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.
The Atlantic Monthly, 176:101–108, Jul 1945.

Open Hypermedia

- 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.

Semantic Web

- World Wide Web is a distributed set of interlinked documents.
- The Semantic Web is 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.”[1]
- 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.

[1] http://c2.com/cgi/wiki?WikiGettingStartedFaq

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

Hypothesis

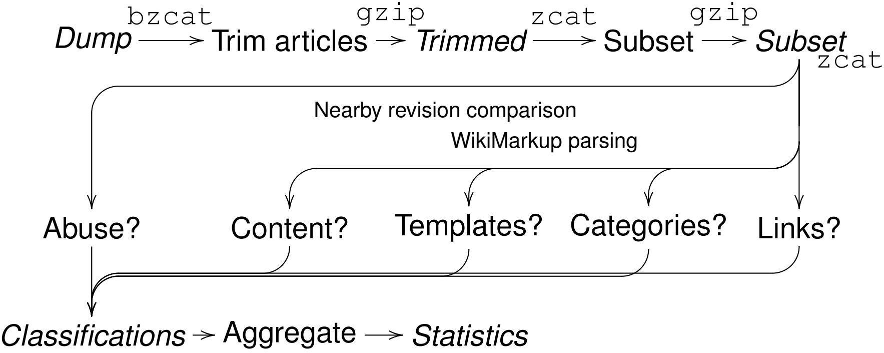
Manual editing of link structure, of a type which richer hypertext features could automate, is a significant overhead versus changes to the text content.

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.

Overview

- Stream processing essential for this much data.



Dump bzcat → Trim articles gzip → *Trimmed* zcat → Subset gzip → *Subset* zcat

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. $2\frac{1}{2}$ 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

Edit type	Proportion
Categories	8.71%
Lists	1.97%
Overhead	10.56%

0.01% subset

Edit type	Proportion
Categories	8.75%
Lists	3.72%
Overhead	12.34%

0.04% subset

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

Link management I

Edit type	Proportion
Links	49.60%
Links only	35.53%
Hyperstruct.	61.65%
Content	17.81%
Edit type	Ratio
Links	2.79
Links only	2.00
Hyperstruct.	3.46

0.01% subset, Levenshtein

Edit type	Proportion
Links	49.60%
Links only	23.36%
Hyperstruct.	61.65%
Content	35.60%
Edit type	Ratio
Links	1.39
Links only	0.71
Hyperstruct.	1.73

0.01% subset, Approximated

Edit type	Proportion
Links	49.56%
Links only	25.24%
Hyperstruct.	61.90%
Content	35.99%
Edit type	Ratio
Links	1.38
Links only	0.70
Hyperstruct.	1.72

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

Edit type	Proportion
Links	49.60%
Links only	35.53%
Hyperstruct.	61.65%
Content	17.81%
Edit type	Ratio
Links	2.79
Links only	2.00
Hyperstruct.	3.46

0.01% subset, Levenshtein

Edit type	Proportion
Links	49.60%
Links only	23.36%
Hyperstruct.	61.65%
Content	35.60%
Edit type	Ratio
Links	1.39
Links only	0.71
Hyperstruct.	1.73

0.01% subset, Approximated

Edit type	Proportion
Links	49.56%
Links only	25.24%
Hyperstruct.	61.90%
Content	35.99%
Edit type	Ratio
Links	1.38
Links only	0.70
Hyperstruct.	1.72

0.04% subset, Approximated

- Twice as many edits change just links as change content.
- Most of these (81%) are internal.

Edit category distribution I

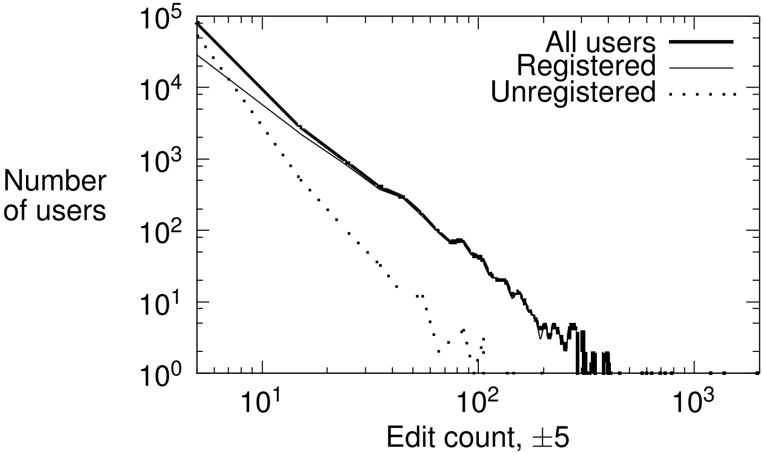
Category	Registered	Unregistered	Total
List of	1,146	453	1,599
Revert	4,069	679	4,748
Category	6,121	954	7,075
URL link	5,548	2,977	8,525
Indexing	7,174	1,397	8,571
Template	7,992	1,330	9,322
Content	10,275	4,182	14,457
Minor	13,776	9,961	23,737
Link only	20,969	7,877	28,846
Page link	27,205	8,871	36,076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

Edit category distribution II

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

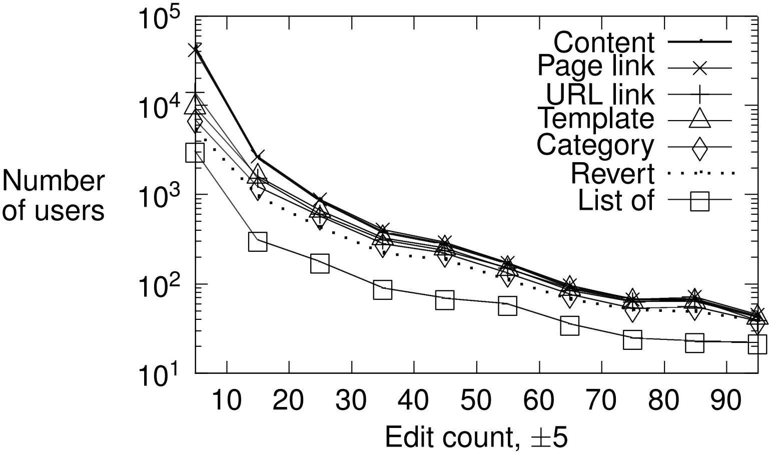
- Over 5% reverts: keeping Wikipedia stationary.
- Over half of edits adjusted the hyperstructure.

User edit counts



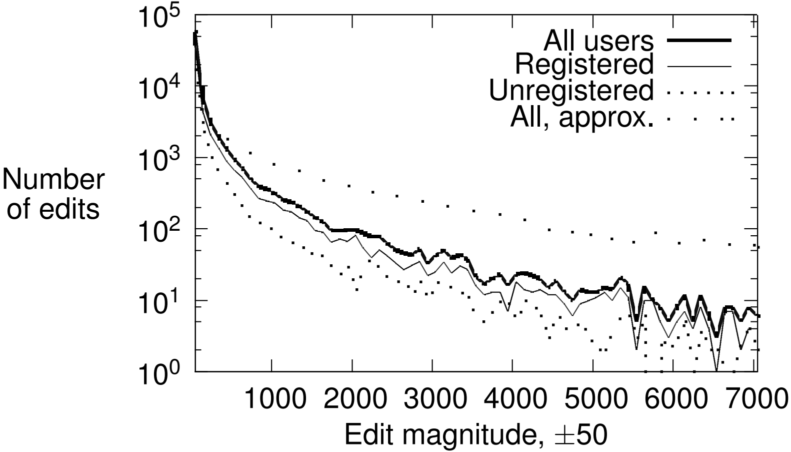
User distribution over total number of edits made; 0.04% subset

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

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.

Experiment

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.
    versus
  - Edits to the navigational structure.
- Compared editing effort expended on each class of edit.

Results

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.

Micro-scale experiment

- 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...

Summary

- (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?