Knowledge-Based Systems 000 (2017) 1–9

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach

Shuo Yang[a,*], Mohammed Korayem[b], Khalifeh AlJadda[b], Trey Grainger[b], Sriraam Natarajan[a]

[a] Indiana University, Bloomington, IN, USA
[b] CareerBuilder, Norcross, GA, USA

ARTICLE INFO

*Article history:*
Received 15 November 2016
Revised 22 August 2017
Accepted 26 August 2017
Available online xxx

*Keywords:*
Recommendation system
Content-based filtering
Collaborative filtering
Statistical Relational Learning
Cost-sensitive learning

ABSTRACT

Recommendation systems usually involve exploiting the relations among known features and content that describe items (content-based filtering) or the overlap of similar users who interacted with or rated the target item (collaborative filtering). To combine these two filtering approaches, current model-based hybrid recommendation systems typically require extensive feature engineering to construct a user profile. Statistical Relational Learning (SRL) provides a straightforward way to combine the two approaches through its ability to directly represent the probabilistic dependencies among the attributes of related objects. However, due to the large scale of the data used in real world recommendation systems, little research exists on applying SRL models to hybrid recommendation systems, and essentially none of that research has been applied to real big-data-scale systems. In this paper, we proposed a way to adapt the state-of-the-art in SRL approaches to construct a real hybrid job recommendation system. Furthermore, in order to satisfy a common requirement in recommendation systems (i.e. that false positives are more undesirable and therefore should be penalized more harshly than false negatives), our approach can also allow tuning the trade-off between the precision and recall of the system in a principled way. Our experimental results demonstrate the efficiency of our proposed approach as well as its improved performance on recommendation precision.



## 1. Introduction

With their rise in prominence, recommendation systems have greatly alleviated information overload for their users by providing personalized suggestions for countless products such as music, movies, books, housing, jobs, etc. Since the mid-1990s, not only new theories of recommender systems have been presented but their application softwares have also been developed which involves various domains including e-government, e-business, e-commerce/e-shopping, e-learning, etc [1]. We consider a specific recommender system domain, that of job recommendations, and propose a novel method for this domain using statistical relational learning. This domain easily scales to billions of items including user resumes and job postings, as well as even more data in the form of user interactions between these items. CareerBuilder, the source of the data for our experiments, operates one of the largest job boards in the world. It has millions of job postings, more than 60 million actively-searchable resumes, over one billion searchable documents, and receives several million searches per hour [2]. The scale of the data is not the only interesting aspect of this domain, however. The job recommendations use case is inherently relational in nature, readily allowing for graph mining and relational learning algorithms to be employed. As Fig. 1 shows, very similar kinds of relationships exist among the jobs that are applied to by the same user and among the users who share similar preferences. If we treat every single job post or user as an object which has various attributes, the probability of a match between the target user and a job does not only depend on the attributes of these two target objects (i.e. target user and target job) but also the attributes of the related objects such as the patterns of the user's previous applied jobs, behaviors of users living in the same city or having the same education level. As we show in this work, richer modeling techniques can be used to determine these relationships faithfully. However, since most of the statistical relational learning approaches involve a searching space exponential to the number of related objects, how to efficiently build a hybrid recommenda-

* Corresponding author.
*E-mail address:* shuoyang@indiana.edu (S. Yang).

http://dx.doi.org/10.1016/j.knosys.2017.08.017
0950-7051/

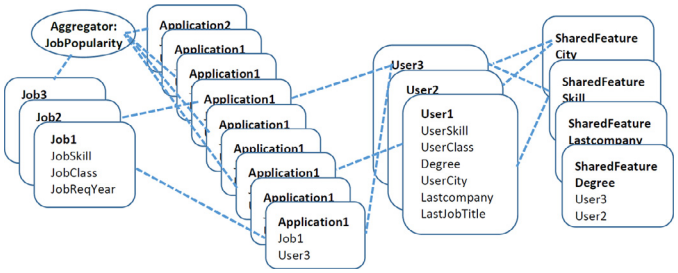


**Fig. 1.** Job recommendation domain.

tion system with statistical relational learning in such a large scale real-world problem remains a challenge in this field.

One of the most popular recommender approaches is *content-based filtering* [3], which exploits the relations between (historically) applied-to jobs and similar features among new job opportunities for consideration (with features usually derived from textual information). An alternative recommendation approach is based on *collaborative filtering* [4], which makes use of the fact that users who are interested in the same item generally also have similar preferences for additional items. Clearly, using both types of information together can potentially yield a more powerful recommendation system, which is why model-based hybrid recommender systems were developed [5]. While successful, these systems typically need extensive feature engineering to make the combination practical.

The hypothesis which we sought to verify empirically was that recent advancements in the fields of machine learning and artificial intelligence could lead to powerful and deployable recommender systems. In particular, we assessed leveraging Statistical Relational Learning (SRL) [6], which combines the representation abilities of rich formalisms such as first-order logic or relational logic with the ability of probability theory to model uncertainty. We employed a state-of-the-art SRL formalism for combining content-based filtering and collaborative filtering. SRL can directly represent the probabilistic dependencies among the attributes from different objects that are related with each other through certain connections (in our domain, for example, the jobs applied to by the same user or the users who share the same skill or employer). SRL models remove the necessity for an extensive feature engineering process, and they do not require learning separate recommendation models for each individual item or user cluster, a requirement for many standard model-based recommendation systems [4,7].

We propose a hybrid model combining content-based filtering and collaborative filtering that is learned by an efficient statistical relational learning approach - Relational Functional Gradient Boosting (RFGB) [8]. Specifically, we define the target relation as *Match*(*User, Job*) which indicates that the user–job pair is a match when the grounded relation is true, hence that job should be recommended to the target user. The task is to predict the probability of this target relation *Match*(*User, Job*)[1] for users based on the information about the job postings, the user profile, the application history, as well as application histories of users that have the similar preferences or profiles as the target user. RFGB is a boosted model which contains multiple relational regression trees with additive regression values at the sink node of each path. Our hypothesis is that these trees can capture many of the weak relations that exist between the target user and the job with which he/she is matched.

In addition, this domain has practical requirements which must be considered. For example, we would rather overlook some of the candidate jobs that could match the users (false negatives) than send out numerous spam emails to the users with inappropriate job recommendations (false positives). The cost matrix thus does not contain uniform cost values, but instead needs to represent a higher cost for the user–job pairs that are false positives compared to those that are false negatives, i.e. precision is preferred over recall. To incorporate such domain knowledge within the cost matrix, we adapted the previous work from [9], which extended RFGB by introducing a penalty term into the objective function of RFGB so that the trade-off between the precision and recall can be tuned during the learning process.

In summary, we considered the problem of matching a user with a job and developed a hybrid content-based filtering and collaborative filtering approach. We adapted a successful SRL algorithm for learning features and weights and are the first to implement such a system in a real-world big data context. Our algorithm is capable of handling different costs for false positives and false negatives making it extremely attractive for deploying within many kinds of recommendation systems, including those within the domain upon which we tested. Our proposed approach has three main innovations: 1. it is the first work which employs probabilistic logic models to build a real-world large-scale job recommendation system; 2. it is the first work which allows the recommender to incorporate special domain requirements of an imbalanced cost matrix into the model learning process; 3. it is the first to prove the effectiveness of statistical relational learning in combining the collaborative filtering and content-based filtering with real-world job recommendation system data.

## 2. Related work

Recommendation systems usually handle the task of estimating the relevancy or ratings of items for certain users based on information about the target user–item pair as well as other related items and users. The recommendation problem is usually formulated as $f$: $U \times I \rightarrow R$ where $U$ is the space of all users, $I$ is the space of all possible items and $f$ is the utility function that projects all combinations of user-item pairs to a set of predicted ratings $R$ which is composed by nonnegative integers. For a certain user $u$, the recommended item would be the item with the optimal utility value, i.e. $u_i^* = argMax_{i \in I} f(u, i)$. The user space $U$ contains the information about all the users, such as their demographic characteristics, while the item space $I$ contains the feature information of all the items, such as the genre of the music, the director of a movie, or the author of a book.

Generally speaking, the goal of *content-based filtering* is to define recommendations based upon feature similarities between the items being considered and items which a user has previously rated as interesting [10], i.e. for the target user-item rating $f(\hat{u}, \hat{i})$, *content-based filtering* would predict the optimal recommendation based on the utility functions of $f(\hat{u}, I_h)$ which is the historical rating information of user $\hat{u}$ on items $(I_h)$ similar with $\hat{i}$. Given their origins out of the fields of information retrieval and information filtering, most content-based filtering systems are applied to items that are rich in textual information. From this textual information, item features $I$ are extracted and represented as *keywords* with respective weighting measures calculated by certain mechanisms such as the *term frequency/inverse document frequency (TF/IDF)* measure [11]. The feature space of the user $U$ is then constructed from the feature spaces of items that were previously rated by that user through various keyword analysis techniques such as averaging approach [12], Bayesian classifier [7], etc. Finally, the utility function of the target user-item pair $f(\hat{u}, \hat{i})$ is calculated by some scoring heuristic such as the cosine similarity [11] between the user profile vector and the item feature vector or some traditional machine learning models [7]. *Overspecialization* is one

[1] Following standard practice inside the machine learning community, we use the terms *relations* and *predicates* interchangeably.

Please cite this article as: S. Yang et al., Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.08.017