LODify: A Hybrid Recommender System based on Linked Open Data

Rouzbeh Meymandpour and Joseph G. Davis

School of Information Technologies
The University of Sydney

{rouzbeh.meymandpour, joseph.davis}@sydney.edu.au

Abstract. We propose LODify, a hybrid recommendation method which measures the semantic similarity of items or resources of interest and combines this with user ratings to make recommendations across diverse domains. The semantic similarity metric draws on information theory and computes the similarity of items based on the information content of their shared characteristics. Detailed semantic analysis of items, considering the special characteristics of Linked Data represented using various kinds of relations, incorporating the relative importance of each relation into the similarity measurement, and successful handling of the item cold-start problem are among the key benefits of the presented approach. We demonstrate how this approach can be successfully applied to provide recommendations and to predict user ratings.

1 LODify and its Innovation

Linking Open Data (LOD) project is one of the successful initiatives of the Web of Data which aims to publish and link public datasets in a wide variety of domains using semantic technologies [1]. Currently, there are more than 295 datasets interlinked in the LOD cloud. Datasets in diverse areas such as media, publications, geography, and life sciences are semantically connected based on LOD standards [1, 2]. However, the increasing size of Linked Open Data and the varying quality levels of the semantic datasets pose new challenges for Semantic Web-based applications.

LODify is a hybrid recommender system that measures the semantic similarity of entities (items or resources) and combines this with user ratings to make accurate recommendations in various application domains. At the heart of LODify is an information theory-based similarity metric, which computes the similarity of items based on the information content (IC) of the shared characteristics among entities as well as their differences. The features of each item are extracted according to their relationships with their neighbors. As Linked Data resources are connected to each other using various types of relations (links or edges), LODify is also able to incorporate the weight of each relation. These weights can be automatically inferred using the proposed semantic ranking method or assigned by domain experts.

In order to reap the benefits of the interlinked semantic knowledge provided by Linked Open Data, LODify retrieves information from heterogeneous data sources on the LOD cloud. This not only facilitates dealing with missing and incorrect information in particular datasets, it also enables the recommender system to make rec-

ommendations in situations when there are a few or even no ratings available for items (known as the item cold-start problem).

Our approach is independent of the structural, taxonomical, and ontological representation of the underlying dataset. As a result, it is capable of making recommendations across a wide range of domains. This improves on previous methods which are generally restricted to specific application areas and require extensive manual effort. Furthermore, LODify allows domain experts to fine-tune the recommender system's parameters to match the specific requirements of the application context.

2 System Architecture and Methodology

LODify consists of three main parts, semantic information retrieval, similarity computation, and recommendation provision. The semantic content related to the items of interest is retrieved from various sources on the Linked Open Data cloud using a sequence of SPARQL queries. Other statistical facts, such as the frequency of relations, required for similarity computation also obtained using SPARQL queries. The core of LODify is a content-based semantic similarity measure, which evaluates the degree of similarity among items. It provides the recommender engine with a list of the most similar items to a given item. Finally, by analyzing the user ratings available for each item in the training set, the recommender engine can predict the missing ratings or provide a list of suggested items depending on the recommendation task.

2.1 Semantic Information Retrieval

In LODify, facts related to items can be obtained from various sources in the LOD cloud. Data sources used for this experiment are DBpedia, Freebase, and YAGO. The required information was retrieved from the latest versions of the datasets using SPARQL queries. The triples were loaded into our local RDF store for further processing. A set of SPARQL queries were also executed in order to extract the features of items and prepare for similarity computation.

2.2 Semantic Similarity Computation

Our proposed semantic similarity measurement approach is based on partitioned information content (PIC) [3], which is a combination of feature-based and information content (IC)-based techniques designed for Linked Data. Through statistics derived from the underlying knowledge base, IC-based approaches rely on the amount of information that entities convey, that is, their informativeness or entropy. Less probable concepts are considered more specific and informative than more common ones, which are more general. In PIC, we applied the concept of information content measurement to Linked Data by formalizing the definition of Linked Data using the notion of features. By defining features of a resource based on its immediate neighbors, and the type and direction of their relationships [3], $LODify_{Basic}$ can be derived as follows:

$$LODify_{Basic}(r_1, r_2) = PIC(F_{r_1} \cap F_{r_2}) = \sum_{\forall f_i \in (F_{r_1} \cap F_{r_2})} -log\left(\frac{\varphi(f_i)}{N}\right) \quad (1)$$

where F_r is the set of features describing an entity, item, or a resource r, the function $\varphi(\cdot)$ computes the frequency (popularity) of a feature, and the denominator N is defined as the frequency of the feature with the maximum number of occurrences in a given Linked Data.

According to the characteristics of PIC, which is based on its probability and information theoretic basis, $\text{LODify}_{\text{Basic}}$ measures the similarity of two resources based on the informativeness of their shared features. For example, it turns out that shared authors contribute more to the similarity score rather than the genres in the books domain due to the fact that probability of two books sharing the same author is lower than that of the books sharing the same genre.

$\text{LODify}_{\text{Basic}}$ similarity metric only considers common features among items to assess their similarity. Therefore, the similarity value increases with more shared features but does not decrease with more differences among the entities. In order to address this problem and to take into account the importance (weight) of various kinds of relations in Linked Data, we propose *LODify*, as follows:

$$LODify(r_1, r_2) = \frac{WPIC\left(F_{r_1} \cap F_{r_2}\right)}{WPIC\left(F_{r_1} \cap F_{r_2}\right) + \alpha\, WPIC\left(F_{r_1} - F_{r_2}\right) + \beta\, WPIC\left(F_{r_1} - F_{r_2}\right)} \quad (2)$$

$$\alpha, \beta > 0$$

where $WPIC(\cdot)$ of a feature is computed by considering weights of each relation [4] and the parameters α and β can be used to adjust the influence of distinctive features of two entities in the similarity score.

2.3 Recommendation Provision

Rating Prediction. The ratings are predicted based on the weighted sum of similarity between the target item and items rated by the target user:

$$\tilde{r}_{ui} = \mu + b_u + b_i + \frac{\sum_{j \in I_u} LODify(i,j)\ (r_{ui} - \mu - b_u - b_j)}{\sum_{j \in I_u} |LODify(i,j)|} \quad (3)$$

where, $\tilde{r}_{ui}$ is the predicted rating of user u on item i, I_u is the set of most similar items to i rated by the user, and r_{ui} is the rating score given to i by the user. The bias terms μ, b_u, and b_i denote the global average of all ratings, the average of ratings given by user u, and the average of ratings given to item i, respectively.

Item Recommendation. After the rating prediction phase, the recommender system ranks all the candidate items based on their predicted rating. Items with the highest scores are selected to be suggested to the user.

3 LODify-based Diversification

In order to improve the diversity of the recommendations, we designed a diversification algorithm to diversify the list of recommendations while preserving the overall accuracy. Because LODify similarity scores are normalized ($LODify(i,j) \in [0,1]$), they can be easily converted to dissimilarity for diversity measurement ($Similarity(i,j) = 1 - Dissimilarity(i,j)$).

For a top-k recommendation task, for each user u, the LODify-based Diversification method (Algorithm 1) starts with measuring the diversity of the first k recommended items ($R(u)_{1\dots k}$). The balance between diversification and the original ranking of items is indicated by the diversification factor, $\mathrm{w}_d \in [0,1]$. A larger diversification factor leads to a higher amount of diversity in the output list of recommendations. The cost of keeping each item in the list ($\forall \mathrm{w}_{i_r} \in \mathrm{W}_{R(u)_{1\dots k}}$) is determined according to its rank in the actual recommendation list and its similarity rank, computed by the $\sigma_{Asc}(\cdot,\cdot)$ function. Based on this value, the candidate item for replacement (c) is the item with the highest similarity with the rest of the items in the list and the lowest score in the recommendations list. It is replaced with an alternative item ($\acute{c}$), which increases the diversity of the list.

In Algorithm 1, the $Diversity(\cdot)$ function refers to the average dissimilarity between all $\frac{2}{n\times(n-1)}$ pairs of items in a given list and $SimilarityDistance(i, I)$ computes the average similarity between an item i and a list of items I. Both of these functions are computed using LODify semantic similarity/dissimilarity measure. Finally, the diversification procedure stops after $iter$ replacements.

Algorithm 1. LODify-Based Diversification

```
procedure Diversify(R(u), k, w_d, iter) {
    d_R(u) ← Diversity(R(u)); w_r ← 1 − w_d;
    do
        D_R(u)1…k : { ∀d_ir ∈ D_R(u)1…k  d_ir ← SimilarityDistance(i_r, R(u)_1…k − {i_r}) };
        W_R(u)1…k : { ∀d_ir ∈ D_R(u)1…k, ∀w_ir ∈ W_R(u)1…k | w_ir
                        ← w_r σ_Asc(i_r, R(u)_1…k) + w_d σ_Asc(d_ir, D_R(u)1…k) };
        w_c ← Max({∀w_ir ∈ W_R(u)1…k});
        for each ć ∈ R(u)_k+1…|R(u)| do
            d_ć ← SimilarityDistance(ć, R(u)_1…k − {c}); D_R(u)1…k ← D_R(u)1…k + {d_ć};
            w_ć ← w_r σ_Asc(ć, R(u)) + w_d σ_Asc(ć, D_R(u)1…k);
            if w_ć < w_c
                R(u) ← R(u) − {c}; replaced ← true;
                break;
            else
                R(u) ← R(u) − {ć}; D_R(u)1…k ← D_R(u)1…k − {d_ć};
        end do
        if not(replaced) then
            break;
        d_R(u) ← Diversity(R(u)_1…k); iter ← iter − 1;
    while iter > 0;
    return R(u);
}
```

4 Evaluation

The following design choices were made:

1. Parameters α and β were set to 1.
2. From the total of 756 relations in our dataset, we selected 157 relations as critical relations based on the automatic ranking provided by GIC [3] and further manual checking. These relations were assigned weights from 1 to 10 based on the perceived importance of each. For example, the 'author' relation was assigned 10

while the 'publisher' relation was weighted 6. The remaining 599 relations were ignored ($w = 0$).

3. For rating prediction (Equation (3)), the number of items selected as the most similar items to the target item was set at 80.
4. We iteratively retrain the model using the predictions until convergence.
5. In LODify-based Diversification procedure (Algorithm 1), we set the maximum number of iterations ($iter$) equal to the size of the recommendation list (k), which was 20 in this experiment.

4.1 Rating Prediction

The evaluation results presented in Table 1 indicates that including more datasets from the Linked Open Data cloud increases the accuracy of LODify. This improves the performance of the recommender system by ameliorating the data quality problems, such as missing facts and incorrect information, in each individual dataset.

4.2 Diversity

Based on the diversity and accuracy evaluation results presented in Fig. 1, the optimal balance achieved at the diversification factor of around 0.6. The final results submitted for the diversity task of the challenge obtained at the diversification factor of 0.7, with ILD of 0.4841 and F-measure of 0.0501.

Table 1. The rating prediction accuracy of LODify and $\text{LODify}_{\text{Basic}}$ using various data sources

	Data sources	RMSE
$\text{LODify}_{\text{Basic}}$	DBpedia	0.8719
LODify	DBpedia (w/o weighting (w = 1))	0.871
	DBpedia	0.8696
	Dbpedia, Freebase	0.8684
	Dbpedia, Freebase, YAGO	0.868

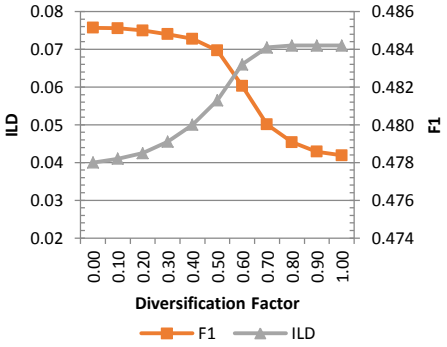

Fig. 1. Impact of the diversification factor on intra list diversity (ILD) and F-measure (F1) values

5 References

[1] Bizer, C., Heath, T. and Berners-Lee, T. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5, 3 (2009), 1-22.

[2] Heath, T. and Bizer, C. 2011. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1, 1 (2011), 1-136.

[3] Meymandpour, R. and Davis, J. G. 2013. Linked Data Informativeness. In Y. Ishikawa, J. Li, W. Wang, R. Zhang and W. Zhang, eds. *Web Technologies and Applications*, 7808, 629-637. Springer Berlin Heidelberg.

[4] Meymandpour, R. and Davis, J. G. 2013. Ranking Universities Using Linked Open Data. In C. Bizer, T. Heath, T. Berners-Lee, M. Hausenblas and S. Auer, eds. *Proceedings of the WWW2013 Workshop on Linked Data on the Web (LDOW2013)* (Rio de Janeiro, Brazil, 14 May, 2013).