The Meaning of Arrows: Diagrams and Other Facets in System Sciences Literature

Jeffrey V. Nickerson
Stevens Institute of Technology
jnickerson@stevens.edu

Abstract

Understanding the use of diagrams and other components of scholarly systems-related papers may inform us about the papers, the field, and the way we think. We analyze 495 papers containing 1899 figures in a system sciences conference proceedings. We code the diagrams into genres. We find that well-specified diagram types occur less than vaguely specified diagram types. Table types, diagram types, and equations in papers appear to correlate to non-topical aspects of the paper such as the research method used. We relate this idea to current ideas about facets and genre. The work has implications for automated document searching; the practical implication for authors is that more explicit consideration of diagram types might both clarify thinking and make later searching by other researchers easier. The implication for publishers is that collecting component information, and presenting thumbnails, may enhance search.

1. Introduction

Diagrams are an integral part of systems literature. This follows from the nature of systems disciplines; systems are connected things; we study systems in the belief that we can abstract ideas from one system and apply them to others. Diagrams are one way we represent this abstraction – when we teach, and when we write scholarly papers.

In the reading activity that is part of our research, some of us seek out diagrams, and even choose which papers to read on the basis of the diagrams contained within. When journals and proceedings are in paper form, this skimming activity is easy – but as more papers and proceedings are published electronically, and as the rate of literature growth increases, we find that we rely more on search engines to find which papers we will even glance it. And the current state of search engines does not provide an easy way of finding papers with a certain sort of diagram.

In order to motivate the later discussion, consider the following search scenarios, derived from the author's recent experience; imagine a researcher looking for:

1. *Papers containing layer diagrams which include a messaging subsystem.*
2. *Papers with models for researching issues of trust.*
3. *Empirical studies related to design education.*
4. *Images which show user interfaces for emergency response simulators.*
5. *Papers with proofs related to the tractability of bicriterion shortest-path problems.*

In all cases, a keyword search will find thousands of papers in the right content domain. But the searches require another criterion to be satisfied – either a match on the method of the paper (3), or the type of a component in the paper (1, 2, 4, 5). Implicit in our search criteria may be a kind of metonymy. For example, a paper with a model in it is likely to be statistical. A paper with a proof in it is likely to be deductive.

All of these are *non-topical characteristics* of the paper – a phrase used to describe everything about a paper except its subject [1, 2].

In generating this research, we have been influenced by two sets of literature – one related to genre, the other to diagrams. Crowston and Kwasnik propose using facets to define genres of documents [2]. They observe that users pay attention to particular clues to identify genres, including tables and figures. They note that one possible facet of a scholarly paper may be the graphics contained within. Genre studies go back to Aristotle [3]; the idea has been more recently discussed in a growing literature [4-8]. The systematic discussion of facets starts with Ranganathan, who described how such a scheme might be used to create flexible classification schemes [9].

Work in the taxonomy of diagrams has been ongoing for many years. Charles Pierce broadly categorized signs into icons, indices, and symbols; in the subcategory of indexical signs he discussed the diagram [10]. Nadin articulates the differences in configurational and sequential communication [11]. Bertin's work analyzes diagrams, including maps, charts, and what he refers to as networks [12]. Larkin and Simon discussed the relationship to computation and cognition [13]. Barbara Tversky, in collaboration with other researchers, has analyzed in depth the way we think with diagrams [14-17]; in particular, the analysis of charts and graphs by Zacks and Tversky has influenced this study [18].

We think this work has two possible applications. The first is in search and retrieval. We would like to see a system built which could answer the motivating search examples we present.

The second application is the improvement of the use of diagrams. Through becoming more conscious of diagrams, we might improve the way they are used, and possibly improve our thinking process.

Our overall research has several steps. First, we want to understand the the nature of diagram usage in information systems literature. We want to establish how this understanding can aid in search and retrieval. We want to relate what we are doing to what is known about faceted classification and genre. We have made the most progress on the first of these steps, and focus on diagram classification for most of the paper. Later, we present some preliminary ideas about search, as well as thoughts about the relation of diagrams to facets and genre.

2. Method

This paper is an exploratory case study; our case is based on the papers in a conference. We are starting with a set of general research questions related to the classification of document components. What are the components? What are the types of the components, particularly diagrams? What do the components tell us about the papers they are contained in? The tracks? The conference?

In order to find answers, we analyze all the 495 full papers in the digital Proceedings of the 37th Annual Hawaii International Conference on System Sciences [19].

We made two passes through the papers, first to make counts of the numbers of the components in the papers, then to extract the figures to separate files, which we printed. We categorized and counted the figures and captions isolated from the papers they were contained in.

How might we categorize these diagrams? Bertin took a formal approach to categorization, based on the visual properties of the diagrams. We went a different route. We think the typing of diagrams stem from a set of social conventions, so the categorization of diagrams ought to, as much as possible, follow the way we normally discuss them. So, for example, we paid attention to the captions of the diagrams. We also paid attention to what is understood in the community surrounding the conference. For example, the information science and computer science literature regularly uses a set of diagram conventions that most readers would be expected to recognize and understand, even with uncaptioned figures.

We refer to these categories of diagrams as *genres*, as they represent the purpose and form of the diagram, and are conventions understood by the community.

We discovered two types of genres. There are some diagrams that are commonly used, and are often taught as part of a curriculum in information systems, computer science, psychology, or engineering. These diagrams have a set of rigorous definitions, which can be found in the literature – or can be used more operationally, as they are often included in common computer drawing packages. For example, flow charts and UML class diagrams are well-understood and often taught in classes.

We call these tight genres. We note that not all authors strictly obey the rigorous definitions – but they could if they wanted to, and reviewers could enforce the definitions if they wanted to.

There are other types of diagrams that are very common, but have appeared in the field without any strict specifications associated with them.

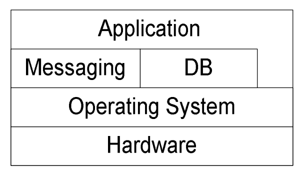

Figure 1. Layer diagram example

For example, consider figure 1, a prototypical layer diagram. It shows how an application talks to a database, which calls the operating system, which is responsible for controlling the hardware. Layer diagrams are used in practice and in academic literature, but, to the author's knowledge, do not have a canonical form – nor are they often explicitly taught in courses. We find them useful, so we emulate them, without much conscious thought. We call types of diagrams which have this ubiquitous but unspecified quality loose genres. To recognize and count them requires knowledge, as the captions and formal aspects of the diagrams will vary.

Our process was the following. We initially scanned for types of diagrams that are well understood and tightly defined, the types that appear in systems analysis textbooks – for example, we expected to see flow charts, and we did.

We also built up other categories from captions and descriptions of diagrams in the text. Then, we identified several categories of diagram which are not explicitly labeled as such, but recurred many times throughout the proceedings. These are the loose genres. The author looked at all the papers and classified all the figures. In future work we plan to use multiple coders and to collect agreement statistics along two dimensions: the categories recognized, and the mapping of diagrams into the categories.

3. Results

3.1 Components of scholarly papers

Previous work has used textual analysis, such as word frequency testing, to classify genre (e.g. [4]). But, to our knowledge, the other components of a paper haven't been as closely analyzed. Here we begin by breaking down scholarly papers into their components.

Scholarly papers in the system sciences contain text, figures, tables, equations, captions, author lists, author affiliations, headings, references, an abstract, and optionally footnotes, acknowledgements, and appendices.

We counted the following in each paper: the number of authors, the number of references, the number of tables, the number of figures, and the number of equations. Our primary interest was in the diagrams; we found, however, that the tables and equations are also important. The number of authors and the number of references we collected out of curiosity, as we have noticed that in different fields these tend to vary; it took little effort to count them, as IEEE conference papers number references.

We found that tables and figures were also easy to count, as they are also numbered. Equation counting was easy when authors separate them out and number them. But many authors embed the mathematics in the text, and it is hard to determine where one equation ends and another begins. As 80% of the papers have no equations at all, the presence of even one equation is useful information.

About 3% of the papers had errors in the labeling of figures – figure numbers were skipped, or duplicated. In several cases, tables were mislabeled as figures. We decided to fix misclassifications and miscounts, in the anticipation that new reviewing systems will over time lessen the frequency of such errors.

3.2 The major components

More than 99% of the papers have at least one equation, figure or table. This in itself is interesting; we doubt this would hold at all academic conferences. On average, each paper contains 7 of these major components. Individual papers tend to cluster toward low numbers of the different elements; all the distributions are unimodal, but skewed toward 0. For this reason, we use Kendall's Tau B throughout the paper in computing correlations.

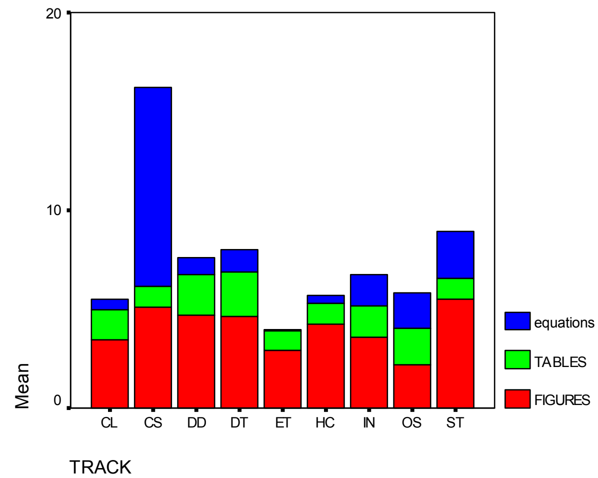

Figure 2. Equations, tables, and figures by track.

In looking at the mix by track in figure 2, we see that the tracks differ. In particular, the complex systems track, which includes minitracks related to research on the electric grid, contains the most equations.

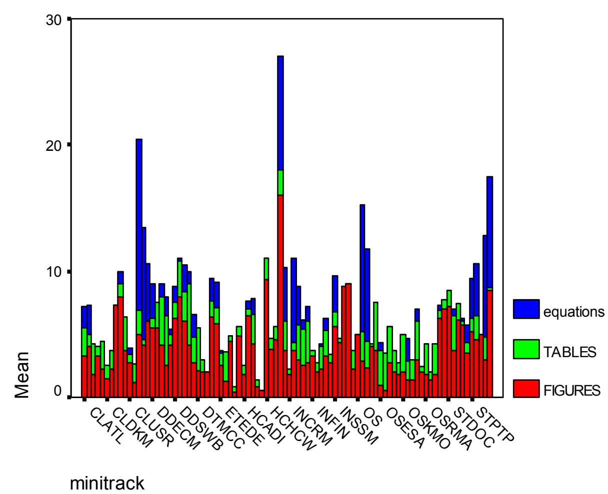

Figure 3. Equations, tables, and figures by mini-track. The order left to right is alphabetical.

We can also look by the 84 minitracks, in Figure 3. The highest peak is the Electronic Commerce minitrack, which contains a concentration of economic papers.

In the spirit of exploratory data analysis, we looked at correlations across the facets we measured, shown in Table 1.

What are we to make of the negative correlation between figures and tables? It could be that given space limitations (10 pages), a lot of one precludes a lot of another.

Table 1. Correlations (Bold numbers are significant at the .01 level.)

	Authors	Refs	Figures	Tables
Refs	-.014			
Figures	**.107**	**-.243**		
Tables	.050	.068	**-.152**	
Equations	.034	**-.202**	.087	.022

And there is also a negative correlation between references and figures. One explanation is again that there is only so much room in a paper – lots of figures means less text, and references might be proportional to text. However, if this were the case, one might expect a similar negative correlation between tables and references, which doesn't occur. One possible explanation is that the figure papers and table papers are written by different communities of researchers, with different norms for citation.

There seems to be interesting information just from a simple look at the components. And now we proceed further, by breaking down the tables and figures into subtypes.

3.3 Types of tables

Tables can contain textual information or numeric information. We counted 794 tables; 61% contained numeric information. 170 papers, or 34%, contain at least one numeric table. Why do we differentiate between numeric and textual tables? We noticed that the statistical, social science-based papers tended to use numeric tables to display the results of correlations and other analysis, and we guessed that the simple presence of a number in a table might correlate with the method of the paper.

3.4 Types of figures

Figure 4 shows a tree structure of classification for the major components we consider.

Looking now at figures, we noted many screenshots. Also, many photographs, several maps, and several realistic computer-generated images.

Photographs, screenshots and realistic computer graphics can be easily detected, as they are displayed in raster, as opposed to vector, formats. Maps may be either raster or vector, and can range from super-realistic aerial images to more abstract, chart-like representations. (Maps are arguably a form of chart – in any event, they constitute less than 1% of the figures).

Next we can distinguish charts and graphs. Distinct from charts and graphs are topological figures showing abstract relationships. We call these diagrams. The definitions of diagrams, charts, and graphs are not stan-

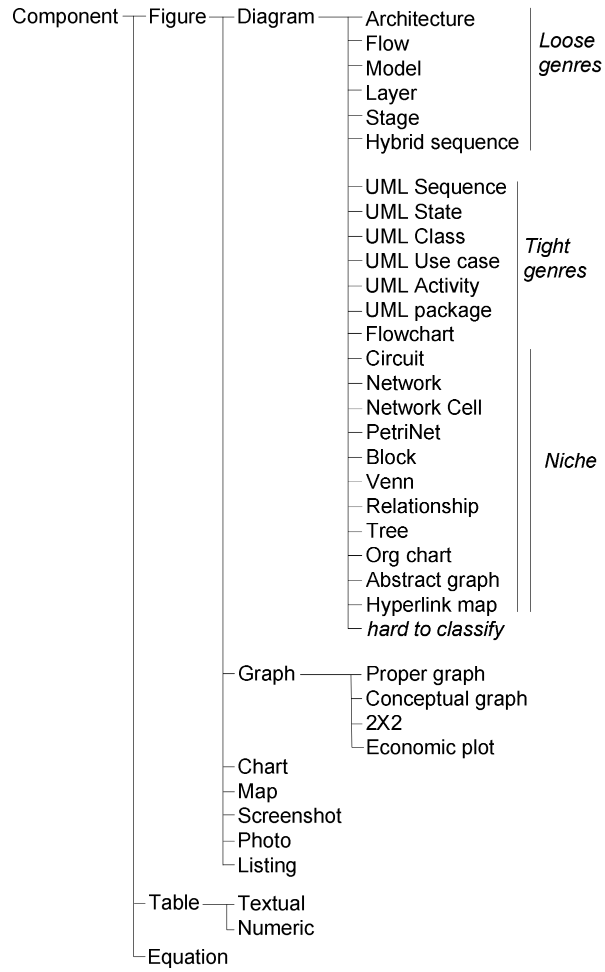

Figure 4. A tree diagram of component types

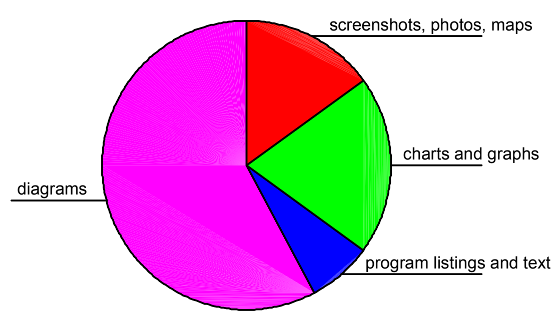

Figure 5. Types of figures

dardized; arguably charts and graphs are diagrams, and Bertin [12] thinks they are the primary diagrams – he goes so far as to put what we call diagrams into another category - networks.

Most important for this study is that we remain internally consistent. Charts and graphs are important to distinguish, as they usually establish that the paper is using some form of statistical or experimental method. Charts and graphs can be further broken down into subtypes, and other authors have done this [12, 18].

We do not go as far as these authors; instead, we make a simple classification into a handful of categories. We separate graphs into the ones that display data points, plots, the ones that are more conceptual and guess at possible relationships (we originally labeled these psuedo-plots, but found some were very useful), and the ones used in economic papers, which show the plots of idealized equations. We noticed several 2 by 2 charts, popular in business. Of the chart and graph category, most are clearly plots (53%) and charts (27%). The remaining 20% is about evenly split between conceptual and economic plots.

Figure 5 shows the breakdown of the 1899 figures we analyzed.

These different classes are easy to distinguish. As we noted, screenshots and photos are pixel-based, not vector based. Charts and graphs are distinguished by axes. Program listings include code, and have a telltale indentation structure. Other text-based displays are also easy to distinguish.

3.5 Types of diagrams

Distinguishing the genres of the 983 diagrams was more difficult. There are several categories of diagrams that are well known and easy to spot. For example, flow charts are always directed graphs, and these graphs often contain loops. The way we normally recognize them is through the telltale diamond shaped decision nodes.

We listed out the recognizable types of diagrams and looked for them. We were surprised to find that genres of diagrams that are tightly specified are far less common than loosely defined genres; of the 983 diagrams, only 37% are in tightly specified genres. We had expected our count to be dominated by diagrams such as those defined in the Unified Modeling Language (UML) standard [20]. But the majority are in the loosely specified genres.

Diagrams labeled as *models* show the relationship between dependent and independent variables. Sometimes the words *influence diagram*, *research design*, or *framework* are used in the caption instead of *model*. There were 101 of these diagrams in the papers. They are loose in that multiple conventions are used to differentiate and label the nodes and edges, although the norm is a directed graph read from left to right, from independent to dependent variables.

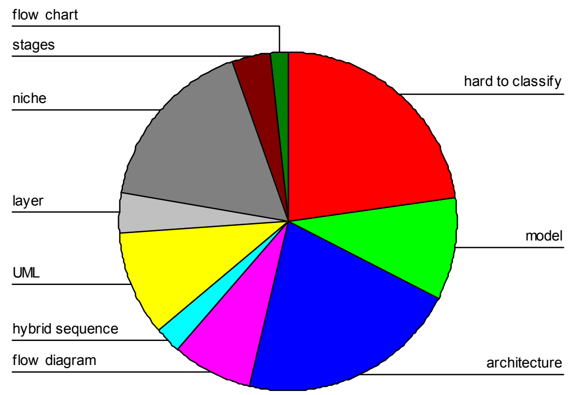

Figure 6. Types of diagrams

The most popular diagrams are architectural diagrams, showing how a system is put together. These are often labeled with *architecture* in the caption – but sometimes the words *framework* or *structure* are used. These diagrams often have undirected edges. Alternatively, the edges are bi-directional. There are 212 of these diagrams.

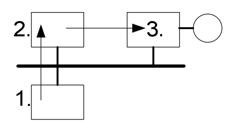

Figure 7. A hybrid sequence diagram

Related to this is a type we called a hybrid sequence diagram, in which architectural diagrams are labeled with numbers in such a way that a sequence of activities can be read off – there are 25 of these, similar to figure 7. Unlike the previous categories, there is no common way of referring to these diagrams – although they are a staple of system designers. This phenomenon might provide an important clue to the creation of new diagram genres; the hybrid sequence diagram is a blend of conventions, of two different aspects – space and time.

Another loose genre type is flow diagrams. There are 80 of these, and they usually indicate the flow of data between entities. Sometimes the architectural and flow diagrams are very close to each other. The general way to differentiate is to look at the arrows. Flow diagrams should have unidirectional arrows, and usually will not have cycles.

We mentioned layer diagrams before. We found them labeled sometimes as *layers*, sometimes as *stacks*, and, confusingly, sometimes as *architectures*.

We also found many diagrams that are designed to show the stages of either the research or a method – *step 1*, followed by *step 2*. Sometimes these are drawn as cycles, sometimes as a sequence of boxes.

We have clumped together all the UML-defined software diagram types. Of these the most popular is the class diagram, followed by the sequence diagram. Interestingly, the UML class diagram seems to have supplanted the entity-relationship diagram in the literature – there appear to be more UML diagrams and less ERD diagrams than a few years ago. They both represent static structure relationships between entities, but represent different approaches to the design of systems. The increased use of UML class diagrams may indicate a shift in systems design education; we have noticed new textbooks emphasize these diagrams.

For convenience in graphing, we include a category called *niche*. This is the aggregation of many well-understood tight-genre diagram types such as circuits, Venn diagrams, block diagrams, network diagrams, and tree diagrams, as enumerated in figure 4. These diagrams were recognizable, but didn't appear very often in this conference; we expect that in other conferences, the frequencies would change.

There is a mass of diagrams that are hard to categorize: 19% of the overall diagrams in the conference.

3.6 Mushy arrows

Why are a large number of diagrams hard to classify? Key to understanding a diagram is knowing the meaning of the conventions used. And usually the problem is understanding the meaning of an arrow. For in the systems world, an arrow is used to indicate several ideas:

1. the next stage in a project
2. the next instruction in a program
3. the movement of a stream of data
4. the transmission of a message
5. influence
6. causation
7. physical movement in space
8. the direction of an abstract relation
9. something important to look at

In well-specified conventions, such as flow charts, we implicitly know that the arrow indicates the flow of control – technically, the positioning of a program counter to the next instruction. If we don't we are guided by explicit templates in our automated drawing tools. But in more loosely specified conventions, our implicit knowledge may be lacking.

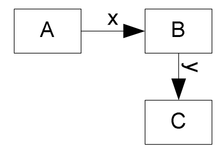

Figure 8. The ambiguity of a diagram.

For example, in one context, the diagram above could mean that A owns B, and B uses C. It could also mean that A sends B an invoice, who passes it on to C. The relative placement of the boxes may not mean anything – or may indicate that process A precedes both B and C, which run at the same time.

So, inadvertently, authors often overload the arrows – an arrow will indicate the next stage in a process on one part of a diagram, and the transmission of a message on the other side. The general idea is that something is moving, but this is a vague and not very useful idea. Most of the diagrams that are hard to classify have this feel – the arrows mean too many things at the same time, which makes it impossible to usefully classify them – and often to understand them. They have the look of meaning, but no meaning. While this might seem discouraging, remember that every paper has on average 7 elements, and multiple diagrams, so that paper-wide classifications can often still be made even if some diagrams are hard to understand.

Other hard-to-classify diagrams are imaginative blends of different types – future research may look at the way different conventions can be successfully combined. Finally, there are a small set of diagrams that appear to be coining new conventions – we found 19 of these, many in the area of mobile communication. New types of diagrams deserve attention, as they may embody new technical ideas or methods.

3.7 Combinations

What happens if we filter on combinations of different facets? For example, might not a combination of at least one model diagram and the presence of at least one numeric table indicate a statistical paper? We tried the filter – it yielded 32 papers. All except one are statistical papers. We find this result encouraging. It does not tell us how many statistical papers the filter missed – the false negatives – but it does suggest that such a filter would be useful in early searching by giving entry into the literature. Future research might confirm this finding, and look at both false positive and false negative error rates.

The exception, [21], is a case study. And it is not so far from what we were looking for – it is a case study with a model, and a mixture of qualitative and quantitative data. Perhaps we might find that case studies are likely to have models and mainly textual tables.

What about filtering on models in combination with graphs and charts? This yields 64 papers, only 8 of which are on the previous list. Most of the non-overlapping papers are different in nature from the first set; many are papers about the results of simulation, as opposed to field studies.

At first glance this doesn't seem to make sense – why shouldn't the types be the same? For the choice of a graph or a table to represent data seems arbitrary. But it may not be. For example, to display multiple correlations, a table may be more efficient than multiple scatterplots. To display the comparison between a simulated and real data, plots may be most efficient. An alternative explanation is that certain disciplines may, by convention, favor certain genres of expression.

Table 2. Correlations of screen shots, equations, charts and graphs (All are significant at the .01 level)

#per paper	Equations	Screen Shots
Screen Shots	-.251	
Charts and Graphs	.254	-.239

Looking for correlations at a more specific level than those shown in table 1, we examined the relationship between equations, screenshots, and our combined category of charts and graphs.

Table 2 shows that screenshots and equations are negatively correlated. But charts and graphs do correlate positively with equations. Charts and graphs are negatively correlated screen shots. We noticed that papers with screenshots are often using a proof-of-concept method. Equations, charts and graphs are used in analytical, evaluative papers, and in formulative papers of a more quantitative bent.

We have not yet explored all possible combinations. A few may be particularly promising. To look for a prototype paper, we might look for a combination of screenshots and code listings. Just as individual diagrams seem to sometimes blend conventions, papers might blend methods, and we might actually be able to formalize the extent of this blending, by examining the combinations, consistent with Fauconnier's ideas [22].

3.8 Limitations

This study is by its nature exploratory, and many different combinations of variables have been examined. In such situations, it is easy for experiment-wise error to occur. Over fitting to the data may have occurred, as well as unconscious anticipatory biasing in the categorizing of data. Any correlations or inferences in this study should be treated as preliminary. More formally defined follow-on studies may produce different results. The diagram categorization was done by a single individual – follow on studies may want to use multiple assessors and evaluate the level of agreement.

The study has looked at one sample set, a particular conference in a particular year. Subsequent studies should vary the year and the venue studied to assess the degree of generalizability of the results.

Since the fields involved are socially constructed, they are changing – the signals we perceive are not stationary, and this will limit our ability to predict. It may also be that observations about this or other conference, through this work or through unrelated work, might affect the behavior of future authors, either on their individual volition or through a changed review process.

4. Implications for search

4.1 The search scenarios

Looking back at the original search scenarios, can the information we gathered help in such searches?

The first search scenario is for layer diagrams. In our classification, we found that there are 40 of these diagrams in the conference. We also found that they are not universally captioned – some use the word *layer*, some the word *stack*, and others the more general terms *architecture* and *framework*. So text search may have difficulty finding them – we need some kind of diagram classification.

The second called for papers with models in them. We found lots of these, and our classification would help find them. Text search which could specify the word *model* to belong to a caption would work most, but not all, the time.

The third search looks for empirical papers. We have shown looking for numeric tables in combination with models works well.

The fourth looks for prototype images. We could search according to the number of screenshots.

Finally, a search for a mathematical proof will want to know that there are equations in the paper, and perhaps that there are also diagrams.

If we find through repeated tests that analyzing the tables and figures of documents does aid us in searching, then what might be the implications for researchers and publishers?

There are three possible strategies that come to mind that might be used to take advantage of such a finding.

4.2 Author classification

First, we might want to somehow classify writing prior to publication. For example, we might insist that authors self-annotate their work, choosing from a list of genres for their overall paper, and also from a list of genres for the tables and diagrams that they use.

While this might seem unlikely, we must remember that authors want to be cited, and that anything which increases the potential for being discovered through search engines may be seen as worth the effort.

Could the author be assisted in this process? Right now, much of electronic publishing works in a way that unintentionally strips out metadata. For example, an author may create a diagram using a tool, within which the author picks out a particular template. The template information is useful metadata. But this data is either lost when the diagram is saved, or lost later when the diagram is composed into a larger document. For example, many journals do not have the capability to handle all the formats of vector files, and ask that diagrams be converted into rasters prior to submission. But we think that vector data, and the metadata associated with the original creation of a diagram, may be useful information to embed in our electronic documents.

4.3 Automatic classification

Second, we might decide to automatically attempt to detect and classify documents. We might want to do this out of a belief that author-dependent data collection will fail, or because we want to classify the backlog of already published literature.

How might we classify diagrams automatically? First of all, with the right tools, we might extract captions and use keywords in captions to infer the type of the diagram. This will work directly in many cases. In other cases, we might need to use text in the document to disambiguate caption terms that are too general, such as the ubiquitous term *framework*.

We might also analyze the diagram itself. If it is in vector form, our task is easier – we can then analyze the graph structure and the labels for clues as to the type of diagram. Formal graph properties correlate with types of diagrams – for example, some types of diagrams, such as flows, have directed edges, and are usually acyclic.

If the diagram is in raster form, we need to perform a raster-to-vector conversion and then analyze the vectors. This will introduce error [23]. We do not need a perfect conversion – we may need only to determine relatively easy features of the diagram, such as whether or not it has arrow heads, and whether these arrow heads are one or two-directional.

At a higher level, understanding diagrams is at least as hard as understanding text, so we expect that any automated attempt to classify diagrams will have high error rates. But even just a better-than-chance classification might contribute positively to the ranking of search results, and therefore be useful.

4.4 Thumbnail-assisted search

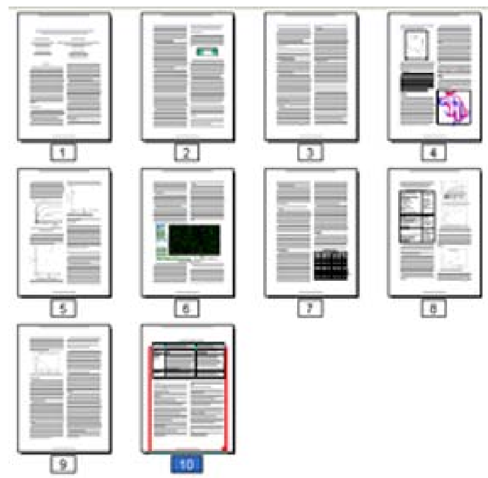

Figure 9. A thumbnail screenshot out of Adobe Acrobat for the conference paper DTMN05 [24].

In scanning the papers, Adobe Acrobat summary views proved valuable – we noticed it was possible to get a fairly good sense, and do a rough count of figures and tables, even from a very low resolution image such as that in figure 9. One can tell that there are tables, graphs, and diagrams. The thumbnail algorithm appears to be sophisticated enough to emphasize the tables and the graphs. However, one often can't tell from the thumbnail that there are equations in the paper, and one can't tell if the tables contain numbers or text.

In order to create more sophisticated thumbnails that would make equations visible, we might consider the work of Woodruff and others who have developed techniques to extract text terms and overlay them on document thumbnails [25, 26]. Using such a technique, we might extract a portion of an equation and overlay it on the thumbnails of papers which are math-intensive. Similarly, terms from tables might be extracted and overlaid to give a sample for the nature of the information contained.

Given such a technology, we might have an inexpensive alternative to the two previous classification methods. Instead of insisting on author self-categorizing, or running analyzers on all previous literature, we might simply serve up thumbnails of papers with the result of a text search. Since the thumbnails can be viewed faster than an abstract can be read, this approach may speed up our search result evaluation process.

Woodruff and others have observed that thumbnails allow us to assess genre quickly. Perhaps thumbnails work because they allow us to instinctively perceive facet characteristics, such as the ratio of diagrams and tables in a paper.

This leads one to ask if it might not be a good idea to include printed thumbnails in the books of abstracts distributed at large conferences. These thumbnails might provide a feel for the paper that would aid conference goers, allowing them the same kind of experience they might get by quickly looking through a full proceedings.

5. Ideas on facets and genre

Let us reconsider our thinking when we analyzed combinations of components in section 3.7. We were using statistics about the components to make predictions about one non-topical aspect of papers, the method. The most general statement we could make is that, given a set *S* of component/type pairs, such as <*figure1, diagram.model*>, <*table1, numeric*>, *table2, numeric*>, we can infer a method, such as *statistical*. In actuality, what we did involved a level of consolidation, as we were in effect looking at the number of a particular type of element.

We described rules based on <*componentType, quantity*>. It might make sense to consolidate further into quantity categories. For example, we might say a particular type is *PRESENT* (with 1 to 5 instances) or *HEAVY* (with > 5 instances). Then we might look for a paper where *S* includes <*diagram.model, PRESENT*>, <*table.numeric, HEAVY*>. We think these quantity categories might better reflect the way we naturally categorize papers.

These pairs of component genre and weight might become facets of the paper. They might be candidates for facets which can be used to constitute the genre of a paper, consistent with the ideas in Crowston and Kwasnik [2].

Do we have much evidence this would work from our own study? It depends how tightly one defines the genres of papers. If we keep to the purpose of a paper, then the proceedings we looked at only had two major genres; the papers, and the minitrack summaries. A reader might have noticed we didn't include the minitrack summaries in our analysis. This is because they didn't include diagrams, tables of formulas. So we at least an example which suggests that different genres may have different mixes of components.

We think the idea will apply more widely. For example, we think we will find that corporate white papers may look similar to academic papers – but have more conceptual graphs and 2 by 2s in place of proper graphs. Many white papers are intended to sell products while many academic papers are intended to demonstrate rigor.

Method may be related to sub-genres in scholarly papers. For example, in conversation we may refer to a paper we are working as being "a case study" or a "conceptual piece". We know that these papers have different formal properties, and are written differently. The ultimate purpose of the paper is the same, to contribute to a field's knowledge. But, at a finer grain level, one may be offering up an example, another an argument, with different rhetorical purposes.

Indeed, classifications of system sciences scholarly papers often focus on method as an important aspect (e. g. [27]). And one journal publisher, Emerald, has recently moved toward structured abstracts, in which method is one of the structured attributes [28].

If one accepts method as related to genre, then the work is this paper suggests a link from components to method and from there to genre. On consideration, such links should not be surprising. For, if we write a paper top down, we consider not only the topic, but the method, and we know for certain methods that we will need to create certain components, such as tables and graphs. If we write a paper bottom-up, we know that we have a result in a certain form, say a screenshot of a system, and we then construct a paper of appropriate content and method around this result.

6. Conclusions

As scholarly literature becomes digital, we have gained an ability to key word search, but lost our ability to flip through pages. Flipping can aid in search, as the components of a paper may be of interest in themselves, or may indicate the research methods used.

In an exploratory data analysis, we have looked at the proceedings of a conference. We have reached several preliminary conclusions. First, that diagrams themselves can be classified according to genre. Second, that the genres of the components of a paper can be used to identify non-topical aspects of the paper as a whole.

The implications of the work are the following. We may want to collect more pre-publication information on the components of papers in order to facilitate search. Alternatively, or in parallel, we may want to attempt to automatically recognize the genres of figures and tables. Finally, we may want to use the power of thumbnail images to recreate the feeling – and the effectiveness - of flipping through the pages of a conference proceeding.

Acknowledgements

The author wishes to thank the anonymous reviewers for their many suggestions. The author also thanks his wife Deborah Morris for assistance in the extraction and analysis of the diagrams.

References

[1] N. J. Belkin, "User modeling in information retrieval," *Sixth International Conference on User Modeling*, 1997.

[2] K. Crowston and B. H. Kwasnik, "A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality," *Proceedings of the 37th annual Hawaii International Conference on System Sciences*, 2004.

[3] Aristotle, *Poetics*. Mineola, N.Y.: Dover Publications, 1997.

[4] B. Kessler, G. Nunberg, and H. Schütze, "Automatic Detection of Genre," *35th Annual Meeting of the Association for Computational Linguistics*, San Francisco CA, 1997.

[5] K. Crowston and M. Williams, "Reproduced and emergent genres of communication on the World-Wide Web," *The Information Society*, vol. 16, pp. 201-216, 2000.

[6] W. J. Orlikowski and J. Yates, "Genre repertoire: The structuring of communicative practices in organizations," *Administrative Sciences Quarterly*, vol. 33, pp. 541-574, 1994.

[7] E. G. Toms, D. G. Campbell, and R. Blades, "Does Genre Define the Shape of Information? The Role of Form and Function in User Interaction with Digital Documents," *American Society for Information Science*, 1999.

[8] T. Yoshioka, G. Herman, J. Yates, and W. Orlikowski, " Genre taxonomy: A knowledge repository of communicative actions," *ACM Transactions on Information Systems (TOIS)*, vol. 19, no. 4, 2001.

[9] S. R. Ranganathan, *Prolegomena to library classification*. London: E. Goldston ltd., 1937.

[10] C. S. S. Peirce, *Collected papers of Charles Sanders Peirce*. Cambridge: Harvard University Press, 1931.

[11] M. Nadin, "On the meaning of the visual: 12 theses regarding the visual and its interpretation," *Semiotica*, vol. 52, no. 3/4, 1984.

[12] J. Bertin, *Semiology of graphics*. Madison, Wis.: University of Wisconsin Press, 1983.

[13] J. Larkin and H. Simon, "Why a diagram is (sometimes) worth ten thousand words," *Cognitive Science*, vol. 11, pp. 65-99, 1987.

[14] B. Tversky, "Cognitive origins of graphic conventions," in *Understanding images*, F. T. Marchese, Ed. New York: Springer-Verlag, 1995, pp. 29-53.

[15] B. Tversky, J. Zacks, P. Lee, and J. Heiser, "Lines, Blobs, Crosses and Arrows: Diagrammatic Communication with Schematic Figures," *Diagrams 2000*, 2000.

[16] M. Suwa and B. Tversky, " What do architects and students perceive in their design sketches? A protocol analysis," *Design Studies*, vol. 18, no. 4, pp. 385-403, 1997.

[17] M. Suwa and B. Tversky, "External representations contribute to the dynamic construction of ideas," in *Diagrammatic Representation and Inference: Proceedings of Diagrams 2002*, M. Hegarty, B. Meyer, and N. H. Narayanan, Eds. London: Springer, 2002, pp. 241-254.

[18] J. Zacks and B. Tversky, "Bars and Lines: A study of Graphic Communication," *Memory and Cognition*, vol. 27, pp. 1073-1079, 1999.

[19] IEEE, "Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences." Los Alamitos, Calif.: IEEE Computer Society, 2004.

[20] J. Rumbaugh, I. Jacobson, and G. Booch, *The unified modeling language reference manual*, 2nd ed. Boston: Addison-Wesley, 2004.

[21] G. K. Roberts and J. B. Pick, "Technology Factors in Corporate Adoption of Mobile Cell Phones: A Case Study Analysis," *Proceedings of the 37th Annual Hawaii International Conference on Systems Sciences*, 2004.

[22] G. Fauconnier, *Mappings in thought and language*. Cambridge, U.K. ; New York, NY: Cambridge University Press, 1997.

[23] A. Chhabra and I. Phiillips, "The Second International Graphics Recognition Contest– Raster to Vector Conversion: A Report," in *Graphics Recognition: Algorithms and Systems, Second International Workshop (LNCS 1389)*, K. Tombre and A. Chhabra, Eds.: Springer, 1998, pp. 390–410.

[24] W. W. Wakeland, E. J. Gallaher, L. M. Macovsky, and C. A. Aktipis, "A Comparison of System Dynamics and Agent-Based Simulation Applied to the Study of Cellular Receptor Dynamics," *Proceedings of the 37th Annual Hawaii International Conference on Systems Sciences*, 2004.

[25] A. Woodruff, R. Rosenholtz, J. B. Morrison, A. Faulring, and P. Pirolli, "A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 2, 2002.

[26] A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrsion, and P. Pirolli, "Using thumbnails to search the Web," *CHI*, 2001.

[27] R. L. Glass, V. Ramesh, and I. Vessey, "An analysis of research in computing disciplines," *Communications of the ACM*, vol. 47, no. 6, pp. 89-94, 2004.

[28] Emerald, http://ninetta.emeraldinsight.com /vl=8058790/cl=76/nw=1/rpsv/literaticlub/editors/editorialadmin/abstracts.htm 2004.