Knowledge graph-based weighting strategies for a scholarly paper recommendation scenario

Rubén Manrique
Systems and Computing Engineering Department, School of Engineering. Universidad de los Andes
Bogotá, Colombia
rf.manrique@uniandes.edu.co

Olga Marino
Systems and Computing Engineering Department, School of Engineering. Universidad de los Andes
Bogotá, Colombia
olmarino@uniandes.edu.co

ABSTRACT

In this paper, we study the effects of different node and edge weighting strategies of graph-based semantic representations on the accuracy of a scholarly paper recommendation scenario. Our semantic representation relies on the use of Knowledge Graphs (KGs) for acquiring relevant additional information about concepts and their semantic relations, thus resulting in a knowledge-rich graph document model. Recent studies have used this representation as the basis of a scholarly paper recommendation system. Even when the recommendation is made based on the comparison of graphs, little has been explored regarding the effects of the weights assigned to the edges and nodes in the representation. In this paper, we present the initial results obtained from a comparative study of the effects of different weighting strategies on the quality of the recommendations. Three weighting strategies for edges (Number of Paths (NP), Semantic Connectivity Score (SCS), and Hierarchical Similarity (HS)) and two for nodes (Concept Frequency (CF) and PageRank (PR)) are considered. Results show that the combination of the SCS and CF outperform the other weighting strategy combinations and the considered baselines.

KEYWORDS

Knowledge-based graph representations, scholarly papers recommendation, semantics-aware recommender system

ACM Reference Format:
Rubén Manrique and Olga Marino. 2018. Knowledge graph-based weighting strategies for a scholarly paper recommendation scenario. In *Proceedings of Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop 2018 (co-located with RecSys 2018) (KaRS'18).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

In recent years and with the advent of the Semantic Web and Linked Open Data (LOD), new semantics-aware content representations have been proposed for the next generation of recommender systems[4, 8]. LOD provides a variety of structured knowledge bases in RDF format that are freely accessible on the Web and interconnected to each other. As a result, there is a large amount of machine-readable data available that could be exploited to build more intelligent systems [8]. From this huge amount of data, Knowledge Graphs (KGs) are particularly important since they concentrate the knowledge of multiple domains, and in general, they specify a large number of interrelationships between concepts [12].

In previous works by the authors, a semantic representation that exploits the structured knowledge present in KGs for the task of recommending scholarly papers was proposed [5–7]. In these works, the representations of both user profiles and scholarly papers are based on the concepts mentions found in the paper's content plus additional processes that consider the semantic relation of the concepts found in the KG. The resulting semantic representation is, in essence, a directed weighted graph whose nodes represent concepts and whose edges express the existence of a semantic linkage between two concepts in the KG. The similarity score between a user profile and a document is computed via a graph similarity algorithm. Compared with standard content-based recommendation systems that look at documents and user profiles as bags of terms (i.e. keyword based representations), results show the superiority of the semantic representation [5–7].

In the proposed semantic representation weights are assigned to both nodes and edges. The weights of the nodes consider the importance of the concept in the document, while the edges' weights capture the degree of associativity between concepts in the KG. Although existing graph similarity measures can exploit these weights, the effect of different weighting strategies in the recommendation has not been fully explored. In this paper, we present initial results of a comparative study on the recommendation performance using different weighting strategies for a graph-based representation. The evaluation is done using a scholarly paper recommendation dataset that contains the user profiles of eleven professors of computer science [7].

The paper is organized as follows: Section 2 reviews some related work in semantics-aware recommender systems. Section 3 presents the semantic representation process and its diverse modules. Sections 4 and 5 present the considered weighting functions. Section 6 describes the evaluation framework and Section 7 the results. Finally, conclusions and directions for future work are discussed in Section 8.

2 RELATED WORK

Recent contributions to semantics-aware recommender systems have focused on exogenous semantic representations that introduce the semantics by linking the recommendation item to a KG [3, 4, 8, 9]. The ESWC 2014 Challenge [2], for example, worked on books that were mapped to their corresponding DBpedia resource. In scenarios where there is not a broad enough open knowledge base that describes items or where most of the important information about the item is encoded via textual content, these approaches

KaRS'18, October 7, 2018, Vancouver, Canada.

cannot be easily used. In such cases, an alternative approach is to build semantic representations by processing textual content and mapping the concept mentions found to a KG via entity linking tools in such a way that the item is represented by the multiple concepts found. Piao et al. explore this approach for personalized link recommendations on Twitter [13, 14] and more recently Manrique et al. for a scholarly paper recommendation task [6, 7]. In this paper, we are also using this type of semantic representation, however instead of using a vector representation of concepts we use a graph-based representation. Therefore, we rely on a graph similarity strategy to rank candidate items.

3 SEMANTIC REPRESENTATION

To build the semantic representation, we begin with the identification of concepts mentions in the text (i.e. annotations). Different tools for entity linking and word sense disambiguation can be used for this task. As an example, consider the short input text in Figure 1. After using an automatic annotation tool a set of URIs corresponding to KG concepts is returned. It is important to mention that no human verification is performed on the set of annotations retrieved by the automatic tool. Then, expansion and filtering processes that consider the semantic relationships found in the KG are applied.

The expansion process is used to enrich the representation with concepts that are not explicitly mentioned in the text or not identified by the annotation tool but are strongly related with annotations. Our previous results show that expanded concepts can be important to reinforce the main topic of the document even if they do not occur in the text. The expansion process adds more discriminative power to the representation. We expand the set of annotations to new related ones following two different approaches: category-based and property-based. The category-based expansion incorporates the hierarchical information of the concepts. For the "Artificial_Intelligence" concept, for example, the category "Category:Computational_neuroscience" is retrieved and incorporated into the representation. For property-based expansion, the KG ontology is navigated and a set of related concepts is incorporated. For example, from the "Robotics" annotation the concept "Cooperation" is retrieved by following the "wikiPageWikiLink" property in the ontology. Only outgoing links from a given annotation are considered.

The filtering strategy seeks to eliminate irrelevant and possible noisy concepts in the representation. The noisy concepts can be the result of incorrect annotations, off-topic concepts found in the text or added in the expansion step [6]. We found that noisy concepts tend to be disconnected (i.e. a low number of connections with other concepts). Property paths[1] between every pair of concepts are analyzed and constitute the base information for the graph edge conformation. Basically, an edge is created if there is a property path between the given pair of concepts. Then, the filtering strategy eliminates concepts with a node degree below or equal to α.

Different property path lengths between concepts can be considered in the filtering process, so different semantic representations can be produced for the same input text. The result of these processes is a graph whose nodes are concepts and edges express the existence of a linkage between two concepts in the KG. Finally,

[1]https://www.w3.org/TR/sparql11-property-paths/

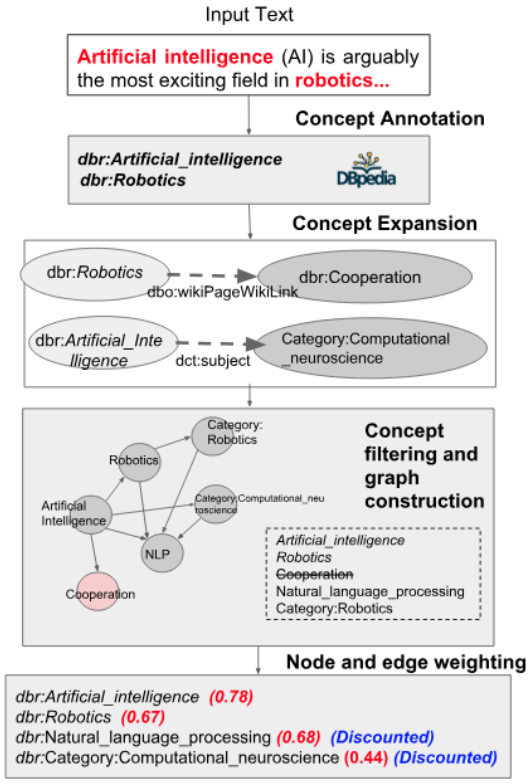

Figure 1: Semantic representation process description

the importance of each concept and the strength of the edges are evaluated via different weighting functions (see Sections 4 and 5). The resulting representation follows Definition 1.

Definition 1. The semantic representation G_i of a profile/document p_i is a directed weighted graph $G_i = (N_i, E_i, w(c), w(e))$, where both nodes and edges have an associated weight defined by the functions $w(c) : N \rightarrow \mathbb{R}+$ and $w(e) : E \rightarrow \mathbb{R}+$. The set of nodes $N_i = \{c_1, c_2, ..., c_k\}$ are entities/concepts belonging to the space of a KG. The node weight $w(c)$ denotes how relevant the node c is for the profile/document. An edge between two nodes (c_a, c_b) represents the existence of at least one statement in the KG that links both concepts. The weight of the edge $w(e)$ denotes how strong this linkage is between the considered concepts.

4 EDGE WEIGHTING FUNCTIONS

Number of paths (NP). NP is defined as $NP(c_i, c_j) = \frac{\sum_{l=1}^{\tau} |paths_{c_i,c_j}^{<l>}|}{MP}$, where $|paths_{c_i,c_j}^{<l>}|$ is the number of paths of length l between concepts c_i and c_j in the KG, and τ is the maximum length of path considered. MP is a normalization parameter that is equal to the maximum number of paths found for a pair of concepts in the representation.

The Semantic Connectivity Score (SCS) [10]. SCS measures latent connections between concept pairs and is computed as: $SCS(c_i, c_j) = 1 - \frac{1}{1+(\sum_{l=1}^{\tau} \beta^l |paths_{c_i,c_j}^{<l>}|)}$. β is a damping factor that penalize longer paths (in our case $\beta = 0.5$). NP and SCS consider

the number of paths between the given concepts as indicative of a strong linkage.

Hierarchical Similarity (HS). We use the hierarchical information of the concepts in the KG to calculate a measure of similarity between the concepts. The higher the similarity, the stronger the link between the concepts. If A is the set of categories for the concept c_i and B is the set of categories for the concept c_j, HS is defined as:

$$HS(c_i, c_j) = \max_{(cat_i \in A, cat_j \in B)} taxsim(cat_i, cat_j) \quad (1)$$

$$taxsim = \frac{\delta(root, cat_{lca})}{\delta(cat_i, cat_{lca}) + \ \delta(cat_j, cat_{lca}) + \ \delta(root, cat_{lca})}$$

Given the hierarchy of categories T, the cat_{lca} of two categories cat_i and cat_j is the vertex of greatest depth in T that is the common ancestor of both cat_i and cat_j. $\delta(cat_a, cat_b)$ is the number of edges on the shortest path between cat_a and cat_b.

5 NODE WEIGHTING FUNCTIONS

Concept frequency (CF) + Discounts. Inspired by TF-IDF, CF analyzes the occurrences of the concept in the input content as well as the frequency in the set of semantic representations. CF is defined as: $CF(c) = w_{cf}(c) \times log\frac{M}{m_c}$, where $w_{cf}(c)$ represents the number of times that c appears in the input content, M is the total number of documents/profiles in the dataset and m_c is the number of documents/profiles with the concept c in their representation. After the expansion process, the representation could be diverted towards frequent properties in the set of instances of the KG or general categories in the hierarchical structure of the KG. For the categories added through the expansion the following discount is applied: $CF_{discount}(cat) = CF(cat) \times \frac{1}{log(SP)} \times \frac{1}{log(SC)}$ where SP is the set of concepts belonging to the category and SC is the set of sub-categories in the category hierarchy. The idea behind this discount strategy is that categories that are too broad and generic are penalizes [11]. Similarly, for property-based expanded concepts, the following discount is applied: $CF_{discount}(c) = CF(c) \times \frac{1}{log(P)}$ where P is the number of occurrences of the property in the KG from which the concept $c \in C$ is obtained.

PageRank (PR): PageRank is a well-known node ranking algorithm. We use the PageRank version for directed weighted graphs (i.e. it considers the edge weights). Therefore, the results obtained with this centrality measure depends on the resulting link structure in the semantic graph and the associated edge weight.

6 EXPERIMENTAL SETUP

Our main goal is to analyze the influence of the different weighting strategies in the context of scholarly paper recommendation. We compare the quality achieved by the same recommendation algorithm when inputting semantic representations for user profiles and documents using the different weighting strategies. In this regard, we embrace the content-based algorithm described in Definition 2.

Definition 2. Recommendation Algorithm: given a user profile u and a set of candidate scholarly papers $SP = \{p_1, ..., p_n\}$, which are represented using the graph-based representation in Definition 1, the recommendation algorithm ranks the candidate items according to their Graph Similarity (GS) to the user profile. For GS we employ the edit distance implemented in [5] and defined in Equation 2.

$$GS(G_i, G_j) = \frac{GS_{nodes}(G_i, G_j) + GS_{edges}(G_i, G_j)}{2} \quad (2)$$

$$GS_{nodes}(G_i, G_j) = 1 - \frac{\alpha_{nodes} + \sum_{c \in N_i \cap N_j} |w_i(c) - w_j(c)|}{\alpha_{nodes} + \sum_{c \in N_i \cap N_j} max(w_i(c), w_j(c))}$$

$$GS_{edges}(G_i, G_j) = 1 - \frac{\alpha_{edges} + \sum_{e \in E_i \cap E_j} |w_i(e) - w_j(e)|}{\alpha_{edges} + \sum_{e \in E_i \cap E_j} max(w_i(e), w_j(e))}$$

where α_{nodes} and α_{edges} are defined as:

$$\alpha_{nodes} = \sum_{c \in N_i} w_i(c) + \sum_{c \in N_j} w_j(c) - \sum_{c \in N_i \cap N_j} w_i(c) - \sum_{c \in N_i \cap N_j} w_j(c)$$

$$\alpha_{edges} = \sum_{e \in E_i} w_i(e) + \sum_{e \in E_j} w_j(e) - \sum_{e \in E_i \cap E_j} w_i(e) - \sum_{e \in E_i \cap E_j} w_j(e)$$

GS evaluates the similarity between two graphs in terms of the weights differences of the common nodes/edges compared to the total weight of the nodes/edges in the two graphs. Therefore, two graphs are similar not only if their nodes/edges coincide but also if their weights are close in magnitude.

6.1 Dataset

We employ the dataset proposed in [6] that contains the user profiles of 11 professors in the area of computer science. The user profiles were built using the full text of the most recent publications found on their Google Scholar web pages. At least a minimum of twelve of each professor's most recent publications were used as input for the semantic representation process. The candidate set is a subset of Core and Arxiv open corpora that contains 5710 different academic documents (i.e. papers, tech reports, thesis, etc.). The ground truth of papers is a subset of the candidate set in which users express an explicit interest via a web-based search system. In the data set, for each user there are at least 10 *relevant* documents. As for the user profile, the full text was used to construct the semantic representation of each document in the candidate set.

For the construction of the semantic representation we use: (i) DBpedia as KG, (ii) DBpedia Spotlight as annotation service, (iii) a maximum path length of 2 for filtering and edge conformation as well as for the τ parameter, (iv) a minimum degree value $\alpha = 1$, (v) categories extracted through `dct:subject` to calculate HS.

7 RESULTS

The performance of the recommender system was evaluated by typical metrics for the evaluation of Top-N recommender tasks: MRR (Mean Reciprocal Rank), MAP@10 (Mean Average Precision), and NDCG@10 (Normalized Discounted Cumulative Gain). We select N=10 as the recommendation objective since it is a common rank and it fits with the minimum number of relevant documents per user in the dataset.

Table 1 presents the results obtained by the different combinations of the weighting strategies. We use a classical content-based recommendation algorithm baseline that ranks candidate items according to their cosine similarity with the user profile. In this case, profiles and documents use a standard bag-of-words Vector Space Model (VSM) representation. According to Beel et al. VSM

Table 1: Performance of scholarly paper recommendation using different weighting strategies for nodes and edges. (*) indicates the improvement over baselines is statistically significant ($p < 0.05$).

Edge Weight	Node Weight	MRR	MAP	NDCG
NP	CF	0.503	0.5*	0.661
NP	PR	0.523*	0.401	0.482
SCS	CF	0.482	**0.523***	**0.716***
SCS	PR	**0.574***	0.502*	0.672
HS	CF	0.421	0.395	0.495
HS	PR	0.442	0.342	0.434
Cosine Baseline		0.401	0.345	0.423
VEO Baseline		0.442	0.427	0.601

with TF-IDF is the most frequent profiling and weighting scheme in research paper recommender systems [1]. As a baseline, we also use VEO (Vertex Edge Overlap [5]). VEO uses an unweighted version of the semantic representation (*Definition 1*). The candidate items, in this case, are ranked based on the number of common nodes and edges they have with the user's profile. VEO was chosen to evaluate whether the consideration of weights in the representation has a positive impact on the recommendation.

We can see that the combination of SCS-CF outperforms other weighting strategies and the baselines in terms of MAP and NDCG. In terms of MRR, the best weighting strategy is SCS-PR. Although SCS and NP consider the number of existing paths, the results show a better performance for SCS. This can be attributed to the fact that in addition to the number of paths, SCS also penalizes the path length through a damping factor. Further experimentation is required to evaluate the effect of the β damping factor. HS presents the worst results among the weighting strategies and only slightly better than the cosine baseline.

Regarding the node weighting strategies, results show that CF is more appropriate. CF is superior to PR in terms of MAP and NDCG. On the other hand, according to the MRR, in average PR ranks higher the first relevant result. The comparison with VEO also shows that the consideration of weights for nodes and edges improves the semantic representation proposed for the task of recommending scholarly papers.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we explored different node and edge weighting strategies for a graph-based semantic model for representing user profiles and papers in the context of the scholarly paper recommendation task. The combination of SCS and CF as weighting strategies for edges and nodes respectively presented the best results. SCS considers the number of existing paths in the KG between the concepts considered, while CF considers the frequency of the concept in the document/profile. In the near future, we plan to explore other measures of centrality for nodes weighting (e.g., betweenness, closeness). We also want to explore the effect of path lengths greater than 2 for the calculation of SCS/NP. Finally, according to our recommendation algorithm and graph similarity measure (Equation 2), the contribution of edges and nodes are considered equivalent. We want to test this assumption by favoring/disfavoring the contribution of nodes and edges.

ACKNOWLEDGMENTS

This work was partially supported by COLCIENCIAS PhD scholarship (Call 647-2014).

REFERENCES

[1] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (2016), 305–338. https://doi.org/10.1007/s00799-015-0156-0

[2] Tommaso Di Noia, Iván Cantador, and Vito Claudio Ostuni. 2014. Linked Open Data-Enabled Recommender Systems: ESWC 2014 Challenge on Book Recommendation. In *Semantic Web Evaluation Challenge*, Valentina Presutti, Milan Stankovic, Erik Cambria, Iván Cantador, Angelo Di Iorio, Tommaso Di Noia, Christoph Lange, Diego Reforgiato Recupero, and Anna Tordai (Eds.). Springer International Publishing, Cham, 129–143.

[3] Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, and Markus Zanker. 2012. Linked Open Data to Support Content-based Recommender Systems. In *Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/2362499.2362501

[4] Tommaso Di Noia and Vito Claudio Ostuni. 2015. *Recommender Systems and Linked Open Data*. Springer International Publishing, Cham, 88–113. https://doi.org/10.1007/978-3-319-21768-0_4

[5] Rubén Manrique, Felipe Cueto, and Olga Mariño. 2018. Comparing Graph Similarity Measures for Semantic Representations of Documents. In *Advances in Computing*. Springer International Publishing, Cham, 3–16.

[6] Rubén Manrique, Omar Herazo, and Olga Mariño. 2017. Exploring the Use of Linked Open Data for User Research Interest Modeling. In *Advances in Computing*, Andrés Solano and Hugo Ordoñez (Eds.). Springer International Publishing, Cham, 3–16.

[7] Rubén Manrique and Olga Mariño. 2017. How Does the Size of a Document Affect Linked Open Data User Modeling Strategies?. In *Proceedings of the International Conference on Web Intelligence (WI '17)*. ACM, New York, NY, USA, 1246–1252. https://doi.org/10.1145/3106426.3109440

[8] Cataldo Musto, Pierpaolo Basile, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2017. Introducing linked open data in graph-based recommender systems. *Information Processing & Management* 53, 2 (2017), 405 – 435. https://doi.org/10.1016/j.ipm.2016.12.003

[9] Cataldo Musto, Pasquale Lops, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2016. Semantics-aware Graph-based Recommender Systems Exploiting Linked Open Data. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. ACM, New York, NY, USA, 229–237. https://doi.org/10.1145/2930238.2930249

[10] Bernardo Pereira Nunes, Besnik Fetahu, Ricardo Kawase, Stefan Dietze, Marco Antonio Casanova, and Diana Maynard. 2015. *Interlinking Documents Based on Semantic Graphs with an Application*. Springer International Publishing, Cham, 139–155.

[11] Fabrizio Orlandi, John Breslin, and Alexandre Passant. 2012. Aggregated, Interoperable and Multi-domain User Profiles for the Social Web. In *Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12)*. ACM, New York, NY, USA, 41–48. https://doi.org/10.1145/2362499.2362506

[12] Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 8, 3 (2017), 489–508. https://doi.org/10.3233/SW-160218

[13] Guangyuan Piao and John G. Breslin. 2016. Analyzing Aggregated Semantics-enabled User Modeling on Google+ and Twitter for Personalized Link Recommendations. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. ACM, New York, NY, USA, 105–109. https://doi.org/10.1145/2930238.2930278

[14] Guangyuan Piao and John G. Breslin. 2016. Exploring Dynamics and Semantics of User Interests for User Modeling on Twitter for Link Recommendations. In *Proceedings of the 12th International Conference on Semantic Systems (SEMANTiCS 2016)*. ACM, New York, NY, USA, 81–88. https://doi.org/10.1145/2993318.2993332