Citeology: Visualizing Paper Genealogy

http://www.autodeskresearch.com/projects/citeology

Justin Matejka
Autodesk Research
210 King St. E.
Toronto, ON Canada
Justin.Matejka@Autodesk.com

George Fitzmaurice
Autodesk Research
210 King St. E.
Toronto, ON Canada
George.Fitzmaurice@Autodesk.com

Tovi Grossman
Autodesk Research
210 King St. E.
Toronto, ON Canada
Tovi.Grossman@Autodesk.com

Abstract

Citeology is an interactive visualization that looks at the relationships between research publications through their use of citations. The sample corpus uses all 3,502 papers published at ACM CHI and UIST between 1982 and 2010, and the 11,699 citations between them. A connection is drawn between each paper and all papers which it referenced from the collection. For an individual paper, the resulting visualization represents a "family tree" of sorts, showing multiple generations of referenced papers which the target paper built upon, and all descendant generations of future papers.

CHI 2012, May 5–10, 2012, Austin, TX, USA.
ACM xxx-x-xxxx-xxxx-x/xx/xx.

Figure 1. Citeology for *Spotlight* [9]. *(Note: high resolution vector version in Appendix A)*

Keywords

Citations, References, Information Visualization

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction and Related Work

Research publications use citations to provide detailed references to previous works which have a bearing on the subject of the current publication. It is simple to see which sources an individual paper has cited (they are generally listed in a section at the end of the paper), and with online article repositories such as the ACM Digital Library (dl.acm.org) and Google Scholar (scholar.google.com) it has become straightforward given a particular paper to see which future articles it was referenced by.

To provide users with an alternative view of citation trends, compelling visualizations of citation trends have also been explored. From the research community, Mackinlay et. al. created the Butterfly system [9] which is an interface for accessing citation databases and visualized the paper search results with scatterplots and 3D piles. More recently, CiteSense [15] was designed to support literature search, selection, and comprehension. Eigenfactor [2] looks at citation data from articles to rank the strength of journals, and Well-formed Eigenfactor [14] has produced compelling visualizations based on this data at the journal level. At Alt.CHI 2009, Kaye presented an analysis of HCI publications [8] focusing on issues at the individual author level, and MacKenzie [10] looked at HCI citations from CHI, TOCHI, and the HCI Journal to rebut the claim that HCI researchers have "little or no impact".

If one considers a research article in a genealogical sense, the papers which an article referenced could be considered the article's "ancestors" or "parents" and the papers which referenced the target article could be considered "descendants" or "children". Traditional methods let you see the *parents* and *children* of a particular paper. Such information can be useful when tracing the history of a piece of work or trying to find related articles. It is however generally presented textually and there is no way to look at multiple generations of ancestors or descendants or get a feel for the overall connection network of the corpus.

Citeology (a portmanteau of *citation* and *genealogy*) is an interactive visualization designed to look at the relationships between publications through their use of citations (Figure 1). The relationships are shown in the context of a collection of papers published in similar venues.

The Citeology System

For the initial deployment of Citeology we used a dataset of all 3,502 papers published at ACM CHI and ACM UIST between 1982 and 2010 and the 11,699 citations made among these papers. The papers are organized in vertical columns by year and the first 25 or so characters of each paper title are displayed. The papers are sorted placing the papers with the most citations for each year in the middle of their respective column, so the papers with most citations in each year can be found along a horizontal band through the middle of the diagram (Figure 2). The years 1984 and 1987 have been omitted as neither of the CHI nor UIST conferences were held in those years.

In the interactive version the titles are rendered in a 2pt font and too small to read, but in the generated PDF files it is possible to zoom in and read the titles. Visible from the arrangement of the papers into yearly columns is the increase in the number of publications (mostly at CHI) starting in 2006.

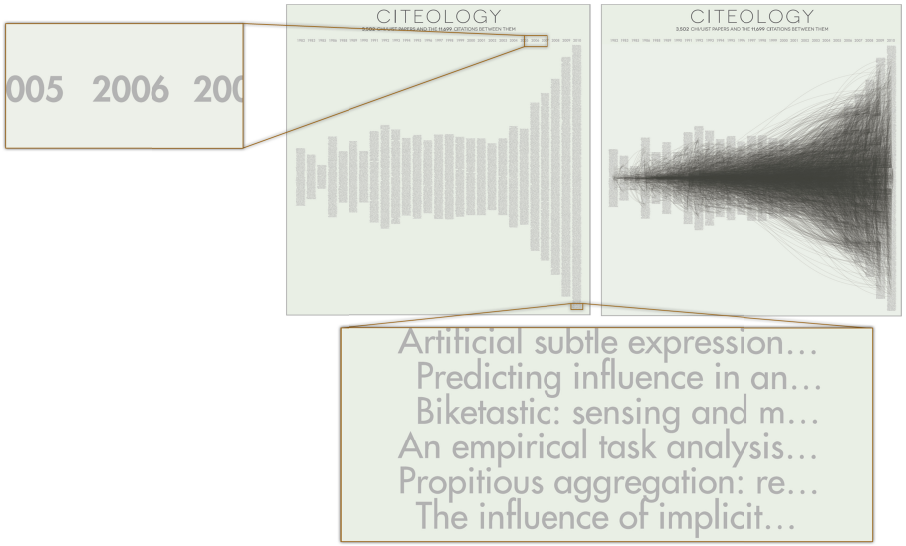

Figure 2. Citeology of all CHI and UIST papers before and after the citation network has been drawn, and a zoomed in view of the years and the paper titles.

Once the paper titles are in place, connections are drawn between each paper and all of the papers from the collection which it referenced (Figure 2). Hovering over an individual paper gives a tooltip of with the paper title (Figure 3) and provides and way to scan the database looking for interesting paper titles.

Figure 3. Feedback when over a paper title.

Clicking on a single paper lets you see the Citeology for that paper. Figure 1 shows the Citeology for *Spotlight: directing users' attention on large displays* [9] from CHI 2005. Descendants of the paper are linked to with blue lines, and ancestors are connected in red. By default all generations of relatives in both directions are shown but controls are available to limit the display to any number between 1 and 8 generations (Figure 4). First generation relatives are connected with relatively thick and opaque curves, and the lines becomes thinner and more transparent as the generational gap increases (Figure 6). The names of all related papers are darkened, with the first generation associations being the darkest and further generations becoming lighter. Papers not connected to the target paper are not darkened and remain the original pale grey colour.

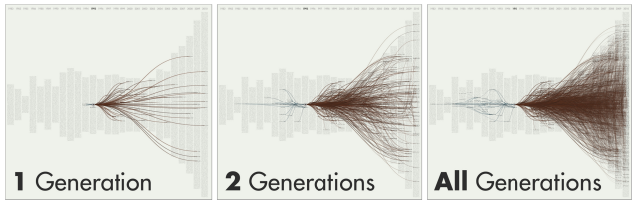

Figure 4. Citeologies showing 1, 2, and All generations from the CHI 1995 paper *Bricks: laying the foundations for graspable user interfaces* [5].

Once a paper has been selected, hovering over other papers will display the shortest connection, if one exists, between the *hovered* paper and the *selected* paper. The path is traced on the visualization and the papers along the shortest path are displayed in the upper left section of the screen (Figure 5).

By clicking on the "Find Paper…" button papers can be found using words in the title or author's, and the official ACM page for a paper can be accessed through the "visit paper page" button.

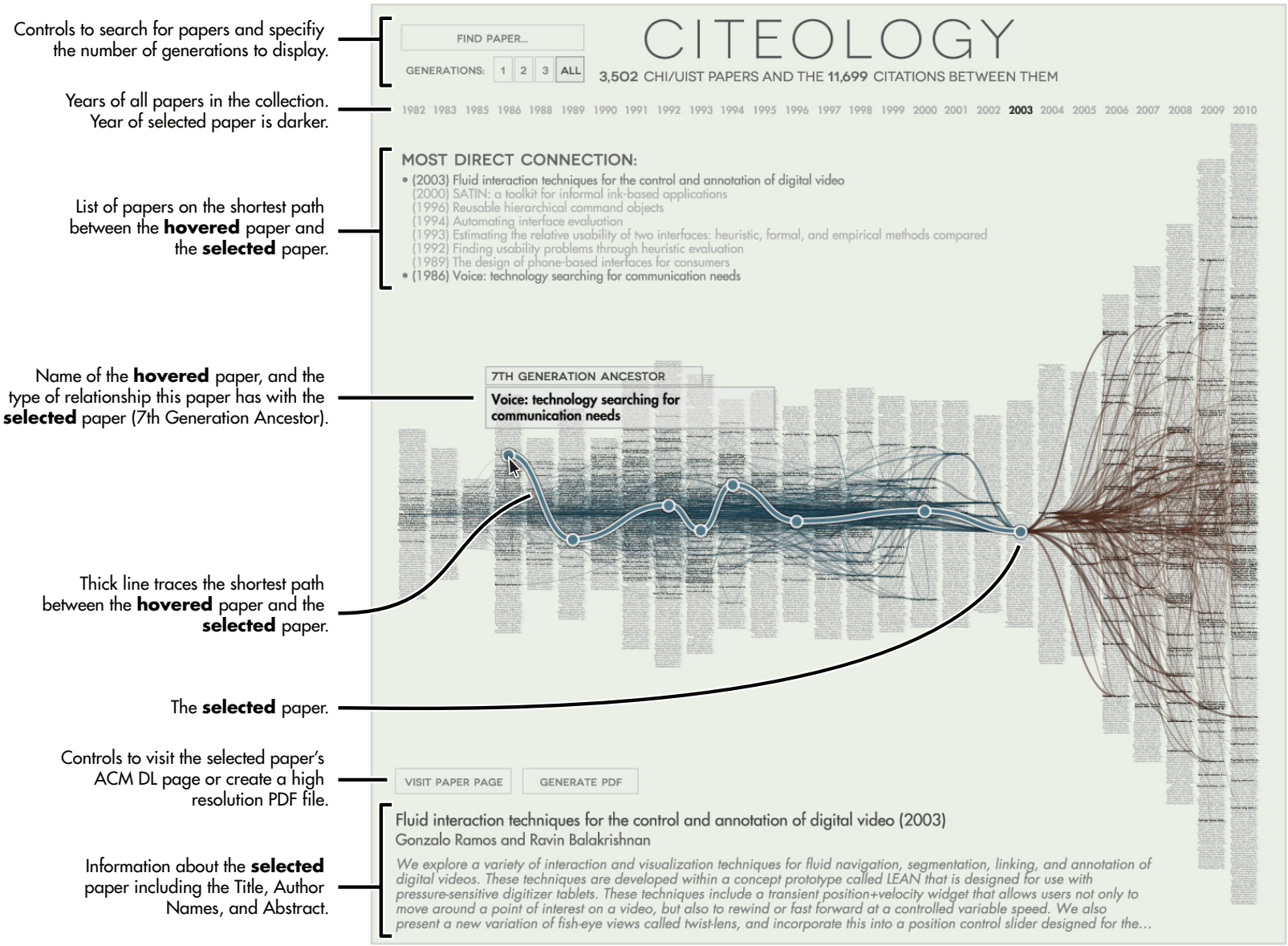

Figure 5. Main Components of the Citeology Interface.

Figure 6. A zoomed in view of a high-resolution exported Citeology.

Implementation

The Citeology system is implemented as a Java applet and makes use of the Processing[1] libraries to simplify the on-screen drawing and pdf-generation procedures. Being a Java applet allows the system to be embedded into a webpage, but a standalone application runnable on Windows, Mac OS, and Linux is able to be built from the same source code base. The applet has been signed allowing for anonymous usage data collection into an Amazon AWS SimpleDB database as well as saving generated PDF files to the user's local file system.

DATA FORMAT

The information for each of the articles in the collection and its associated citation information is stored in a tab-delimited text file using a relatively straightforward format (Table 1). All of the data needed to create the visualization is stored in this file, making it possible to create Citeologies for other collections of conferences by simple creating a new data file.

For the fields which can have more than one distinct entry (`Authors` and `References`) a tilde character (~) is used as a secondary delimiter. The Digital Object Identifier (DOI) system is used as the unique identifier for each article since it is widely adopted for academic publishing and gives an easy way to refer back to the official web page for the publication. This can be accomplished by entering a URL of the form: `http://dx.doi.org/<doi>` which will resolve to, in the case of CHI and UIST papers, the ACM Digital Library entry for the given paper.

Column	Example
Conference	CHI
Year	2005
Title	Spotlight: directing users' attention on large…
Abstract	We describe a new interaction technique…
Authors	Azam Khan ~ Justin Matejka ~ George Fitzmauric…
DOI	10.1145/1054972.1055082
References	10.1145/108844.108845 ~ 10.1145/275519.275520…

Table 1. Citeology data file format.

Interesting Findings

It is straightforward to see how many 1st generation descendants a paper has using traditional paper reference websites like the ACM Digital Library and Google Scholar so we thought it would be interesting to see which papers have the most descendants over multiple generations (Figure 7).

The paper with the most 1st generation descendants is *Generalized Fisheye Views* [6] from CHI 1986 with 89 direct citations from within the CHI/UIST collection. This paper remains in the top spot for up to five generations of descendants with 1,663 but when we look at all descendants up to six generations, *A Study in Two Handed Input* [4], also from CHI 1986 becomes the most prolific with 1,760 descendants. From seven generations onward, the paper with the most total descendants is the CHI 1983 paper *Evaluation and analysis of users' activity organization* [1] which has 1,887 descendants of 7th generation or less, and 2,120 overall. An amazing 62% of all papers published after 1993, and 84% of the papers published in 2010 at CHI and UIST are descendants of this paper.

[1] www.processing.org

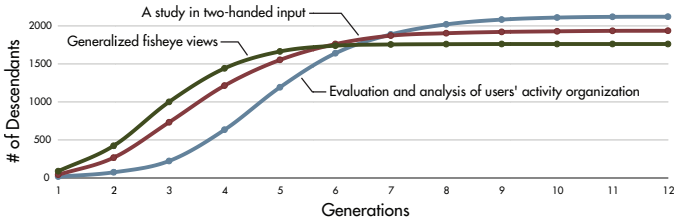

Figure 7. Plot of the number of total descendants up to a given generation for each of the three papers with the most descendants.

When hovering over a paper, we calculate the shortest path between that *hovered* paper and the currently *selected* paper. Out of curiosity we looked for the longest direct connection between any two papers in our collection and it turned out to be an 18 generation gap between the CHI 1985 paper *A theory of stimulus-response compatibility applied to human-computer interaction* [7] and *The effects of empathetic virtual characters on presence in narrative-centered learning environments* [12] from CHI 2008.

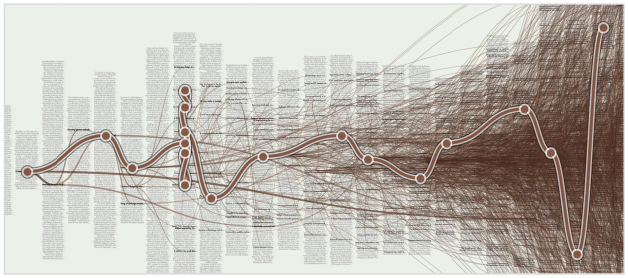

Figure 8. Longest most direct path between two CHI papers. (18 generations)

When exploring the visualization we see that the majority of citations are bundled along the middle row of the papers. This is of course a result of the papers being sorted with the most cited papers being placed in the middle, but there are a surprising number of CHI and UIST papers which have never been referenced by a paper in the collection. In fact, 934 of the 3159 papers published between 1982 and 2009 have not been referenced by another CHI or UIST paper. This could be because those papers have not been of interest to any future researchers, but perhaps there are also some good papers in the un-referenced collection which have just been forgotten and would be a fruitful place to look for relatively unexplored research topics.

Usage Observations

Prior to the deadline for this paper, Citeology has been deployed publically for approximately 3 weeks. In that time the applet has been run 3,681 times from 2,977 unique host computers. During those sessions, the user community selected 24,378 papers for viewing. The results of which papers were selected the most can be seen in Figure 9. The darker red labels indicate papers with higher click counts, and from the diagram we can see that people were most interested in the papers at the very start and end dates of the collection, as well as the band of most popular papers which run through the middle of the diagram. The majority of papers had at least one click, with only 207 of the 3502 (5.9%) of the papers receiving 0 selections, and the most selected paper has been Richard A. Bolt's CHI 1982 paper, *Eyes at the Interface* [3], which is most frequently paper cited in the collection from 1982.

During these sessions the search functionality was used 1295 times, but strong patterns for common search terms are not apparent in the data.

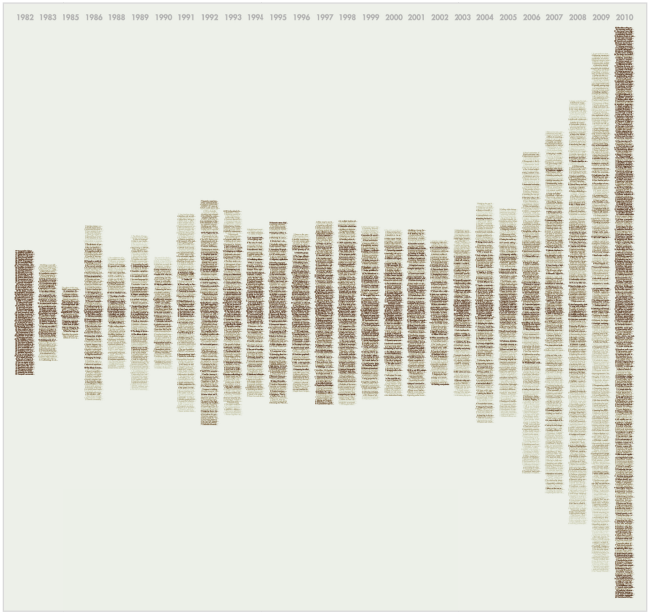

Figure 9. Heatmap of click counts per paper during 3 week deployment. Ligher yellow names were clicked less, and dark red ones were clicked more.

Discussion and Future Work

Since the database used in this initial deployment of Citeology uses only the papers published at CHI and UIST, not all dependency relationships can be captured. For example, if paper ***A*** (at CHI) is referenced by paper ***B*** (at AVI) which is referenced by paper ***C*** (at CHI), paper ***C*** will not show up as a descendant of paper ***A*** in the current visualization. Expanding the breadth of the articles in the database would help create a more complete picture of the genealogy of a paper.

We started out with a Human Computer Interaction focused collection of papers since that is our primary research interest. Given the simple text-based data format that the tool takes as input we hope to be able to expand the system to work for different conferences and journals in completely different areas. It would also be interesting to look at other areas where references are used such as patents or legal cases and see what kinds of differences reveal themselves in these different contexts.

Citeology could be a useful tool for finding related work when working on a research article. However, it would be interesting to broaden the collection of searchable papers to include topics outside of the main focus of interest to encourage the broadening rather than narrowing of influences on research. A system like this could have the effect of exposing researchers to completely new conferences or topic areas they were not previously aware of.

Since the references associated with academic papers do not typically have any contextual information associated with them, Citeology currently treats all citations as equal. But papers can be referenced not only because they are good examples of prior work which the new paper builds upon, but papers are often referenced for obtaining background information, or maybe the author disagrees with their finding. There are many other reasons that work can be cited which are described in the CiTO system [13] and it would be useful to be able to categorize (manually, or automatically) citations based on how they are used in context.

The lack of zooming capability in the applet is something we would like to address in a future version. Currently the work around is to generate a PDF file for

closer inspection. Implementation wise, it would be nice to redo the application as an HTML5 Canvas based page to remove the current Java installation dependency.

Acknowledgements

The authors thank all of our former interns for beta testing and sharing the site with their colleagues. We would also like to thank the over 500 Twitter and Google+ users who have posted about Citeology, especially the following:

@hiroshi_ishii Hiroshi Ishii

Citeology by Autodesk Research is great! Visualization of the research paper citation network. CHI & UIST in 1982~2010.

@benbendc Ben Shneiderman

Effective visualization for CHI & UIST papers from Autodesk – Citeology. #infovis

@gtsakonas Giannis Tsakonas

This is sheer beauty. Citeology visualizes citations of ~30 yrs of CHI and UIST Human Computer Interaction confs.

@cuanbela Adrian Giddings

Beautiful live infovis work by Autodesk and bonus points for actually being useful.

Max Wilson (on Google+)

Extremely interesting visualization of CHI and UIST cross citation data. Very fun to play with and explore.

↳  Mary Czerwinski *+1'd this*

References

[1] Bannon, L., Cypher, A., Greenspan, S., and Monty, M.L. (1983). Evaluation and analysis of users' activity organization. *ACM CHI*. 54-57.

[2] Bergstrom, C.T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, Vol. 68, No. 5, pp. 314-316.

[3] Bolt, R.A. (1982). Eyes at the interface. *ACM CHI*. 356-362.

[4] Buxton, W., and Myers, B. (1986). A study in two-handed input. *ACM CHI*. 321-326.

[5] Fitzmaurice, G., Ishii, H., and Buxton, W. (1995). Bricks: laying the foundations for graspable user interfaces. *ACM CHI*. 442-449.

[6] Furnas, G.W. (1986). Generalized fisheye views. *ACM CHI.* 16-23.

[7] John, B.E., Rosenbloom, P.S., and Newell, A. (1985). A theory of stimulus-response compatibility applied to human-computer interaction. *CHI.* 213-219.

[8] Kaye, J. (2009). Some statistical analyses of CHI. *Alt.CHI 2009 (CHI EA).* 2585-2594.

[9] Khan, A., Matejka, J., Fitzmaurice, G., and Kurtenbach, G. (2005). Spotlight: directing users' attention on large displays. *ACM CHI*. 791-798.

[10] MacKenzie, I.S. (2009). Citedness, uncitedness, and the murky world between. *Alt.CHI 2009 (CHI EA).* 2545-2554.

[11] Mackinlay, J.D., Rao, R., and Card. S.K. (1995). An organic user interface for searching citation links. *ACM CHI.* 67-73.

[12] McQuiggan, S.W., Rowe, J.P., and Lester, J.C. (2008). The effects of empathetic virtual characters on presence in narrative-centered learning environments. *ACM CHI*. 1511-1520.

[13] Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics.* 1:S6.

[14] Well-Formed Eigenfactor: Visualizing information flow in science. *http://well-formed.eigenfactor.org/*.

[15] Zhang, X., Qu, Y., Giles, C.L., and Song, P. (2008). CiteSense: supporting sensemaking of research literature. *ACM CHI.* 677-680.

Continue reading to see:
Appendix A – High Resolution Vector Citeology
⬇

Appendix A – High Resolution Vector Citeology

This is a sample of the type of PDF output generated by the Citeology applet. Zoom in to see more detail.

Citeology is available to try here: *http://www.autodeskresearch.com/projects/citeology*