Systematic retrieval of scientific literature based on citation relations: Introducing the CitNetExplorer tool

Nees Jan van Eck and Ludo Waltman

Centre for Science and Technology Studies, Leiden University, The Netherlands

{ecknjpvan,waltmanlr}@cwts.leidenuniv.nl

Abstract. We consider the problem of scientific literature search, and we suggest that citation relations between publications can be very helpful in the systematic retrieval of scientific literature. We introduce a new software tool called CitNetExplorer that can be used for citation-based scientific literature retrieval. To demonstrate the use of CitNetExplorer, we employ the tool to identify publications dealing with the topic of community detection in networks. Citation-based scientific literature retrieval can be especially helpful in situations in which one needs to obtain a comprehensive overview of the literature on a certain research topic, for instance in the preparation of a review article.

1 Introduction

Searching for scientific literature can be a time-consuming activity. This is especially the case when comprehensive search results are needed. For instance, performing a literature search for a review article that intends to provide a complete overview of all literature on a certain research topic often takes a lot of time. Likewise, it may require a substantial amount of effort to systematically search through a body of literature in order to determine in an accurate way the origins of a certain scientific idea or concept.

In many cases, citation relations between publications are helpful when searching for scientific literature. For instance, when one has found a relevant publication, additional relevant publications may be identified by checking the publications cited by the publication that has already been found. When additional relevant publications have been found, one may in turn check the publications cited by these publications. In this way, citation relations may help to identify significant numbers of relevant publications. Citation relations may also be used in a forward rather than a backward direction in time. In that case, when one has found a relevant publication, additional relevant publications may be identified by looking for publications citing the publication that has already been found.

Although citation relations can be very helpful in the retrieval of scientific literature, following large numbers of citation links is quite time consuming. To some degree, bibliographic databases such as Web of Science, Scopus, and Google Scholar can be used to increase the efficiency of citation-based literature retrieval. In Web of Science and Scopus, for instance, one can easily look up both the publications that cite and the publications that are cited by a given publication. Nevertheless, bibliographic databases offer only limited support to researchers who want to use citation relations for systematic literature retrieval. These databases do not give a visual overview of the citation relations within a set of publications. (Web of Science does provide a so-called 'citation map' visualization, but the information offered by this visualization is limited.) They also do not support sophisticated citation-based search queries. For instance, when one has available a set of relevant publications, there is no easy way to identify all publications that cite or are cited by at least three publications in the set.

In this paper, we introduce a software tool for analyzing and visualizing citation networks. The tool is called CitNetExplorer and is available at www.citnetexplorer.nl. We have developed CitNetExplorer primarily for the purpose of studying the evolution of the scientific literature in a research field [1,2]. However, CitNetExplorer can also be used for systematic literature retrieval based on citation relations. The tool may serve as a prototype for citation-based information retrieval functionality that could be implemented in bibliographic databases such as the ones mentioned above.

The organization of this paper is as follows. In Section 2, we discuss CitNetExplorer in more detail. In Section 3, we demonstrate the use of CitNetExplorer for the purpose of systematic literature retrieval. Finally, in Section 4, we summarize our conclusions.

2 CitNetExplorer

CitNetExplorer, which is an abbreviation of 'citation network explorer', offers the following functionality:

- Searching for publications based on author, journal, title, etc. This is the traditional way of searching for scientific literature.
- Visualization of the citation network of a set of selected publications.
- Identification of sets of related publications by identifying connected components, clusters, or core sets in a citation network.
- Identification of predecessors or successors of a set of selected publications. Predecessors are publications cited by the selected publications. Successors are publications citing the selected publications. The minimum number of citation relations required for a publication to be identified as a predecessor or successor is determined by the user.
- Identification of publications located on a citation path between selected publications. These publications are referred to as intermediate publications.
- Drilling down into a citation network. This means reducing the number of selected publications.

- Expanding a citation network. This means increasing the number of selected publications.

A screenshot of CitNetExplorer is presented in Fig. 1. The user interface of CitNetExplorer consists of a left and a right panel. The right panel shows the selected publications, either by displaying the citation network of the publications or simply by displaying the publications in a list. The left panel offers some general information on the selected publications (i.e., number of publications, number of citation relations, and time period), it provides access to a number of important parameters related to CitNetExplorer's drill down functionality, and it displays detailed bibliographic information on a single highlighted publication. In the top part of CitNetExplorer's user interface, a number of buttons are available. These buttons can be used to perform some key operations, in particular the drill down and expand operations. They can also be used to navigate back and forth between different sets of selected publications and the corresponding citation networks. This is somewhat similar to the back and forward buttons in a web browser.

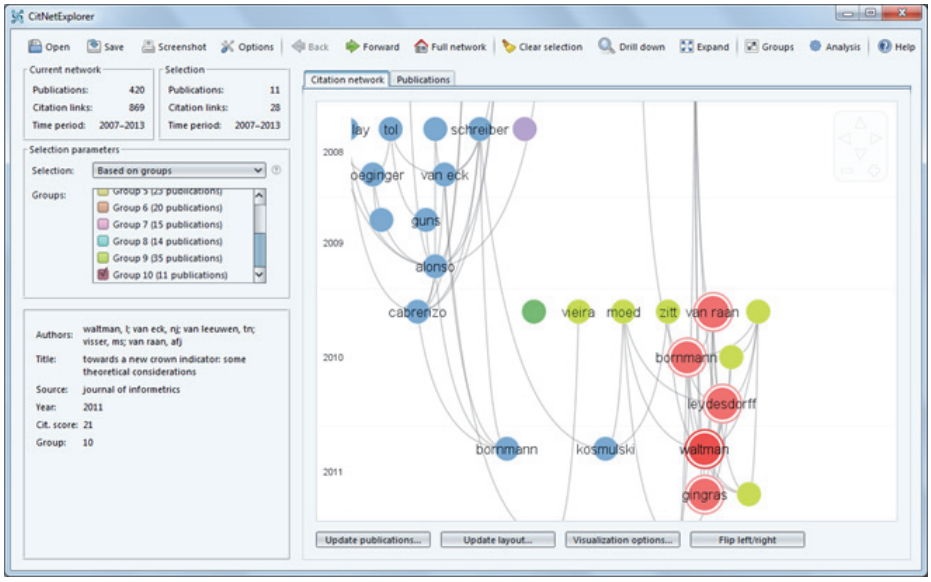

Fig. 1. Screenshot of CitNetExplorer.

An example of a citation network visualization produced by CitNetExplorer is shown in Fig. 2. In this visualization, the circles indicate publications, labeled by the last name of the first author, and the curved lines indicate citation relations. The vertical dimension represents time, with more recent years being located below earlier years. The location of a publication in the vertical dimension is determined by the year in which the publication appeared. In the horizontal dimension, publications are located close to each other if they are closely connected in the citation network. Publications that do not have a close connection in the citation network are located further

away from each other. Similar to software tools for exploring geographical maps (e.g., Google Maps) and also to our own VOSviewer tool [3], CitNetExplorer offers zooming and panning (scrolling) functionality. This can be used to explore specific areas in the visualization of a citation network in more detail. CitNetExplorer can handle very large citation networks, with millions of publications and tens of millions of citation relations. However, in the visualization of a citation network, only a selection of all publications are displayed. By default, the 40 most frequently cited publications in a citation network are included in a visualization.

The idea of a software tool for visualizing citation networks of scientific publications is not new. This idea can also be found in the HistCite tool for 'algorithmic historiography' developed by Eugene Garfield and colleagues [4] (see www.histcite.com). However, in addition to visualizations of citation networks, CitNetExplorer also offers extensive functionality for exploring and analyzing citation networks. Given the large size of the citation networks that can be handled by CitNetExplorer, a crucial element of CitNetExplorer is the functionality for drilling down into a citation network. This functionality for instance makes it possible to start at the level of a full citation network consisting of several millions of publications and to then gradually drill down into this network until a small subnetwork has been reached including no more than, say, 100 publications, all dealing with a specific topic of interest. While drilling down reduces the number of publications that are being considered, expansion has the opposite effect and increases the number of publications under consideration. When one finds oneself in a given subnetwork, CitNetExplorer makes it possible to expand the subnetwork by adding publications cited by or citing to one or more publications in the subnetwork. In addition to drill down and expand functionality, CitNetExplorer also provides a number of options for analyzing citation networks. Examples include options for clustering the publications in a citation network and for identifying the core publications in a citation network. As we will demonstrate in the next section, CitNetExplorer's functionality for exploring and analyzing citation networks enables a more sophisticated approach to citation-based retrieval of scientific literature compared with the possibilities provided by bibliographic databases such as Web of Science, Scopus, and Google Scholar.

CitNetExplorer has been implemented in Java and therefore works on any system that offers Java support. The tool can be downloaded from www.citnetexplorer.nl. Ideally, one would like a tool such as CitNetExplorer to be linked directly to a bibliographic database. Although this is technically feasible, the proprietary nature of bibliographic databases does not allow us to create such a direct link. Users of CitNetExplorer therefore need to take care themselves of obtaining the publication and citation data required by CitNetExplorer. They can then use their data as input to CitNetExplorer. To support users in these steps, CitNetExplorer is able to directly process data that has been downloaded through the web interface of Web of Science. However, because of restrictions imposed by the Web of Science web interface (i.e., data can be downloaded for at most 500 publications at a time), only small and medium-sized citation networks can be handled in this way. To work with large citation networks, for instance including several millions of publications, one needs to have direct and unrestricted access to Web of Science or another bibliographic database. Most re-

searchers do not have this type of access to a bibliographic database. In that case, the use of CitNetExplorer is limited to small and medium-sized citation networks, perhaps including at most several tens of thousands of publications. Of course, even with this limitation, CitNetExplorer can still serve as a prototype for citation-based information retrieval functionality that could be implemented at a large scale in bibliographic databases.

3 Demonstration

To demonstrate the use of CitNetExplorer for citation-based retrieval of scientific literature, we take as an example the topic of community detection in networks (for a review of the literature on this topic, see [5]). We use this topic because we are familiar with it and because we are therefore able to assess which publications are relevant and which are not. Our focus is on the community detection literature in the field of physics. We do not consider publications from other fields, but we do take into account publications in multidisciplinary journals such as *Nature* and *Science*.

For the purpose of this demonstration, CitNetExplorer receives as input a citation network of publications in physics journals and multidisciplinary journals in the period 1998–2012. The data is taken from our institute's in-house version of the Web of Science database. The citation network includes about 1.8 million publications and about 15.1 million citation relations.

To retrieve in a systematic way the literature on community detection, we take the following five steps in CitNetExplorer:

1. First, we search for all publications whose title matches '*communit* detect*' or '*detect* communit*'. This yields 113 publications. We drill down to the subnetwork consisting of these 113 publications.
2. Some of the publications in our subnetwork are false positives and do not deal with community detection. An example is a publication with the title 'Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities'. We assume that false positives have no citation relations with publications that have been correctly identified as being relevant to the topic of community detection. We therefore identify the largest connected component in our subnetwork and remove from the subnetwork the seven publications not included in this component.
3. We now have a subnetwork consisting of 106 publications. The visualization of this subnetwork produced by CitNetExplorer is shown in Fig. 2. Recall from the previous section that only the 40 most frequently cited publications are included in the visualization. We observe that some important early publications on community detection are not included in our subnetwork. We therefore expand the subnetwork with 33 publications that are each cited by at least ten publications already included in the subnetwork.
4. Within our subnetwork of 139 publications, some of the early publications deal with networks but not specifically with community detection. An example is a publication with the title 'Statistical mechanics of complex networks'. We manually

identify six publications not dealing specifically with community detection, and we remove these publications from the subnetwork.

5. Finally, we expand our subnetwork of 133 publications with 439 publications that each cite at least four publications already included in the subnetwork. A visualization of the subnetwork that is obtained after the expansion is presented in Fig. 3.

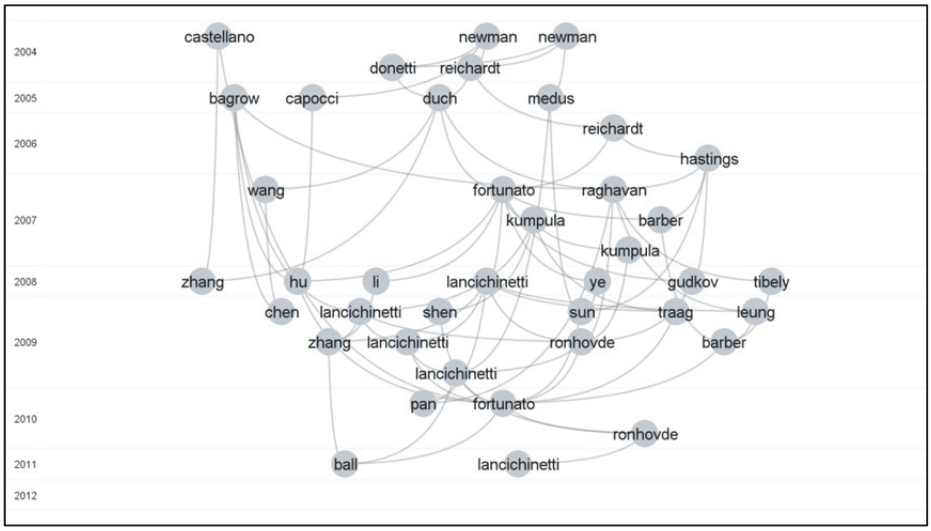

Fig. 2. Visualization of a citation network of 106 publications on the topic of community detection in networks. Only the 40 most frequently cited publications are included in the visualization. To reduce the number of lines in the visualization, some citation relations are not displayed. For instance, if publication A cites both publication B and publication C and if publication B also cites publication C, then the citation relation from publication A to publication C is not displayed. Only citations relations included in the so-called transitive reduction of the citation network are displayed.

After taking the above five steps, we end up with a set of 572 publications. The large majority of these publications turn out to deal with community detection, although some less relevant publications are included as well. These are for instance review articles that cite a number of publications on community detection even though this topic is not their main focus. Some further manual work would be needed to get rid of these less relevant publications. What is important is that we may expect a very large share of all publications dealing with community detection to be included in our set of 572 publications. Relevant publications may have been missed only if they cite no more than three publications out of the 133 publications that we start with in step 5. Since these 133 publications include a number of the most important publications on community detection, this does not seem very likely.

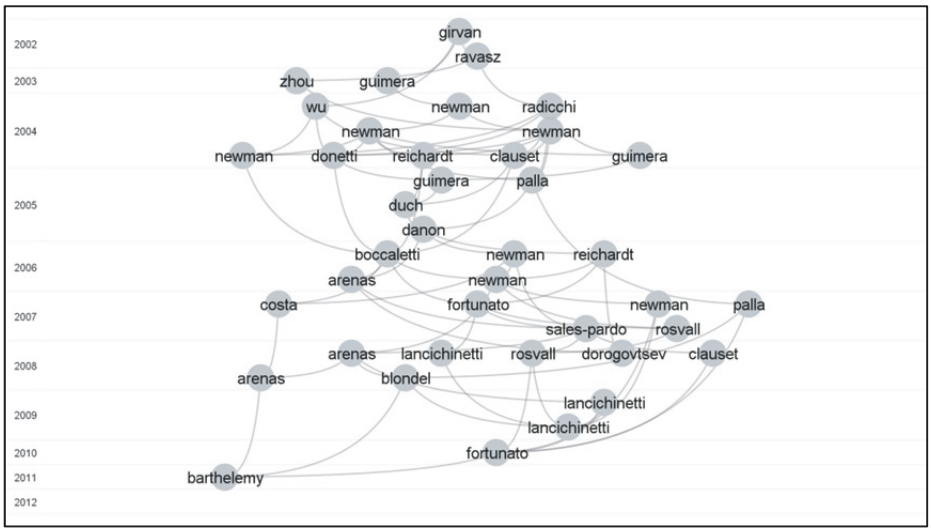

Fig. 3. Visualization of a citation network of 572 publications on the topic of community detection in networks. Only the 40 most frequently cited publications are included in the visualization. Like in Fig. 2, to reduce the number of lines in the visualization, only citations relations included in the so-called transitive reduction of the citation network are displayed.

4 Conclusions

We have discussed the idea of using citation relations between publications to support the systematic retrieval of scientific literature. We have introduced the CitNetExplorer software tool that can be used for this type of citation-based information retrieval. We believe that our approach is especially helpful in situations in which one needs to get a comprehensive overview of the literature on a certain research topic, for instance when working on a review article or when tracing the origins of a scientific idea or concept. Although our focus has been on the retrieval of scientific literature, we note that our approach may also be useful in the retrieval of patents.

The retrieval performance of our approach has not yet been evaluated in a formal way, for instance based on precision and recall measures. However, during the development of the CitNetExplorer tool, we have used the tool extensively to test its usefulness in systematic literature search. We consider the results of these tests to be encouraging. In a number of cases, for instance, CitNetExplorer has been helpful in the identification of interesting publications that so far we had not been aware of. On the other hand, one should of course keep in mind the limitations of the use of citation relations in scientific literature retrieval. Obviously, when two scientific communities both work on the same topic but do not cite each other's publications, the use of citation relations will be of no help in identifying the connection between the work of the two communities.

Finally, we briefly mention some related work. There are at least two areas in which related work can be found. On the one hand, there is the area of information

visualization and human-computer interaction. Examples of related work in this area include work on the visualization of citation networks to support literature reviewing [6] and work on the combined use of citation network visualization and natural language processing to support researchers in the exploration of a body of scientific literature [7]. On the other hand, there is the area of information retrieval. In this area, researchers study the problem of citation recommendation (e.g., [8]). This is about the situation in which someone is working on a publication and is looking for related publications to cite. These related publications can be identified based on textual similarity or based on citation relations, and especially in the latter case there is a relationship to our work.

References

1. Van Eck, N.J., Waltman, L.: Visualizing bibliometric networks. Manuscript in preparation (2014)
2. Van Eck, N.J., Waltman, L.: CitNetExplorer: A new software tool for analyzing and visualizing citation networks. Manuscript in preparation (2014)
3. Van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538 (2010)
4. Garfield, E., Pudovkin, A.I., Istomin, V.S.: Why do we need algorithmic historiography? Journal of the American Society for Information Science and Technology 54, 400–412 (2003)
5. Fortunato, S.: Community detection in graphs. Physics Reports 486, 75–174 (2010)
6. Chou, J.-K., Yang, C.-K.: PaperVis: Literature review made easy. Computer Graphics Forum 30, 721–730 (2011)
7. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B.: Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. Journal of the American Society for Information Science and Technology 63, 2351–2369 (2012)
8. Strohman, T., Croft, W.B., Jensen, D.: Recommending citations for academic papers. In: 30th Annual International ACM SIGIR Conference, pp. 705–706. ACM, New York (2007)