Identification of Hidden Structures in the Reference Network of E-Assessment Systems

Viola Weickenmeier[1] and Sven Strickroth[2]

Abstract: Using e-assessment systems, such as automated graders or automated feedback systems, is quite common in programming courses. Various tools have been developed to support students and teachers in learning and teaching programming. For overviews, comparisons of tools, and the identification of categories, a number of literature surveys and reviews have been carried out manually. This study does not try to find new systems, but uses (social) network analysis and citation data to identify important/influential systems as well as connections and similarities between systems in an existing corpus. The references were automatically extracted from the scientific publications related to these systems. Using these analyses, different types of communities and influential systems could be identified. Furthermore, there seem to be two types of references, those that simply mention a system and those that discuss specific features in more detail.

Keywords: reference analysis; e-assessment systems; (social) network analysis; modularity clustering

1 Motivation

Systems to automatically assess and provide feedback on student submissions are often used in programming courses. Over the past decades, many systems with different features and goals have been developed to support instructors and students by many researchers [SS22].

Literature reviews are a common practice for gaining insights into current approaches and topics. Typical goals of such reviews are to identify relevant literature and to analyze the described approaches and results often regarding very focused aspects. Sometimes there are meta-data analyses such as “How many papers were published per year?” or attempts to identify different categories of such systems (e. g., [SFB16; SS22]).

There is, however, no analysis of the reference network resulting from linking citations between scientific publications or associated systems for e-assessment systems. A network analysis can provide a holistic view, considering the relationships and contextual information surrounding citations [WVN10]. By identifying clusters of closely connected systems, central systems, influential authors, hidden patterns, and latent research communities can be uncovered. Such insights can foster dialogue, and promote the exchange of ideas.

In this paper preliminary results on a reference network analysis are presented to answer the following research questions: What are the most influential systems?, What (type of) clusters can be identified?, and Can regional communities be identified by their citations?

[1] LMU Munich, Institut für Informatik, Oettingenstr. 67, 80538 München, viola.weickenmeier@campus.lmu.de
[2] LMU Munich, Institut für Informatik, Oettingenstr. 67, 80538 München, sven.strickroth@ifi.lmu.de, https://orcid.org/0000-0002-9647-300X

doi:10.18420/abp2023-5

This paper is organized as follows: First, related work is reviewed and the research gap identified. Second, the methodology of this research is described and the results are presented. The paper concludes with a discussion of the results, a conclusion, and an outlook.

2 Related Work

There are several literature reviews on e-assessment systems, automatic graders, ITS, and hint systems that analyze student submissions, provide feedback, and/or grade them ([SS22] provides an overview): There are reviews such as Keuning et al. [KJH16] which categorizes 69 tools based on the type of feedback (cf. [Na13]). Saito et al. [Sa17] reviewed 43 tools and proposed a pedagogical taxonomy. De Souza et al. [SFB16] reviewed 30 systems and provided a classification on (semi-)automatic assessment type, student or teacher centered approaches and speciality (contest, quiz, software testing). There are also reviews focusing on specific aspects such as usage of AI in assessment systems [Le13]. All categories are developed manually and might not unveil hidden ones. Several approaches exist to discover topics or categories such as an analysis of the used keywords or topic modelling (e. g., [GS04]). These text mining methods, however, rely on the wordings used in the papers.

All these approaches ignore the reference network. Commonly used for linking data are author-based analyses such as co-authorship [GS05]. Here, two authors are linked if they published at least one paper together. Major limitations are, however, that collaborations are not represented sufficient enough through co-authorships [GS05] especially in this domain as most e-assessment systems seem to be developed independently (often as part of theses) [SS22], and are not suitable for finding cross-country connections between systems [Li05].

With the increasing use of (Social) Networking Analysis for investigating structures, there are different methods tested and validated for comparing scientific papers such as citation analysis [BK10]. This method focuses on references instead of content or authorship and is capable of finding the most important papers of a research topic [Wh04] as well as clustering or mapping of bibliographic data [WVN10].

Until now, there has been no work published that focuses on citation analysis to investigate structures and connections between e-assessment systems. Hence, there is no analysis of the reference network to get deeper insights, e. g. to find (new) classifications or to identify influential systems. It also remains unclear whether existing categorizations also reflect research communities or whether there are different research communities working on similar topics without knowing from each other. This research gap is addressed in this paper.

3 Method

The goal of this research is to analyze the reference-network of the papers and systems included in the corpus of [SS22]. To build the reference network, a tool was developed that

automatically analyzes PDF files and extracts the references. These data are matched with the meta-information in form of a BibTeX-file that contains all publications for all systems of in the corpus. The data used here, is based on an optimized version of the tool developed in [We23], and will be made available for download (also the data).[3] The nodes are the systems (no paper deals with multiple systems). The network is a directed graph, however, it was interpreted as an undirected graph if not stated otherwise. The graph analysis was conducted using Gephi.[4] References to the mentioned systems can be found in [SS22].

For the first part of the analysis, techniques from graph theory and (social) network analysis are used. Apart from general characteristics such as degrees, centrality is measured. Centrality indicators assign a ranking to nodes within a graph corresponding to their position in the graph [Ko05]. Two centrality measures are calculated: First, the Closeness Centrality (CC) is calculated, which is the inverse of the sum of the lengths of all shortest paths between a node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes. Second, the Betweeness Centrality (BC) is calculated, which is the number of shortest paths from every node to every other node that a specific node is on. Hence, the BC indicates how often a node acts as a bridge along the shortest paths between different systems. For the second part of the analysis, the modularity method was used for clustering based on the works [Bl08; LDB08] as implemented in Gephi.

4 Results

The reference graph consists of 178 nodes and 340 edges (i. e., references; self-loops are excluded). The graph is not fully connected and consists of 32 components. There is one component comprising 146 systems, one component consisting of the two systems `HackerRank` and `Senecode`, and 30 unconnected systems. The diameter (longest distance between two connected nodes) is 8, the density is 0.02, and the average path length is 3.3. Median in and out degrees are both 1 and average in and out degrees are both 1.9. The most referenced systems (i. e., most incoming edges) in the graph are `WebCat` with 30 references, followed by `Coursema(rk|st)er` with 29 references (cf. Tab. 1). 62 systems have no outgoing edges and 81 no incoming edges.

Tab. 1 shows the top-10 systems ordered by their reference frequency (left) and betweeness centrality (middle). The top-3 systems are the same in both cases (namely `WebCat`, `CourseMarker` and `Singh`'s system), however, only three further systems are in both top-10 (namely `AutoLEP`, `Mooshak`, and `Fitchfork`). All closeness centralities are below .5 except for Senecode and Hackerrank (the cluster of two, hence CC=1). On the right side Tab. 1 shows the top-10 systems ordered by the directed betweeness centrality. Again, WebCat is the system with the highest value, however, the following two systems are of German origin (`JACK` and `ASB`). Overall, `Praktomat` is the German system with the most references (11) and `JACK` (referenced 6x) with the largest closeness (.38) and betweeness (703) centralities.

[3] `https://systemscorpus.strickroth.net`
[4] `https://gephi.org/`, version 0.10

System	deg	CC	BC	System	BC	System	dir. BC
WebCat (2003)	30	0.47	3468	WebCat	3468	WebCat	1291
CourseMarker (1998)	29	0.43	1821	CourseMar...	1821	JACK	1238
Singh-name (2013)	19	0.38	1648	Singh-name	1648	ASB	982
Mooshak (2001)	16	0.38	796	Fitchfork	799	CourseMar...	976
AutoLEP (2004)	12	0.38	607	Mooshak	796	eduComp...	947
Praktomat (1999)	11	0.32	342	JACK	703	Mooshak	711
JITS (2003)	10	0.33	498	Ask-Elle	641	PABS	468
Fitchfork (2006)	9	0.40	799	AutoLEP	607	eduJudge	245
Progtest (2011)	7	0.39	386	DsLab	595	Fitchfork	211
GAME (2004)	7	0.33	299	Galan-name	505	Progtest	207

Tab. 1: Overview of the top-10 systems (first usage year in parenthesis, cf. [SS22]) sorted by the in degree (left), betweeness centrality (middle), and directed betweeness centrality (right)

The modularity method for clustering formed 39 clusters (modularity: 0.53). The largest of the 32 graph components consists of 8 clusters. In total, 9 out of the 39 clusters have at least two nodes, the biggest cluster consists of 29 nodes. The average number of systems is 4.6, the median is 1. Without the clusters consisting of exactly one system the average number of systems is 9. In the following three of these clusters are examined in more detail.

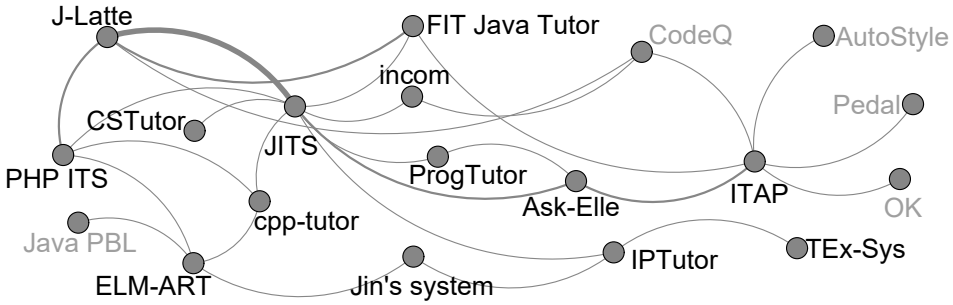

Fig. 1: Cluster 15 with 19 systems of which 14 are ITS (the nodes are systems)

In cluster 15 (cf. Fig. 1) out of the 19 systems 14 characterize themselves as Intelligent Tutoring Systems (ITS) and all but three use "intelligent tutor(ing system)" as a keyword in at least one paper. The most important system seems to be `JITS` here. An interesting case is `OK`, as this system self-characterizes as a hint system (the only one in the corpus) and not as an ITS. In the data set 21 systems self-characterize as ITS and 18 systems have at least one publication with the keyword "intelligent tutor(ing system)". The other systems are all distributed to different clusters. The filtered graph for self-characterized ITS is not fully connected. It contains, however, a subgraph of 14 connected systems (all black labelled nodes in Fig. 1) and 7 isolated systems (`COALA`, `iList`, `Burke`'s system, `Prutor`, `AWAT`, `M-PLAT`, and `WebIDE`). When the keyword "intelligent tutor(ing system)" is used as a additional filter,

J-Latte (no keywords), ITAP (“programming tutors”), ProgTutor (“tutoring”), and IPTutor (“programming tutor”) “vanish”, hence TEx-Sys is not connected any more to the other ITS.

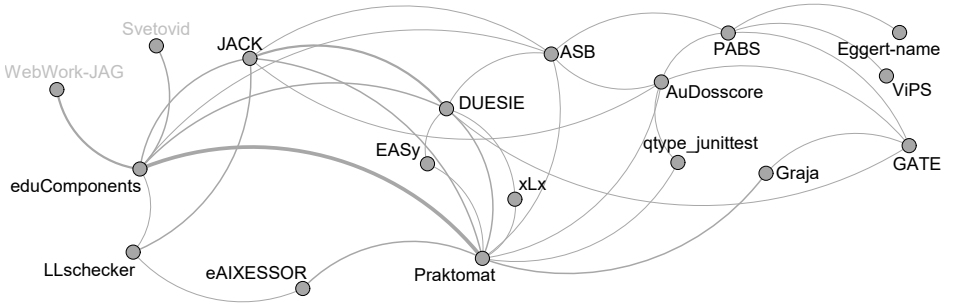

Fig. 2: Cluster 26 with 18 systems of which 16 have German origin (the nodes are systems)

Cluster 26 (Fig. 2) consists of 18 systems out of which 16 originate from German authors. There is also one system from Serbia (Svetovid) and WebWork-JAG from the USA included in this cluster. The most important system seems to be Praktomat. In the full data set there are 32 systems with a German origin. Half of the systems are included in this cluster. The German systems are included in 12 different clusters, five with more than one system (cluster 0: 1 system of 14 systems, 2: 2/29, 15: 3/18, 19: 1/17, 21: 2/24, 26: 16/18). There are three German ITS (ELM-ART, FIT Java Tutor, and incom) that are in the ITS cluster (cf. Fig. 1). Interestingly, the reference network of the 32 German systems is not connected if filtered. Unconnected (directly) to to other German systems are 14 systems including six systems Subato, OnExSy, IT4all, Burke’s system, ViPLab, and AutoTool that have no connection to any other system.

Cluster 28 consist of four systems namely MOE, Pythia, Code-Hunt, and Pex4Fun that form a “linked list” (in this order). Interestingly, the papers of Code-Hunt and Pex4Fun often share the very same main author and cover a similar approach. Code-Hunt and MOE share the keyword “(learning|educational) platform”.

Apart from the clusters, there are 15 systems in the corpus originating from Spain and 7 from Portugal which are the largest communities in Europe after Germany. The Spanish systems can be found in mainly three clusters (cluster 0: 4 systems of 14 systems, 21: 6/24, and 22: 3/21). Interestingly, the filtered graph contains a connected subgraph of 12 systems and three isolated systems (SAC, M-PLAT, and Munoz’s system). Four of the Portuguese systems are in cluster 21 (consisting of 24 systems) and the other three in three different clusters (2: 1/14, 6: 1/1, 22: 1/21). Only the systems code.org (isolated node, cluster 6) and CodeInsights are isolated, the others are all connected to the system Mooshak.

5 Discussion

A major limitation of this study is the restriction to papers and systems included in the corpus. The reason is the otherwise lack of annotated meta-data of publications. The way in which the corpus was compiled (cf. [SS22]) is probably also the reason for the quality of the clusters and the large number of isolated systems, as the vast majority of papers contain multiple references. However, by being strictly based on the corpus, the analysis can be reproduced easily. The extraction of the references was done automatically using an optimized version of a self-written tool. The accuracy was evaluated for the first prototype by randomly selecting 60 papers and checking their references manually (cf. [We23]). These papers contained 934 references in total of which 197 are to papers that are within the corpus of which 143 were correctly assigned. Out of the 737 other references, 2 were falsely assigned to papers within. The optimized version improved the recognition, but still no 100 % accuracy can be guaranteed as references itself contain mistakes or typos quite frequently (cf. [Ni07]). Finally, only the undirected graph was analyzed. The difference between the undirected and directed betweeness centrality indicates that the link direction also carries information (potentially e. g., evolution of ideas or systems that connect subcommunities).

Using the modularity method, three different types of reasonable clusters could be identified: First, a characteristic-based cluster that is heavily based on the system type (ITS). Here, closely related systems that do not identify themselves as ITS but use similar techniques are included. However, some ITS are not – maybe these are not well received by the ITS community or a "glue" system is missing in the corpus. Second, an origin-based cluster that is heavily based on the origin (Germany) of the authors. Still, there are half of the German systems in other clusters indicating that there are different communities or some researchers are not part of the community. Interestingly there is a fundamental difference between the German community and the Spanish and Portuguese ones. The German one forms its own cluster, having only 2 non-German systems in it. This is, however, not the case for the Spanish or Portuguese communities. While their systems are still well connected, they do not form their own clusters but instead are sub-groups in cluster 24 and 26. They refer to each other and also seem to be better embedded and referenced at the international level. This becomes evident in the most important Portuguese system `Mooshak` for which there are also papers reporting on usage from Spain, Finland, and India as well as a cooperation with authors from Greece. To some extend, the difference in the reference distribution is caused by the composition of the corpus which includes papers written in German but not papers written in Spanish or Portuguese. With more of such papers, the composition of the clusters may change, shifting the focus more on the language and nationality. Nevertheless, this difference shows that there seems to be a certain preference for German authors to reference German systems. Third, clusters that could not be explained based on their keywords, characteristics such as developed for MOOCs, or (self-)characterization. Especially keywords turned out not to be a good classifier as "assessment" or "automatic grading" can be found for nearly every cluster as top keywords. Additionally, keywords do not seem to be consistently used (cf. [SS22], e. g, the self-characterizes e-assessment `JACK` also used ITS in a publication).

One might think that the older a system is the more citations it gains. TEx-Sys and ELM-ART are the oldest systems in the corpus (both from 1992), but are not in the top-10 of the most cited systems. A reason could be that these two are ITS and not generic e-assessment systems. Most references are for systems such as CourseMarker (2002) and WebCat (2003). These two systems, seem to have founded a new trend and, therefore, might be referenced more often. Still, Praktomat (1999) seems to be the oldest German system and has indeed the most references followed by JACK (2008) that seems to have the most associated publications.

The references to the top 10 systems can be divided into two different groups, which we call quantitative and qualitative. CourseMarker was one of the first e-assessment systems. Most of the references to CourseMarker target two out of its four corresponding papers. The first is an introduction to the system, the other an usage report which also highlights the advantages of the system. Other systems that reference the first instead of the second paper are likely not interested in the properties of CourseMarker, instead referencing it just as an example for e-assessment systems in general. Additionally, CourseMarker has more links to other clusters as links within. That also means that many systems reference it without having much similarities, as they then would be in the same cluster instead. References to Mooshak tend to be more focused on the system and its properties. It is one of the first programming judges and is often cited for that reason. But instead of only mentioning it as one of the already existing systems like CourseMarker, it is mostly referenced with a longer and more detailed description. Also, most of its references are from within its own cluster.

6 Conclusion & Outlook

In this paper the reference network of the corpus established by [SS22] was analyzed. This preliminary study provides a first insight into the results. Based on a modularity clustering, three types of clusters could be identified: one that mainly contains ITS (characteristic-based cluster), one that mainly contains systems of German origin (origin-based cluster), and clusters that are not easily explainable based on common properties. Hence, no new categorizations could be unveiled. Additionally, for the most referenced systems an analysis of how the references look like in the paper was conducted. Two types of references could be identified: references just mentioning existing systems (quantitative) and references that intensively discuss system properties (qualitative). A further conclusion might be that “the” German community should write more in English to be reference-able from non-native speakers and network internationally.

Further research should try to use different clustering techniques such as clique percolation or fuzzy clustering where a node can be in multiple clusters and compare those with the clusters found here. This might address issues arising through the fact that systems might change their character over time. Also other characteristics such as the type of feedback or type of reference could be analyzed. Additionally, the analysis here is strictly bound to the system and papers included on the said corpus. It would be interesting to see whether external data on references could be used to extend the data set and search for clusters there.

Bibliography

[BK10] Boyack, K. W.; Klavans, R.: Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? JASIST 61/12, pp. 2389–2404, 2010.

[Bl08] Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E.: Fast unfolding of communities in large networks. JSTAT/10, P10008, 2008.

[GS04] Griffiths, T. L.; Steyvers, M.: Finding scientific topics. Proc. National Academy of Sciences 101/suppl_1, pp. 5228–5235, 2004.

[GS05] Glänzel, W.; Schubert, A.: Analysing scientific networks through co-authorship. In: Handbook of quantitative science and technology research. Springer, pp. 257–276, 2005.

[KJH16] Keuning, H.; Jeuring, J.; Heeren, B.: Towards a systematic review of automated feedback generation for programming exercises. In: ITiCSE. Pp. 41–46, 2016.

[Ko05] Koschützki, D. e. a.: Centrality Indices. In: Network Analysis. Springer, pp. 16–61, 2005.

[LDB08] Lambiotte, R.; Delvenne, J.-C.; Barahona, M.: Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv:0812.1770/, 2008.

[Le13] Le, N.-T.; Strickroth, S.; Gross, S.; Pinkwart, N.: A Review of AI-Supported Tutoring Approaches for Learning Programming. In: Proc. ICCSAMA. Springer, pp. 267–279, 2013.

[Li05] Liu, X.; Bollen, J.; Nelson, M. L.; Van de Sompel, H.: Co-authorship networks in the digital library research community. Information processing & management 41/6, pp. 1462–1480, 2005.

[Na13] Narciss, S.: Designing and evaluating tutoring feedback strategies for digital learning. Digital Education Review/23, pp. 7–26, 2013.

[Ni07] Nicolaisen, J.: Citation analysis. ARIST 41/1, pp. 609–641, 2007.

[Sa17] Saito, D. e. a.: Program learning for beginners: survey and taxonomy of programming learning tools. In: Proc. ICEED. IEEE, pp. 137–142, 2017.

[SFB16] de Souza, D. M.; Felizardo, K. R.; Barbosa, E. F.: A Systematic Literature Review of Assessment Tools for Programming Assignments. In: Proc. CSEET. Pp. 147–156, 2016.

[SS22] Strickroth, S.; Striewe, M.: Building a Corpus of Task-based Grading and Feedback Systems for Learning and Teaching Programming. iJEP 12/5, pp. 26–41, 2022.

[We23] Weickenmeier, V.: Identifying hidden structures among Research Papes for E-Assessment Systems, MA thesis, LMU Munich, Germany, 2023.

[Wh04] White, H. D.: Citation analysis and discourse analysis revisited. Applied linguistics 25/1, pp. 89–116, 2004.

[WVN10] Waltman, L.; Van Eck, N. J.; Noyons, E. C.: A unified approach to mapping and clustering of bibliometric networks. Journal of informetrics 4/4, pp. 629–635, 2010.