Artificial Intelligence 230 (2016) 224–245

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Semantic sensitive tensor factorization

Makoto Nakatsuji [a,*], Hiroyuki Toda [b], Hiroshi Sawada [b], Jin Guang Zheng [c], James A. Hendler [c]

[a] Smart Navigation Business Division, NTT Resonant Inc., Japan
[b] NTT Service Evolution Laboratories, NTT Corporation, Japan
[c] Tetherless World Constellation, Rensselaer Polytechnic Institute, USA

ARTICLE INFO

Article history:
Received 18 August 2014
Received in revised form 28 August 2015
Accepted 1 September 2015
Available online 14 September 2015

Keywords:
Tensor factorization
Linked open data
Recommendation systems

ABSTRACT

The ability to predict the activities of users is an important one for recommender systems and analyses of social media. User activities can be represented in terms of relationships involving three or more things (e.g. when a user tags items on a webpage or tweets about a location he or she visited). Such relationships can be represented as a tensor, and tensor factorization is becoming an increasingly important means for predicting users' possible activities. However, the prediction accuracy of factorization is poor for ambiguous and/or sparsely observed objects. Our solution, Semantic Sensitive Tensor Factorization (SSTF), incorporates the semantics expressed by an object vocabulary or taxonomy into the tensor factorization. SSTF first links objects to classes in the vocabulary (taxonomy) and resolves the ambiguities of objects that may have several meanings. Next, it lifts sparsely observed objects to their classes to create augmented tensors. Then, it factorizes the original tensor and augmented tensors simultaneously. Since it shares semantic knowledge during the factorization, it can resolve the sparsity problem. Furthermore, as a result of the natural use of semantic information in tensor factorization, SSTF can combine heterogeneous and unbalanced datasets from different Linked Open Data sources. We implemented SSTF in the Bayesian probabilistic tensor factorization framework. Experiments on publicly available large-scale datasets using vocabularies from linked open data and a taxonomy from WordNet show that SSTF has up to 12% higher accuracy in comparison with state-of-the-art tensor factorization methods.

1. Introduction

The ability to analyze relationships involving three or more objects is critical for accurately predicting human activities. An example of a typical relationship involving three or more objects in a content providing service is one between a user, an item on a webpage, and a user-assigned tag of that item. Another example is the relationship between a user, his or her tweet on Twitter, and the locations at which he or she tweeted. The ability to predict such relationships can be used to improve recommendation systems and social network analysis. For example, suppose that a user assigns a thriller movie the tag "romance" and another user tags it with "car action". Here, methods that only handle bi-relational objects ignore tags and find these users to be similar because they mentioned the same item. In contrast, multi-relational methods conclude

* Corresponding author.
E-mail address: nakatsuji.makoto@gmail.com (M. Nakatsuji).

http://dx.doi.org/10.1016/j.artint.2015.09.001

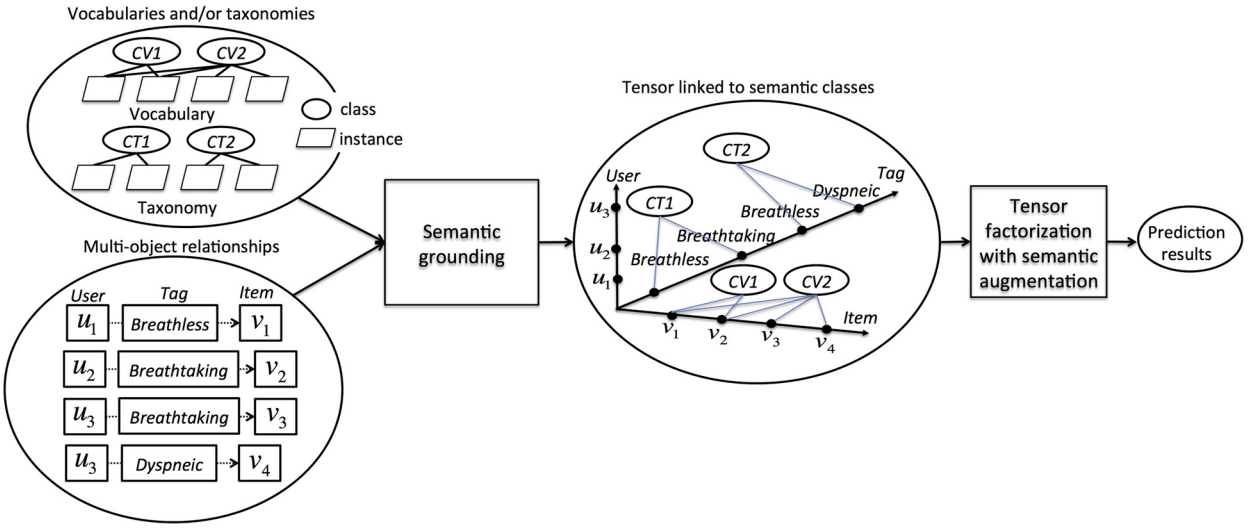

Fig. 1. Flowchart of SSTF.

that such users are slightly different because they have different opinions about the item. The quality of recommendations would be higher if they reflected these small differences [1].

Tensors are useful ways of representing relationships involving three or more objects, and tensor factorization is seen as a means of predicting possible future relationships [2]. Bayesian probabilistic tensor factorization (BPTF) [3] is especially promising because of its efficient sampling of large-scale datasets and simple parameter settings. However, this and other tensor factorization schemes have had poor accuracy because they fail to utilize the semantics underlying the objects and have trouble handling ambiguous and/or sparsely observed objects.

Semantic ambiguity is a fundamental problem in text clustering. Several studies have used Wikipedia or WordNet taxonomies [4] to resolve semantic ambiguities and improve the performance of text clustering [5] and to compute the semantic relatedness of documents [6]. We show in this paper that taxonomy-based disambiguation can improve the prediction accuracy of tensor factorization.

The *sparsity problem* affects accuracy if the dataset used for learning latent features in tensor factorizations is not sufficient [7]. In an attempt to improve prediction accuracy, generalized coupled tensor factorization (GCTF) [8,9] uses social relationships among users as auxiliary information in addition to user–item–tag relationships during the tensor factorization. Recently, GCTF was used for link prediction [10], and it was found to be the most accurate of the current tensor factorization methods that use auxiliary information [7,11,12]. However, so far, no tensor methods have used semantics expressed as taxonomies to resolve ambiguity/sparsity problems even though taxonomies are present in real applications as a result of the spread of the Linked Open Data (LOD) vision [13] and the growing knowledge graphs used in search.[1]

In this paper, we propose semantic sensitive tensor factorization (SSTF), which uses semantics expressed by vocabularies and taxonomies to overcome the above problems in the BPTF framework. Vocabularies and taxonomies, sometimes called "simple ontologies" [14], are collections of human-defined classes with a hierarchical structure and classification instances (i.e., items or words). We will disambiguate the objects (items or tags) by using "vocabularies" for graph structures and "taxonomies" for tree structures. Fig. 1 overviews SSTF. The factorization has two components, *semantic grounding* and *tensor factorization with semantic augmentation*, which respectively resolve the semantic ambiguity and sparsity problems.

Semantic grounding resolves semantic ambiguities by linking objects to vocabularies or taxonomies. It first measures the similarity between objects and instances in the vocabularies (taxonomies). It then links each object to vocabulary (taxonomy) classes via the instance that is most similar to the object. Consequently, it can construct a tensor whose objects are linked to classes. For example, in Fig. 1, the tag "Breathless" can be linked to the class *CT*1, which includes word instances such as "Breathtaking" and "Breathless" or *CT*2, which includes the word instances such as "Dyspneic" and "Breathless". *CT*1 and *CT*2 have different meanings, and a tensor factorization trained on observed relationships that include such ambiguous tags can degrade prediction accuracy. SSTF extracts the properties that are assigned to item entry v_1 in LOD sources and those assigned to instances in WordNet. Then, it links the tag "Breathless" to *CT*1 if the properties of "Breathless" in *CT*1 are more similar to the properties of v_1 than to those of "Breathless" in *CT*2. We describe this process in detail in Section 4.2.

Tensor factorization with semantic augmentation solves the sparsity problem by incorporating semantic biases based on vocabulary (taxonomy) classes into tensor factorization. The key point of this idea is that it lets sparse objects share semantic knowledge with regard to their classes during the factorization process. It lifts sparse objects to their classes to create augmented tensors. To do this, it determines multiple sets of sparse objects according to the degree of sparsity to create

[1] http://www.google.com/insidesearch/features/search/knowledge.html.