CHI DATABASE VISUALIZATION

Niko Vegt

Introduction

The CHI conference is leading within the field of interaction design. Thousands of papers are published for this conference in an orderly structure. These publications can be found rather easily if you know the right query words, but how about providing inspiration and supporting exploration?

The CHI papers are implemented in a database, which provides a good opportunity to explore possibilities of visualizing the data. What kind of information is meaningful to visualize? And how can it support the exploration of information? The database contains of several parameters for content description per paper: general terms, keywords and a hierarchical categorization system.

The general terms are the simplest form of descriptive parameters of the CHI papers. There are 16 options, and every paper can be specified with several general terms. The keywords of a paper are unsuitable for this module, because they are unstructured. They can be freely used by the author. The categorization system is interesting because it is formalized and has a hierarchical structure. This hierarchy is also implemented in the database. There are more papers specified with general terms (±3500) than with categories (±3000) though. The categories serve a more detailed image, with a large amount of unique options. It seems more appropriate to start on a generic level, i.e. with the general terms, to provide an overview.

Analysis

The Access database of CHI papers contains 6298 unique papers. For 3524 of these papers general terms are specified, this is 56% of the total amount of papers. It appears that the database isn't complete with respect to the general terms, because a random search on the ACM online database shows that a paper with no general terms in the Access database does contain the specification with general terms online.

The coverage of general terms changes significantly per year. The first ten years (1981-1991) of the conference is almost complete, whereas the coverage after that period differs between 100% (1995, 1998) and 4% (2006). All years with coverage rates lower than 90% were left out for data visualization. They only represent the incompleteness of the Access database. So 12 of 27 year records were removed. This is a data reduction of 29%.

copyright year	articles	with general terms	coverage	general terms	general terms per paper
1981	79	77	97%	77	1,0
1982	75	75	100%	108	1,4
1983	60	60	100%	168	2,8
1985	41	41	100%	243	5,9
1986	55	54	98%	128	2,4
1987	58	58	100%	325	5,6
1988	48	48	100%	140	2,9
1989	73	73	100%	190	2,6
1990	71	71	100%	169	2,4
1991	98	98	100%	209	2,1
1992	200	111	56%	250	2,3
1993	218	110	50%	289	2,6
1994	358	69	19%	137	2,0
1995	313	313	100%	735	2,3
1996	300	298	99%	1308	4,4
1997	280	77	28%	77	1,0
1998	292	292	100%	1108	3,8
1999	273	78	29%	78	1,0
2000	277	72	26%	72	1,0
2001	321	69	21%	69	1,0
2002	277	61	22%	61	1,0
2003	301	58	19%	58	1,0
2004	372	208	56%	370	1,8
2005	402	84	21%	158	1,9
2006	495	18	4%	36	2,0
2007	408	401	98%	666	1,7
2008	553	550	99%	550	1,0
Total	*6298*	*3524*	*56%*	*7779*	*2,3*

Table 1: Overall general terms coverage and distribution over time

The change in use of general terms over time is of first interest for visualization. To achieve a real picture, the distribution of every single general term (like Human Factors) was calculated over all published papers per year. Resulting in a general term distribution per year and the sum of all distributions resembled the amount of general terms used per paper.

General terms	1981		1982		1983		1985		1986		1987		1988		1989		1990		1991	
Algorithms	1	*0,013*		*0*		*0*	2	*0,049*		*0*	3	*0,052*	2	*0,042*	1	*0,014*	8	*0,113*	3	*0,031*
Design	2	*0,025*	17	*0,227*	24	*0,400*	40	*0,976*	38	*0,691*	58	*1,000*	31	*0,646*	50	*0,685*	46	*0,648*	67	*0,684*
Documentation	1	*0,013*	6	*0,080*	6	*0,100*	7	*0,171*	7	*0,127*	6	*0,103*	3	*0,063*	5	*0,068*	2	*0,028*	3	*0,031*
Economics		*0*		*0*		*0*		*0*		*0*		*0*	2	*0,042*		*0*		*0*	2	*0,020*
Experimentation		*0*	3	*0,040*	3	*0,050*	12	*0,293*		*0*	8	*0,138*		*0*	7	*0,096*	4	*0,056*	9	*0,092*
Human Factors	6	*0,076*	39	*0,520*	55	*0,917*	40	*0,976*	49	*0,891*	57	*0,983*	48	*1,000*	71	*0,973*	68	*0,958*	89	*0,908*
Languages	3	*0,038*	6	*0,080*	18	*0,300*	6	*0,146*	8	*0,145*	6	*0,103*	11	*0,229*	11	*0,151*	8	*0,113*	14	*0,143*
Legal Aspects		*0*		*0*		*0*		*0*		*0*		*0*		*0*	1	*0,014*		*0*	1	*0,010*
Management	27	*0,342*	7	*0,093*	10	*0,167*	31	*0,756*	20	*0,364*	57	*0,983*	26	*0,542*	28	*0,384*	16	*0,225*	8	*0,082*
Measurement		*0*		*0*	1	*0,017*	22	*0,537*	1	*0,018*	11	*0,190*	2	*0,042*	1	*0,014*	3	*0,042*		*0*
Performance	13	*0,165*	22	*0,293*	45	*0,750*	39	*0,951*	1	*0,018*	58	*1,000*	6	*0,125*	3	*0,041*	8	*0,113*	2	*0,020*
Reliability		*0*		*0*		*0*	4	*0,098*	2	*0,036*	3	*0,052*		*0*		*0*	1	*0,014*		*0*
Security	1	*0,013*		*0*		*0*		*0*		*0*		*0*		*0*		*0*		*0*	1	*0,010*
Standardization		*0*		*0*		*0*		*0*		*0*	1	*0,017*		*0*	1	*0,014*		*0*	2	*0,020*
Theory	23	*0,291*	8	*0,107*	3	*0,050*	40	*0,976*	2	*0,036*	57	*0,983*	8	*0,167*	11	*0,151*	4	*0,056*	8	*0,082*
Verification		*0*		*0*	3	*0,050*		*0*		*0*		*0*	1	*0,021*		*0*	1	*0,014*		*0*
Total articles	**79**	***0,975***	**75**	***1,440***	**60**	***2,800***	**41**	***5,927***	**55**	***2,327***	**58**	***5,603***	**48**	***2,917***	**73**	***2,603***	**71**	***2,380***	**98**	***2,133***

Table 2: Number of general terms used and distribution over all published papers per year (first 10 years)

A second point of interest is the relation between general terms. To which extend are they related? Do specific general terms co-occur more often than others? Are there general terms that are never used together? Queries in the Access database gave the numbers for all possible combinations, resulting in a matrix with the amount of papers with co-occurring general terms.

General terms	Algorithms	Design	Documentation	Economics	Experimentation	Human Factors	Languages	Legal Aspects	Management	Measurement	Performance	Reliability	Security	Standardization	Theory	Verification
Algorithms	**108**	74	4	0	17	77	20	0	9	5	22	4	0	0	24	1
Design	74	**1681**	74	5	167	1298	136	1	532	116	636	24	7	9	657	8
Documentation	4	74	**145**	0	13	101	29	0	45	16	57	3	0	2	78	0
Economics	0	5	0	**10**	0	5	0	0	4	0	0	0	1	0	0	0
Experimentation	17	167	13	0	**315**	217	19	1	45	55	88	12	3	1	75	3
Human Factors	77	1298	101	5	217	**2241**	172	3	562	142	684	24	11	10	686	14
Languages	20	136	29	0	19	172	**225**	0	45	18	65	3	0	1	58	1
Legal Aspects	0	1	0	0	1	3	0	**7**	0	1	0	0	1	0	0	0
Management	9	532	45	4	45	562	45	0	**628**	76	432	12	1	5	436	2
Measurement	5	116	16	0	55	142	18	1	76	**190**	110	12	1	0	101	3
Performance	22	636	57	0	88	684	65	0	432	110	**775**	20	1	2	550	6
Reliability	4	24	3	0	12	24	3	0	12	12	20	**33**	2	1	16	4
Security	0	7	0	1	3	11	0	1	1	1	1	2	**30**	0	3	0
Standardization	0	9	2	0	1	10	1	0	5	0	2	1	0	**16**	3	0
Theory	24	657	78	0	75	686	58	0	436	101	550	16	3	3	**1340**	5
Verification	1	8	0	0	3	14	1	0	2	3	6	4	0	0	5	**35**

Table 3: Counted papers of co-occurring general terms

General terms	Algorithms	Design	Documentation	Economics	Experimentation	Human Factors	Languages	Legal Aspects	Management	Measurement	Performance	Reliability	Security	Standardization	Theory	Verification
Algorithms		4,4%	2,8%	0,0%	5,4%	3,4%	8,9%	0,0%	1,4%	2,6%	2,8%	12,1%	0,0%	0,0%	1,8%	2,9%
Design	68,5%		51,0%	50,0%	53,0%	**57,9%**	60,4%	14,3%	84,7%	61,1%	82,1%	72,7%	23,3%	56,3%	49,0%	22,9%
Documentation	3,7%	4,4%		0,0%	4,1%	4,5%	12,9%	0,0%	7,2%	8,4%	7,4%	9,1%	0,0%	12,5%	5,8%	0,0%
Economics	0,0%	0,3%	0,0%		0,0%	0,2%	0,0%	0,0%	0,6%	0,0%	0,0%	0,0%	3,3%	0,0%	0,0%	0,0%
Experimentation	15,7%	9,9%	9,0%	0,0%		9,7%	8,4%	14,3%	7,2%	28,9%	11,4%	36,4%	10,0%	6,3%	5,6%	8,6%
Human Factors	**71,3%**	**77,2%**	**69,7%**	**50,0%**	**68,9%**		**76,4%**	**42,9%**	**89,5%**	**74,7%**	**88,3%**	**72,7%**	**36,7%**	**62,5%**	**51,2%**	**40,0%**
Languages	18,5%	8,1%	20,0%	0,0%	6,0%	7,7%		0,0%	7,2%	9,5%	8,4%	9,1%	0,0%	6,3%	4,3%	2,9%
Legal Aspects	0,0%	0,1%	0,0%	0,0%	0,3%	0,1%	0,0%		0,0%	0,5%	0,0%	0,0%	3,3%	0,0%	0,0%	0,0%
Management	8,3%	31,6%	31,0%	40,0%	14,3%	25,1%	20,0%	0,0%		40,0%	55,7%	36,4%	3,3%	31,3%	32,5%	5,7%
Measurement	4,6%	6,9%	11,0%	0,0%	17,5%	6,3%	8,0%	14,3%	12,1%		14,2%	36,4%	3,3%	0,0%	7,5%	8,6%
Performance	20,4%	37,8%	39,3%	0,0%	27,9%	30,5%	28,9%	0,0%	68,8%	57,9%		60,6%	3,3%	12,5%	41,0%	17,1%
Reliability	3,7%	1,4%	2,1%	0,0%	3,8%	1,1%	1,3%	0,0%	1,9%	6,3%	2,6%		6,7%	6,3%	1,2%	11,4%
Security	0,0%	0,4%	0,0%	10,0%	1,0%	0,5%	0,0%	14,3%	0,2%	0,5%	0,1%	6,1%		0,0%	0,2%	0,0%
Standardization	0,0%	0,5%	1,4%	0,0%	0,3%	0,4%	0,4%	0,0%	0,8%	0,0%	0,3%	3,0%	0,0%		0,2%	0,0%
Theory	22,2%	39,1%	53,8%	0,0%	23,8%	30,6%	25,8%	0,0%	69,4%	53,2%	71,0%	48,5%	10,0%	18,8%		14,3%
Verification	0,9%	0,5%	0,0%	0,0%	1,0%	0,6%	0,4%	0,0%	0,3%	1,6%	0,8%	12,1%	0,0%	0,0%	0,4%	
Average	**15,9%**	**14,8%**	**19,4%**	**10,0%**	**15,2%**	**11,9%**	**16,8%**	**6,7%**	**23,4%**	**23,0%**	**23,0%**	**27,7%**	**6,9%**	**14,2%**	**13,4%**	**9,0%**
Max	**71,3%**	**77,2%**	**69,7%**	**50,0%**	**68,9%**	**57,9%**	**76,4%**	**42,9%**	**89,5%**	**74,7%**	**88,3%**	**72,7%**	**36,7%**	**62,5%**	**51,2%**	**40,0%**

Table 4: Percentage of co-occurrence over the total count of a general term (vertical)

Interesting to see is that Reliability co-occurs most with other general terms, followed by Management, Measurement and Performance. Legal Aspects and Security have the least relation with other general terms.

Visualization

NodeXL

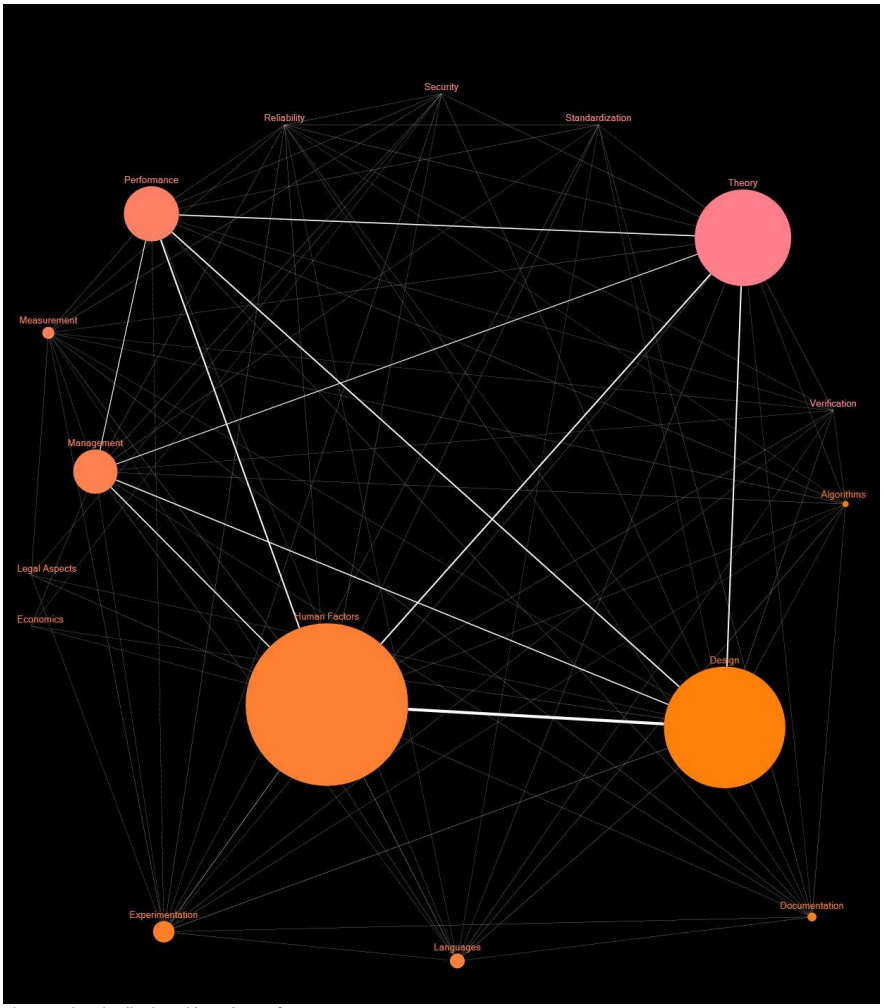

Figure 1: First visualization with NodeXL of paper count

The first visualization that was generated showed the amount of papers for every particular general term and the amount of papers counted with a co-occurrence. This picture is a tautology, because it is obvious that Human Factors is the largest general term for papers that are about computer-human interaction. And naturally it has the largest amount of co-occurrence with other general terms. More interesting to see would be how many Human Factors papers co-occur with other general terms.

Excel

To explore the data and see what kind of data would be meaningful I used quick Excel visualizations. They showed that it was important to calculate percentages and distributions over the total amount of papers. Figure 2 revealed the gaps in general term coverage for example. The excel visualizations also helped to specify the characteristics of the final visualization (the streamgraph).

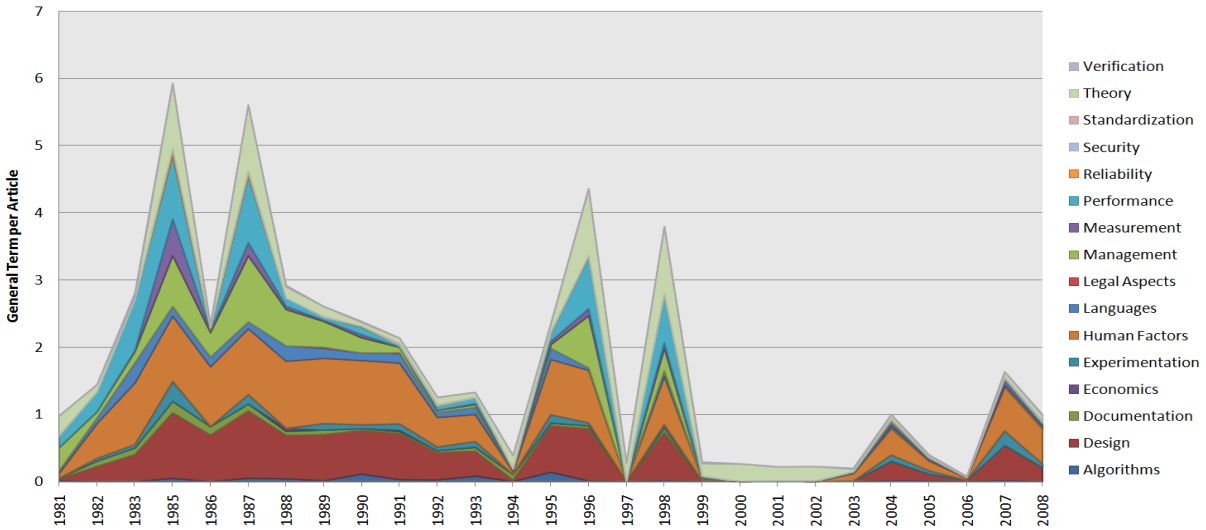

Figure 2: Average distribution of general terms per paper over time

Python

The visualization in Python also showed that it is important to not just count the amount of general terms used. It did provide more insight in the programming of visualizations.

Figuur 3: General terms used per conference, where the different colored general terms are different from the first conference in 1981

Streamgraph

The previously mentioned visualizations gave inspiration for the final data visualization. I wanted to visualize the use of general terms over time and how they are related to each other. The years with bad coverage were removed and the coloring shows the relations between the general terms as calculated in table 4.

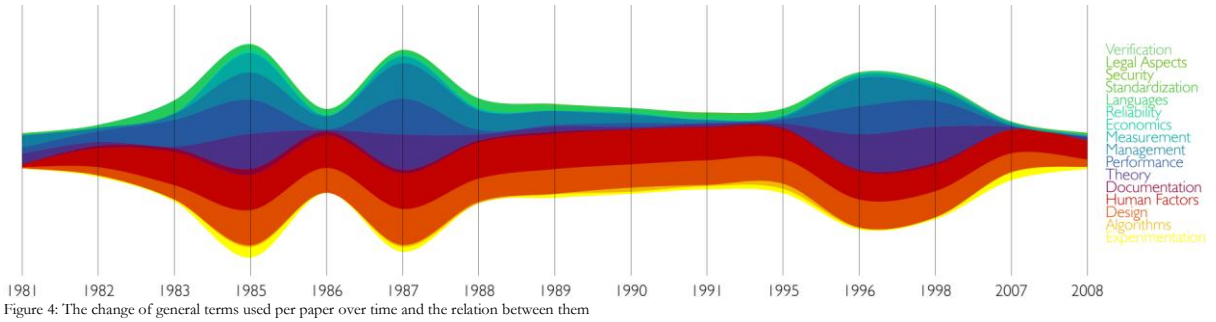

Figure 4: The change of general terms used per paper over time and the relation between them

The streamgraph shows that there are peaks in adding general terms to papers in 1985, 1987. There also appears to be an increase of general terms in the mid 90's. It also seems that the more recent papers go back to the level of 1981, although the distribution has clearly changed. In the relation between general terms two subdivisions can be identified. There is a cluster of Human Factors and Design (warm colors) and a cluster of Theory, Performance and Management (cold colors). The peaks are clearly generated by the cold colors.

Reflection

The goal for this module was to gain insight in the process of data visualization and get inspiration for my final master project. I also wanted to learn about databases and how to deal with them concerning graphic visualizations.

The main skill that I learned is to be critical of what the visualization communicates. To create a meaningful graphic it is important to consider the context of the data. In this case the database was about papers, so it is important to relate every parameter to the amount of papers. It didn't make sense to just count the general terms. It was interesting to see that the visualizations significantly changed in the process. This especially worked well with the Excel visualizations because they automatically changed with the different methods of calculation.

This made me aware of the impact I, as a designer, have on the story that a graphic tells. Graphic design is the communication tool and I become the journalist. Just as journalists should always carefully validate what they write about, I have to do the same for data visualizations. An academic attitude is a necessity to create data visualizations that model 'reality'. They simplify the real world, but the simplification process is complex and delicate.

I found a clear relation between this module and a research project that I work on as a student assistant. In the research project I learned how important visualizations are for finding interesting relations in the data. Plotting the data in several different ways worked well in finding directions for drawing conclusions and performing deeper analysis. In this module I started exploring the database in Access to understand the relationships between the different tables. I generated queries that might be of interest. I did miss visualization elements in this exploration though. It took me some time to get the right SQL code and generate the queries that I needed.

Generating the right data clusters is just the start for the data visualization process. Selecting the right clusters depends on what I want to visualize and how this will be done. The data has to be prepared for the visualization method that I want to use. Than several iteration steps are needed to generate visualizations that are meaningful and interesting. It is important to verify how complete the database is. I see a database as a collection of parameters that cover a small part of the 'real' world. The question is to which extend this coverage is complete. And a broader issue: what is real?

This insight will be very useful for my final master project. I want to design a system that generates visualizations of qualitative data, resulting in an information platform. It is important to build in verification steps and determine what the 'real' world is for the particular project. I very much like the role of journalist and clearly see the relation with my design work. The opportunity that I see for my final master project can be seen as 'automated journalism'.