ROMAI J., v.8, no.2(2012), 13–25

PROBLEMS OF VISUALIZATION OF CITATION NETWORKS FOR LARGE SCIENCE PORTALS

Z. V. Apanovich

A.P. Ershov Institute of Informatics Systems, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia

apanovich@iis.nsk.su

Abstract A generally accepted way to facilitate understanding of large and complex data sets is graph visualization. In this paper we present three different methods of visualization for the citation networks, one based on hierarchical edge bundles algorithm, one implementing dynamic layered drawing, and one utilizing a geometry-based edge bundling. As test sets, we make extensive use of citation networks designed from the data sets of Open Linked Data portals.

Acknowledgement. This research is supported by RFBR grant Nr. 11-07-00388-a and SBRAS project 15/10.

Keywords: science portal, information visualization, hierarchical edge bundles, ontology, citation networks, Open Linked Data.

2010 MSC: 68P15, 68P20.

1. INTRODUCTION

Due to the fast progress of Semantic Web and its new branch of Linked Open Data, large amounts of structured information on various scientific fields are getting available. The main part of the content of scientific digital libraries and specialized portals constitute research publications, the most reliable source of information dedicated to any research area. The most active and influential researchers, organizations in which they work, and geographic locations of the research units – all this information is currently available in the rdf / xml format. This information evolves over time and rapidly grows in volume. To optimize the science management, new tools for investigation and analysis of these data are needed. A generally accepted way to facilitate understanding of large and complex data sets is graph visualization. The topic of our paper consists in several visualization methods for citation networks. Previously, we considered methods of visualization of information on scientific cooperation, represented by co-authorship networks derived from small information portals [1-2]. Our current work is a further development of this research. The data under consideration has significantly greater volume, and newly developed algorithms are presented to analyze and visualize this data.

A citation network is a network in which the vertices represent documents and the edges between them represent reference of one document to another. Citation

networks are directed: citations go from one document to another. Citation networks evolve over time as new documents are created. The citation network analysis started with the paper of Garfield et al. [10] and has been studied by many authors [9, 12]. Force-directed methods of visualization used to be the main tool of investigation for these networks.

In this paper we present three different methods of visualization for the citation networks, one based on hierarchical edge bundles algorithm, one implementing dynamic layered drawing, and one utilizing a geometry-based edge bundling. As test sets, we make extensive use of citation networks designed from the data sets of Open Linked Data portals. The paper is organized as follows. Section II discusses extracting citation networks from the content of Linked Open Data portals. Section III demonstrates some problems of the citation networks visualization by the hierarchical edge bundles method. Section IV describes some results of visualization of the citation networks by a layered dynamic method. Section V demonstrates the citation networks visualization with a geometry-based edge bundling method. Finally, section VI presents conclusion and perspectives for further work.

2. OPEN LINKED DATA AND CITATION NETWORKS GENERATION

The datasets of Linked Open Data (LOD) portals such as DBLP, Citeseer, CORDIS, NSF, EPSRC, ACM, IEEE, [4-7]. etc. have been used as a test data. These datasets are described in RDF format and have a very impressive size. For example, the data provided by the Citeseer portal consists of 8,146,852 triples, ACM portal data comprises 12,402,336 triples, and DBLP portal has granted 28,384,790 triples. A user can either download the files in RDF format, or generate data using a sparql query. All datasets of these portals are described according to a single ontology AKT Reference Ontology [5], which is the union of several ontologies (Support Ontology, Portal Ontology, Extensions Ontology and RDF Compatibility Ontology).

Portal Ontology is the main one among these ontologies, it describes such concepts as organizations, persons, projects, publications, geographic data, etc. AKT Ontology has a rather deep hierarchical structure (Fig.1). For example, to describe the publications, there exist two root classes ”Information-Bearing-Object” and ”Abstract-Information”. Subclasses of ”Information-Bearing-Object” are the classes ”Recorded-Audio”, ”Recorded-Video”, ”Publication”, ”Edited-Book”, ”Composite-Publication”, ”Serial-Publication”, ”Periodical-Publication ”and ”Book”. All individuals of the class ”Information-Bearing-Object” have a relationship ”has-publication-reference”, pointing to an object of the class ”Publication-Reference”, which is a subclass of the class ”Abstract-Information”. In turn, the class ”Publication-Reference” has as subclasses the classes ”Web-Reference”, ”Book-Reference”, ”Edited-Book-Reference”, ”Conference-Proceedings-Reference”, ”Workshop-Proceedings-Reference”, ”Book-Section-Reference”, ”Article-Reference”, ”Proceedings-Paper-Reference”, ”Thesis-

Reference" and "Technical-Report-Reference". The individuals of the class "Publication-Reference" have such relationships as: "has-date", "has-title ", "has-place-of-publication", "cites-publication-reference", etc. There exists the class "Organization", which is a subclass of the class "Legal-Agent", and the class "Legal-Agent" is a subclass of the class "Generic-Agent". The class "Person" is a subclass of the class "Generic-Agent".

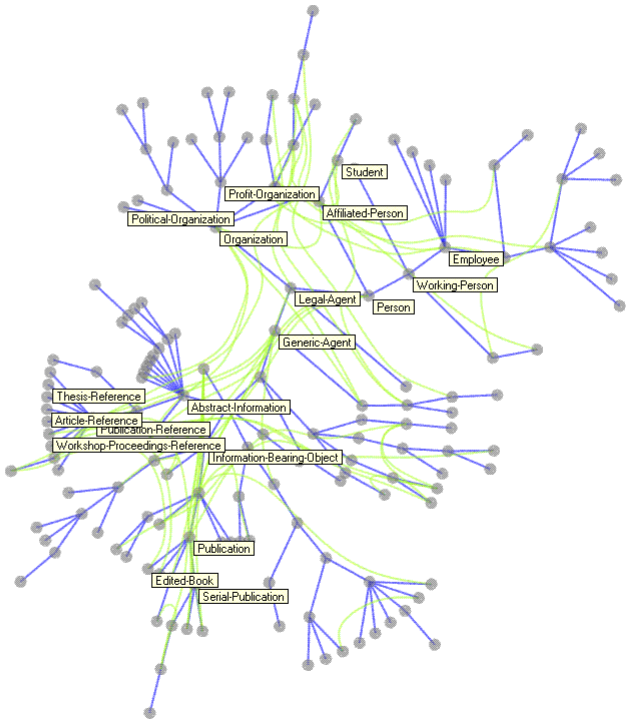


Fig. 1. AKT Reference ontology.

There are several problems of using the LOD datasets. Although all bibliographic datasets of the LOD cloud use as common vocabulary AKT Ontology, the contents of these sets are very heterogeneous and are based on very narrow subsets of this vocabulary. To describe real objects, classes of the highest level of hierarchy are normally used. For example, the classes "Publication-Reference" and "Article-Reference" are used for the description of publications while such classes as "Proceedings-Paper-Reference" are not used at all. This feature makes difficult generation of the hierarchical structure needed for applying the hierarchical edge bundles method. Also, the data sets are not complete and many attributes remain to be filled. Besides, the cita-

tion relationship (*akt: cites-publication-reference*) existing in AKT Reference Ontology, is described explicitly only for several datasets such as Citeseer and ACM [6]. However, the common mechanism of access simplifies working with these data. It is easy enough to generate a simple citation network for any storage of the LOD cloud if the publications described in these datasets have the relationship "cites-publication-reference". An example of user interface and SPARQL 1.0 query intended for citation networks generation from ACM dataset is shown in Fig. 2.

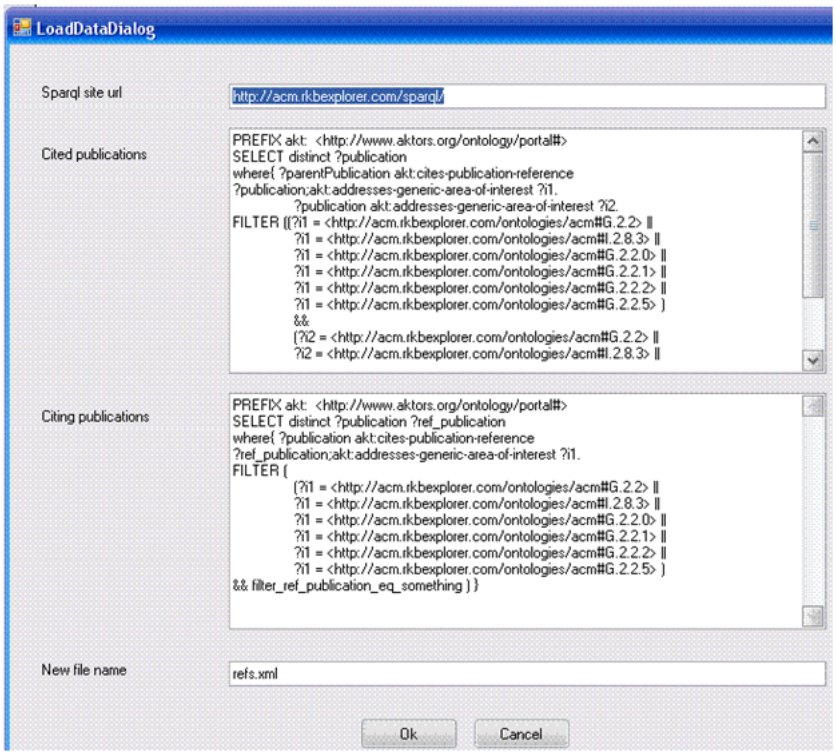


Fig. 2. An example of user interface and SPARQL 1.0 query intended for citation networks generation.

To select the desired volume of data, the query modifier LIMIT N was used. We could relatively easy extract citation networks of 20-30 thousand vertices.

3. VISUALIZATION OF CITATION NETWORKS USING THE HIERARCHICAL EDGES BUNDLES

We have started our experiments by applying already implemented hierarchical edge bundles method [11] for the citation networks visualization. This method allows a drawing of a citation network to be combined with drawings of other elements

of the portal content. It is implemented as follows. Some predefined hierarchical structure is drawn as a tree whose leafs are research papers. Then each link of the citation network is modeled as a single B-spline [14] using the control points along the shortest path in the tree layout from one leaf point to another. A test set of 561 publications on information visualization for 10 years is shown in Fig. 3. A three-level hierarchy consisting of years, conferences, and publications is depicted with balloon tree method (Fig 3(a)), and the citation links are drawn with hierarchical edge bundles method. Research papers are shown as black circles. Scientific conferences and periodical issues are shown as yellow circles. The paper's publishing years constitute the upper level of hierarchy and are shown as purple circles. The edges of the tree are shown in blue (a year includes conferences, a conference includes publications). The direction of a link from a citing publication to a cited publication is shown by progression of color from purple to green. When looking at this drawing we can easily identify the years with the largest number of publications (the years 1995 and 1996). We have slightly improved the drawing comprehensibility by depicting the citation index of papers by the radius of nodes. Since we do not want the area of drawing to grow up due to the node size enlargement, the nodes overlap is permitted. Hence the nodes visibility also depends on the citation index, as it is shown in Fig. 3(b), where the number of visible nodes and the number of the node overlaps has been reduced. Further on, users can also change the width of reference links and their opacity as a function of the citation index of the incident nodes.

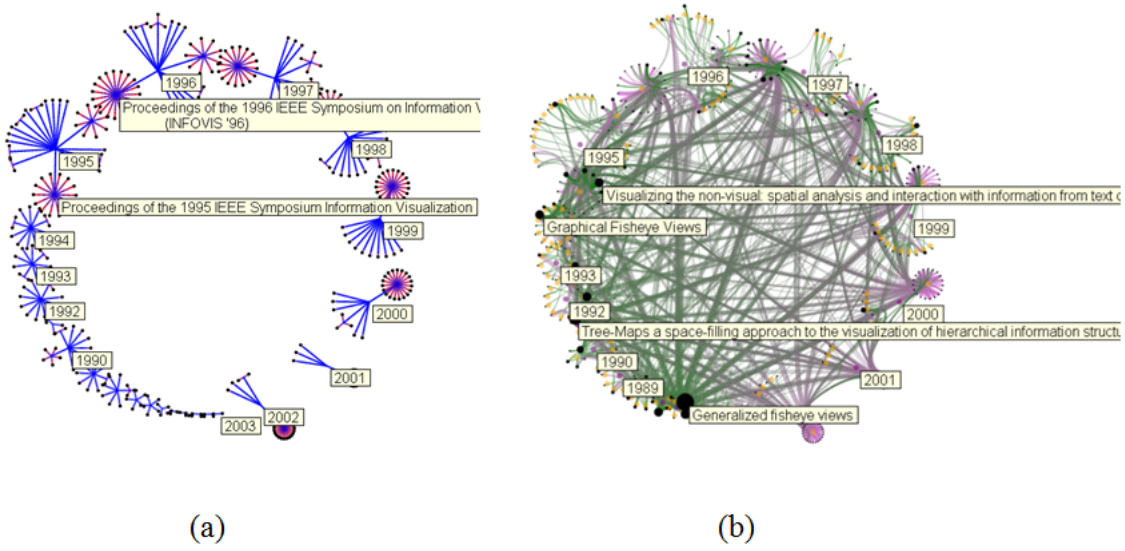


Fig. 3. Hierarchical structure and citation network. (a) A three-level hierarchy consisting of years, conferences, and publications. (b) Hierarchical edge bundles drawing of a citation network.

Some possible functions for these parameters calculation are:

$$y = (o_{\max} - o_{\min})\frac{I - I_{\min}}{I_{\max} - I_{\min}} + o_{\min} \tag{1}$$

$$y = (o_{\max} - o_{\min}) \cdot \left(1 - \sqrt{\frac{I_{\max} - I}{I_{\max} - I_{\min}}}\right) \tag{2}$$

Where I – citation index, I_{max} and I_{min} – the largest and the smallest citation index in the citation network under consideration, o_{max} and o_{min} – upper and lower bounds of values for y.

The formula (1) helps to identify the group of the most cited publications, since the node sizes are proportional to their citation indexes. The formula (2) helps to find the most cited publication since it assigns a much larger radius to the node with the highest citation index.

After the most cited papers are identified, user can choose such a node with a mouse pointer and examine its name, list of its authors and all the papers citing it as is shown in Fig. 4.

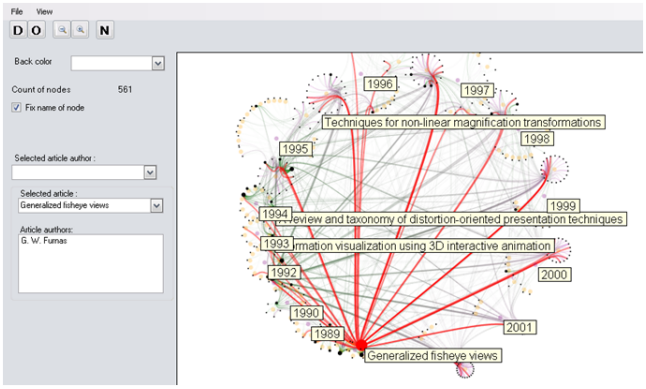
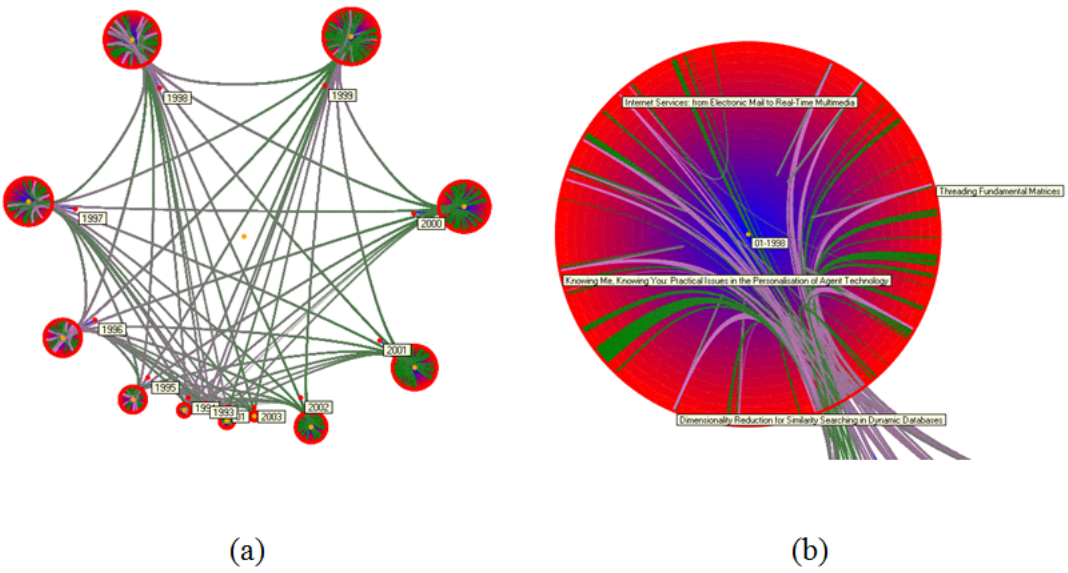


Fig. 4. The most cited paper and links citing it (shown in red).

When the size of citation network increases, the hierarchical edge bundles method gets difficult to use. For example, a drawing of a citation network of 20 000 vertices, retrieved from the Citeseer database is shown in Fig. 5. We have only managed to create a two-level hierarchy for the Citeseer dataset: the year of publication – the month of publication. That results in a drawing, rather sparse in the center (Fig. 5(a)) and very dense at the periphery (Fig. 5(b)). The time interval of these publications dataset covers the period from 1993 to 2003. The drawing permits to compare the number of publications by year: the largest number of publications of the test set falls on the years 1998 and 1989 while publications of 2003 are not numerous. Unfortunately, it is not possible to get any detailed information from this drawing. The central part of the drawing is complete graph stating that there exist citing links from

any year to any posterior year in this network. And the publications of every year are that numerous, that it is very difficult to select a vertex by the mouse pointer for further investigation.



(a) (b)

Fig. 5. A citation network of 20 000 vertices retrieved from Citeseerportal. (a) A global view, (b) one-month publications of the 1998 year.

Since it is not always possible to extract a deep hierarchy allowing the hierarchical edge bundles to be applied, we have implemented two alternative strategies for citation networks visualization:

1 To emphasize the directed nature of links in the citation networks, a dynamic layered method of visualization was implemented.

2 To reduce the visual density of drawings, a geometry-based edge bundling method was developed.

4. DYNAMIC LAYERED DRAWING OF THE CITATION NETWORKS

A citation network is a directed graph, so it is desirable that all edges are directed to one side. The direction of the edges corresponds to the chronological order of publications. Also, the citation networks are assumed to be acyclic, even if it is not always the case. For example, if a scientific paper sometimes cite work that is forthcoming but not yet published, the resulting network will have a closed loop. However, such loops are rare and short.

The construction of a layered graph drawing [13] proceeds in a sequence of standard steps:

1 **Layer assignment**. The vertices of the directed acyclic graph are assigned to layers, such that each edge goes from the left to the right. In the current implementation each layer corresponds to a publishing year, i.e. the papers, published in the same year are assigned to the same layer. We are going to parameterize the length of the time intervals in the nearest future. Edges that span multiple layers are replaced by paths of dummy vertices so that, after this step, each edge in the expanded graph connects two vertices on adjacent layers of the drawing.

2 **Crossing minimization.** The vertices within each layer are permuted in an attempt to reduce the number of crossings among the edges connecting it to the previous layer. Since finding the minimum number of crossings is NP-complete, we place each vertex at a position determined by the average of the positions of its neighbors on the previous level and then permuting adjacent pairs as long as that improves the number of crossings.

3 **Coordinate assignment.** To each vertex is assigned a coordinate within its layer, consistent with the permutation calculated in the previous step. The dummy vertices are removed from the graph and the vertices and edges are drawn.

Figure 5 shows the drawing of a citation network generated by the layered method of placement. Publishing years of papers in the citation network are shown as rectangles of different colors at the top of the image. All papers published in the same year are placed in a vertical column corresponding to this year. The edges of the network correspond to the citations. The color of each edge is identical to the color of label of the year of the citing publication. The more citation links has some publication, the more input edges has the corresponding vertex, and the greater is its radius. As a result, the citation links of publications form highly visible bundles. Four buttons at the top of the screen are used to track the dynamics of the citation network year by year. The buttons "<" and ">" are designed to move through the drawing and observe the evolution of the citation network over time. Technically, this feature is implemented by filtering vertices and edges of the citation network. Pressing the ">>" button displays the entire citation network, and the "<<" button is used to clean the drawing.

The evolution over time of a citation network of papers devoted to the graph theory is shown in Fig. 5. The four fragments of this figure show different intervals of time between 1965 and 2005. In the period from 1965 to 1989 (Fig. 6(a)) the test set of publications is dominated by the "Linear-time algorithm for isomorphism of planar graphs" paper. The corresponding vertex has the largest radius and a large brown tail of input edges. In 1993 (Fig. 6 (b)) the number of references to the papers "A data

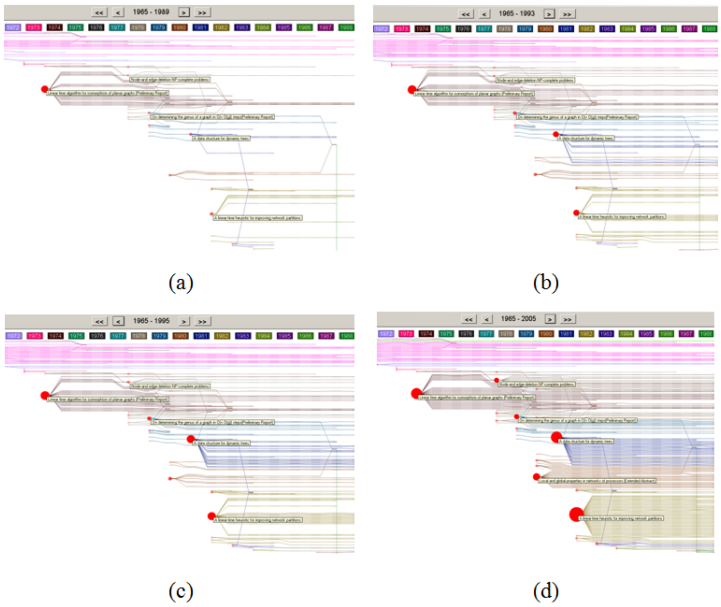
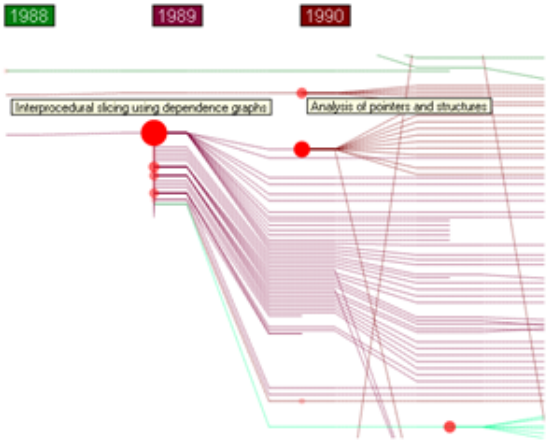


Fig. 6. The evolution of a citation network over time.

structure for dynamic trees" and "A linear-time heuristic for improving network partition" increases. In 1995 (Figure 6 (c)), these two papers have the same level of citation index as the paper "Linear-time algorithm for isomorphism of planar graphs". Finally, in 2005 (Fig. 6 (d)) the paper "A linear-time heuristic for improving network partition" gets the most cited. Hence, data sets are better comprehended due to the dynamic visualization.

It is also possible to observe growing interest to the paper "Node-and-edge-deletion NP-complete problems" that refers to the previously dominating paper "Linear-time algorithm for isomorphism of planar graphs", i.e. a chain of highly cited related publications arises.

Besides, this visualization method helps to detect errors and inaccuracies in bibliographic data. Fig. 6(a) shows a fragment of a citation network generated for the ACM dataset on the time interval from 1988 to 1990. A brown link connects the node of the "Analysis of pointers and structures" paper published in 1990 and the "Interprocedural slicing using dependence graphs" paper published in 1988. Since the color of the link corresponds to the year 1990 it should mean that the arc is oriented backward and a paper published in 1988 cites a paper published in 1990. By checking the ACM dataset (Fig. 6(b)) we have discovered that the paper "Interprocedural slicing using dependence graphs" has several dates of publication and the corresponding node is placed in the layer of the earliest date of publication.



(a)

Analysis of pointers and structures	akt:has-date	1990-01-01
Interprocedural slicing using dependence graphs	akt:cites-publication-reference	Analysis of pointers and structures
Interprocedural slicing using dependence graphs	akt:has-date	1988-01-01
Interprocedural slicing using dependence graphs	akt:has-date	1988-07-01
Interprocedural slicing using dependence graphs	akt:has-date	1990-01-01

(b)

Fig. 7. Datasets inaccuracies. (a) Backward link representing a paper published in 1988 citing a paper published in 1990. (b) Multiple dates of publishing for a paper.

The main problem with the conventional layered method is that drawings get overloaded very quickly and using the filtering removes irrelevant papers but distorts reality: irrelevant publications are the major contributors in determining the significance of other publications. The hierarchical edge bundles method is not applicable in the absence of external hierarchical structure. Therefore we have implemented an algorithm , which can reduce the drawing density by forming bundles of edges based on their own geometry, and not introduced from outside.

5. GEOMETRY-BASED EDGE BUNDLING METHOD.

The main idea of the geometry based edge bundling method [8] is to reduce the visual clutter of the image by bending the edges through a special control grid without changing the original locations of graph vertices. This method proceeds as follows:

1 Generate a rectangular NxN grid and put it over a graph drawing constructed in any way.

2 For each grid cell, calculate the main direction of the edges crossing the cell.

3 Merge into zones the adjacent cells having directions that differ by no more than the threshold value .

4 Calculate the basic direction and normal vector to the main direction of each zone.

5 Calculate the points of intersection of the normal segments with zones' boundaries .

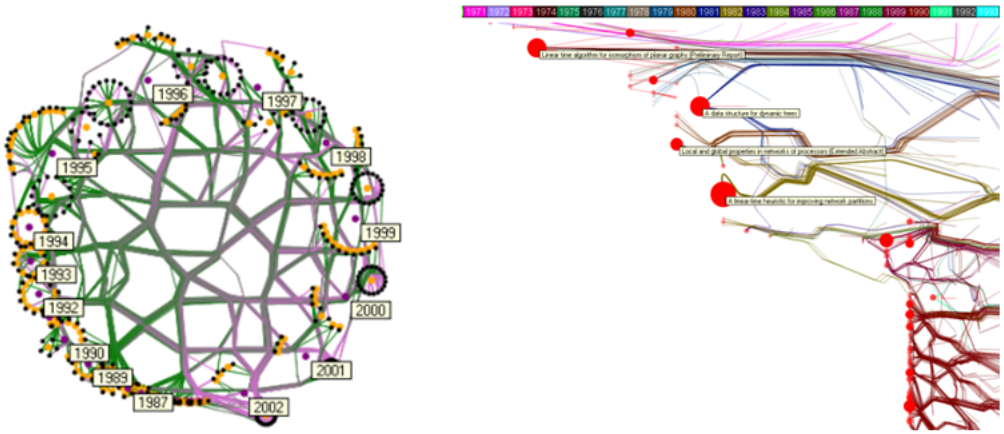
6 Use the resulting points to construct a triangulation.

7 Find for each edge of the constructed triangulation the point of intersection with the edges of the original graph drawing. Calculate the centers of these points.

8 Use the resulting points as control points of b-splines.

Fig. 8(a) demonstrates applying the geometry based edge bundling strategy to the drawing obtained by circular drawing method from Fig.2(b). Fig. 8(b) shows applying the geometry based edge bundling strategy to the drawing obtained by layered visualization method from Fig.5(d).

No doubt, due to this methodology the drawing congestion is reduced. But at this stage, there are more questions with this method than answers. How to choose the best direction for a rectangular grid? How does the direction of the edge bundles depend on the size of the grid? How to choose the best edge direction within each zone in function of the underlying visualization method? Nevertheless, we hope to develop this method to the point where it can be used to examine trends in a research field.



(a) (b)

Fig. 8. The application of the geometry based edge bundling strategy to the drawing obtained by layered visualization method. (a) Applying the geometry based edge bundling strategy to the circular drawing. (b) Applying the geometry based edge bundling strategy to the layered drawing.

6. CONCLUSION

In this paper we have demonstrated three visualization methods of citation networks generated for datasets of Linked Open Data portals. These drawings are rather helpful for understanding of datasets of large volumes. Also they enable users to observe the evolution of datasets over time. In the nearest future we are going to apply the previously developed clustering methods for the citation networks analysis and to compare the results obtained by the two groups of methods.

References

[1] Z. V. Apanovich, T.A. Kislicyna, *Extending the subsystem of content visualization of informational portal by visual analytics tools*, Complex systems control and modeling problems: Proceedings of the XII International Conference. (Samara, Russia, 2010), 518-525.

[2] Z. V. Apanovich, P. S. Vinokurov, *Ontology based portals and visual analysis of scientific communities*, First Russia and Pacific Conference on Computer Technology and Applications(Vladivostok, Russia, 2010), 7-11.

[3] C. Bizer, T. Heath, T. Berners-Lee, *Linked Data - The Story So Far*, Int. J. Semantic Web Inf. Syst., **5**, 3 (2009), 1-22.

[4] *Linked Open Data datasets: http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets*.

[5] *AKT ontology description: http://www.aktors.org/ontology*.

[6] *CiteSeer dataset: http://citeseer.rkbexplorer.com/*.

[7] *DBLP dataset: http://dblp.rkbexplorer.com/.*

[8] W. Cui, H. Zhou, H. Qu, P. C. Wong, X. Li, *Geometry-Based Edge Clustering for Graph Visualization*,IEEE Transactions on Visualization and Computer Graphics, **14**, 6 (2008).

[9] Ch. Chen, I-Y.Song, Zhu W.Weizhong, *Trends in conceptual modeling: Citation analysis of the ER conference papers (1979-2005)*, Proceedings of the 11th International Conference on the International Society for Scientometrics and Informatrics. CSIC, (Madrid, Spain, 2007), 189-200.

[10] E. Garfield, I.H. Sher, R.J.Torpie, *The Use of Citation Data in Writing the History of Science*, Philadelphia: The Institute for Scientific Information, (1964). http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf.

[11] D. Holten, *Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data*, IEEE Transactions on Visualization and Computer Graphics, **12**, 5(2006), 741-748.

[12] H. Small,*Visualizing Science by Citation Mapping*, Journal of the American Society for Information Science, **50**, 9 (1999), 799-813.

[13] K. Sugiyama, S.Tagawa, M.Toda, *Methods for Visual Understanding of Hierarchical System Structures*, IEEE Trans. Systems, Man, and Cybernetics, (1981), 109-125.

[14] *http://ru.wikipedia.org/wiki/B-spline*.