. Goldberg, D. Nichols, et.al., Using collaborative filtering to weave an information tapestry. Commun. ACM 35, 12 (Dec. 1992), 61–70
4. H. Wang, Collaborative Filtering is Wrong and Here is Why, CCCE 2024
5. A. Jakubowski, M. Gasic and M. Zibrowius, Topology of Word Embeddings: Singularities Reflect Polysemy, ACL 2020
6. H. Wang, Human Language is Non-manifold, ICNLP 2024
7. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, Item-based collaborative filtering recommendation algorithms, WWW, 2001
8. H. Guo, et al., DeepFM: a factorization-machine based neural network for CTR prediction, arXiv preprint arXiv:1703.04247 . 2017

 doi:10.20944/preprints202510.1025.v1

9. H. Cheng, et al., Wide & deep learning for recommender systems, Proceedings of the 1st workshop on deep learning for recommender systems 2016
10. T. Mikolov, K. Chen, G. Corrado, et al. Efficient Estimation of Word Representations in Vector Space, International conference on learning representations 2013
11. J. Pennington, R. Socher, C.D.Manning, et al. Glove: Global Vectors for Word Representation, Empirical methods in natural language processing 2014
12. J.Devlin, M.Chang, et.al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019