Hyperstructure Maintenance Costs in Large-scale Wikis

Philip Boulain

Department of Electronics and Computer Science
University of Southampton

22nd April 2008
Outline

1. Introduction
   - Context
   - Relationships
   - Suggestions

2. Experiment
   - Hypothesis
   - Dataset
   - Procedure

3. Results
   - Overheads
   - Conclusions
   - Future
Broader project

- Looking at the potentially beneficial relationships between
  - Open Hypermedia
    - Interconnected documents
  - Semantic Web
    - Interconnected databases
  - Wiki Wiki Web
    - Communal editing systems
Hypermedia

- Long-standing field of research.
- How can documents expand beyond limitations of paper?
  - Cross-referencing (hyperlinks).
  - Sharing and re-use (composition and transclusion).
- “Essential feature” is “the process of tying two items together” (linking).

V. Bush.
As We May Think.
Many types of links have been developed since.
- First-class: has identity distinct from content linked.
- N-ary: more than one source, one target.
- Typed: specifies why or how documents related.
- Generic: endpoints selected by document criteria.
- Functional: endpoints from arbitrary functions.

Open Hypermedia focuses on interoperation.
- Sharing hyperstructure with other systems.
- Applying hyperstructure to non-hypermedia resources.
- Both require non-embedded links.
World Wide Web is a distributed set of interlinked documents.

The Semantic Web is a distributed set of interlinked data.

Discover and combine data from disparate sources as one may discover and browse web pages.

Core technology is the Resource Description Framework (RDF).

RDF describes things using triples: subject has predicate value object.

Subjects, predicates, and non-literal objects named using URIs.
Wiki Wiki Web

“A collection of Web pages which can be edited by anyone, at any time, from anywhere.”

Users can create pages with much of the structure and styling of HTML web pages.

- But usually in a bespoke ad-hoc markup.

Linking is web-style, like `<a href=""></a>.

- But some wikis offer the ability to show backlinks.
- They can only do this because they are not distributed.
- The wiki system is aware of the entire document space.

Massively successful.

- Wikipedia: over 2,000,000 articles.
- Low barrier to entry.


Philip Boulain

Hyperstructure Maintenance Costs in Large-scale Wikis
Suggested benefits

- We have devised a mapping between hypermedia and semantic wikis.
  - Semantic wikis can be treated as simple hypermedia systems.

- Highlights the gaps in wiki capabilities.
  - Anything but embedded, binary links!

- Generic links offer well-defined version of WikiWords.

- First-class links make backlinks easy, even when editing.

- Transclusion allows content re-use and composition.
  - MediaWiki templates are not transcluded at edit time; they are not suitable for this.

- Adaptive hypermedia allows multiple contents per node, selected by e.g. level of detail.
Hypothesis

Manual editing of link structure, of a type which richer hypertext features could automate, is a significant overhead versus changes to the text content.
English Wikipedia

- Large, varied dataset: crosses many domains.
- Articles are heavily interlinked.
- Keeps complete history of editing process.
- Socially significant: widely-used resource with active community.
- Very, very big.
  - 84.6GB compressed; 2TB estimated uncompressed.
Stream processing essential for this much data.

- **Dump** → Trim articles → **Trimmed** → Subset → **Subset**
- Nearby revision comparison
- WikiMarkup parsing
- Abuse?
- Content?
- Templates?
- Categories?
- Links?
- **Classifications** → Aggregate → **Statistics**
Measuring text changes

- Dump contains the page sources in wikimarkup.
- Need to parse this, at least approximately.
  - Parse links to identify hyperstructure changes.
  - Extract body text to identify content changes.
  - Best-effort 42-state $LL(k)$ ‘parser’.
- Need some metric of ‘how much this text has changed’.
  - Best-effort $\Omega(n)$, $O(n \times m)$ string distance algorithm.
- Sliding window algorithm ran 0.01% subset in eight minutes, vs. $2^{1\frac{1}{2}}$ hours for Levenshtein.
  - 0.04% subset took 27 hours with sliding window, and was infeasible otherwise.
- Error could be large in some cases, but not near boundary of major/minor thresholding.
Categories

- Revert: Edit which simply undoes a previous edit.
- Content: Major (nontrivial) edit of the page content.
- Minor: Minor (trivial) edit of the page content.
- Category: Edit to the categories of a page.
- List of: Edit to a page which is an index to other pages.
- Indexing: Edit to categories or listings, possibly both.
- Template: Edit to the templates used by a page.
- Page link: Edit to an internal page link.
- URL link: Edit to a WWW URL link; usually external.
- Links: Edit to page or URL links.
- Link only: As ‘links’, but excluding major edits.
- Hyperstruct: Indexing, linking, or template use changes.
### Index management

<table>
<thead>
<tr>
<th>Edit type</th>
<th>Proportion</th>
<th>Edit type</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories</td>
<td>8.71%</td>
<td>Categories</td>
<td>8.75%</td>
</tr>
<tr>
<td>Lists</td>
<td>1.97%</td>
<td>Lists</td>
<td>3.72%</td>
</tr>
<tr>
<td>Overhead</td>
<td>10.56%</td>
<td>Overhead</td>
<td>12.34%</td>
</tr>
</tbody>
</table>

- 0.01% subset
- 0.04% subset

- Serious effort spent on organising, not just providing/updating, information.
- Semantic wiki: lists could be query results.
## Introduction

## Experiment

## Results

<table>
<thead>
<tr>
<th>Edit type</th>
<th>Proportion</th>
<th>Edit type</th>
<th>Proportion</th>
<th>Edit type</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links</td>
<td>49.60%</td>
<td>Links</td>
<td>49.60%</td>
<td>Links</td>
<td>49.56%</td>
</tr>
<tr>
<td>Links only</td>
<td>35.53%</td>
<td>Links only</td>
<td>23.36%</td>
<td>Links only</td>
<td>25.24%</td>
</tr>
<tr>
<td>Hyperstruct.</td>
<td>61.65%</td>
<td>Hyperstruct.</td>
<td>61.65%</td>
<td>Hyperstruct.</td>
<td>61.90%</td>
</tr>
<tr>
<td>Content</td>
<td>17.81%</td>
<td>Content</td>
<td>35.60%</td>
<td>Content</td>
<td>35.99%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Edit type</th>
<th>Ratio</th>
<th>Edit type</th>
<th>Ratio</th>
<th>Edit type</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links</td>
<td>2.79</td>
<td>Links</td>
<td>1.39</td>
<td>Links</td>
<td>1.38</td>
</tr>
<tr>
<td>Links only</td>
<td>2.00</td>
<td>Links only</td>
<td>0.71</td>
<td>Links only</td>
<td>0.70</td>
</tr>
<tr>
<td>Hyperstruct.</td>
<td>3.46</td>
<td>Hyperstruct.</td>
<td>1.73</td>
<td>Hyperstruct.</td>
<td>1.72</td>
</tr>
</tbody>
</table>

0.01% subset, Levenshtein 0.01% subset, Approximated 0.04% subset, Approximated

- Ratios of hyperstructure edits over content edits.
- Results consistent between 0.01% and 0.04%, so use 0.01% Lev.
### Link management II

<table>
<thead>
<tr>
<th>Edit type</th>
<th>Proportion</th>
<th>Edit type</th>
<th>Proportion</th>
<th>Edit type</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links</td>
<td>49.60%</td>
<td>Links</td>
<td>49.60%</td>
<td>Links</td>
<td>49.56%</td>
</tr>
<tr>
<td>Links only</td>
<td>35.53%</td>
<td>Links only</td>
<td>23.36%</td>
<td>Links only</td>
<td>25.24%</td>
</tr>
<tr>
<td>Hyperstruct.</td>
<td>61.65%</td>
<td>Hyperstruct.</td>
<td>61.65%</td>
<td>Hyperstruct.</td>
<td>61.90%</td>
</tr>
<tr>
<td>Content</td>
<td>17.81%</td>
<td>Content</td>
<td>35.60%</td>
<td>Content</td>
<td>35.99%</td>
</tr>
<tr>
<td>Edit type</td>
<td>Ratio</td>
<td>Edit type</td>
<td>Ratio</td>
<td>Edit type</td>
<td>Ratio</td>
</tr>
<tr>
<td>Links</td>
<td>2.79</td>
<td>Links</td>
<td>1.39</td>
<td>Links</td>
<td>1.38</td>
</tr>
<tr>
<td>Links only</td>
<td>2.00</td>
<td>Links only</td>
<td>0.71</td>
<td>Links only</td>
<td>0.70</td>
</tr>
<tr>
<td>Hyperstruct.</td>
<td>3.46</td>
<td>Hyperstruct.</td>
<td>1.73</td>
<td>Hyperstruct.</td>
<td>1.72</td>
</tr>
</tbody>
</table>

- 0.01% subset, Levenshtein
- 0.01% subset, Approximated
- 0.04% subset, Approximated

- Twice as many edits change just links as change content.
- Most of these (81%) are internal.
# Edit category distribution I

<table>
<thead>
<tr>
<th>Category</th>
<th>Registered</th>
<th>Unregistered</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of</td>
<td>1,146</td>
<td>453</td>
<td>1,599</td>
</tr>
<tr>
<td>Revert</td>
<td>4,069</td>
<td>679</td>
<td>4,748</td>
</tr>
<tr>
<td>Category</td>
<td>6,121</td>
<td>954</td>
<td>7,075</td>
</tr>
<tr>
<td>URL link</td>
<td>5,548</td>
<td>2,977</td>
<td>8,525</td>
</tr>
<tr>
<td>Indexing</td>
<td>7,174</td>
<td>1,397</td>
<td>8,571</td>
</tr>
<tr>
<td>Template</td>
<td>7,992</td>
<td>1,330</td>
<td>9,322</td>
</tr>
<tr>
<td>Content</td>
<td>10,275</td>
<td>4,182</td>
<td>14,457</td>
</tr>
<tr>
<td>Minor</td>
<td>13,776</td>
<td>9,961</td>
<td>23,737</td>
</tr>
<tr>
<td>Link only</td>
<td>20,969</td>
<td>7,877</td>
<td>28,846</td>
</tr>
<tr>
<td>Page link</td>
<td>27,205</td>
<td>8,871</td>
<td>36,076</td>
</tr>
<tr>
<td>Links</td>
<td>29,671</td>
<td>10,606</td>
<td>40,277</td>
</tr>
<tr>
<td><strong>Hyperstruct.</strong></td>
<td><strong>38,358</strong></td>
<td><strong>11,701</strong></td>
<td><strong>50,059</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>57,463</strong></td>
<td><strong>23,733</strong></td>
<td><strong>81,196</strong></td>
</tr>
</tbody>
</table>
## Edit category distribution II

<table>
<thead>
<tr>
<th>Category</th>
<th>Registered</th>
<th>Unregistered</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of</td>
<td>1,146</td>
<td>453</td>
<td>1,599</td>
</tr>
<tr>
<td>Revert</td>
<td>4,069</td>
<td>679</td>
<td>4,748</td>
</tr>
<tr>
<td>Links</td>
<td>29,671</td>
<td>10,606</td>
<td>40,277</td>
</tr>
<tr>
<td>Hyperstruct.</td>
<td>38,358</td>
<td>11,701</td>
<td>50,059</td>
</tr>
<tr>
<td>Total</td>
<td>57,463</td>
<td>23,733</td>
<td>81,196</td>
</tr>
</tbody>
</table>

- Over 5% reverts: keeping Wikipedia stationary.
- Over half of edits adjusted the hyperstructure.
User edit counts

User distribution over total number of edits made; 0.04% subset

Philip Boulain  
Hyperstructure Maintenance Costs in Large-scale Wikis
User edit categories

User distribution over total number of edits made, by category; 0.04% subset
Edit magnitudes

Edit distribution over magnitude of edit; 0.01% subset

Number of edits

Edit magnitude, ±50

10^0 10^1 10^2

10^3 10^4 10^5

1000 2000 3000 4000 5000 6000 7000

All users
Registered
Unregistered
All, approx.
Limitations

- Cannot reasonably detect some overheads automatically:
  - Template substitution indistinguishable from adding markup.
  - Same fix applied to multiple pages.
  - Reverts further back than the immediately preceding version.
Hypothesis

Manual editing of link structure, of a type which richer hypertext features could automate, is a significant overhead versus changes to the text content.

- Took random sample from English Wikipedia dataset.
- Processed edits into non-mutually-exclusive classes.
  - Edits to the information content.
  - Edits to the navigational structure.
- Compared editing effort expended on each class of edit.
### Main observations

- Twice as many edits changed links alone than content.
- 10% of edits maintained manual indexes of pages.

- Richer hypermedia features can automate some of this burden.
  - Generic links can automate linking to articles.
- Semantic wiki features can also help.
  - Lists can be generated by query on metadata.
- Hyperstructure matters!
  - It is a major part of editor activity.
  - It is worth working on improving it.
This has been a macro-scale experiment.

We are now finishing an experiment looking at the micro-scale.

- By what processes do editors make changes?
- Why do editors make changes?
- How can we prioritise improvements to support this?

Initial results have been promising.
Open Semantic Hyperwiki Model

- Defines a wiki with open hypermedia features.
  - First-class links, with edit-time embedding.
  - Transclusion, including while editing.
  - Generic and functional links.
  - Parametric nodes and links.
  - Full versioning, including links.

- Defines how this operates as a semantic wiki.
  - From hyperlinks to RDF relations.

- Implementing in Weerkat: a highly flexible and modular wiki system.
  - Orthogonal with regards to storage, accounts, markup, policy, rendering...
(Semantic) wikis are simple hypermedia systems.

Analysis of edit type across large sample of Wikipedia: hyperstructure or content?

Link editing alone is two times more common than content editing.

We can improve this situation by applying the lessons of decades of hypermedia research to these new hypermedia systems.
Thanks for listening

Questions?