Who are you working with?
- Visualizing TEL Research Communities -

Marco Fisichella   Eelco Herder   Ivana Marenzi   Wolfgang Nejdl
L3S Research Center
Hannover, Germany
{fisichella, marenzi, herder, nejdl}@L3S.de

Abstract: Author Co-Citation Analysis (ACA) provides a principled way of analyzing research communities, based on how often authors are cited together in scientific publications. In this paper, we present preliminary results based on ACA to analyze and visualize research communities in the area of technology-enhanced learning, focusing on publicly available citation and conference information provided through CiteseerX and DBLP. We describe our approach to collecting, organizing and analyzing appropriate data, as well as the problems which have to be solved in this process. We also provide a thorough interpretation of the TEL research clusters obtained, which provide insights into these research communities. The results are promising, and show the method's potential as regards mapping and visualizing TEL research communities, making researchers aware of the different research communities relevant for technology enhanced learning, and thus better able to bridge communities wherever needed.

Introduction and Motivation

Technology Enhanced Learning (TEL) is a fascinating field, with lots of different research questions and aspects to focus on. Researchers in TEL can focus on learning infrastructure to support the re-use of learning objects or personalization, on intelligent tutoring systems, on mobile learning, or on collaborative learning in teams. They can also focus on professional learning and knowledge management infrastructures, learning in universities (computer science, engineering or other disciplines) and on learning in schools, with a lot of interesting research questions and results. Many different conferences and journals are devoted to different aspects of technology enhanced learning, providing a variety of forums through which to publish TEL research results.

The downside of this variety is, however, that TEL is a much more fragmented area than most other research areas, making it difficult to gain an overview of recent advances in the field. Even for experienced TEL researchers answering the questions: "*What communities and sub-communities can be identified in TEL", "what research topics/specialties can be identified in a field of studies"* and "*what conferences are the most relevant for what topic and for which community"* is a difficult task, and for beginners it is obviously an impossible one.

Being aware of this fragmentation and of the various sub-communities which make up the TEL area is an important pre-requisite towards overcoming this fragmentation, increasing synergies between different sub-areas and researchers, and, last but not least, providing funding agencies with evidence of new research results, innovative applications and promising new approaches for technology enhanced learning.

This paper provides a first step towards this goal, by employing the technique of Author Co-citation Analysis (ACA) on the large subset of TEL conferences related to computer science as indexed by DBLP[1] and CiteseerX[2] - the latter provides citation information for each indexed paper. ACA relies on the insight, that if two authors are cited together very often in scientific articles, their work must be related to the same research field.

We will describe our methodology for data collection, solutions for problems that we encountered, and the techniques of author-co-citation and factor analysis for detecting communities in a given research area. We will

[1] http://www.informatik.uni-trier.de/~ley/db/

[2] http://citeseerx.ist.psu.edu/

further describe and discuss our results, which provide an interesting insight into some important TEL research clusters, and close with a summary and discussion of next steps and future work.

Related work

Co-author analysis and citation analysis is an important method when analyzing scientific communities. Ochoa et al. (2009) provides a very nice example of how such analysis can help provide greater insight into TEL research communities and collaborations, through visualizing and intuitively describing research community structure, focusing on TEL publications presented at recent EDMedia conferences. They focus on co-author analysis and visualization of these relations and provide interesting insights into collaboration networks in the TEL area. Wild et al (in press) used the same data corpus for a trend analysis in the EDMEDIA conference. By applying clustering techniques to the paper titles, they showed how certain technologies and approaches gained importance – including, among others, mobile learning, blended learning, portfolios, podcasts, game-based learning and assessment.

Similar introspective analyses have been applied to other research fields in the past. Henry et al (2007) provide an analysis of the area of human computer interaction, based on the four major HCI conferences, focusing on citation analysis that use data relating to these conferences (between conferences, articles and authors), word cloud visualizations to characterize the four conferences, and other visualizations that characterize collaboration and other networks. This paper does not rely on sophisticated mathematical network analysis modes but is a very good example of the power of visualization to make the structure of these networks explicit.

The approach we build upon in this paper, author co-citation analysis, has not yet been used widely despite its potential for detecting and clustering scientific communities based on the mathematical notion of factor analysis. One of the best papers and a good introduction to this approach is the paper by White et al, (White, H. D. and McCain, K. W. 1998). This study presents an extensive domain analysis of a discipline – information science – in terms of 120 top-cited authors, based on their papers from 1975 to 1995, with citations retrieved from Social Scisearch via DIALOG. Tables and graphics reveal the specialist nature of the discipline over 24 years, based on author co-citation analysis. The results show an interesting split of the field into two main specialties, which barely overlap, namely experimental retrieval/information retrieval and citation analysis. Included is also a dynamic analysis of the field, based on three 8-year-periods, which shows changes of authors and areas. The analysis is based on journal citations, but neglects important conferences such as the ACM SIGIR conference, the most relevant conference for the IR community. In contrast, the citation database used in our paper, CiteseerX, includes all important computer science conferences and workshops, providing a broad overview of computer science as it relates to TEL.

Using similar techniques, Chaomei Chen and Les Carr (1999) present an analysis of hypertext research based on the ACM Hypertext conference series, with papers included from 9 conferences over 10 years. About half of the citations in this series refer to papers from the same series, which points to a very homogeneous research community. Again, dynamic analysis using three time periods is included. Only citations within these conference series were considered, while we include citations from all conferences. Due to their restricted focus, the factors discovered represent a finely grained view of the hypertext research area (including subareas such as design models, hypertext writing, open hypermedia and information visualization), while our factors represent broader research communities, centered around one or a few community-centered conferences such as Adaptive Hypermedia or AIED.

Collecting Co-Citation Data

Following White et al. (1998), we assume that citing practices in a research community reflect the judgments as to which works by which authors are the most influential − for the field in general and for specific sub-themes. Aggregated over time, a definite structure emerges that can be considered the current state of the field. *Co-citation* is a very good way of establishing relations between authors that correspond to specific sub-themes and research areas in a research community − even though they do not directly reference each other. We consider author A and B to be co-cited, if they are both cited by an author C − that is, both names appear at least once in the reference section of C's paper. The more co-citations, the stronger the relationship is.

Our data sets were obtained from CiteSeerX and DBLP. CiteSeerX is a digital library focusing on the literature in computer and information science, being fairly complete. The articles are crawled automatically from the Web and then metadata and citations are extracted from these articles, again automatically. The CiteSeerX dataset contains more than 1.4 million paper records correlated with about 28 million citations. Due to the automatic data collection process, metadata in CiteSeerX are not always prefect, which leads to considerable problems that have to be solved before analysis starts. We will describe these problems and our solutions in the following subsections. In addition, DBLP is a computer science bibliography database, which relies more on human input (the maintainer of DBLP is Michael Ley, from the University of Trier), which covers about the same field as CiteSeerX, and currently contains about 1.3 million bibliographical records. DBLP metadata does not include citations, but has been used in our project to contribute high-quality metadata, to cope with ambiguous author names and to provide reliable conference statistics.

Data collection

While it was not the goal of our research to determine the most relevant authors in TEL –such a goal would involve a more elaborate discussion on how "most relevant authors" should be defined – a good sample of highly cited authors in TEL covering as many areas of TEL as possible was obviously necessary. Obtaining such a sample for a diverse area such as TEL is no trivial matter. The following paragraphs discuss our approach and the steps needed to gather such a sample. Our data collection focused on data available through the CiteseerX and the DBLP databases, both covering all computer science related research, and will extend this through additional databases covering educational and psychological research for TEL in the future.

Obtaining a first sample. To obtain a first sample of TEL conferences, we collected the lists of TEL conferences and journals to which a small sample of 13 well-known researchers submit their papers (Duval, Scott, Brusilovsky, Koper, Kieslinger, Klamma, Nejdl, Balacheff, Sharples, Davis, Zimmermann, Wolpers, Sutherland). From these conferences and journals (as identified in DBLP[3]), we extracted the 100 most prolific researchers. In a second iteration, we collected the list of top-100 conferences and journals to which these 100 most prolific authors submit their papers. Our final sample of authors represents the most prolific authors from the 20 conferences and journals in the latter list that have a specific focus on TEL[4]. These conferences and journals cover 13.557 publications in total.

For these authors we created a co-citation matrix. This first step resulted in a rather sparse matrix (with some authors not co-cited with any other authors) and consequently a set of clusters extracted through our SPSS factor analysis which was difficult to interpret. Thus, subsequent iterations were designed to extend and refine the set of authors, as discussed in what follows; in addition they included other conferences such as Adaptive Hypermedia, User Modeling or Artificial Intelligence, which provide techniques for TEL infrastructures and algorithms.

Adding more authors, increasing co-citations. As regards extending and refining the set of authors, in the second iteration we first included more authors: the 50 most prolific authors from ED-MEDIA[5] and ECTEL[6], 15 new authors from the IEEE TLT Board and Steering Committee[7], and 5 more authors from the Telearn archive[8]. We also included the top-15 cited papers or books from EDMedia 2005 – 2008 (Ochoa et al. 2009). Second, after merging these sets, we selected the authors with at least 20 publications in CiteceerX DB and with at least 10 co-citations in our co-citation matrix. We also experimented with a threshold of 20 and 30 co-citations, but finally kept the 10-co-citation threshold, as the clusters obtained were of similar quality.

Disambiguating authors. At this point we realized there was a problem of disambiguation for some names, so we decided to check the name occurrences in DBLP (where author names are manually disambiguated by the DBLP maintainer, Michael Ley) and to keep only the author strings, that unambiguously identified the TEL authors we wanted to include). For example, we deleted John Cook because we found 269 occurrences of his surname in DBLP but, when queried by his full name we found only 12 publications in DBLP and 8 publications in CiteceerX. We

[3] The detailed procedure is described in the Stellar deliverable D7.1: http://www.stellarnet.eu/d/7/1/Investigating_two_silos

[4] Other topics are computer science (27 venues), artificial intelligence (26), human-computer interaction (22) and databases (5).

[5] http://ariadne.cs.kuleuven.be/edmedia/rankings.html

[6] http://ariadne.cs.kuleuven.be/ectel/rankings.html

[7] http://www.computer.org/portal/web/tlt/edboard

[8] http://telearn.noe-kaleidoscope.org/

deleted John Black as well, because the occurrences both in DBLP and CiteceerX were too ambiguous to correctly attribute publications or citations (John Black, John A. Black, John B. Black, John D. Black, John E. Black, John R. Black, John A. Black Jr). Based on this disambiguation, we kept the full name of each author, and the initials when this did not result in duplicates or ambiguity in DBLP. This left us with 77 authors for our analysis.

Adding and checking more conferences. To better characterize the clusters found through Component analysis, we checked the top 4 venues for each author. This had to be done using DBLP, as CiteseerX does not contain complete references for all papers, but sometimes only refers to them as technical reports. We then used DBLPVis[9], to check for the five most prolific authors in all these TEL conferences covered DBLP and CiteseerX (AIED, CSCW , EC-TEL, Edutainment, ICALT, ICCE, ICWL, ITiCSE, ITS (Intelligent Tutoring Systems), SIGCSE, Wissensmanagement, WMTE), to make our final co-citation matrix more complete, in total 55 authors. Using a threshold of 50 DBLP publications, we kept 30 of them. 25 of them were already in our matrix, which was an encouraging sign that our previous iterations had already produced a good sample for these TEL conferences. We added 5 new authors to our matrix, for a final matrix of 82 highly cited and co-cited TEL authors.

Data processing – Problems and Solutions

We conducted our analysis on CiteseerX dataset. The following paragraphs discuss our approach and give an overview about the relevant tables considered from the database, as well as the problems encountered during data processing and our solutions for these problems.

Tables. CiteseerX is organized in terms of three main tables: Papers, Authors and Citations. The *Papers* table contains all the papers, unequivocally retrieved through an identifier. Every paper can be a different version of the same publication, each associated to a single value of the attribute *cluster*, e.g. one cluster ID is coupled with several paper IDs. In addition, the papers are connected with their authors. A single author can have multiple occurrences in the *Authors* table, one for each paper s/he wrote. Thus, the data set contains duplicated author identifiers, a common problem when dealing with publication data. Finally, the references for each paper are stored in the table *Citations* with the following information: paper identifier *cited_paperID* of the paper which the reference is cited by, *citation title*, *venue*, *year* and the *authors* of the cited paper (a string field, with all authors concatenated).

Processing. To compute the co-citation matrix , we collected the subset of the paper citations corresponding to the references to papers written by the relevant authors, selected for our analysis. The lack of a paper identifier of the citation made our mining task more complex: to retrieve the cited papers of our author list, we had to search for our authors within the value of the attribute *authors* in the *Citations* table. This was possible after processing the dataset in three steps: 1) drop all the foreign keys inside the *Citations* table; 2) change dataset engine from InnoDB to MyISAM to enable efficient full-text search; 3) create a full-text index for the attribute *authors*. All the results were stored within a new *citations_TopAuthors* table so as to provide reasonable processing time for our queries (the size of the new table is about 50,000 records compared to the 28 million in the original *Citations* table. Finally, to further increase processing time, we built another full-text index on authors.

Multiple author aliases. Since a single author can have multiple occurrences in the *Authors* table, we had to cope with the problem that author names may be misspelled or use initials instead of full first names; authors may also change their names or use different combinations of formal and informal names and initials in different papers, producing multiple identifiers we call *aliases* for a single person. The author “Wolfgang Nejdl” appears more than two hundreds time with his complete name, for example, and about ten times as “W. Nejdl”.

Unique author identifier. We then collected all the paper citations which had at least one previously computed alias in the *authors* attribute. For each of these circa fifty thousand records, we added one *firstAuthor* attribute in the new table to describe a single author with aliases with one identifier, e.g. we put “Nejdl” as identifier of “Wolfgang Nejdl” and “W. Nejdl”. Thanks to the fact that *firstAuthor* contains only one identifier, we were able to solve the problem of keeping information about the identifiers of a possible second or third author who wrote the same cited paper. We therefore duplicated, for each author of interest, the corresponding citation in the new table *citations_TopAuthors* with the identifier for a second and subsequent author.

Paper multi versioning. Another issue we encountered was paper multi-versioning. Because the same paper can have several versions each of which has been crawled from the Web and given that each of these publications keep information about their references in the *Citations* table, we had to remove from our table the duplicate citations

[9] http://dblpvis.uni-trier.de/help/overview.html

related to different editions of the same paper. To achieve this goal, we exploited the attribute *cluster*, as described before, of the table *Papers*.

Matrix creation

For subsequent analysis, we then created a quadratic, symmetric matrix containing the listing of our selected authors as rows and columns, to be filled by co-citation data: for the *j-th* row and the *i-th* column, the retrieved value in this cell refers to the number of times the *j-th* author was co-cited with the *i-th* one.

For *i* equal to *j* we included a null value because it corresponds to the cell representing the number of co-citations of one author with her/himself.

Our matrix construction process includes three main steps:,

- Select the identifier of all cited papers we collected in our table *citations_TopAuthors*.
- For each of these identifiers, gather distinct authors, i.e. the values of the attribute *firstAuthor*.
- Whenever this previously computed result set carried more than one author, for each possible author pair, we incremented the corresponding values <*i*,*j*> and <*j*,*i*> in the matrix.

These steps lead to the following algorithm, described in pseudo-code and relevant SQL statements:

```
Select distinct cited_paperID from citations_TopAuthors;
For each cited_paperID
    Select distinct firstAuthor from citations_TopAuthors where cited_paperID = current cited_paperID
    If more than 1 firstAuthor
        Compute all possible author pairs
        For each author pair <i,j>
            Update matrix cell <i,j> and <j,i>
```

Listing 1. Pseudo code for the matrix computation.

Cluster Analysis and Discussion

We then proceeded to analyze our data, using principal component analysis, to detect appropriate clusters / areas in TEL research, and then visualize and interpret these clusters.

Using Principal Component Analysis to Detect TEL Research Areas

Principal Component Analysis. “In the social sciences we are often trying to measure things that cannot directly be measured (so-called latent variables)”, as Andy Field states in his book (Field, 2009). In our case, the interest in different topics or research areas of different authors in TEL cannot easily be measured. We could not measure motivation and interest directly, but we tried to analyze a possible underlying variable (collaboration in the form of co-citations among the major authors), to detect different sub-communities and possible trends. To do so, we used the statistical application SPSS to perform the Principal Component Analysis (PCA): a technique for identifying groups or clusters of variables and reduce the data set to a more manageable size while retaining as much of the original information as possible. Often, its operation can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data.

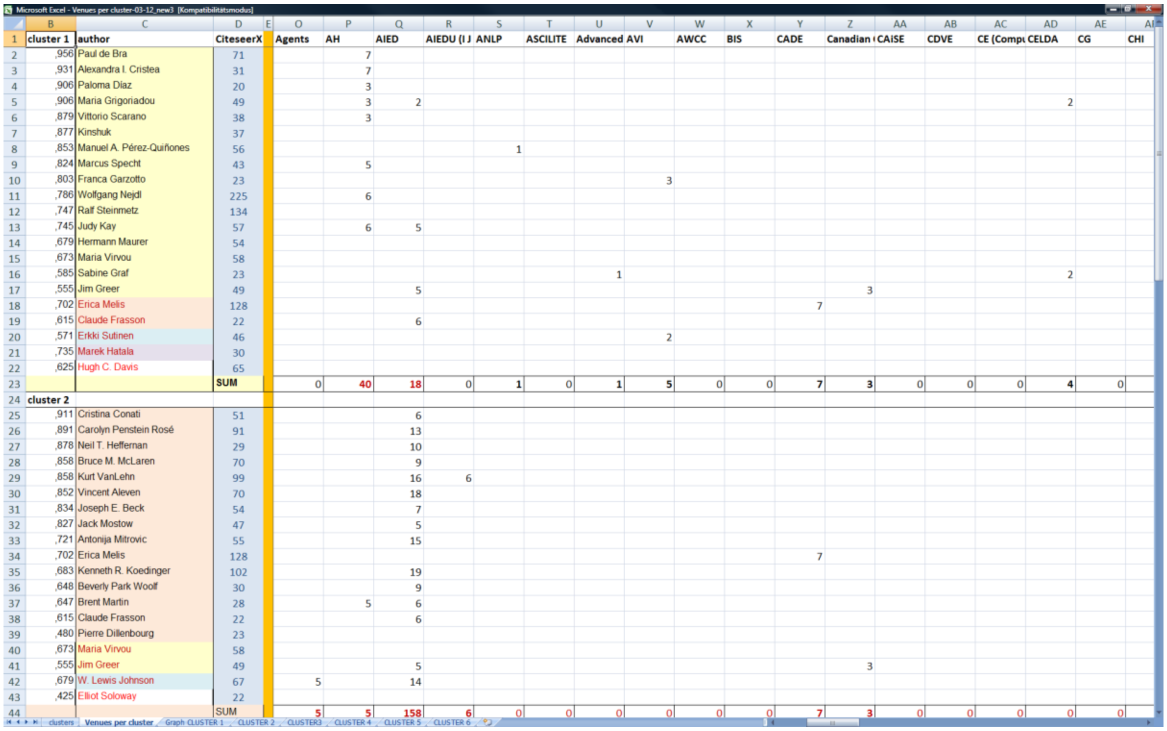
PCA vs FA. Principal Component Analysis is similar to Factor analysis, but merely has the goal of finding linear components within the data and how a variable might contribute to these components (which basically means, finding some meaningful clusters within the data). Factor analysis uses the same techniques, but the aim is to build a sound mathematical model from which factors are estimated. The choice of PCA vs. FA depends on what we hope to do with the analysis: whether we want to generalize the findings from your sample to a population, or whether we want to explore our data or test specific hypotheses. In our specific research, we used PCA because we wanted to explore the data with a descriptive method and apply our findings to the collected sample.

Correlation determinant. When we measure several variables with the PCA, the correlation between each pair of variables can be arranged in what is known as an R-matrix: a table of correlation coefficients between variables. The existence of clusters of large correlation coefficients between subsets of variables, suggests that those variables could be measuring aspects of the same underlying dimensions. These underlying dimensions are known as factors (or latent variables). In Factor analysis we strive to reduce this R-matrix to its underlying dimensions by looking at which variables seem to cluster together in a meaningful way. This data reduction is achieved by looking for variables that correlate highly with a group of other variables, but do not correlate with variables outside that group. Because our main aim is PCA, we did not have to worry about the correlation matrix determinant. Strictly speaking, the determinant or correlation matrix should be checked only in factor analysis: in pure principal component analysis it is not relevant (Field 2009), so that we could leave all our authors in the sample.

Defining factors. Not all factors are retained in an analysis, but only the most relevant and meaningful one for the research. In our case, we used Varimax orthogonal rotation[10] to discriminate between factors (to rotate the factor axes such that variables are loaded maximally to only one factor and we could better calculate the loading of the variable on each factor). We sorted the variables by size ordering them by their factor loadings, to displayall the variables which load highly onto the same factor together. As a result we obtained a Rotated Component Matrix which shows the variables listed in order of size of their factor loadings. For interpretation purposes, we also suppressed absolute values which were less than 0,4.

We obtained 15 factors in total, which explain 78% of the variance; for this paper we focus on the first six factors, explaining 59%. Compared to (White and McCain 1997), where the first eight factors alone explain 78% of the variance, our lower value reflects the different disciplines that come together in TEL, producing many more sub-communities, while Information Science has some well-established communities that focus on a particular topic.

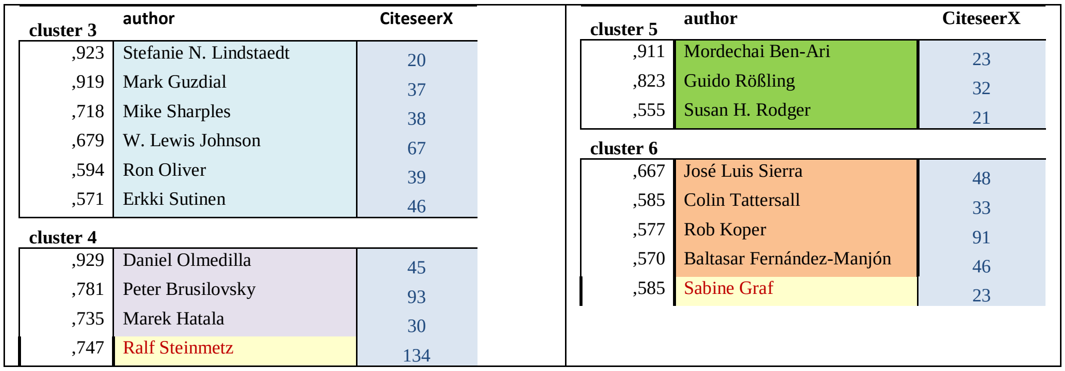
To describe the meaning of each factor more precisely we also added information regarding the conferences where our sample authors usually publish. For this paper, we included the top 4 venues for each author, as well as the number of papers published. Figure 1 shows the first two clusters, with a (small) subset of conferences displayed, Figure 2 clusters 3-6.



	B	C	D	E	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AI
1	cluster 1	author	CiteseerX		Agents	AH	AIED	AIEDU (I J	ANLP	ASCILITE	Advanced	AVI	AWCC	BIS	CADE	Canadian (	CAiSE	CDVE	CE (Compu	CELDA	CG	CHI
2	,956	Paul de Bra	71			7																
3	,931	Alexandra I. Cristea	31			7																
4	,906	Paloma Díaz	20			3																
5	,906	Maria Grigoriadou	49			3	2													2		
6	,879	Vittorio Scarano	38			3																
7	,877	Kinshuk	37																			
8	,853	Manuel A. Pérez-Quiñones	56						1													
9	,824	Marcus Specht	43			5																
10	,803	Franca Garzotto	23									3										
11	,786	Wolfgang Nejdl	225			6																
12	,747	Ralf Steinmetz	134																			
13	,745	Judy Kay	57			6	5															
14	,679	Hermann Maurer	54																			
15	,673	Maria Virvou	58																			
16	,585	Sabine Graf	23								1									2		
17	,555	Jim Greer	49				5									3						
18	,702	Erica Melis	128												7							
19	,615	Claude Frasson	22				6															
20	,571	Erkki Sutinen	46									2										
21	,735	Marek Hatala	30																			
22	,625	Hugh C. Davis	65																			
23			SUM		0	40	18	0	1	0	1	5	0	0	7	3	0	0	0	4	0	
24	cluster 2																					
25	,911	Cristina Conati	51				6															
26	,891	Carolyn Penstein Rosé	91				13															
27	,878	Neil T. Heffernan	29				10															
28	,858	Bruce M. McLaren	70				9															
29	,858	Kurt VanLehn	99				16	6														
30	,852	Vincent Aleven	70				18															
31	,834	Joseph E. Beck	54				7															
32	,827	Jack Mostow	47				5															
33	,721	Antonija Mitrovic	55				15															
34	,702	Erica Melis	128												7							
35	,683	Kenneth R. Koedinger	102				19															
36	,648	Beverly Park Woolf	30				9															
37	,647	Brent Martin	28			5	6															
38	,615	Claude Frasson	22				6															
39	,480	Pierre Dillenbourg	23																			
40	,673	Maria Virvou	58																			
41	,555	Jim Greer	49				5									3						
42	,679	W. Lewis Johnson	67		5		14															
43	,425	Elliot Soloway	22																			
44			SUM		5	5	158	6	0	0	0	0	0	0	7	3	0	0	0	0	0	

Figure 1: Authors and top 4 venues for each author, for the first two clusters

[10] The Varimax rotation attempts to maximize the dispersion of loadings within factors. It tries to load a smaller number of variables highly onto each factor resulting in more interpretable clusters of factors.



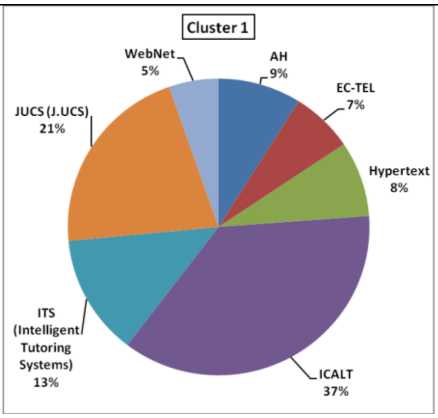
cluster 3	author	CiteseerX
,923	Stefanie N. Lindstaedt	20
,919	Mark Guzdial	37
,718	Mike Sharples	38
,679	W. Lewis Johnson	67
,594	Ron Oliver	39
,571	Erkki Sutinen	46
cluster 4		
,929	Daniel Olmedilla	45
,781	Peter Brusilovsky	93
,735	Marek Hatala	30
,747	Ralf Steinmetz	134

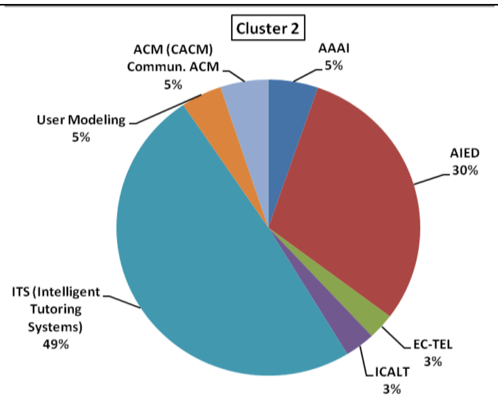
cluster 5	author	CiteseerX
,911	Mordechai Ben-Ari	23
,823	Guido Rößling	32
,555	Susan H. Rodger	21
cluster 6		
,667	José Luis Sierra	48
,585	Colin Tattersall	33
,577	Rob Koper	91
,570	Baltasar Fernández-Manjón	46
,585	Sabine Graf	23

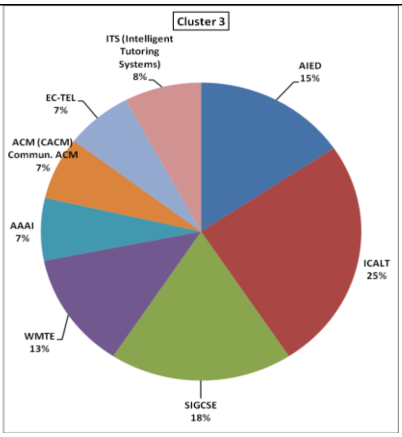
Figure 2: Authors, factor loadings and CiteseerX publications for cluster 3-6

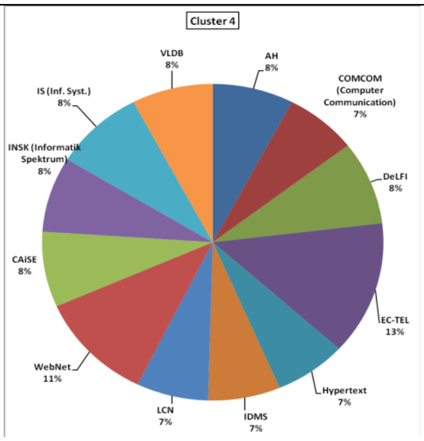
Visualizing TEL research clusters

Visualization based on conferences. Based on this analysis, the following figures provide a visualization of the TEL research clusters obtained, first based on pie charts relating to the most relevant conferences for each cluster. To produce the conference-based charts, for each author we collected his/her four most frequented conferences according to DBLP (names of conferences as well as number of papers published by this author), added the number of papers for each conference and cluster, and then produced the following pie-charts including the most representative conferences for each cluster. For Clusters 1 and 2, conferences were selected if they included more than 20 publications (for Cluster 1) and 15 publications (for Cluster 2) from the cluster authors, for Clusters 3-6, we used a threshold of 5-7 publications to select the representative conferences.









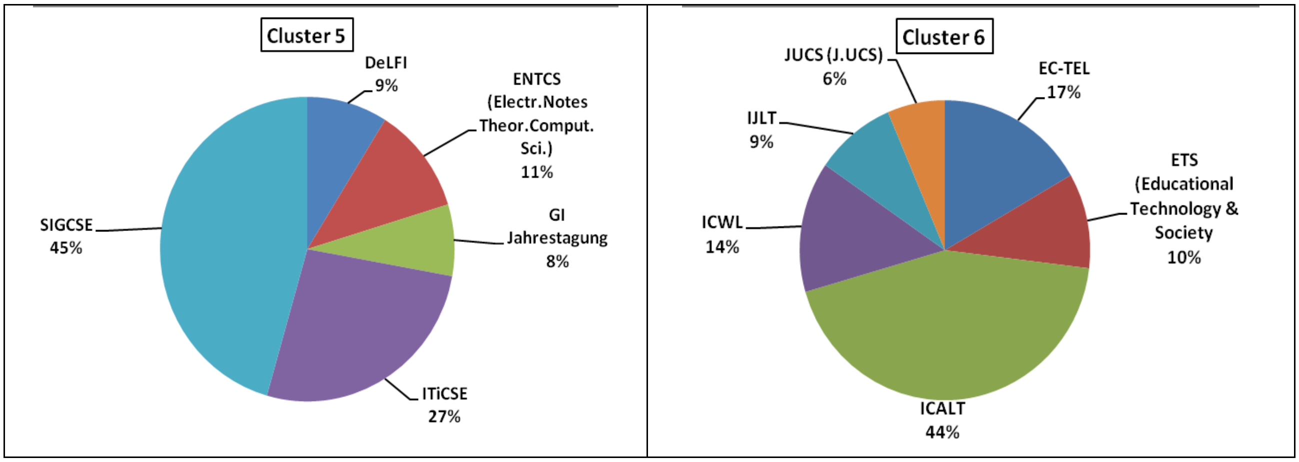


Figure 3: a visualization of the TEL research clusters, based on relevant conferences

Visualization based on Tag Clouds. Based on the clusters we retrieved, we selected form the CiteseerX dataset all the paper titles whose authors were in the cluster of interest. From the extracted paper titles we removed the words with less than 2 characters and the words consisting of numbers because these were not useful when determining the topic of a paper; for those words containing punctuation marks such as \-"\?" \%" and \/", we removed the punctuation marks and combined the remaining parts. We also removed stop words and applied stemming, as well as duplicate words inside a paper´s title. We then assigned a counter to each distinct word, counting the number of occurrences of the word inside the titles. Last, we sorted all words in increasing order based on the counters and visualized the first 150 words.

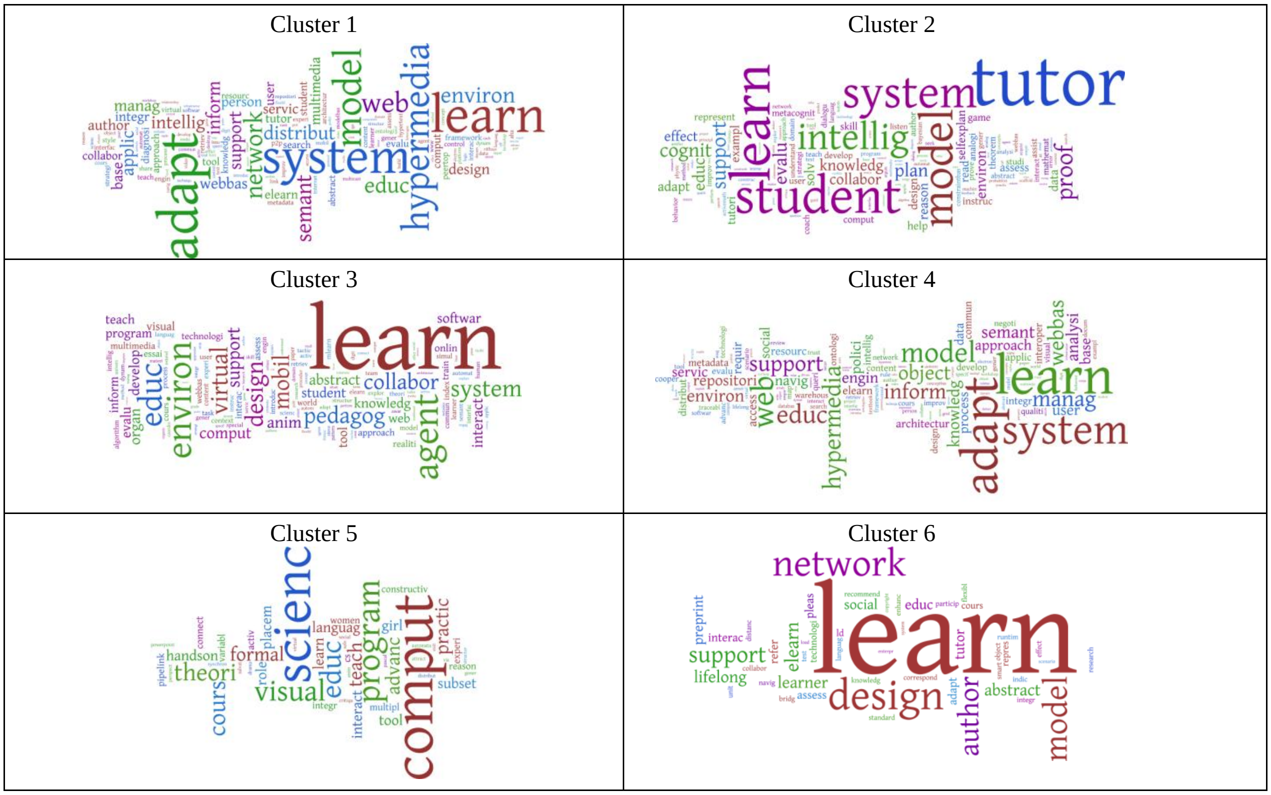


Figure 4: a visualization of the TEL research clusters, based on paper titles (created using Wordle.net)

Discussion

The combined information from the clusters of researchers, the main conferences and journals that they address and the most often used keywords in their publications clearly show the differences in focus in the community – in terms of research as well as in terms of publications and connections. In this section, we discuss the main findings from the visualizations presented before.

The main publication venues (Figure 3) of the *first cluster* of researchers (Figure 1) include − besides main TEL conferences such as ITS and ICALT and the general journal JUCS − Adaptive Hypermedia, Hypertext and ECTEL. From the word cloud (Figure 4) of this cluster – with “Adapt”, “Model” and “Hypermedia” as distinctive words –, a clear focus on *adaptive hypermedia systems* can be observed. This cluster contains authors like Paul de Bra (his four most frequent conferences are Hypertext, WebNet, AH and EC-TEL), Marcus Specht (EC-TEL, AH, WebNet), Hugh Davis (ICALT, Hypertext) and Wolfgang Nejdl (AH and many non-TEL conferences focusing on the Web and Information Systems). The cluster also includes personalization as represented in other relevant conferences listed (Judy Kay, for example, publishes most in ITS, AH and AIED).

Most authors in the *second cluster* have their roots in the field of artificial intelligence − as shown from the main publication venues AAAI and AIED. The conference on Intelligent Tutoring Systems is – in terms of quantity – the most important conference of this cluster. Authors in this cluster include Carolyn Penstein Rose (ITS and AIED), Bruce McLaren (ITS, AIED and EC-TEL) and Kurt Van Lehn (ITS and AIED). Jim Greer is included in the first two clusters, publishing most in ITS and AIED, but also in the EC-TEL and UM conferences, which are closer to the first cluster. Whereas the focus of the first cluster is on personalization and adaptation, the second cluster mainly focuses on understanding learners’ needs, by applying reasoning techniques to the models of the learner – this can also be observed from the word clouds – “Learn(-er/-ing)”, “Student”, “Model” and “Cogni(tion)” are the most significant words for this cluster.

The differences in terms of background and focus between the first two clusters are striking, given the similarity in research goals. Learner or user modeling is the first step in the process of adapting a system to the learner (Paramythis and Weibelzahl 2005). It is to be expected that these clusters will become more related with one another, as the targeted conferences AH (first cluster) and UM (second cluster) have merged into the UMAP conference in 2009.

Terms that show up in the *third cluster* are “Environment”, “Mobile”, “Pedagogy”, “Agent” and “Design”. Researchers in this cluster have more diverse backgrounds than in the first two clusters, but with the common denominator that they focus on the application of specific technologies to learning. These focuses include mobile technologies (Mike Sharples, Erkki Sukinen − WMTE), computer science education (SIGCSE, Mark Guzdial) and knowledge management.

The *fourth cluster* is an interesting cluster, related to Cluster 1 (“Personalization”), with Peter Brusilovsky as most prominent author. However, this cluster is more focused on learning objects than the first cluster, as witnessed by Erik Duval, as another prominent author. Apart from “Adaptation” and “Hypermedia”, the word clouds of this cluster include “Object”, “Semantic”, “Repository” and “Metadata”. As the first cluster, it also includes authors publishing not only in TEL, but in other areas (Ralf Steinmetz and Matthias Jarke), which (because of the smaller cluster size) has a bigger impact on the pie chart, which now includes several non-TEL related conferences relevant to information systems and communications as an explicit hint as to how other computer science related areas often influence TEL research.

The *fifth cluster* is a very application oriented cluster, with two TEL conferences mostly relating to computer science education (SIGCSE, ITiCSE, Mordechai Ben-Ari as prominent author), and an interesting non-TEL conference on Theoretical Computer Science showing the background of Guido Rößling (ENTCS, otherwise publishing mainly in ITiCSE and DeLFI, the German eLearning conference).

In terms of number of publications, Rob Koper is the most prominent researcher in the *sixth cluster*. An online search on these researchers shows that all of them have contributed to the theory of Learning Design (Koper and Tattersall 2005) and related technologies and standards, such as SCORM (Dodds 2007) – as exemplified by Baltasar Fernández-Manjón. Not surprisingly, “Learning Design” is the leading term of this cluster’s word cloud.

It is apparent that the lists of most popular conferences and journals for each cluster do not only contain TEL-specific conferences: they also contain conferences with a focus on artificial intelligence (AAAI) and human-computer interaction (AH, UM). On the one hand, this shows the importance of these areas to TEL – which matches the numbers of non-TEL venues that we identified during our data collection, as explained earlier in this paper – but also shows that TEL-related work is presented at other venues. This can be interpreted as evidence for the multidisciplinary character of TEL research.

From these six clusters, the *building blocks* of the computer-science related research in TEL can be observed as:

- human-computer interaction, most prominently (adaptive) hypermedia systems (cluster 1)
- artificial intelligence and (reasoning techniques for) user modeling (cluster 2)
- semantics, repositories and metadata (cluster 4)

Cluster 3 and 6 represent the more TEL-specific innovative areas. The terms in their word clouds overlap to a large extent with the 'new terms' in EDMEDIA, as identified by Wild et al (in press).

Conclusions and Future Work

In this paper, we used author co-citation analysis to analyze and visualize research communities in the area of technology-enhanced learning, focusing on publicly available citation information provided through CiteseerX and conference information available through DBLP. The results are visualized based on relevant conferences and themes for each cluster, providing a first important step to provide a structured overview over research in technology enhanced learning and make TEL researchers aware of the different research communities relevant for their work.

As an important next step, we will extend our dataset with additional publication and citation data relevant for TEL, most importantly education and psychology, as relevant for example for computer supported collaborative learning.[11] These steps are currently performed, together with other project partners, in the context of the STELLAR Network of Excellence.

We hope, that this work as well as future work building on it, will help overcome TEL research fragmentation, by making TEL researchers aware of the different research communities relevant for technology enhanced learning, and thus more able to bridge communities wherever needed.

References

Chen, C., Les Carr (1999). *Trailblazing the Literature of Hypertext: Author Co-Citation Analysis* (1989-1998). Hypertext 1999.

Dodds, P. *SCORM Primer*. Retrieved from http://adlcommunity.net/mod/resource/view.php?id=458. Last modified: April, 2007.

Henry, N., Howard Goodell, H., Elmqvist N., Fekete J-D. *20 Years of Four HCI Conferences: A Visual Exploration*. Intl. Journal of Human-Computer Interaction, Volume 23/3, December 2007.

Ochoa, X., Mendez, G., Duval, E. (2009). *Who we are: Analysis of 10 years of the ED-MEDIA Conference*. ED-MEDIA 2009.

Field A., (2009). *Discovering Statistics Using SPSS*. SAGE Publications

Koper, R. & Tattersall, C. *Learning Design – A Handbook on Modelling and Delivering Networked Education and Training*. Springer.

Paramythis, A., & Weibelzahl, S. (2005). *A decomposition model for the Layered Evaluation of Interactive Adaptive Systems*. 10th International Conference on User Modeling, 2005, Springer.

White, H.D. and McCain, K.W. (1998). *Visualizing a discipline: an author co-citation analysis of information science*, 1972–1995. J. Am. Soc. Inf. Sci. 49, 4, Apr. 1998.

Wild, F., Valentine, C. and Scott, P. *Shifting Interests: Changes in the Lexical Semantics of ED-MEDIA*. Intl. Journal of Elearning, in press.

Acknowledgments

Work on this paper has been partially sponsored by the STELLAR Network of Excellence.

[11] The CSCL conference, for example, is not indexed in DBLP and therefore missing in our analysis.