Improving Semantic Relatedness in Paths for Storytelling with Linked Data on the Web

Laurens De Vocht[1(✉)], Christian Beecks[2], Ruben Verborgh[1], Thomas Seidl[2], Erik Mannens[1], and Rik Van de Walle[1]

[1] Multimedia Lab, Ghent University - iMinds,
Gaston Crommenlaan 8 Bus 201, 9050 Ghent, Belgium
{laurens.devocht,ruben.verborgh,erik.mannens,rik.vandewalle}@ugent.be
[2] Department of Computer Science 9, Data Management and Data Exploration Group, RWTH Aachen University, 52056 Aachen, Germany
{beecks,seidl}@cs.rwth-aachen.de

Abstract. Algorithmic storytelling over Linked Data on the Web is a challenging task in which many graph-based pathfinding approaches experience issues with *consistency* regarding the resulting path that leads to a story. In order to mitigate arbitrariness and increase consistency, we propose to improve the semantic relatedness of concepts mentioned in a story by increasing the relevance of links between nodes through additional domain delineation and refinement steps. On top of this, we propose the implementation of an optimized algorithm controlling the pathfinding process to obtain more homogeneous search domain and retrieve more links between adjacent hops in each path. Preliminary results indicate the potential of the proposal.

1 Introduction

Algorithmic storytelling can be seen as a particular kind of querying data. Given a set of keywords or entities, which are typically, but not necessarily dissimilar, it aims at generating a story by explicitly relating the query context with a path that includes semantically related resources. Storytelling is utilized for example in entertaining applications and visualizations [4] in order to enrich related Linked Data resources with data from multimedia archives and social media [1] as well as in scientific research fields such as bio-informatics where biologists try to relate sets of genes arising from different experiments by investigating the implicated pathways [3].

The most frequently encountered algorithm to determine a path between multiple resources is the A* algorithm [2]. This algorithm, which is based on a graph representation of the underlying data (i.e., resources and links between them define nodes and edges, respectively) determines an optimal solution in form of a lowest-cost traversable path between two resources. The optimality of a path, which is guaranteed by the A* algorithm, does not necessarily comply with the users' expectations. By considering for instance large real-world semantic graphs, such as Linked Data graphs, where links between nodes are semantically

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9341, pp. 31–35, 2015.
DOI: 10.1007/978-3-319-25639-9_6

annotated, users are able to directly interpret the transitions between nodes and thus the meaning of a path. Caused by the inevitable increasing number of nodes and sometimes loosely related links among them in nowadays online datasets on the Web, optimal paths frequently show a high extent of *arbitrariness*: paths appear to be determined by chance and not by reason or principle and are often affected by resources that share many links.

In order to mitigate arbitrariness of a story, we propose a control algorithm that wraps an optimized version of our original core algorithm, which is embedded in the Everything is Connected Engine (EiCE) [1]. In fact, our contribution is twofold: (i) We outline the control algorithm which reduces arbitrariness by increasing the relevance of links between nodes through additional domain delineation and refinement steps; and (ii) we optimized the original core algorithm to support two-hop paths rather than directly linked nodes, this allows to define heuristics and weights on a broader context than a pair of directly linked nodes and predicates between them. We discuss how paths consisting of two-hop node steps are presented as building blocks for storytelling. We conclude our paper with preliminary results and an outlook on future work.

2 Pathfinding for Storytelling

Each path that contributes to a story is determined within a query context comprising both start and destination resources. Our algorithm reduces the arbitrariness of a path between these resources by increasing the *relevance* of the links between the nodes using a domain delineation step. The path is refined by controlling the iteratively application of the A* algorithm and with each iteration attempting to improve the overall semantic relatedness between the resources until a fixed number of iterations or a certain similarity threshold is reached.

2.1 Domain Delineation

Instead of directly initializing the graph as-is by including all links between the resources, we identify the relevance of predicates with respect to the query context. This is done by extracting and giving higher preference to the type of relations (predicates) that occur frequently in the query context. In this way, links leading to a story are of relevance since each predicate that describes the semantics of a link also occurs in the direct neighborhood of the query context. The goal is to determine a more homogeneous subgraph with more potentially relevant nodes to the user.

2.2 Core Algorithm

Determining a path between two nodes is carried out by means of the A* algorithm, because this algorithm provides an optimal solution, i.e., a (shortest) cost-minimal path between two nodes with respect to the weights of the links contained in the path. While the A* algorithm is able to compute an optimal

solution within a computation time complexity of $\mathcal{O}(|E|)$, where E denotes the number of links to be examined, heuristics are able to reduce the runtime of this algorithm significantly and, thus, to achieve an improvement in efficiency when computing the lowest-cost path between two nodes. Our algorithm utilizes a bidirectional variant of the A* algorithm which turns out to have higher efficiency.

2.3 Refinement

After a path is determined by the A* algorithm, we measure the semantic relatedness between all resources occurring in the path with respect to the query context. This done by counting the number of overlapping predicates (i) among each other combined with those in the start and destination resources; and then (ii) averaging and normalizing this count over all resources. Depending on the threshold and the maximum number of iterations configured, typically between 3 and 10 times, the core algorithm is repeated excluding the middle hop in the path. Different outputs after each iteration are guaranteed by forcing the core algorithm to find a path without the excluded node. Finally, the path with the highest similarity score is selected for the story.

3 Presenting the Story

Obviously a set of paths is not a presentable story yet. We note that even if a path comprise just the start and destination (indicating they are linked via common hops or directly to each other), the story will contain interesting facts. This is because each step in the path is separated with at least one hop from the next node. For example, to present a story about *Carl Linnaeus* and *Charles Darwin*, the story could start from a path that goes via *J.W. von Goethe*. The resulting statements serve as basic facts, which are relation-object statements, that make up the story. It is up to the application or visualization engine to present it to end-users and enrich it with descriptions, media or further facts. Table 1 exemplarily explicates the idea of statements as story facts.

4 Preliminary Results

To determine whether the arbitrariness of a story is reduced, we validated that our optimization increased the semantic relatedness of the concepts mentioned

Table 1. The statements as story facts

About	Relation	Object
Carl Linnaeus and Charles Darwin	are	scientists
J.W. von Goethe	influenced	Carl Linnaeus and Charles Darwin
J.W. von Goethe and Charles Darwin	influenced	Karl Marx and Sigmund Freud

Table 2. The comparison between the original and optimzed algorithm shows that the semantic relatedness can be improved in all cases except for the last two when the entities were already closely related.

Query context	Original	NGD	Optimized	NGD
C._Linnaeus - C._Darwin	C._H._Merriam	0.50	J._W._Von_Goethe	**0.43**
C._Linnaeus - A._Einstein	Aristotle	0.70	J._W._Von_Goethe	**0.45**
C._Linnaeus - I._Newton	P._L._Maupertuis	0.48	D._Diderot	**0.40**
A._Einstein - I._Newton	Physics	0.62	D._Hume	**0.45**
C._Darwin - I._Newton	D._Hume	**0.38**	Royal_Liberty_School	0.40
C._Darwin - A._Einstein	D._Hume	**0.43**	B._Spinoza	0.44

in a story. To this end, we computed stories about the four highest ranked DBpedia scientists, according to their PageRank score[1], and have determined their pairwise semantic relatedness by applying the Normalized Google Distance (NGD). The results are shown in Table 2.

Table 2 shows that the entities *Aristotle* and *Physics* are included in the story when applying the original algorithm. These entities are perfect examples of *arbitrary* resources in a story which decrease the consistency. Except that they are related to science, it is unclear to the user why the algorithm 'reasoned' them to be in the story. When utilizing the optimized algorithm these entities are replaced by *J._W._Von_Goethe* and *D._Hume*.

In order to verify our results, we also include the total semantic similarity of a path by computing the semantic relatedness between all neighboring node pairs in that path. As can be seen in Table 2, the optimized algorithm is able to improve the semantic relatedness of the resulting paths.

5 Conclusions and Future Work

We proposed an optimized pathfinding algorithm for storytelling that reduces the number of arbitrary resources popping up in paths contained in the story. We added an additional resource pre-selection and a post-processing step that increases the semantic relatedness of resources. Preliminary evaluation results using the DBpedia dataset indicate that our proposal succeeds in telling a story featuring higher semantic relatedness, especially in cases where the previous algorithm did not make seemingly optimal choices in terms of semantic relatedness. Future work will mainly focus on developing the algorithm and making it available for use in applications. We will extend and verify our findings with additional semantic similarity measures besides the NGD and by investigating different weights and heuristics within the core algorithm. We will validate the correlation between the increased semantic relatedness and the impact on the

[1] http://people.aifb.kit.edu/ath#DBpedia_PageRank.

story consistency as perceived by users. Additionally, we will evaluate the scalability of our approach in a distributed client/server architecture.

Acknowledgment. This work is partially Funded by the Excellence Initiative of the German federal and state governments; Flanders (IWT, FWO); and the European Union.

References

1. De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Discovering meaningful connections between resources in the web of data. In: Proceedings of the 6th Workshop on Linked Data on the Web (LDOW 2013) (2013)
2. Hart, P., Nilsson, N., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE Trans. Syst. Sci. Cybern. **4**, 100–107 (1968)
3. Kumar, D., Ramakrishnan, N., Helm, R.F., Potts, M.: Algorithms for storytelling. IEEE Trans. Knowl. Data Eng. **20**(6), 736–751 (2008)
4. Vander Sande, M., Verborgh, R., Coppens, S., De Nies, T., Debevere, P., De Vocht, L., De Potter, P., Van Deursen, D., Mannens, E., Van de Walle, R.: Everything is connected: using linked data for multimedia narration of connections between concepts. In: Proceedings of the 11th International Semantic Web Conference Posters and Demo Track, November 2012