
Internet's impact on publishing

How to solve the Web printing problem by cooling down the Web as a medium

Giordano Beretta

**Hewlett-Packard Laboratories
Imaging Technology Department**

**1501 Page Mill Road
Palo Alto, CA 94303**

http://www.hpl.hp.com/personal/Giordano_Beretta/

Recent events

2



- 1990: Print-on-demand (POD)
- Color POD: workflow problem
- Paradigm shift facilitated by Internet:
 - *individual printing*
 - *many inexpensive printers*
 - *idle most of the time*
 - *obsolete before broken*
 - *insensitivity viz. consumable prices*
 - *possible purchase funding by advertisers*
- Challenge: simple workflow

The Bottom Line

3



- World Wide Web is the hot new publication medium
- Paper is best medium to present written information
- To own the digital printing market you have to be the *best* in printing information off the Web

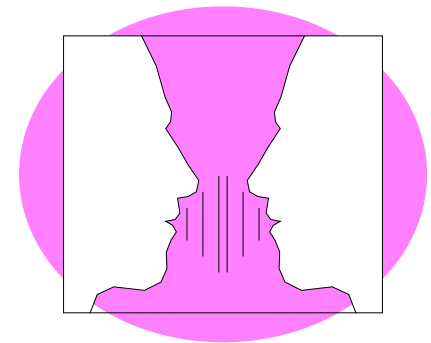


Main Points

4



- Web printing is hard because Web pages are linked but disconnected: poor structure
- Web printing is hard because in traditional printing
 - *the author decides contents, structure, and appearance*
- ... while on the web
 - *the author decides contents and structure*
 - *the reader decides the appearance*



Holy Grails



5

Desktop publishing:

- WYSIWYG — ability to see during document creation the formatted page as it will be printed

Web publishing:

- a multi-dimensional multimedia communications means
- Holy Grail of Web printing is to create the hard copy *facet* for printed output

Limitation:

- We are not talking about Internet printing

Example: Printing a Manual

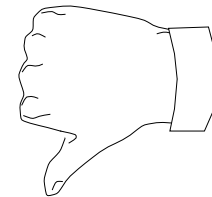
6



... we are in the middle of a session ...

Suppose a customer needs a hard copy of an oscilloscope manual

- Typically the manual will be stored as a number of HTML pages, one per section
- Typically the user has to perform the following steps:
 - 1 find first section
 - 2 click URL of section
 - 3 click on print button
 - 4 click OK in print dialog
 - 5 click back in browser
 - 6 increment section
 - 7 if endOfDocument exit else goto 2



Facets of Web Sites

7



The user is required to perform many tedious and repetitive steps to perform one action

- Ideally the user just clicks on the printer icon and the browser instructs the printer to negotiate with the server printing the manual
- Style sheets do not solve this problem
- A Web site can have many facets
 - *fast or slow link view*
 - *PDA view*
 - *hard copy*
 - *read or browse*
 - *different environment capabilities*
 - *different audiences*
 - ...

The Hotting-Up of the Web

8



- Mass medium—hot medium
- Leveling of information
- Nothing is too trivial or too important
- Everybody is interested in everything
- Accumulate information—postpone decisions indefinitely
- Aesthetic sphere of existence

1996: standardization efforts to control appearance

- Style sheets
- Fonts
- ...



1997: the Web meets Occam's Razor

9



- Hybrid magazine/news-service/television
 - *Web sites organized like TV channels*
 - *Web sites organized like magazines*
 - *Same with TV tie-ins*
 - *Ephemeral (paper as archiving medium)*
- High-concept designs
- Alliances with Hollywood
- Buy contents in Silicon Alley
- Very high entry price
- Maturity of Internet allows big players to enter the market directly
 - *No pioneer phase*

Intranets & Commercial

10



Copyrights

- Corporations migrate from proprietary to Internet protocols
- Real cost of creating and maintaining a Web site
- Willingness to invest in emerging technologies
- Print-on-demand
 - *forms*
 - *manuals, procedures*
 - *bookstores*
- Copyrights
 - *ideas & facts*
 - *expression of ideas*
 - *fair use*

what can be done with an idea



Social Changes and Knowledge

11



The post-industrial society

20 years of social polarization

- Career self-reliance, virtual corporations
- Think locally — act globally
 - *Authority* → *Power*
 - *Achievement* → *Celebrity*
 - *Doubt* → *Certainty*
 - *Science* → *Magic*
- Opportunity: Empower individuals to emancipate from levelled world to strong identities
- Provide tools to distill information into knowledge
- Knowledge: justified true belief
Mastery

New Holy Grail: Structure

12



Empower for the Web

- Quality of knowledge can be measured by how effectively it is communicated
- Effective communication requires clear organization
- Clear organization is achieved by introducing good structures

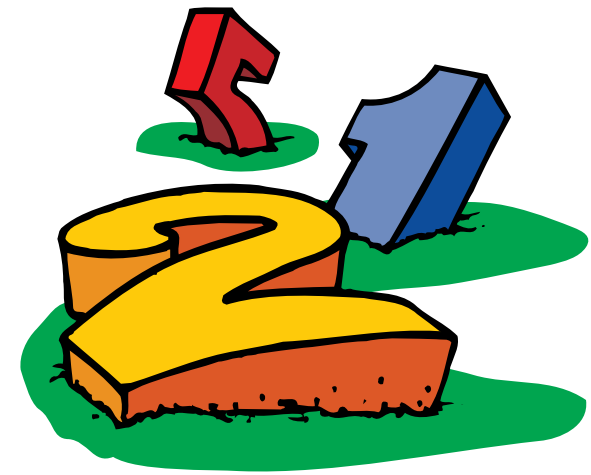


Modern Mathematics is About Structure

13



- New interpretation of mathematics after 1935
- Relational construct: a set with a relation
- System of axioms represents properties of constructs
- Mathematical creativity: find new constructs by defining maps that preserve the relations
- Two-step approach
 - *find a good system of axioms*
 - *find a good isomorphic construct*



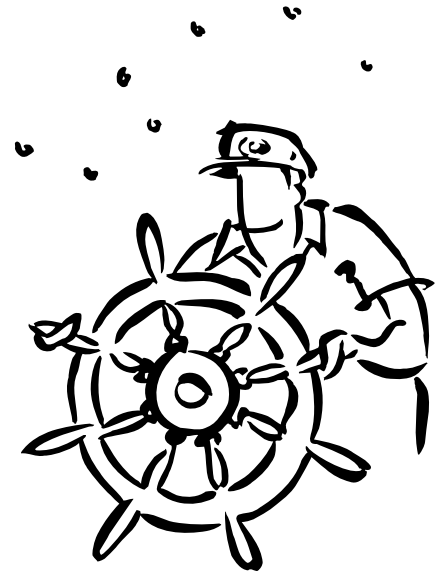
Defining a Construct Set

14



Hypertext System

- W^3 is like HyperCard with the data stored on the Internet instead in a file
- User wanders around by clicking on hot spots representing links
 - *Acquisition of knowledge*
 - *Surfing the net, channels, cruising vs. goal oriented travel*
- Many roads can lead to the destination
- Several other hypertext systems have been conceived and implemented over the years



Introducing Relations

15



- Not all links are equal
- How long does it take to get there ?
- How many words do I have to read to reach the hot spot ?
- How difficult are the words ?
- ... the sentences ?
- ... the concepts ?

Categorization



16

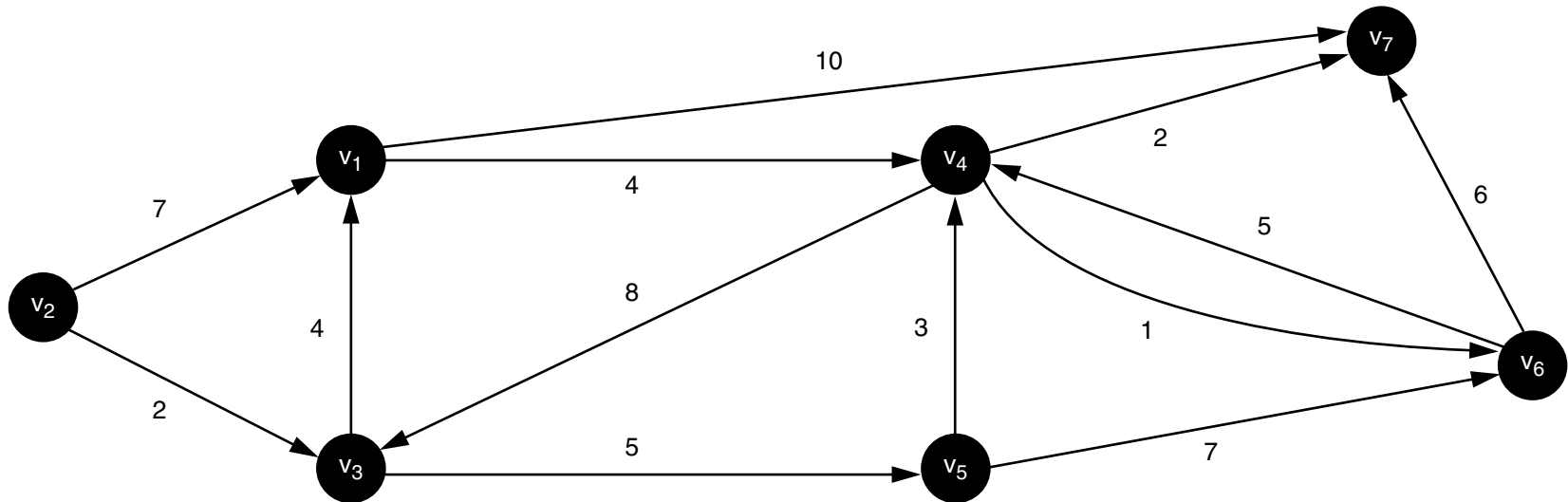
The most difficult step in distilling knowledge from information is categorization

- Key words — *image processing, JPEG, rate-control*
- Difficulty — *beginner, intermediate, advanced*
- Audience — *student, scientist, buyer, family*
- Discourse level — *main thread, note, comment, reference, detail, source*
- Presentation medium — *computer monitor, TV set, printer, communications speed*
- Many more ...

Isomorphism



Weighted digraph



Plurality of Graphs

18



There are many possible graphs for a fixed set of nodes

- **Example: One graph for each category or set of categories**
 - *graph by out medium: PC monitor, printer, TV, PDA*
 - *graph by difficulty: pupil, student, scholar*
 - *graph by audience: family, buyer, developer, user*
 - *graph by interest level: curious, information seeker, desperate*
 - *graph by intellectual challenge: tabloid, encyclopedia, treatise*
 - *graph by spin: republican, democrat, green, tory, socialist*
 - *graph by ethnicity: Afro-, Native-, Asian-, Hispanic-American*
 - *graph by subculture: jet-set, VC, transcendental*
 - ...

Lessons from CAI Systems

19



- Navigational information is valuable only if data is structured systematically
- Users get lost in generic graphs
- Cycles make it most difficult to stay on course



Trees



20

- Hierarchical
- Root
- No ambiguity
 - *exactly one path between two vertices*

Conclusion:

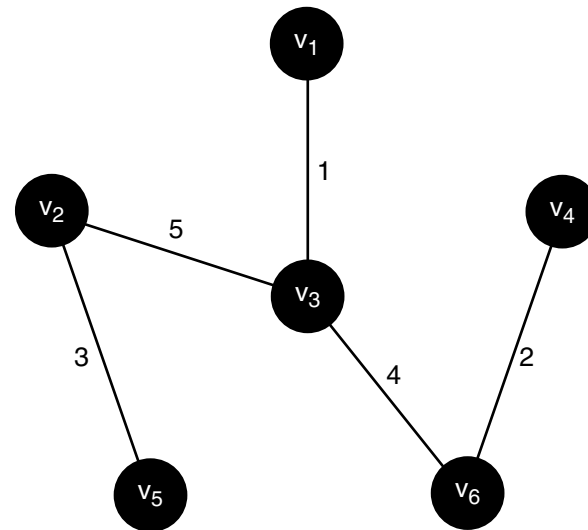
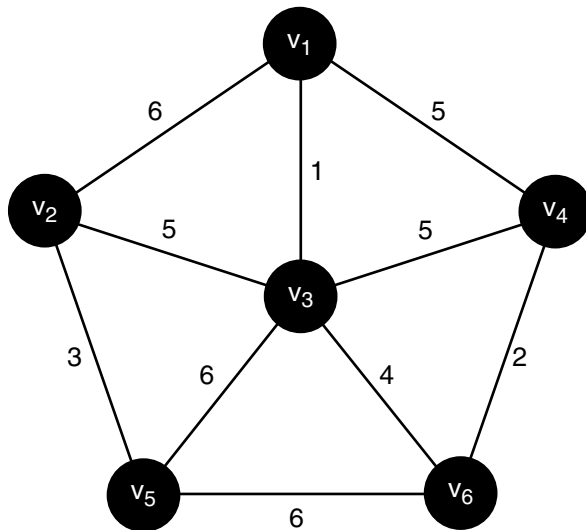
Each collection of Web pages should be organized as a tree

Note: Each graph represents a collection

Spanning Trees



A tree that is a subgraph of G and contains every vertex of G is called a **spanning tree** of G



Exploiting the Isomorphism

22



- What good properties are known about spanning trees?
- The spanning tree with minimum total edge weight is called a *minimum spanning tree* (MST)
- Good algorithms are readily available to compute an MST
 - *Prim's algorithm* — $O(n^2)$
 - *Kruskal algorithm* — $O(e \log e)$, better if $e \ll n^2$

What Have We Done ?

23



- We have designed a methodology where a user collects information and links it
- With help from a standard lexical analysis program the user can assign weights to the links
- The authoring tool finds the MST
- The MST publishes the knowledge at the lowest intellectual cost
- The method scales well

Is it Limiting ?



24

Have we limited the author's creativity ?

- No !
- There are many graphs for each Web site, predicated by the categorization
- Authors can create rich knowledge by interconnecting MSTs
- Rich \leftrightarrow compelling

Role of the Tool

25



- The authoring tool does not necessarily automatically build a Web site (but it can)
- The tool proposes good structures (e.g., one per category) that a human editor can interweave
- Important property: it scales !
- Example: the comments thrown out by Manutius can be reinstated

Requirement for Browsers

26



Need ability to disambiguate links by category

- Proposal: encode using color
- Color is related to appearance
- Category to structure



Other proposed syntactic changes:

- Use symbols to tag each paragraph by category
- In each HTML file's header section, store a representation of the current subtree
- For each node of such subtree, list all attributes

Benefits for Web Printing

28



- Tree is easily linearized into a print job, e.g., by traversing it depth-first
- Size of print job can easily be estimated when file size attributes are stored with a category related to file implementation details
- Contents rendition for hard copy can easily be assembled from categories
 - [http://frank:secret1@www.hp.com/~smith#8751\\$facet=print](http://frank:secret1@www.hp.com/~smith#8751$facet=print)
- Same for cellular palmtop devices
- Asynchronous down-loading and printing enabled by subtree availability (look-ahead)

Conclusions



- Print-on-demand
→ individual printing
- Search engines: good to find information, not to deliver knowledge
to be or not to be
- Empower users, let them build knowledge
- Partition work Human ↔ Machine
- Color fidelity as a goal: oxymoron
→ color integrity

