Cerovsek, T. and B. Martens, 2004, On the Extended Use of Citations in CAAD, In: Van Leeuwen, J.P. and H.J.P. Timmermans (eds.) *Developments in Design & Decision Support Systems in Architecture and Urban Planning*, Eindhoven: Eindhoven University of Technology, ISBN 90-6814-155-4, p. 1-17.

On the Extended Use of Citations in CAAD

Making the network between related publications visible

Tomo Cerovsek and Bob Martens
University of Ljubljana
Vienna University of Technology

Keywords: Scientific Knowledge Management, Retrospective CAAD Research, Graph Theory, CAAD-Related Publications, Web-Based Bibliographic Database

Abstract: This paper discusses the extended use of the Cumulative Index of CAAD (CUMINCAD) - a digital library set up in 1998 serving the CAAD-community as an important source of scientific information with over 6.000 recorded entries published on-line. The aim of this paper is to elaborate a related Citation Index to CUMINCAD - with over 20.000 references - and to provide information on entries with an exceptional high impact in the CUMINCAD database. The importance is determined through its use (citing) in the framework of afterwards published scientific materials. By utilizing graph theory methods extensive citation analyses will be presented illustrating the impact of particular contributions in different research topics.

1. INTRODUCTION

CUMINCAD – an acronym for CUMulative INdex of CAD / cumincad.scix.net – was created in 1998 as response to limited, difficult access to scientific information in the field of CAAD. It was clear from the very beginning that the success of a digital library like CUMINCAD would strongly depend on its contents. Several associations made their materials available and the filling of the CUMINCAD-repository could be realised on a shoestring budget.

Within the European Union's SciX project ("Open, self organising repository for scientific information exchange", http://www.scix.net, 2002-2004), resources for further development and extension of CUMINCAD -

such as a Citation Index - could be realised. CUMINCAD serves now as a significant indicator in terms of digital library technology. Nearly 2.000 individuals are registered as users for this service. SciX is meanwhile also providing hosting for digital libraries of other fields of science.

In the framework of the SciX-project, all contributions from the early days of DDSS - which were archived on paper - have been digitized to produce e-prints and to extract metadata. Digitization from paper-based materials by means of scanning and OCR is much more labour-intensive than "reusing" archived digital data sources. This applies to the DDSS-conferences 1992, 1994 and 1996. From 1998 on digital data was archived and this facilitated the easier creation of metadata. At the moment 334 entries from DDSS-conferences are available in CUMINCAD; with the exception of the 1994 conference also full paper e-prints are attached to the records.

On completion of the above mentioned SciX-project, CUMINCAD will be carried on in a similar way as in its very infancy (1998-2001). Thus integration of the new conference proceedings will play a major role. As the required metadata usually is issued by the conference organisations in a predefined quality, this will only amount to slight efforts regarding subsequent editing and indexing the full-papers in pdf-format. Subsequent editing of the references to be linked with the bibliographic entries can be more time-consuming, as datasets are to be divided into four information modules (author, title, source and year of publication).

1.1 Related efforts

The concept of citation indexing is neither new nor very common in most scientific associations. It started in the early fifties and until recently, citation analysis was solely accommodated in the domain of specialized institutions like Thomson ISI, which offers a multidisciplinary collection of bibliographic information mainly from printed and especially peer-reviewed journals. The disadvantage of services like ISI is that they do not provide information on domain specific collections like CUMINCAD.

A historical overview of citation analysis can be found at the ISI site (Thomson, 2004). The extensive use of citation is justified through the following statement: "*Citations symbolize the conceptual association of scientific ideas as recognized by publishing research authors.*" (Garfield, 1997). Opponents to the relevancy of citation analysis claim that the real value of citations is questionable. Usual critique addresses different aspects of scientific publishing, for example: "*Often not the discoverer or originator is cited but a later author, often a reviewer who only redescribed or mentioned the finding in a more understandable way*" (for details see Hauffe, 1994).

With the evolution of digital libraries and open scientific exchange, citation analysis has also become a domain of open access libraries. The use of scientific contributions is not only determined through citation analysis, but also through the web usage statistics (views of abstracts, downloads of full texts, etc.). An example of an extended initiative dealing with infrastructure for reference linking and citation analysis for open archives is the Open Citation Project (2002), which developed solutions for citation impact analysis, and reference linking in large-scale open-access archives (OAI, 2004). Furthermore, this project also began to explore the critical relationship between usage and citations. Some similar or complementary analyses were covered in our previous efforts (Martens et al, 1999) and (Turk et al, 2001), where combined text and web mining was demonstrated.

1.2 Methodological aspects

The work presented is guided by the following basic principles:

- *Gradual development*. The goal was to enable gradual development of tools and immediate use of results. The results should be organized in a way appropriate for further use, or complementary analysis.
- *Paradigms.* The analysis was influenced by several paradigms: process-oriented model, graphical models, IR (information retrieval) concept of bag of stem words, as well as systemic view to similarity – defined by canonical form. The terms citation and reference are used interchangeably.
- *Types of analysis.* Three main types of analysis have been used to produce results presented in this paper: (1) Text mining, (2) Statistical analysis of citations (3) Graphical models as a way of interpretation of relationships between elements subjected to analysis.
- *Use of multiple techniques*. Since the methods used in the analysis do not offer sufficient, or one-and-only interpretation of the contents and relationships, multiple techniques might be applied.
- *Use of voting mechanisms*. If there are alternative methods available, which sometimes deliver different results – a mechanical voting mechanism may be used to automatically select the best result.

1.3 Structure of the paper

The paper is structured into three related parts covering the main aspects of citation analysis. First of all, input data for the analysis is presented, giving a brief description of main characteristics of input data and analysis procedures. Second, the most important results are presented including

statistical figures and data, as well as graphical models. Finally, interpretation, discussion and conclusion are presented.

2. ANALYSIS OF RELATED PUBLICATIONS

In the framework of the SciX-project a substantial extension of full text e-prints could be realised for the CUMINCAD-repository. The extraction of references from these e-prints and storage in another digital library - called CUMINCAD.REFS - defined the next working stage. The citation data was divided into four parts and tagged with the CUMINCAD-ID of the specific publication, in order to provide a link to the display. A substantial number of references can be extracted for the archived e-prints. These papers (with references) were subjected to analysis. In this section we first give a brief description of the input data that was used for further analysis. Afterwards, methods and tasks that were performed are presented more in detail.

2.1 Input data for the analysis

The data is represented in figures relevant to IR (Information Retrieval). As shown in table 1, there was a significant growth of records and vocabulary in the CUMINCAD database in the period 2002-2004 after the SciX project was initiated (more than 3000 new records were added). As indicated by the average term vector length (dimensions represent stemmed words) – these newly added records do have more elaborate descriptions; on the other hand the standard deviation indicates that the length of descriptions varies very much.

Table 1. Collection of statistics in the CUMINCAD database

Year	Total number of records (in whole collection)	Number of stemmed terms (in whole collection)	Average term usage (in whole collection)	Average term vector length	Standard deviation of vector length	Average freq. of terms in vector	% of terms with freq. 1
CUMINCAD repository							
2002	3042	11762	14.76	57.39	30.32	1.39	78.66
2004	6399	21255	18.41	61.32	57.00	1.40	78.47
CUMINCAD references							
2004	23988	8550	14.36	5.12	2.22	1.02	98.30

Term vectors are much shorter because CUMINCAD references do not include abstracts, which were taken into account in the case of CUMINCAD records. The next two figures show a number chronological distribution of records (or records per year of publishing) in both databases.

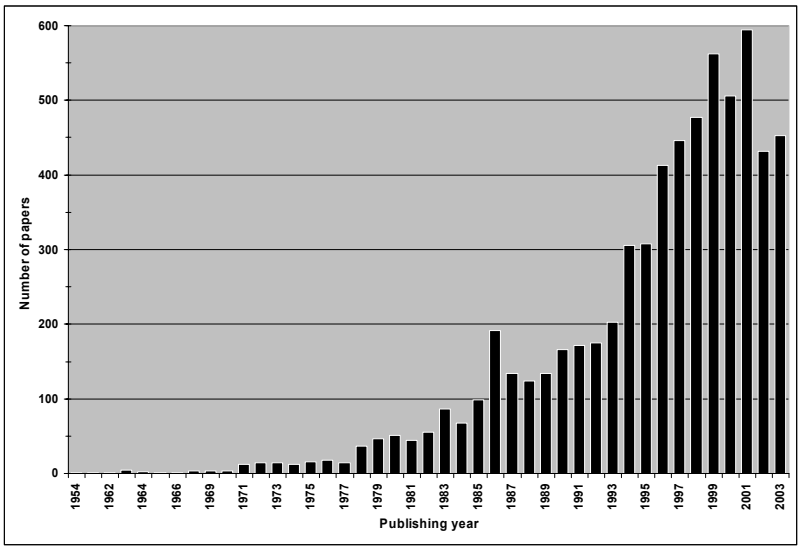

Figure 1. Chronological distribution of CUMINCAD papers

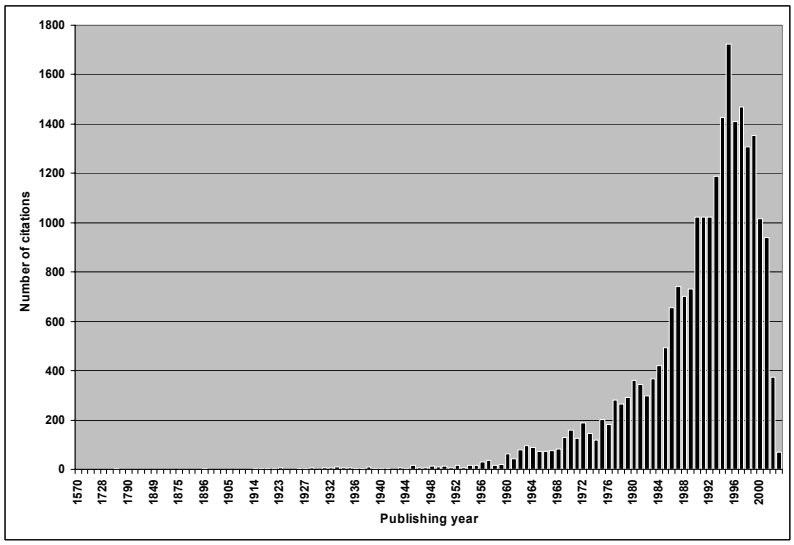

Figure 2. Chronological distribution of CUMINCAD references

A rough evaluation of the 20.000+ references demonstrates that approximately 30% evaluated citations are dating from 1996 or earlier. Slightly more than the half date from the period 1981-1995, while approx. 10% were created in the period of 1971-1980, i.e. fewer than 10% of the entries were published before 1971. The entries in CUMINCAD are comparably “younger” - no big surprise. Only 3% of the CUMINCAD entries concern publications prior to 1980. Two-thirds of the current available citations were published before 1995. The remaining 30% refer to publications dating from 1981-1995. Further assessment, such as to ranking according to authors and publications most quoted is useful and will be presented. For example: “Which papers with respect to which authors have been most often cited?”. Such queries, however, have to be entered "manually". In the framework of this submission, therefore, graph-theory will be utilized in order to make a wider net of relationships visible.

2.2 Pre-processing of raw data

In order to provide relevant results, several issues had to be solved first. If the input data are not correct and/or properly prepared, or if they are analysed under wrong assumptions, the analysis will not be relevant. Having in mind that the same items (articles) have several records, and/or have different descriptions that were typed by authors who used the references, the following two issues were the main concern:

(1) How to identify parts of the fields, whole fields, and whole records that are (or should be) identical in one of the two databases?
(2) How to identify typing errors that are not spelling errors, and how to identify, add, or/and change missing or wrong data automatically?

Both issues are somehow related and require a similar approach: (1) they cannot be solved with just one single solution in one step – the process is iterative; (2) more than one measure of similarity must be used. The first step in the analysis procedure was creation of a similarity matrix based on different length measures of field values, for example Euclidian distance, angle based distances, asymmetrical Levenstein or edit distance.

The problem of identifying errors is more complex since it includes more details, and it depends on human factors. Most of the problems were related to the database of references, which does include duplicates of records that have quite different descriptions. In order to be able to decipher paths through references it is extremely important to map and assign the same identifiers to the entries in the database of references as well as to the entries in the main CUMINCAD database.

The selection of canonical form that would uniquely identify authors was quite straightforward:

- For example, variations of descriptions: "Kalay, Yehuda E.", "Kalay Y E ", and "Kalay, Y.E. " are identified as Kalay.Y.E.
- Multiple authors like “Zilles, S.N., Lucas, P., Linden, T.M., Lotspiech, J.B. and Harbury, A.R.” would be transformed into array with five elements: (1) “Zilles.S.N.”; (2) “Lucas.P.”; (3) “Linden.T.M.”; (4) “Lotspiech.J.B.”; (5) “Harbury.A.R.”.
- Missing or incomplete author data was automatically detected from related records based on associated fields. Identified inconsistent use of authors’ names was additionally improved through mapping of names to main (most frequent) identifier.

The next important issue in the context of citation analysis of sources is the identification of two identical or similar articles. For example, two articles may be considered the same or similar if: (1) title, year and author(s) and other are exactly the same; (2) author(s), title, etc. are almost the same; (3) they have the same textual content; (4) they have the same citations; (5) they have the same usage. Based on preliminary analysis and available metadata, the following combinations were used to determine paper identfier:

- Canonical form of the author(s) + year + similar title, different distance measures were used to determine similarity;
- Publishing year + extracted words from the title, extraction was based on simple elimination of stop words;
- Canonical form of the author(s) + extracted words from the title
- Missing or incomplete title data was replaced with data from related records if at least first author and year matched and if the normalized distance measure was relatively high (at least 80 %).

2.3 Processing of references

One of the main purposes of automated analysis is to detect similarities. The similarity depends on the definition of equality, which is related to the definition of canonical form. The aim was to create a canonical form that can also serve as the identifier. The processing of data was primarily focussed on the relationships between references, as well as preparation of the results in the form, which would be useful for further analysis. Therefore, the following major steps were carried out:

1. Select analysis parameters based on preliminary statistical data;
2. Eliminate sources of errors and assure that assumptions are satisfied;
3. Establish relationships between papers;
4. Evaluate relationships;
5. Visualize the results in the form of graphical models.

3. RESULTS

In this section the results of the analysis will be presented. First, we give an overview of quantitative analysis where we address two interesting questions about referencing habits: (1) How many references do authors use, and (2) How "old" are the references. Second, some qualitative results cover specific relations between chronological developments, with regard to topics and authors illustrating the growth of the scientific community and specific citation related topics.

3.1 Quantitative analysis of referencing habits

The paper with the highest number of references (70 - in words: seventy) is (Bermudez et. al., 2000). However, the average number of references per paper is around 10.4. A Distribution of the number of papers related to the corresponding number of references is presented on the figure below.

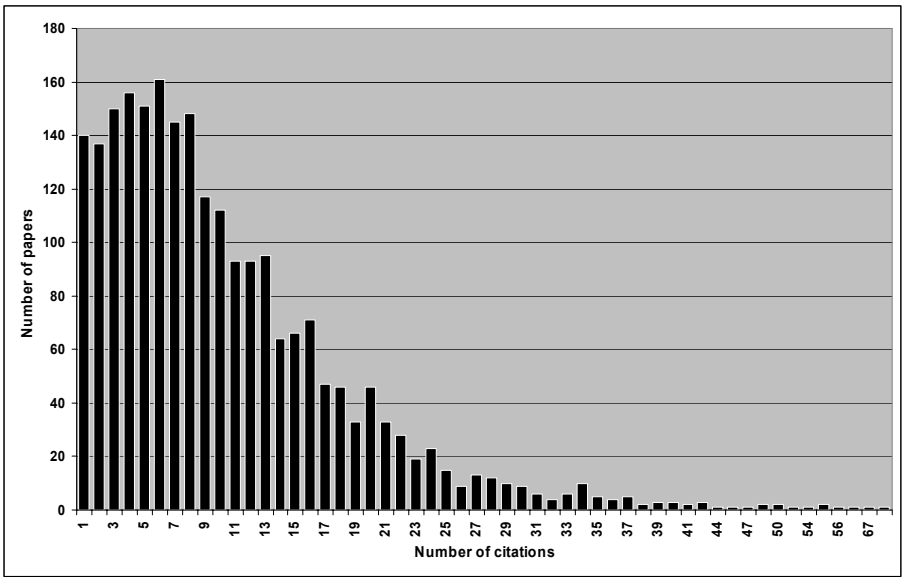

Figure 3. Frequency of citations

The analysis shows that 60% of the papers provide 1-10 references; around 30 % of the papers have 10-20 references; slightly more than 7 % have 20-30 references and less than 3% attach more than 30 references. A distribution of citations in Figure 4 is presented in such a way as to be comparable with analysis presented at (Open Citations project, 2002).

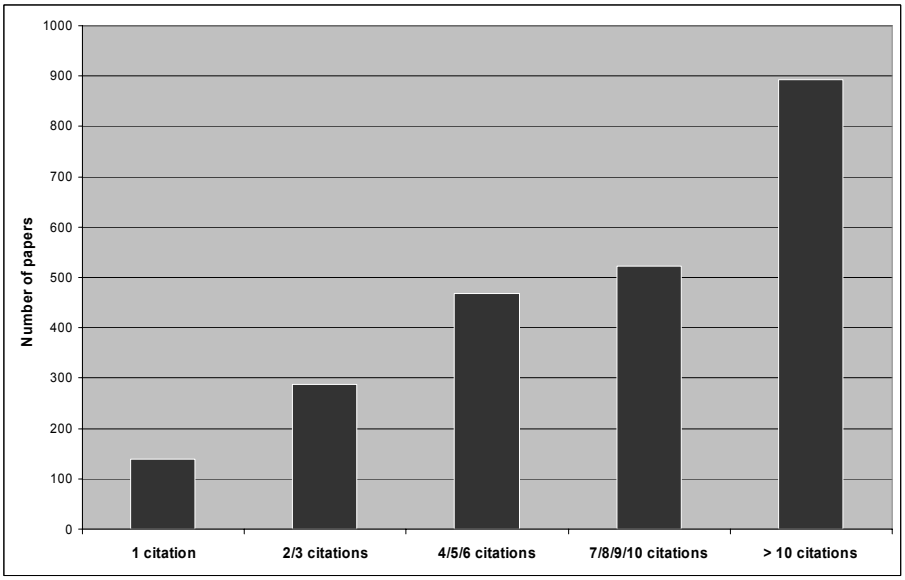

Figure 4. Distribution of citations

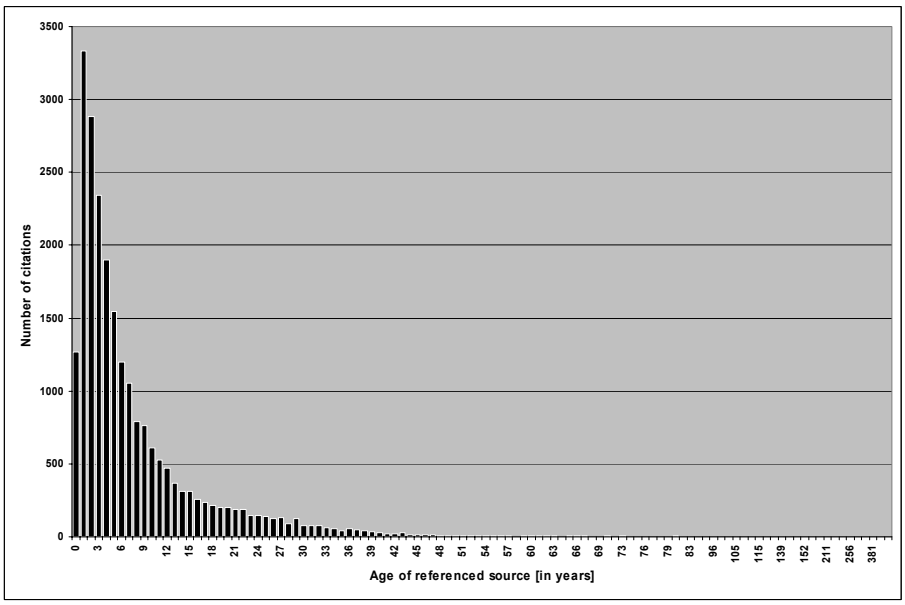

Figure 5. Age of citations

The age of citations may show the growth of interest in a particular topic, or in general – the speed of knowledge exchange. The relationships between citations in the chronological sense may be represented in the form of year-to-year citation networks. An example of such network representing citations younger than seven years is presented in the next figure.

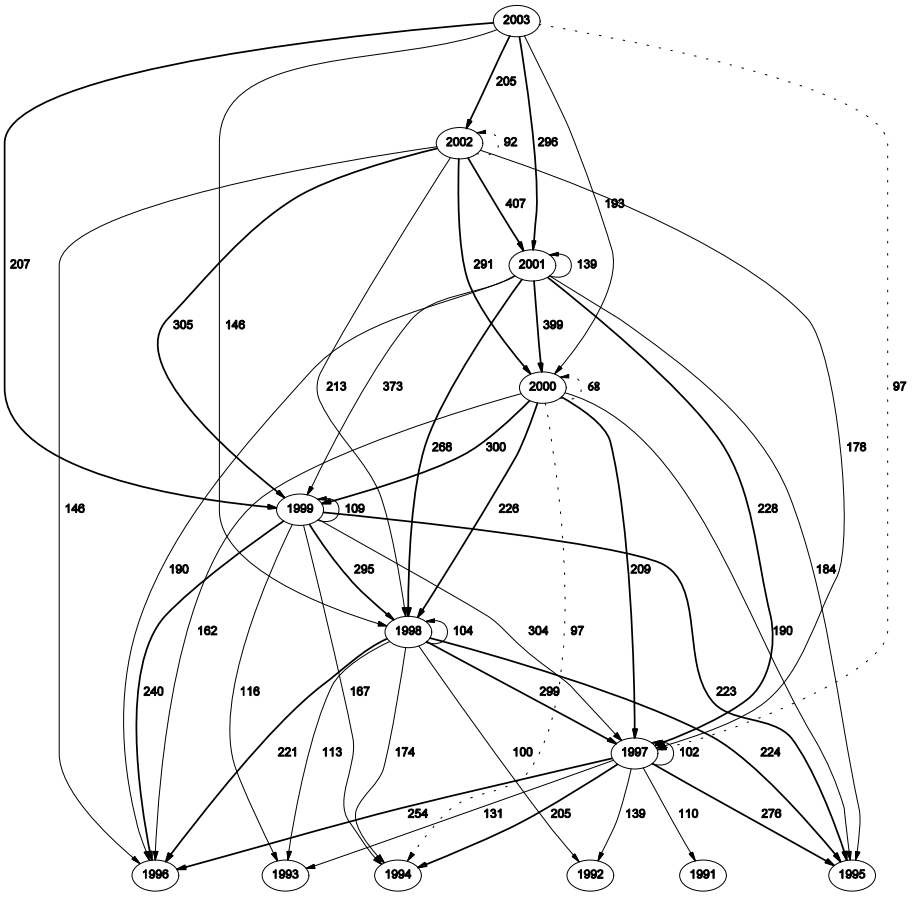

Figure 6. Year-to-Year relationships between citations

The vertices in the graph (Figure 6) represent all papers that were published in a certain year. Directed edges from vertex *i* to vertex *j* denote that papers published in year labelled on vertex *i* have referenced papers that were published in year labelled on vertex *j*. These edges are visualised in three styles and reflect a number of year-to-year citations (dotted < 100, thin > 100, and thick lines > 200). Labels on edges represent the exact number of references.

A normalized view on the age of references is shown in the figure below.

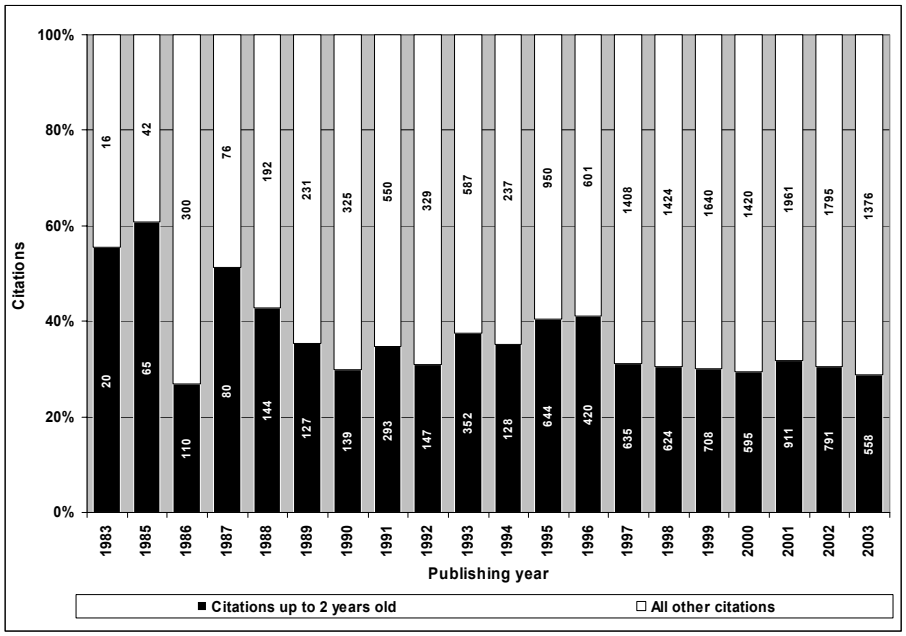

Figure 7. Distribution of citations used in papers published in the last two decades

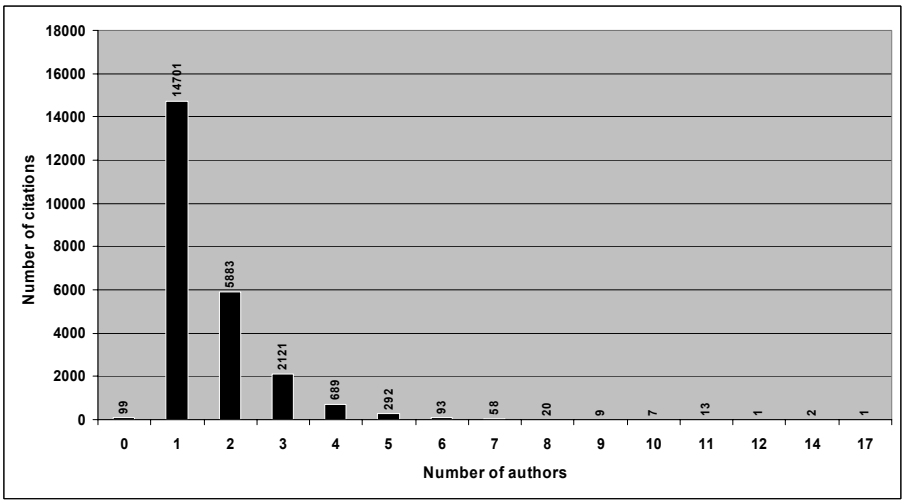

Figure 8. Distribution of citations per number of authors

3.2 Analysis on authors of cited material

There is just a small number of unknown sources (99); from the remaining part most of the references (60 % or 147001) were written by a single author, 25 % of the referenced papers were written by 2 authors, a bit less than 10 % by three, and 4 or more persons wrote less than 5 % of cited papers. The cited paper with the highest number of authors (17) is (Akin et al, 1997).

About 12500 different individual authors were authors of 24000 cited sources, precisely: in average each author was cited 1.92 times. A bar chart illustrating number of citations of twenty-five most popular authors is shown on Figure 9. These authors cover about 20 % of citations in the CUMINCAD database of references. About half of the most cited papers were written individually, and equal shares (25 %) represent number of papers that were written by most cited authors whether as first authors or as co-authors.

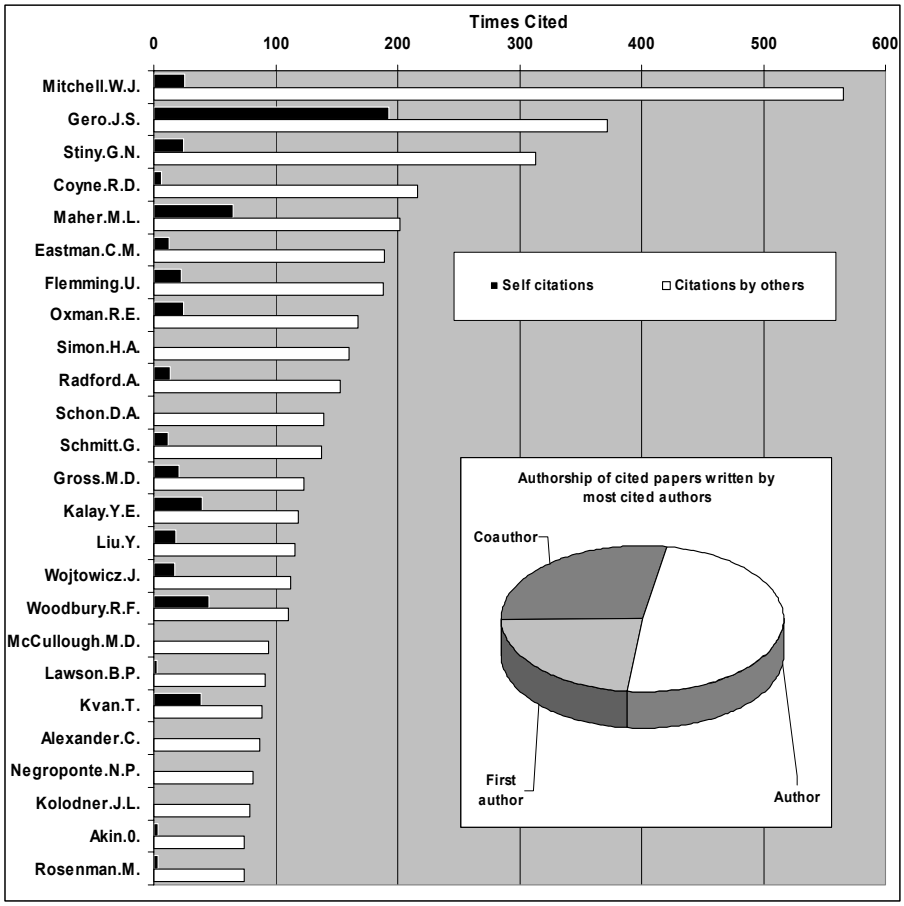

Figure 9. Top 25 most cited authors

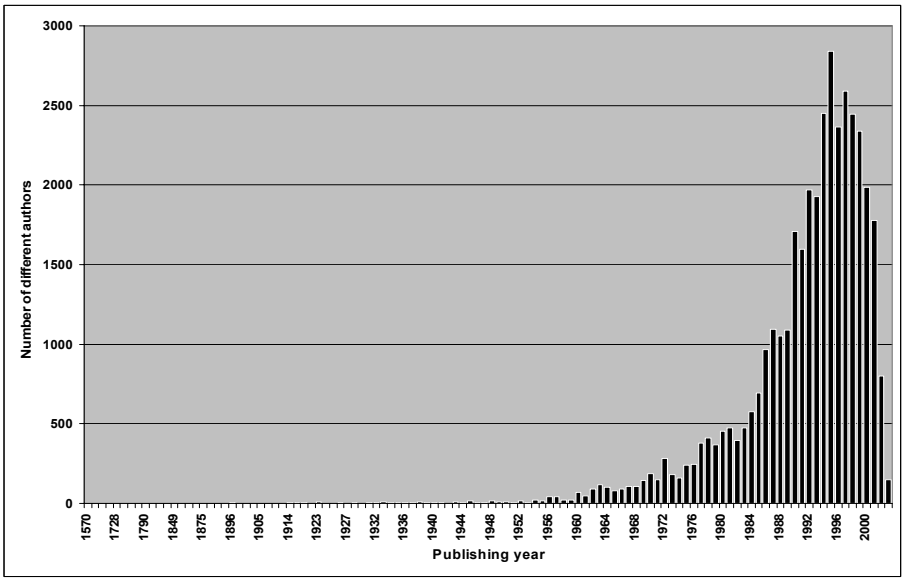

Figure 10. Number of authors per year

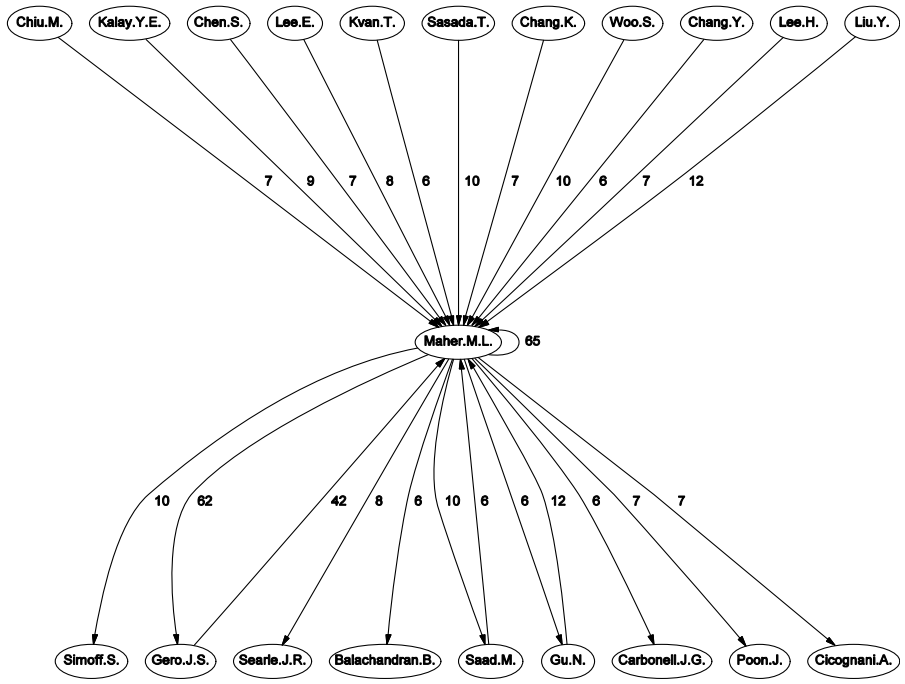

Figure 11. Example of citation-network

The following two graphs illustrate related scientific efforts, and collaboration, as a further exploration of authors' activities that were included in the network visualized in Figure 11.

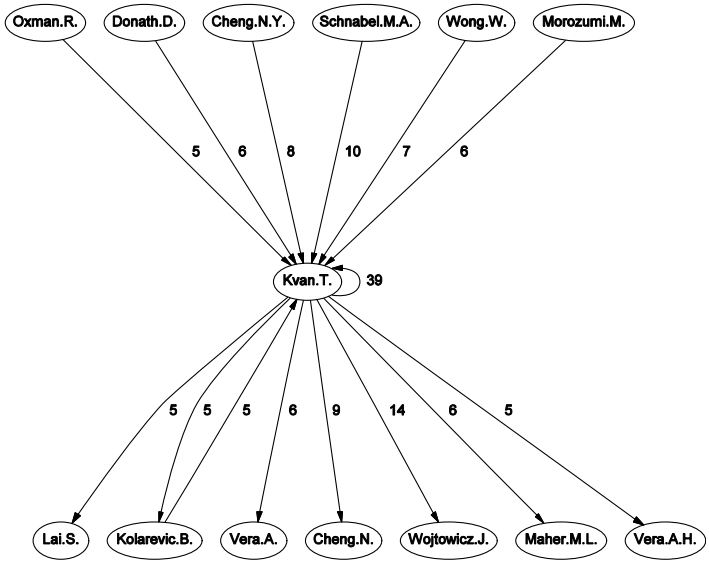

Figure 12. Example of author-citations network

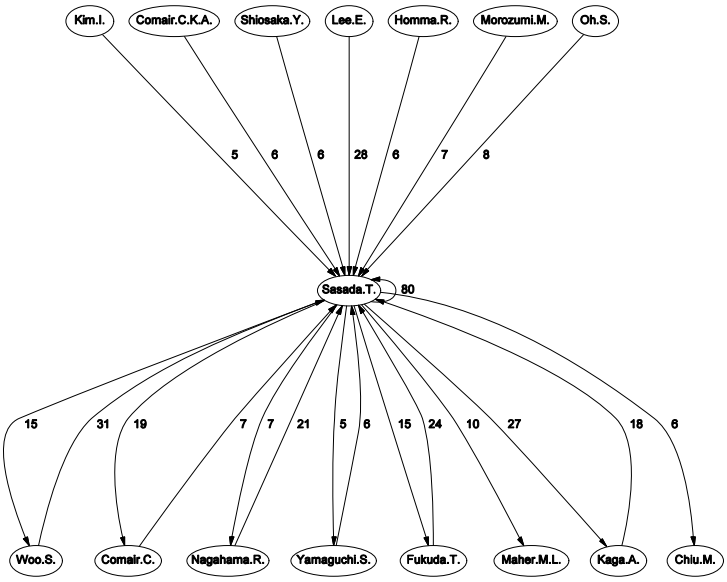

Figure 13. An example of a well established network

The results could be integrated into the CUMINCAD profile (see Figure 14).

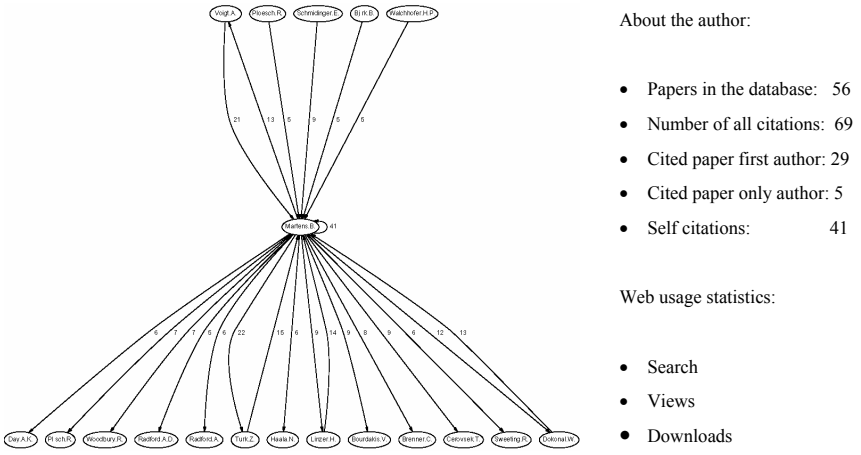

Figure 14. An example of authors' profile details as possible CUMINCAD feature

4. DISCUSSION AND CONCLUSION

This section gives an overview of the analysis with comments on interesting findings. It covers citing habits, scientific information exchange, community networks, as well as exploitation of the results and future work.

Citing habits. The paper gives an insight into citing habits illustrated on Figure 3. Frequency of citations, and Figure 4. Distribution of citations. A comparative analysis of citing habits clearly indicates that CAAD community uses many more references compared to other fields presented in (Open Citation project, 2002). The conclusions could be twofold: CAAD research is much broader - multidisciplinary, and that it therefore requires a broader overview on studied topics - since citing reflects literature study. A complementary analysis on web usage statistics would offer additional information on specific articles. Namely, citations only provide information about the articles that were cited and not about the content that wasn't. That is made possible through the use of digital libraries, since we can gain more information on papers' usage – i.e.: they have been read, abstracted viewed, or their full-texts have been downloaded but they were never cited. Embedding such relationships into the digital library's functionality would give an added value and completely new view to the impact of scientific works.

Chronological view of scientific information exchange. - From Figure 5: Age of citations, we can conclude that most scientific ideas evolve over a period of 2 years. Although the specimen might be considered too small, from observation of Figure 6. Year-to-Year relationships between citations, and Figure 7. Distribution of citations used in papers published in the last

two decades, as well as additional analysis of papers containing citations, we can claim that importance of digital libraries is evident and that they highly influence the evolution and directions of research topics, and their transfer.

Scientific collaboration networks. From the Figure 9. Top 25 most cited authors, it is evident that the most influential authors are most frequently the first authors or write papers solely. This paper describes how networks between papers (and especially their authors!) in the area of Computer Aided Architectural Design can be identified with the help of graph theory. Often two or more authors are connected as illustrated on Figure 11, and Figure 12 (people that are in focus of the networks or have connections, happen to be on the list of most cited authors), or they build self-sustained networks as shown on Figure 13. From Figure 1. Chronological distribution of CUMINCAD papers, one could claim that number of authors that are cited is constantly falling since 1996, but it is not so – the number depends on age of cited papers (Figure 5).

Exploitation of the results. First of all, an "enabler" for scientific collaboration has to be mentioned as the developed mechanisms heavily support the task of finding right persons (as potential collaborators in the future). This is facilitated already by conferences, but a machine-identified procedure would offer a complementary way. In this respect tools are needed to improve efficient knowledge transfer. There is such a rich body of research in CAAD and therefore a clear need to allow researchers ready access so that they can minimize these side effects. Second, research by means of a well developed Citation Index and measurement of Impact Analysis would work out well for CAAD as well. Repeating research already undertaken by others and the failure to evaluate is nothing new. And last, but not least, it is important to emphasise that possible use of the results in the form illustrated on Figure 14 would significantly improve the scientific digital-library end-user experience. With the help of smart user interfaces it would enable easier literature study as well as establishing of social networks. This will enhance representations (keywording system, classification of papers, for optimal access, browsing and search routines), which will be more easily generated with the help of text mining, clustering and usage analysis techniques. The results could significantly improve the precision of the search results by embedding automated, low-cost procedures - handling on a shoestring budget defines the focus for future work. The paradigm is not just to try to copy ISI or related services, but to keep "science in the hands of scientists".

ACKNOWLEDGEMENTS

The presented work has been conducted in the context of the SciX project, funded by the European commission under the contract IST - 2001 - 33127. SciX is a 24-month project with an EU funding of €1.000.000. Co-ordinated by the University of Ljubljana (Slovenia), the partners include Swedish Business School of Finland, Icelandic Building Research Institute, an e-Business company Indra/Atlante (Spain), Vienna University of Technology (Austria), FGG Institute (Slovenia) and the University of Salford (UK). The home page of the SciX project is at http://www.scix.net/.

REFERENCES

Akin, O., Aygen, Z., Chang, T.-W., Chien, S.-F., Choi, B., Donia, M., Fenves, S.J., Flemming, U., Garrett, J.H., Gomez, N., Kiliccote, H., Rivard, H., Sen, R., Snyder, J., Tsai, W.-J., Woodbury, R. and Zhang, Y., 1997, SEED: A Software Environment to support the Early phases of building Design, The International Journal of Design Computing. http://www.arch.usyd.edu.au/kcdc/journal/index.html

Bermudez, J., Agutter, J., Westenskow, D., Foresti, S., Zhang, Y., Gondeck-Becker, D., Syroid, N., Lilly, B., Strayer, .D. and Drews, F., 2000, Data Representation Architecture: Visualization Design Methods, Theory and Technology Applied to Anesthesiology, Eternity, Infinity and Virtuality in Architecture [Proceedings of the 22nd Annual Conference of the Association for Computer-Aided Design in Architecture] Washington D.C. 19-22 October 2000, pp. 91-102.

Garfield, E., 1997, Concept of Citation Indexing: A Unique and Innovative Tool for Navigating the Research Literature. Far Eastern State University Vladivostok - September 4, 1997. http://www.garfield.library.upenn.edu/papers/vladivostok.html

Hauffe, H., 1994, Is Citation Analysis a Tool for Evaluation of Scientific Contributions? 13th Winter workshop on Biochemical and Clinical Aspects of Pteridines, St.Christoph/Arlberg, Feb.25, 1994. http://www.uibk.ac.at/sci-org/voeb/texte/vhau9402.html

Martens, B. and Turk, Ž., 1999, Working Experiences with a Cumulative Index on CAD: "CUMINCAD", Proceedings, ECAADE99 Conference, Turing to 2000, Liverpool, Spetember 15-17, 1999. 327-333.

Open Citation Project - Reference Linking and Citation Analysis for Open Archives, 2002, http://opcit.eprints.org/

OAI - Open Archives Initiative, 2004, http://www.openarchives.org/

Thomson, 2004, Thomson ISI – Citation Indexing. History of Citation Indexing http://www.isinet.com/essays/citationindexing/

Turk, Z., T. Cerovsek and B. Martens, 2001, "The Topics of CAAD: A Machine's Perspective", in: *Proceedings CAAD Futures 2001, Eindhoven, The Netherlands*, Kluwer Academic Publishers, p. 547-560.