Artisanal Datasets: Guides for Massively Online Emergent Systems

Rosemary Michelle Simpson
Information Programming
Providence, RI
1-401-751-8853
rms@cs.brown.edu

Mark Anderson
Shoantel Limited
Southsea, Hampshire, UK
023 92816085
mac@yeardley.demon.co.uk

ABSTRACT
Hypertext is a system that represents relationships. Historically, hypertext has used many approaches, including spatial clustering, simple pointers in classic HTML webpages, probability-based Petri nets, and directed graphs. Indexes with traditional see-also entries are a form of hypertext that represents associational relationships implicitly but doesn't explicitly describe the semantic relationships of the text's domain. In addition, index see-also trails can reveal structure, and capture different points of view and levels of detail.

Many current areas of investigation, such as large and dynamically growing MOOC (Massively Open Online Course) user forum datasets, must accommodate users who have an urgent need to extract information, discover relationships, and develop understanding in the face of incomplete and inconsistent data. Like MOOC user forum datasets, unprocessed index-entry datasets are incomplete and inconsistent, and thus present an opportunity to develop strategies and insights for working with such massively online emergent systems.

The work reported in this paper uses an index-entry dataset of 8000+ entries to extract patterns and relationships between abstractions and concrete instances. Many of the index terms have a small set of see-also values to which we add metadata that converts implicit associations into explicit relations. The system is scalable because it works locally with these see-also sets, while the results reflect the global nature of the text domain.

Based on our experience with visualizing topical semantic structures, we raise questions about how to extend strategies developed in a closed application system with a moderately-sized dataset to big data. We propose that our bottom-up search and semantic visualization strategies, which discover and develop useful structure and semantics, provide a guide for users and webmasters dealing with large systems.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *frames and scripts, relation systems, semantic networks*

General Terms
Design, Experimentation, Human Factors

Keywords
emergent structure and semantics, MOOC (Massively Open Online Course), user forum datasets, semantic visualization.

1. INTRODUCTION
This paper describes work in progress on a system for extracting, discovering, and visualizing topical semantics in a specific domain dataset. The first author is a hypertext system designer and user who is currently working with the Tinderbox personal content management toolkit and the second author is a Tinderbox expert who has created an extensive online system reference for Tinderbox[1]. We collaborated on building the approaches that the first author uses for topical semantic structure visualization. In this paper, "I" refers to Rosemary, the first author.

Hypertext is a system that represents relationships. Historically, hypertext has used many approaches, including spatial clustering[10], simple pointers in classic HTML webpages, probability-based Petri nets [9], and directed graphs [18]. Book indexes with traditional see-also entries are a form of hypertext that represents associational relationships implicitly but doesn't explicitly describe the semantic relationships of the text's domain. For example, the index entry "programming, See also prototypes" associates programming with the index entries for prototypes, but doesn't explicitly explain what the relationship is. Index see-also entries can be linked together into trails and networks that reveal structure and capture different points of view and levels of detail.

In this paper we describe our experience with a general strategy for semantic visualization in dynamically changing datasets. Initially, we discuss the visualization strategy used in editing a rich index . To convey the very different results that can emerge from the general strategy, depending on the context of the domain, we then describe three different index-editing use cases. These three use cases elicit term-specific semantic structures, which we call topical semantic structures, for three specific index terms - "ensemble", "programming", and "example" - and their associated "see-also" references. Finally, we briefly summarize the results and discuss the issues and possibilities for extending this strategy to the domain of Coursera user forums.

2. EMERGENT SEMANTICS IN AN INDEX DATASET
2.1 Search and the Making of Meaning
People use indexes and search engines for many different purposes. Sometimes it is for direct access - to find a specific thing that they are interested in. Many times, however, they are

trying to articulate a question that is not well-formed. Rich indexing is a hypertext tool that helps users explore, discover, and in some instances, create meanings for things they had vaguely sensed but had been unable to explicitly articulate. Rich indexes that include annotated cross references and detailed context specifications provide a pattern discovery and exploration resource for this fuzzy search process in addition to being directed information retrieval tools. Some of the purposes that such indexes can serve include:

2.1.1 Directed Search
The user is searching for occurrences of a known term in context

2.1.2 Fuzzy Search
The user has a vague sense of what they are looking for but either doesn't know the specific technical term used by the book/domain, or has a just general sense of the topic

2.1.3 Browsing
The user is exploring a topic or set of related topics of interest

2.1.4 Alternative Points-Of-View
By its nature a book is a linear thread through a domain; the domain will have multiple alternative points of view that can link the topics together

2.1.5 Alternative Domain Structures
Some of the alternative points of view are minor alternative threads or topical groupings, while others are completely different structurings of the domain topics

2.2 Index Creation and Editing
The process of creating an index is a bottom-up process of first generating the entries from the immediate sentences and paragraphs of the text that is being indexed. An index ordering is alphabetical because people understand intuitively, as with an encyclopedia, how to find alphabetically ordered things [5].

When entry creation is finished, then the editing process must work with the necessarily inconsistent and incomplete corpus of index entries. "Necessarily", because the generation of entries is done by a human over what can be a long period of time. Interest, fatigue, mood, and simply different understandings affect both the depth and wording of index entries Part of editing is working with - and creating - the "see-also" associations to identify the major themes in the text, which may be spread out throughout the domain of the book. These "see-also" structures are very different from a table of contents, which partitions the domain of the text; index see-alsos provide implicit topical semantic structures. They are discovery tools for the author of the index who is editing the unprocessed mass of entries and for the user of the index who is engaged in searching for something felt to be in the text.

Tinderbox is a hypertext toolkit for personal content generation and management that facilitates the development of strategies for relationship exploration and visualization. Over the last nine months I've been developing a comprehensive rich index over the domain of "The Tinderbox Way, Second Edition" [4], Mark Bernstein's book on strategies for working with Tinderbox.

The database is very large for an index of a 400-page book - 8000+ entries - but considerably smaller and more structured than the dynamically changing, growing, and relatively unstructured world of the Coursera [6] user forums. During the process of taking several Coursera courses I've found myself frustrated by their primitive string search capability for working with the rigid yet chaotic forum structure. (Section 6.1 - Coursera Critique includes suggestions for improvement that I wrote in a post-course student survey for one of the courses.) As a result I've wondered if the work I'm doing in crafting an artisanal[1] hypertext from *The Tinderbox Way* index might provide strategic guidance for developing semantic discovery and visualization strategies for the much larger collaborative forums.

A large index that evolves over a period of months has similar characteristics to a user forum database - it is inconsistent, incomplete, and dynamically changing. Editing the unprocessed entries requires a different set of tools from the flat file database used to generate the entries. This paper describes some illustrations of how I've been using Tinderbox in conjunction with the index entries to craft a hypertext that represents the Tinderbox world and way of working as reflected in Mark's book. My goals are to develop a strategy for exploring and visualizing the high level semantics, structure, and dynamics of the text through developing the semantic structures embedded in the relationships captured by the see-also entries.

3. THREE USE CASES - DETAILS
The primary Tinderbox objects [3], called notes, are named collections of attribute-value pairs. Notes may be viewed as outlines, 2D node maps, timelines, charts, and trees. The particular view used depends on what the user is trying to do, what kind of structural and semantic representation is needed.

Agents are notes with scoped and faceted queries that create sets of aliases for notes found by the queries. Agents also provide methods, called agent actions, which act on attribute values of found notes.

Map views show notes on one hierarchical document level and may have adornments, which are notes that visually provide context for collections of notes. Adornments can have queries, whose scope is the outline level shown in the map view. Such "smart adornments" physically gather notes found by the query onto the space of the adornment.

The three use cases described below arise from the editing of the unprocessed *The Tinderbox Way* index dataset, specifically the see-also relationship indicators. Each Filemaker index entry record becomes a Tinderbox note whose name is the index entry. The page number for an entry is a user attribute, Page, on the note whose name is the entry. See-also entries do not have page numbers. For example, the index entry "ensembles, pairs as components of" is the name of one Tinderbox note. In this case the Page attribute has the value C7-93-4 (Chapter 7, page 93, paragraph 4).

Three use cases demonstrate the use of agents, lists, alphabetically-ordered outlines (indexes), spatial clusters, link trails, directed graphs, and appearance attributes to edit the index, by exploring and visualizing the structure and semantics of the domain. We use a semantic discovery and visualization strategy, which is a general bottom-up strategy for discovering and visualizing topical semantic structures that uses 2D spatial clustering and metadata relationship annotations.

---

[1] The term "artisanal" refers to hand-crafted objects as distinct from machine-made objects. Mark Bernstein uses the term with respect to the crafting of software.

3.1 Use Case One - simple analysis of associations - clusters

First, I did a search on the index entry term "ensemble" and its "see-also" I found see-also entries for agents, adornments, containers, clusters. Then I looked at the see-alsos for each of those terms.

I noticed that both agents and adornments *gather* notes. With agents, aliases, which are pointers, are gathered from the scope of the query and presented as a list, which may be sorted on any system or user attribute. With smart adornments, notes are gathered from the level of the adornment. Ordinary adornments can have the notes be hand placed on the adornment.

Then, exploring further about the meaning of the term ensemble, I came up with the following classification of relationships for ensemble within the context of this book: Types, Elements, Tools, Creation, Uses

So, an example of "see-also" metadata for the term "ensemble" is: "ensemble, See also composites *[type of ensemble]"*

This simple clustering visualization and exploration gave me insight into the semantics of the term ensembles in the context of Tinderbox and *The Tinderbox Way* book; it allowed me to generate both a focus and context I hadn't been aware of before.

3.2 Use Case Two - generating paths from clusters

I did a search on the index entry term "programming" and its "see-alsos". Uncovering the "programming" relationship structure involved both abstract to concrete entries such as "programming, See also agents" and concrete to abstract entries such as "prototypes, See also programming".

Since programming is a major theme of both the Tinderbox tool and this book, it threads its way into most areas of the book. The clustering process involved quite a bit of work to uncover commonalities and, unlike the disjoint clusters for the term "ensemble", in Use Case One, it was possible to create a network graph with typed links from the clusters once established.

3.3 Use Case Three - creating families of clusters from hundreds of concrete examples

I did a search on the index entry term "example" and its "see-alsos" and found only three entries. This result showed a serious disconnect between what I expected to find and what was actually there. I had expected find trails from concrete cases to abstract categories, but the index dataset didn't yet capture that because I had just been tagging specific instances of examples in the text, such as attributes and operators, as well as uses and strategies. There were 503 entries tagged as examples, such as: "cameras, See also examples" and "assignment, attribute values, [examples]; C10-154-6".

What I did next was to sort the entries and create a micro index, which I linked to from the Example map. While the trail-basis wasn't there, the raw data was, and I was able to extract meaningful trails using the different structures and tools. The result was a set of families of clusters that reflected major themes in the book as well as capturing dominant uses of Tinderbox as a content creating and organization tool. The cluster families elicited the semantic structures implicit in the text, and subsequent index entries, for the book.

4. THREE USE CASES - SUMMARY

The striking thing about the three use cases described above is the unexpected and divergent semantic structures that emerged from the general strategy of first clustering see-also terms and then developing metadata descriptions that turn implicit associations into explicit relations.

5. BACKGROUND AND RELATED WORK

For the last 30 years I have been working with various ways to represent semantic, structural, and dynamic relationships among different points of view (POV) and levels of detail (LOD) within specific domains, starting with the Gateway project at LMI (Lisp Machines Inc.) in the middle 1980s. Gateway was a documentation system that ran on the Lisp Machine and provided different views over the documentation, giving authors and readers the capability of seeing different contexts for the same material. The user interface and the underlying database were tightly integrated so that readers were also authors and could change and annotate the documents they were reading.

Later, at Brown I developed the Memex project [14] and ConceptLab [15] to provide multiple views over a document set that consisted of the first ten years of the ACM Hypertext proceedings. The Memex version was Web-based and list-oriented while the ConceptLab version was an application that ran on a 2D unbounded plane. Both applications had the same set of attributes that accessed the underlying database; Memex was read-only, while ConceptLab was read/write. The attributes included co-authors, institutions, publications, conferences, keyword concepts, and URLs. Querying one attribute value, such as author Catherine Marshall, would access all the publications, co-authors, institutions, conferences, and keyword concepts associated with her within the domain. The key difference between this earlier work on Memex and ConceptLab and the work reported here is the replacement of the closed set of attributes (institutions, people, publications, conferences) used in the Memex and ConceptLab projects by the SeeAlso attribute, which is a set of values that are open-ended and heterogeneous in type and content.

5.1 Related Work

* Top-down semantic networks. Taxonomies, such as many of the Semantic Net projects [7], are tree-structured not network-structured and are more limited in their purposes and domains. Strand Maps [13] require expert users to develop semantic networks that are general

* General search vs. domain-constrained search. Peter Norvig [12] points out that for Google search metadata no longer has value because it has been co-opted by people gaming the system. However, in constrained search over specific domains, such as an index or Coursera user forums, metadata is essential.

* Visual thinking. In 1969 Rudolph Arnheim's "Visual Thinking" [2] provided early insights about how the brain uses visual thinking to represent concepts and associations. In recent years Colin Ware [17]has both grounded visual thinking investigations in cognitive and neuro-science research and applied its results to understanding the nature of software design.

* Focus + context. In his 1986 SIGCHI paper [8] George Furnas described strategies for combining the ability to work both with details and with their context. Subsequent studies have explored different approaches to combining different levels-of-detail, and research areas such as augmented reality are using the insights to unobtrusively provide context while maintaining a particular

focus [16]. The strategy described in this paper provides users with a way of exploring both the context and details of semantic relationships in a way that reflects their personal point-of-view within the domain context.

6. ISSUES AND POSSIBILITIES

MOOC (Massively Open Online Course) user forum datasets typically involve thousands of minimally-structured threads created by tens - or in some cases hundreds - of thousands of students [11]. The students come from an exceedingly diverse range of backgrounds, mindsets, experience, needs, ages, and goals. The search facilities must accommodate urgent needs to extract information, discover relationships, and develop understanding in the face of this incomplete, inconsistent, and constantly changing data.

The work reported in this paper uses an index-entry dataset of 8000+ entries to extract patterns and relationships between abstractions and concrete instances. Many of the index terms have a small set of see-also values to which we add metadata that converts implicit associations into explicit relations. The system is scalable because it works locally with these see-also sets, while the results reflect the global nature of the text domain.

How might this be useful to the visualization of the significantly larger Coursera user forum search results? The forum thread components are not in a compact structure such as index entries are, but tags or metadata combined with textual search could be used in the same way. The general strategy of search, spread out, see patterns, and capture semantic structure is characterized by bottom-up, emergent, combination of personal interests and external domain information.

The specific tool suite doesn't matter here since this is a general strategy but we can identify two key requirements: (1) a two-dimensional plane for spreading out the results of a search, clustering related groups, and drawing the resulting topical semantic structure, and (2) a means of saving and restoring the results of the exploration and visualization. A third option that would be very useful in a diverse user forum setting is the ability to iteratively collaborate on annotating user-generated visualizations.

6.1 Coursera Critique

Suggestions for improvement submitted in a Coursera post-course student survey:

What single thing would you most want to change about this course?

Structure, specifically search and forum structures: the forum structure was both rigid and lacking in intermediate structure, e.g., the threads were a chaotic mess, while search was incredibly primitive, lacking any structure at all. The result was a hit-or-miss chaos. What is needed is a combination of fully-faceted search plus an evolving forum structure with multiple-points-of-view.

Faceted search:

Scope specification

- whole website
- all forums
- specific forum
- thread titles
- thread contents
- tag cloud
- transcriptions of the lectures
- people, by name and community TA identifier
- Booleans
    * NOT this
    * this OR that NOT BOTH
    * this OR that POSSIBLY BOTH
    * string
    * this AND this
- regular expressions

Related topics

- similarity
    * sounds like
    * looks like
- see also semantic relationships

Searches should be able to be saved and then used for search refinements. The same automatic visualization tools that should illustrate the evolving forum graph structure could be used to visualize the results of searches and sub-searches. Structure/relationship visualization is a key tool for gaining deep understanding.

Forums Structure:

1. It is currently impossible to track all threads. Need to automatically assign author-editable tags to entries, and from that develop an emergent substructure among the threads. Threads should be sortable by tag, creation and modification date, author, and title,

2. The current structure is like a rigid class hierarchy and needs cross-cutting views. The structure needs to be a graph structure to reflect the emerging multiple POVs and LODs.

3. The community TAs need a tool for effectively traversing the forums and adding intermediate structure as needed beyond the automatic evolution suggested in Point 1.

4. There needs to be a topics forum that is independent of lecture and assignment. The topics forum could have automatic links into relevant lectures and other forum threads. Obviously, the topics forum needs to evolve deep structure as the course proceeds.

5. An evolving, linked visualization of the interacting threads graph would be extremely valuable.

7. CONCLUSION

To summarize the topical semantic visualization strategy:

7.1 Process

- search the dataset and extract related topics
- iteratively explore, discover, and visualize relationships

7.2 Context

- personal mindset and background

- specific domain
- immediate need
- changing information base

7.3 Examples

- unprocessed index entries over a specific document
- Coursera user forum search results, tagged as generated with metadata by forum webmasters (proposed)

7.4 Uses

- maps for users to edit with and collaboratively annotate
- visualizations as well as textual threads in user forums
- meme evolution over time
- emergent semantic structures for specific topics

7.5 Strengths

- simple general strategy
- bottom up development
- iterative interaction
- personal + domain context

8. ACKNOWLEDGMENTS

Our thanks to Mark Bernstein (Eastgate Systems) for Tinderbox and so much else; Hypertext pioneer Andy van Dam (Brown University), and Bob Zeleznik and Emanuel Zgraggen of the Brown Computer Graphics Lab for support, inspiration, and challenges; and Coursera professors Keith Devlin and Al Filreis, Coursera designers, and Coursera forum users for serving as superb exemplars.

9. REFERENCES

[1] Anderson, Mark. 2012. *aTbRef - A Tinderbox Reference* File. http://www.acrobatfaq.com/atbref5/index.html.

[2] Arnheim, Rudolph. 2004. *Visual Thinking*, 35th Anniversary Edition. University of California Press.

[3] Bernstein, Mark. 2003. Collage, composites, construction. In *Proceedings of the fourteenth ACM conference on Hypertext and Hypermedia (HYPERTEXT '03).* pp. 122-123.

[4] Bernstein, Mark. 2012. *The Tinderbox Way, Second Edition eBook*. http://www.eastgate.com/Tinderbox/TinderboxWay.html.

[5] Blair, Ann. 2010. *Too Much to Know: Managing Scholarly Information Before the Modern Age*. New Haven: Yale University Press.

[6] Coursera. 2012. http://www.coursera.org/.

[7] Dogac, Asuman, Gokce Laleci, Yildiray Kabak, and Ibrahim Cingil. 2002. Exploiting Web Service Semantics: Taxonomies vs. Ontologies. in *IEEE Bulletin of the Technical Committee on Data Engineering*, December 2002 Vol. 25 No. 4.

[8] Furnas, George W. 1986. Generalized fisheye views. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '86),* pp. 16-23.

[9] Furuta, Richard and P. David Stotts. 1989. Programmable browsing semantics in Trellis. In *Proceedings of the second annual ACM conference on Hypertext (HYPERTEXT '89).* pp. 27-42.

[10] Marshall, Catherine C., Frank M. Shipman, III, and James H. Coombs. 1994. VIKI: spatial hypertext supporting emergent structure. In *Proceedings of the 1994 ACM European conference on Hypermedia technology (ECHT '94).* pp. 13-23.

[11] MOOC (Massively Open Online Courses). 2012. http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?pagewanted=all.

[12] Norvig, Peter. 2005. *Semantic Web Ontologies: What Works and What Doesn't.* http://www.alwayson-network.com/comments.php?id=P7480_0_3_0_C.

[13] NSDL (National Science Digital Library) Science Literacy Maps. 2007. *Strand Maps* - http://strandmaps.nsdl.org/

[14] Simpson, Rosemary Michelle. 2001. Memex and Beyond. http://www.cs.brown.edu/memex/home.html.

[15] Simpson, Rosemary Michelle. 2001. Requirements, Characteristics, and Issues for an Information Structures Spatial Hypermedia Environment. Position paper for *ACM Hypertext 2001 Spatial Hypermedia Workshop*. http://www.cs.brown.edu/~rms/SpatialHypertextPosition.pdf

[16] Veas, Eduardo E., Erick Mendez, Steven K. Feiner, and Dieter Schmalstieg. 2011. Directing attention and influencing memory with visual saliency modulation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '11*). 1471-1480.

[17] Ware, Colin. 2008. *Visual Thinking: for Design*. Morgan Kaufmann Series in Interactive Technologies.

[18] Yankelovich, Nicole, Norman Meyrowitz, and Andries van Dam. 1985. Reading and Writing the Electronic Book. In *IEEE Computer* 18, 10 (October 1985), pp. 15-30.