RADON2: A buffered-Intersection Matrix Computing Approach To Accelerate Link Discovery Over Geo-Spatial RDF Knowledge Bases

OAEI2018 Results

Abdullah Fathi Ahmed[1] Mohamed Ahmed Sherif[1,2] and Axel-Cyrille Ngonga Ngomo[1,2]

[1] Paderborn University, Data Science Group, Pohlweg 51, D-33098 Paderborn, Germany
E-mail: {firstname.lastname}@upb.de
[2] Department of Computer Science, University of Leipzig, 04109 Leipzig, Germany
E-mail: {lastname}@informatik.uni-leipzig.de

Abstract. Geospatial data is at the essence of the Semantic Web, where a knowledge base such as *LinkedGeoData* consists of more than 30 billions facts. Reasoning on these considerable amounts of geospatial data lacks efficient methods for the computation of links between the resources contained in these knowledge bases. In this paper, we present the participation of the extension of RADON algorithm (dubbed RADON2) in the OAEI 2018 campaign. The OAEI results show that RADON2 outperforms the other state of the art in most of the cases.

1 Presentation of the System

we present the extension of RADON algorithm [8, 6] (dubbed RADON2), where we, compute all topological relations of DE9-IM in order to accelerate the topological relation discovery among geospatial resources.

1.1 State, Purpose and General Statement

In the following, we start by formally defining the general link discovery problem. Thereafter, we formally define the link discovery of topological relations problem, which we takeld by RADON2.

Link Discovery. Let K be a finite RDF knowledge base. K can be regarded as a set of triples $(s, p, o) \in (\mathcal{R} \cup \mathcal{B}) \times \mathcal{P} \times (\mathcal{R} \cup \mathcal{L} \cup \mathcal{B})$, where $\mathcal{R}$ is the set of all resources, $\mathcal{B}$ is the set of all blank nodes, $\mathcal{P}$ the set of all predicates and $\mathcal{L}$ the set of all literals. The Link Discovery (LD) problem can be expressed as follows: Given two sets of resources S and T (for example hotels and water bodies) and a relation r (e.g., `:touches`), find all pairs $(s, t) \in S \times T$ such that $r(s, t)$ holds. The result is produced as a set of links called a *mapping*: $M_{S,T} = \{(s_i, r, t_j) | s_i \in S, t_j \in T\}$. Optionally, a similarity score ($sim \in [0, 1]$) calculated by an LD tool can be added to the entries of mappings to express assurance of a computed link. Finding solutions for the LD problem is challenging due to the typically the large volume of current datasets as well as its semantic heterogeneity. The main purpose of LD approaches is to meet the main requirements of *(1) high effectiveness* (i.e maximize a fitness function such as F-measure) and *(2) high efficiency* (i.e., minimize runtime).

Link Discovery of Topological Relations. The Dimensionally Extended nine-Intersection Model (DE-9IM) [3] is a topological model and a standard used to describe the spatial relations of two geometries in two-dimensional space. Since the spatial relations expressed by DE-9IM are topological, they are invariant to rotation, translation and scaling transformations [4]. The DE-9IM model is based on a 3×3 intersection matrix with the form:

$$DE9IM(g_1, g_2) = \begin{bmatrix} dim(I(g_1) \cap I(g_2)) & dim(I(g_1) \cap B(g_2)) & dim(I(g_1) \cap E(g_2)) \\ dim(B(g_1) \cap I(g_2)) & dim(B(g_1) \cap B(g_2)) & dim(B(g_1) \cap E(g_2)) \\ dim(E(g_1) \cap I(g_2)) & dim(E(g_1) \cap B(g_2)) & dim(E(g_1) \cap E(g_2)) \end{bmatrix} \quad (1)$$

where dim is the maximum number of dimensions of the intersection $\cap$ of the $interior(I)$, $boundary(B)$, or $exterior(E)$ of the two geometries g_1 and g_2. The domain of dim is $\{-1, 0, 1, 2\}$, where -1 indicates no intersection, 0 stands for an intersection that results in a set of one or more points, 1 indicates an intersection made up of lines and 2 stands for an intersection that results in an area. A simplified binary version of $dim(x)$ with the binary domain $\{true, false\}$ is obtained using the Boolean function $\beta(dim(I(g)) = false$ iff $dim(I(g)) = -1$ and $true$ otherwise. There is only a subset of the topological relations obtainable through DE-9IM that reflects the semantics of the English language [3] [2] including `equals, within, contains, disjoint, touches, meets, covers, coveredBy, intersects, crosses` and `overlaps`.

1.2 Specific Techniques Used

in this section, we discuss the main idea behind our new extension of RADON.

RADON2 *vs.* RADON. The basic idea behind the original RADON approach [8] for topological relation discovery is to provide an indexing method combined with space tiling that allows for efficient computation of topological relations between geospatial resources. In particular, RADON presents a novel sparse index for geospatial resources. Then, based on bounding boxes of the indexed geospatial resources, RADON applies a strategy for discarding unnecessary computations of DE-9IM relations. In RADON2, our concerns is focused on optimizing the computing of intersection matrix (IM) used in DE9-IM standard. In the original RADON, the intersection matrix is computed for each topological relation, while in RADON2 we compute the IM once for all relations among the same pair of resources. We then apply the mask for each relation to the the computed IM. In particular, we buffer the IM of each pair of geometries so that all topological relations of same pair can be retrieved with no need to recompute their respective IM again. By applying this strategy, we can save the time for recomputing the IM for each individual topological relation. Moreover, calculating IM at once for each pair of geometries for all topological relations does not affect the completeness of the linking result. i.e., the F-measure of RADON2 is the same as the F-measure of RADON, which is always 1.

1.3 Adaptations Made for the Evaluation

No specific adaptations were made to the original Radon algorithm, we only provide a Java `SystemAdapter` according to the campaign guidelines[3].

[3] https://project-hobbit.eu/challenges/om2017/om2017-tasks/

1.4 Link to the System

Both RADON and RADON2 are implemented in the link discovery framework LIMES. LIMES is available under the *GNU Affero General Public License v3.0*[4]. RADON2 source code is available online from the project website[5]. The project web site also provide a user manual[6] as well as a developer manual[7].

2 Results

RADON2 has been evaluated only in the Hobbit Link Discovery Track Task 2 (Spatial). The basic idea behind this task was to measure how well the systems can identify DE- 9IM (Dimensionally Extended nine-Intersection Model) topological relations. The supported spatial relations were: `equals`, `within`, `contains`, `disjoint`, `touches`, `meets`, `covers`, `coveredBy`, `intersects`, `crosses` and `overlaps`. The geospatial resources traces were represented in Well-known text (WKT) format as LineStrings. The result is produced as a set of links called a *mapping*: $M_{S,T} = \{(s_i, r, t_j)|s_i \in S, t_j \in T\}$. All the systems were tested against two datasets: (1) the sandbox dataset, with a scale of 10 instances, and (2) the mainbox dataset with a scale of $5K$ instances. The other participants to this task in addition to Radon were Agreement Maker Light(AML) and Silk.

The systems were judged on the basis of precision, recall, F-Measure and run time. The final results are shown in Figures 1, 2, 3 and 4. Note that we are only presenting the time performance and not precision, recall and F-Measure as all were equal to 1.0.

From these results we can see that RADON2 outperforms the other systems in all relations for the sandbox and mainbox (linestrings –polygons) (see Figures 3 and 4) dataset as well as the for the the mainbox dataset (linestrings–linestrings) (Figure 2). For the sandbox dataset (linestrings–linestrings) (Figure 1), RADON achives a better performance in most of the relations (e.g., `overlaps`, `crosses`, `covered by`, `covers`, `within`, `contains`, `disjoint and equal`. Only for the `touches and intersects` AML was able to outperform RADON2 for the TomTom dataset of the sandbox (linestrings–linestrings). The differences in performance between `touches` and `intersects`, where AML outperforms RADON cannot be explained from an implementation point of view, as these two relations share the exact optimizations. However, due to the datasets consisting exclusively of `LineStrings`, it is apparent that *touches* and *intersects* are much more likely to hold between any two geometries than other relations. Therefore, the benchmarks on these relations are the hardest in this task.

3 Conclusions and Future Work

We present RADON2, a simple strategy for scaling the original RADON approach by computing the intersection matrix for each pair of resources once and use it for computing all possible topological relations associated with such resources at hand. The presented evaluation during the OAEI 2018 showed that, in addition to being complete and correct (i.e. achieving an F-Measure of 1.0), RADON2 also outperforms the other participating systems in most of the cases

[4] https://github.com/dice-group/LIMES/blob/master/LICENSE
[5] https://github.com/dice-group/LIMES
[6] https://dice-group.github.io/LIMES/user_manual/
[7] https://dice-group.github.io/LIMES/developer_manual/

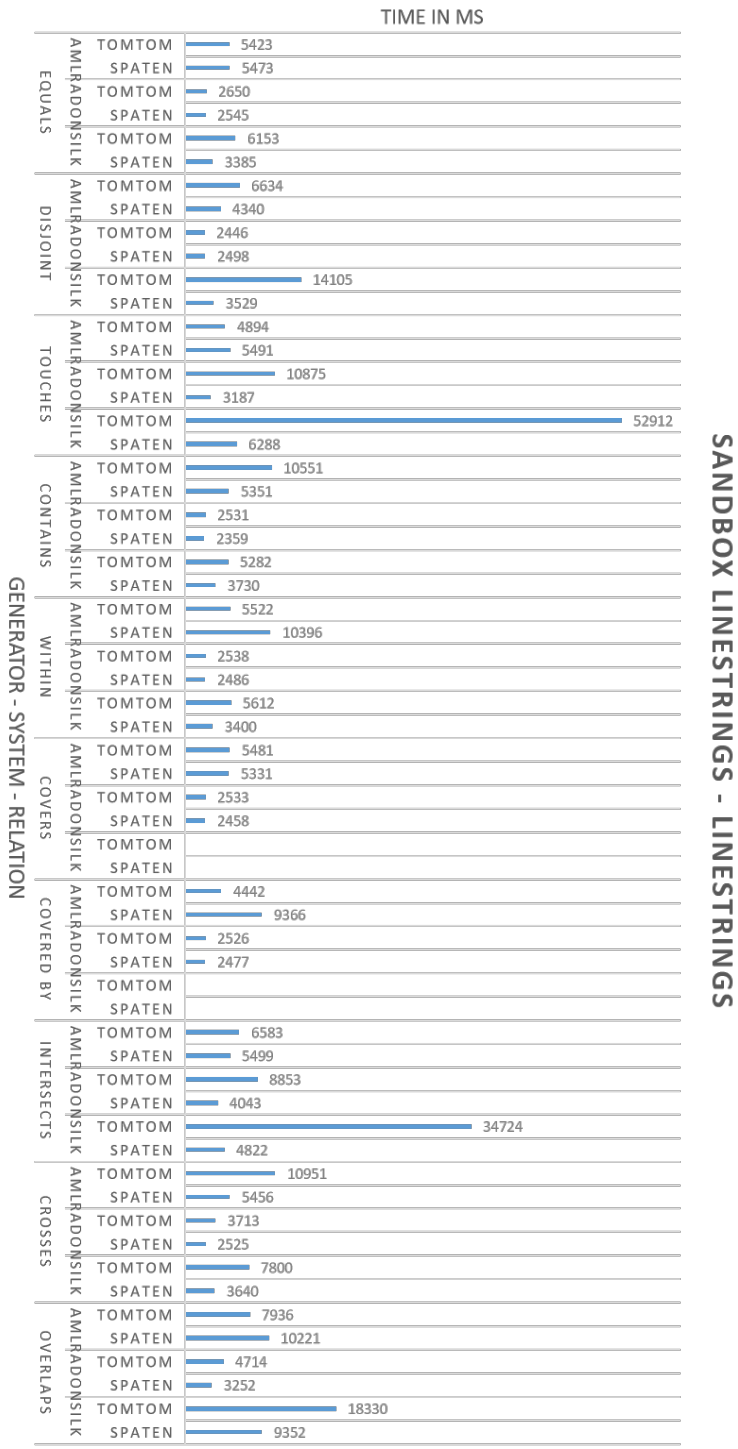

Fig. 1: Runtime results of linestrings-linestrings *Sandbox* Dataset

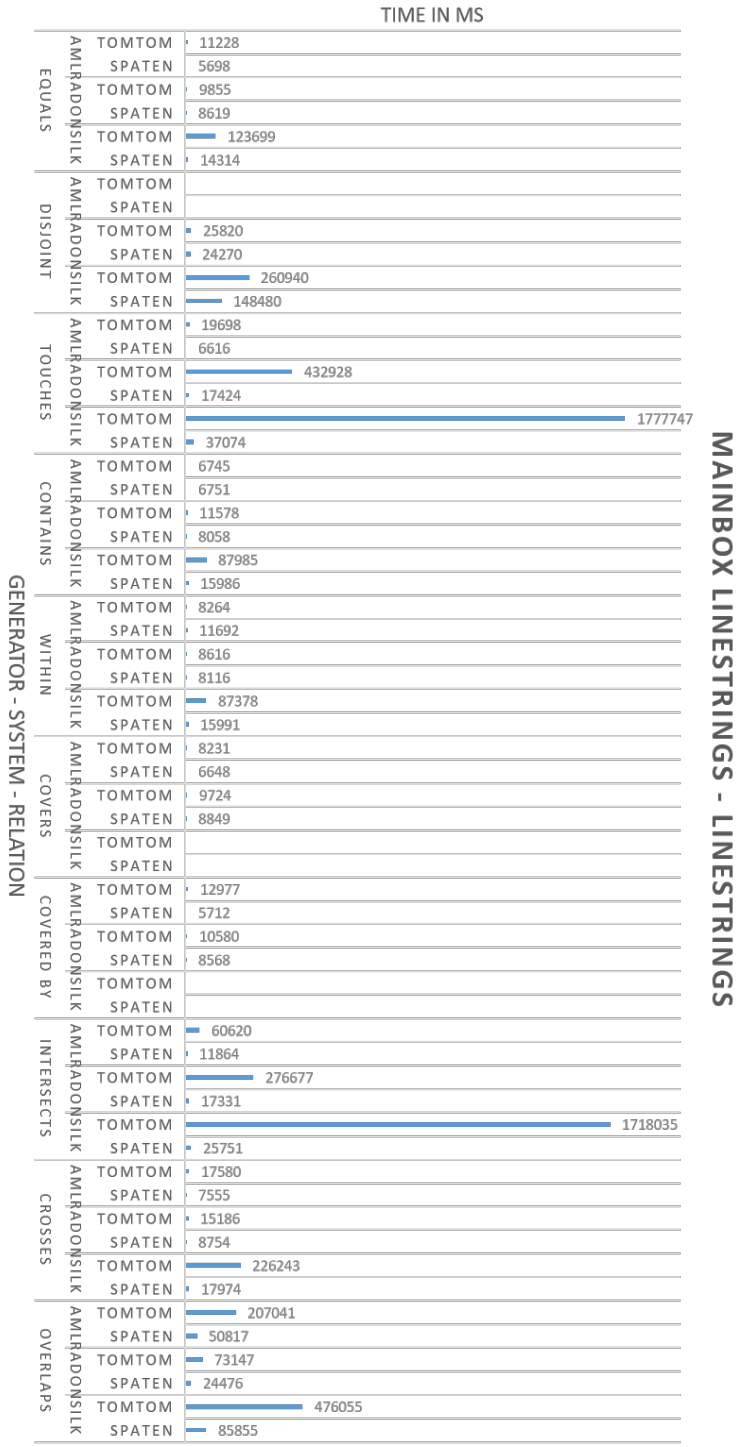

Fig. 2: Runtime results of linestrings-linestrings *Mailbox* DataSet

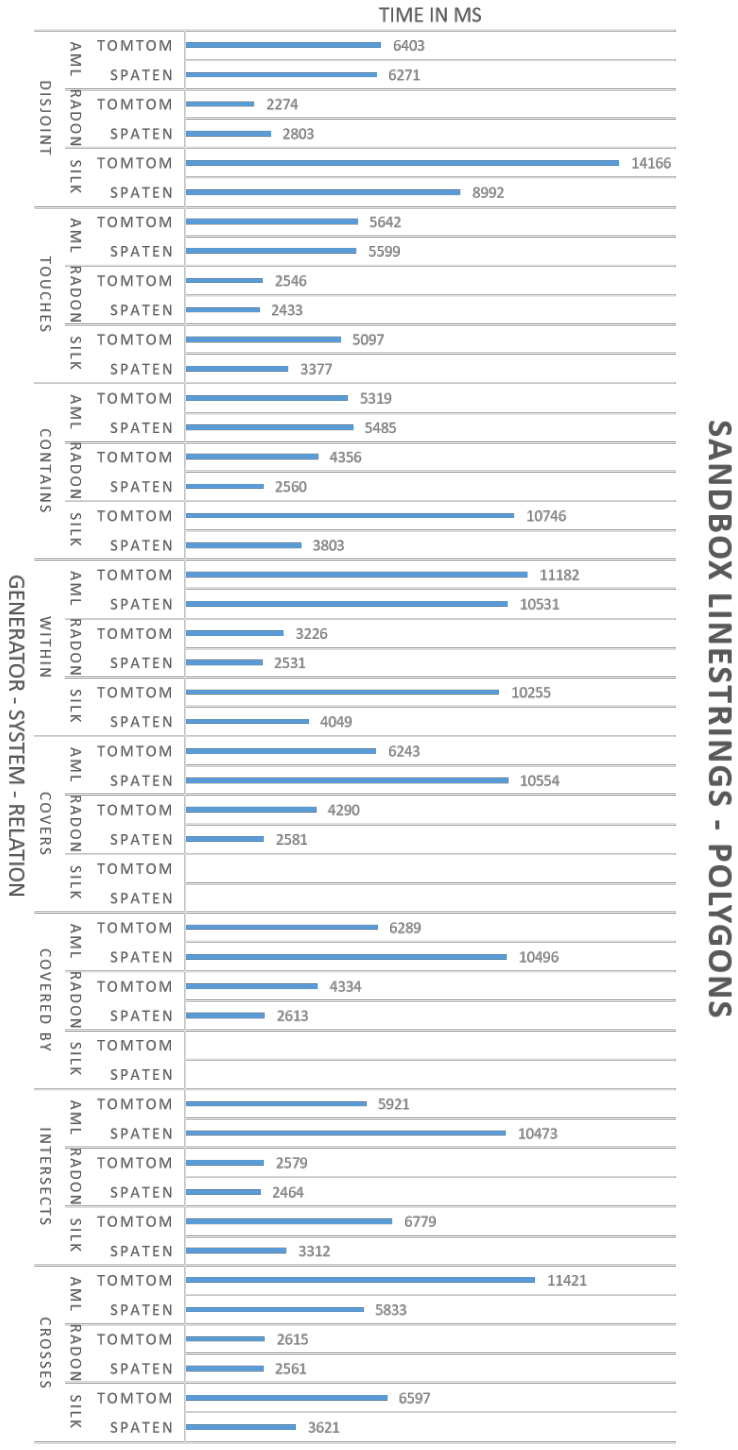

Fig. 3: Runtime results of linestrings-polygons *Sandbox* Dataset

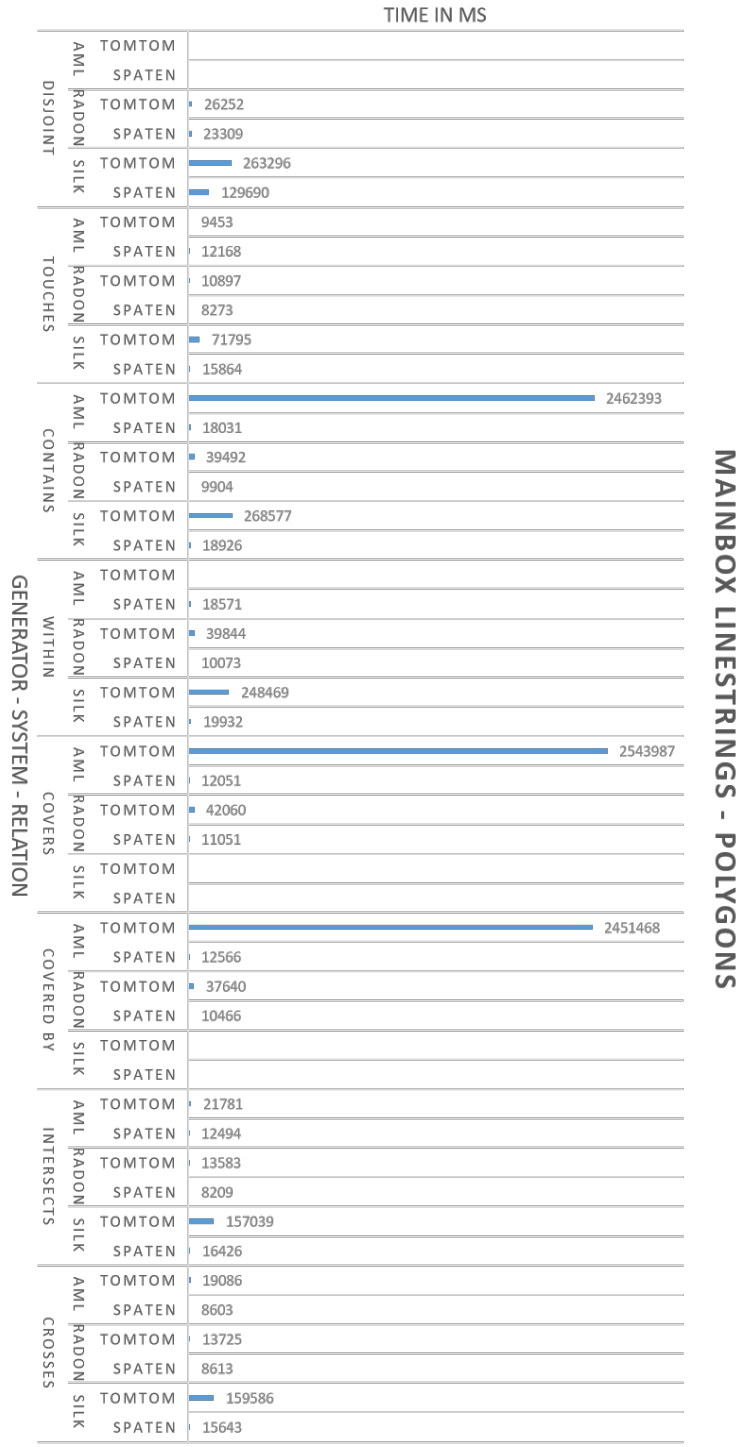

Fig. 4: Runtimes results of linestrings-polygons *Mailbox* DataSet

In future work, we will apply this strategy on a larger datasets with more resources and more points per resource, where we will implement more sophisticated parallelization techniques. For enabling automatic configuration of RADON2, we will combine RADON2 with the machine learning algorithm WOMBAT [7] implemented in LIMES. Also, we will extend RADON2 for discovering spatial-temporal relation by integrating it with [5]. Moreover, we intend to combine RADON2 with the simplification algorithms introduced in [1] in order to achieve even better speedup.

Acknowledgments This work has been supported by Eurostars Project SAGE (GA no. E!10882), the BMVI project the LIMBO (GA no. 19F2029C), the DFG project LinkingLOD (project no. NG 105/3-2), the BMWI Project GEISER (project no. 01MD16014) as well as the H2020 projects SLIPO (GA no. 731581) and HOBBIT (GA no. 688227).

References

1. A. F. Ahmed, M. Sherif, and A. Ngonga Ngomo. On the effect of geometries simplification on geo-spatial link discovery. In *Proceedings of SEMANTiCS 2018*, 2018.
2. E. Clementini, P. Di Felice, and P. Van Oosterom. A small set of formal topological relationships suitable for end-user interaction. In *International Symposium on Spatial Databases*, pages 277–295. Springer, 1993.
3. E. Clementini, J. Sharma, and M. J. Egenhofer. Modelling topological spatial relations: Strategies for query processing. *Computers & graphics*, 18(6):815–822, 1994.
4. M. J. Egenhofer and R. D. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information System*, 5(2):161–174, 1991.
5. K. Jha, M. Röder, and A.-C. Ngonga Ngomo. All That Glitters is not Gold – Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking. In *The Semantic Web. Latest Advances and New Domains: 14th International Conference, ESWC 2017, Proceedings*. Springer International Publishing, 2017.
6. M. A. S. a.-C. N. Kevin Dreßler. Radon results for oaei 2017. In *Proceedings of Ontology Matching Workshop 2017*, 2017.
7. M. Sherif, A.-C. Ngonga Ngomo, and J. Lehmann. WOMBAT - A Generalization Approach for Automatic Link Discovery. In *14th Extended Semantic Web Conference, Portorož, Slovenia, 28th May - 1st June 2017*. Springer, 2017.
8. M. A. Sherif, K. Dreßler, P. Smeros, and A.-C. Ngonga Ngomo. RADON - Rapid Discovery of Topological Relations. In *Proceedings of The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.