rv you're dumb

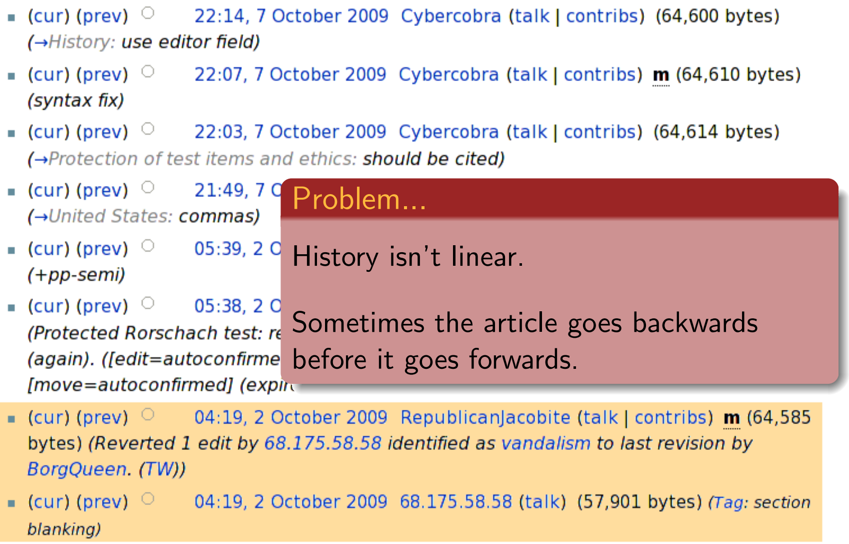
Identifying Discarded Work in Wiki Article History

Michael Ekstrand John Riedl
{ekstrand,riedl}@cs.umn.edu

GroupLens Research
Department of Computer Science
University of Minnesota

WikiSym 2009

Wiki article history



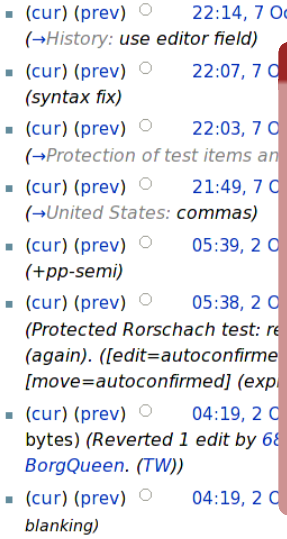
- (cur) (prev) ○ 22:14, 7 October 2009 Cybercobra (talk | contribs) (64,600 bytes) (→History: use editor field)
- (cur) (prev) ○ 22:07, 7 October 2009 Cybercobra (talk | contribs) m (64,610 bytes) (syntax fix)
- (cur) (prev) ○ 22:03, 7 October 2009 Cybercobra (talk | contribs) (64,614 bytes) (→Protection of test items and ethics: should be cited)
- (cur) (prev) ○ 21:49, 7 O (→United States: commas)
- (cur) (prev) ○ 05:39, 2 O (+pp-semi)
- (cur) (prev) ○ 05:38, 2 O (Protected Rorschach test: re (again). ([edit=autoconfirme [move=autoconfirmed] (expir
- (cur) (prev) ○ 04:19, 2 October 2009 RepublicanJacobite (talk | contribs) m (64,585 bytes) (Reverted 1 edit by 68.175.58.58 identified as vandalism to last revision by BorgQueen. (TW))
- (cur) (prev) ○ 04:19, 2 October 2009 68.175.58.58 (talk) (57,901 bytes) (Tag: section blanking)

Problem...

History isn't linear.

Sometimes the article goes backwards before it goes forwards.

Wiki article history (cont.)



- (cur) (prev) ○ 22:14, 7 October 2009 Cybercobra (talk | contribs) (64,600 bytes) (→History: use editor field)
- (cur) (prev) ○ 22:07, 7 O (syntax fix)
- (cur) (prev) ○ 22:03, 7 O (→Protection of test items an
- (cur) (prev) ○ 21:49, 7 O (→United States: commas)
- (cur) (prev) ○ 05:39, 2 O (+pp-semi)
- (cur) (prev) ○ 05:38, 2 O (Protected Rorschach test: re (again). ([edit=autoconfirme [move=autoconfirmed] (exp
- (cur) (prev) ○ 04:19, 2 O bytes) (Reverted 1 edit by 68 BorgQueen. (TW))
- (cur) (prev) ○ 04:19, 2 O blanking)

Questions

1. What is the lineage of the article's current state?
2. What revisions got thrown away? Did mine get thrown away? How much work in general is being discarded?
3. At a particular time, what revision is most representative of the article's "true" state?
4. If there is an edit/revert war, who is involved? Who (if anyone) violated the 3RR?

What's coming

1. Building history trees
2. Visualizing history trees
3. Computing on history trees
4. Improving tree building
5. Conclusion

Where we're at

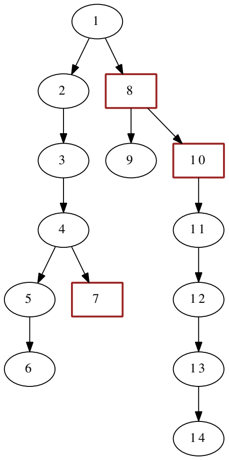
1. Building history trees
2. Visualizing history trees
3. Computing on history trees
4. Improving tree building
5. Conclusion

Building a history tree – example

1. “Hello”
2. “Hello, world”
3. “Good afternoon, world”
4. “Good afternoon, Orlando”
5. “Hello, world”
6. “Hello, world$^{\text{\{citation needed\}}}$”

The file `blfull.eps` hasn’
We attempted to create it
‘`dot -Tps2 -o blfull.e`
but that seems not to have
the ‘`-shell-escape` option

History tree algorithm



We build a tree rooted at the first revision in the article’s history.

Add each revision in chronological order:

- If a revision is different from all prior revisions, it is an edit node whose parent is the previous revision.
- If a revision is identical to some prior revision, it is a revert node whose parent is the most recent identical revision.

Previous work on history structuring

[Sabel, 2007] used text retention to find parents.

How our work differs:

- More conservative
- Basic algorithm has unambiguous mappings to editor actions
- Designed to support specific tasks in a well-defined manner

Where we're at

1. Building history trees
2. Visualizing history trees
3. Computing on history trees
4. Improving tree building
5. Conclusion

End-user use cases

How might a wiki user want to use history trees?

Analyzing content disputes
Who was involved in this edit/revert war? How did it play out?

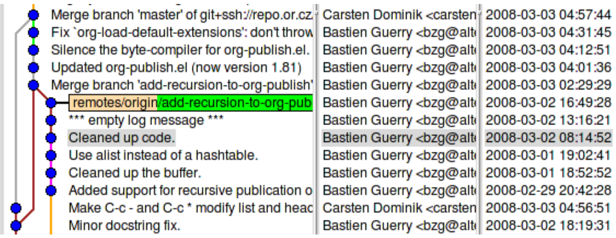
Satisfying curiosity
How did the article get to its current state? What work was used in that development?

Visualization objectives

To make these trees accessible to users, we developed a graphical tree view that meets four design criteria.

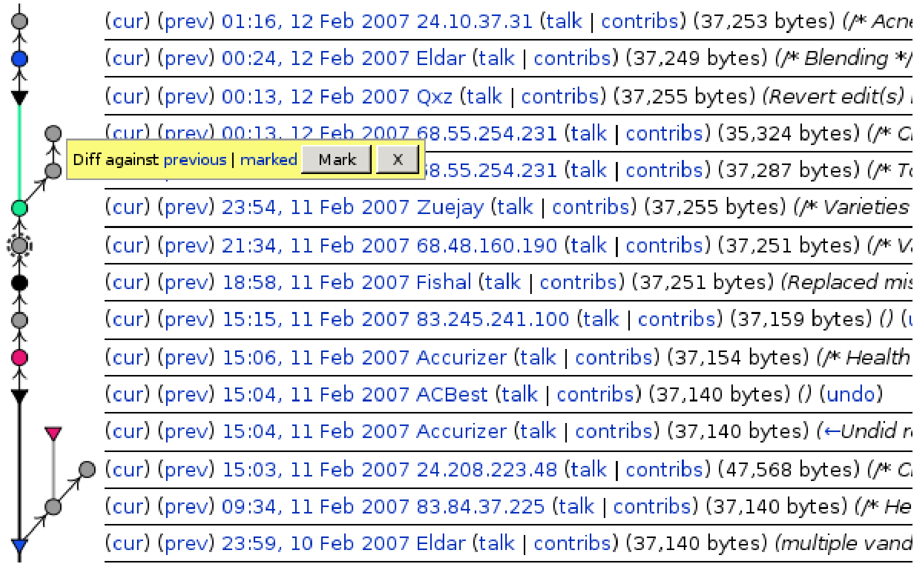
1. Indicate accepted and rejected revisions
2. Clearly mark reverts
3. Indicated shared authorship
4. Integrate with existing history views

`gitk` provided our inspiration for making this happen.



```
Merge branch 'master' of git+ssh://repo.or.cz  Carsten Dominik <carsten  2008-03-03 04:57:44
Fix `org-load-default-extensions': don't throw  Bastien Guerry <bzg@alt  2008-03-03 04:31:45
Silence the byte-compiler for org-publish.el.  Bastien Guerry <bzg@alt  2008-03-03 04:12:51
Updated org-publish.el (now version 1.81)      Bastien Guerry <bzg@alt  2008-03-03 04:01:36
Merge branch 'add-recursion-to-org-publish'    Bastien Guerry <bzg@alt  2008-03-03 02:29:29
  remotes/origin/add-recursion-to-org-pub      Bastien Guerry <bzg@alt  2008-03-02 16:49:28
  *** empty log message ***                    Bastien Guerry <bzg@alt  2008-03-02 13:16:21
  Cleaned up code.                             Bastien Guerry <bzg@alt  2008-03-02 08:14:52
  Use alist instead of a hashtable.            Bastien Guerry <bzg@alt  2008-03-01 19:02:41
  Cleaned up the buffer.                       Bastien Guerry <bzg@alt  2008-03-01 18:52:52
  Added support for recursive publication o    Bastien Guerry <bzg@alt  2008-02-29 20:42:28
  Make C-c - and C-c * modify list and heac    Carsten Dominik <carsten 2008-03-03 04:56:51
  Minor docstring fix.                         Bastien Guerry <bzg@alt  2008-03-02 18:19:31
```

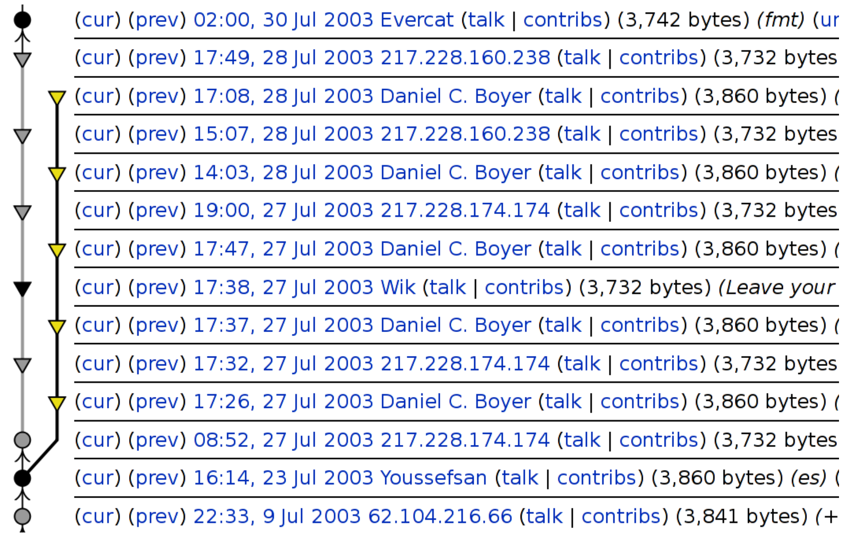
Visualization design



```
(cur) (prev) 01:16, 12 Feb 2007 24.10.37.31 (talk | contribs) (37,253 bytes) (/* Acn
(cur) (prev) 00:24, 12 Feb 2007 Eldar (talk | contribs) (37,249 bytes) (/* Blending */
(cur) (prev) 00:13, 12 Feb 2007 Qxz (talk | contribs) (37,255 bytes) (Revert edit(s)
(cur) (prev) 00:13, 12 Feb 2007 68.55.254.231 (talk | contribs) (35,324 bytes) (/* C
8.55.254.231 (talk | contribs) (37,287 bytes) (/* T
(cur) (prev) 23:54, 11 Feb 2007 Zuejay (talk | contribs) (37,255 bytes) (/* Varieties
(cur) (prev) 21:34, 11 Feb 2007 68.48.160.190 (talk | contribs) (37,251 bytes) (/* V
(cur) (prev) 18:58, 11 Feb 2007 Fishal (talk | contribs) (37,251 bytes) (Replaced mi
(cur) (prev) 15:15, 11 Feb 2007 83.245.241.100 (talk | contribs) (37,159 bytes) () (
(cur) (prev) 15:06, 11 Feb 2007 Accurizer (talk | contribs) (37,154 bytes) (/* Health
(cur) (prev) 15:04, 11 Feb 2007 ACBest (talk | contribs) (37,140 bytes) () (undo)
(cur) (prev) 15:04, 11 Feb 2007 Accurizer (talk | contribs) (37,140 bytes) (←Undid r
(cur) (prev) 15:03, 11 Feb 2007 24.208.223.48 (talk | contribs) (47,568 bytes) (/* C
(cur) (prev) 09:34, 11 Feb 2007 83.84.37.225 (talk | contribs) (37,140 bytes) (/* He
(cur) (prev) 23:59, 10 Feb 2007 Eldar (talk | contribs) (37,140 bytes) (multiple vand
```

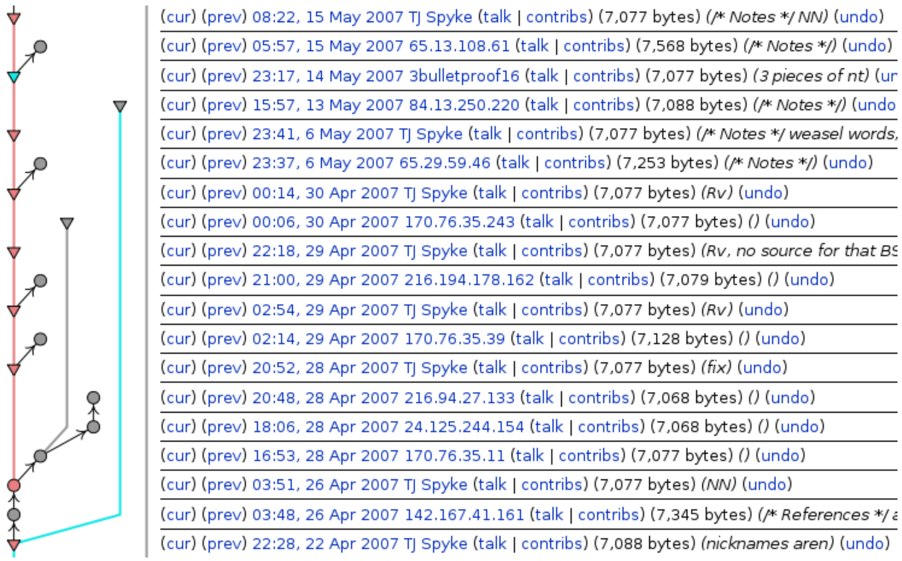
Case study: Chocolate [Viégas et al., 2004]

The revert war examined in “Studying cooperation and conflict”:



(cur) (prev) 02:00, 30 Jul 2003 Evercat (talk | contribs) (3,742 bytes) *(fmt)* (ur
(cur) (prev) 17:49, 28 Jul 2003 217.228.160.238 (talk | contribs) (3,732 bytes
(cur) (prev) 17:08, 28 Jul 2003 Daniel C. Boyer (talk | contribs) (3,860 bytes) (
(cur) (prev) 15:07, 28 Jul 2003 217.228.160.238 (talk | contribs) (3,732 bytes
(cur) (prev) 14:03, 28 Jul 2003 Daniel C. Boyer (talk | contribs) (3,860 bytes) (
(cur) (prev) 19:00, 27 Jul 2003 217.228.174.174 (talk | contribs) (3,732 bytes
(cur) (prev) 17:47, 27 Jul 2003 Daniel C. Boyer (talk | contribs) (3,860 bytes) (
(cur) (prev) 17:38, 27 Jul 2003 Wik (talk | contribs) (3,732 bytes) *(Leave your*
(cur) (prev) 17:37, 27 Jul 2003 Daniel C. Boyer (talk | contribs) (3,860 bytes) (
(cur) (prev) 17:32, 27 Jul 2003 217.228.174.174 (talk | contribs) (3,732 bytes
(cur) (prev) 17:26, 27 Jul 2003 Daniel C. Boyer (talk | contribs) (3,860 bytes) (
(cur) (prev) 08:52, 27 Jul 2003 217.228.174.174 (talk | contribs) (3,732 bytes
(cur) (prev) 16:14, 23 Jul 2003 Youssefsan (talk | contribs) (3,860 bytes) *(es)* (
(cur) (prev) 22:33, 9 Jul 2003 62.104.216.66 (talk | contribs) (3,841 bytes) (+

Case study: WrestleMania III



(cur) (prev) 08:22, 15 May 2007 TJ Spyke (talk | contribs) (7,077 bytes) *(/* Notes */ NN)* (undo)
(cur) (prev) 05:57, 15 May 2007 65.13.108.61 (talk | contribs) (7,568 bytes) *(/* Notes */)* (undo)
(cur) (prev) 23:17, 14 May 2007 3bulletproof16 (talk | contribs) (7,077 bytes) *(3 pieces of nt)* (ur
(cur) (prev) 15:57, 13 May 2007 84.13.250.220 (talk | contribs) (7,088 bytes) *(/* Notes */)* (undo
(cur) (prev) 23:41, 6 May 2007 TJ Spyke (talk | contribs) (7,077 bytes) *(/* Notes */ weasel words,*
(cur) (prev) 23:37, 6 May 2007 65.29.59.46 (talk | contribs) (7,253 bytes) *(/* Notes */)* (undo)
(cur) (prev) 00:14, 30 Apr 2007 TJ Spyke (talk | contribs) (7,077 bytes) *(Rv)* (undo)
(cur) (prev) 00:06, 30 Apr 2007 170.76.35.243 (talk | contribs) (7,077 bytes) *()* (undo)
(cur) (prev) 22:18, 29 Apr 2007 TJ Spyke (talk | contribs) (7,077 bytes) *(Rv, no source for that BS*
(cur) (prev) 21:00, 29 Apr 2007 216.194.178.162 (talk | contribs) (7,079 bytes) *()* (undo)
(cur) (prev) 02:54, 29 Apr 2007 TJ Spyke (talk | contribs) (7,077 bytes) *(Rv)* (undo)
(cur) (prev) 02:14, 29 Apr 2007 170.76.35.39 (talk | contribs) (7,128 bytes) *()* (undo)
(cur) (prev) 20:52, 28 Apr 2007 TJ Spyke (talk | contribs) (7,077 bytes) *(fix)* (undo)
(cur) (prev) 20:48, 28 Apr 2007 216.94.27.133 (talk | contribs) (7,068 bytes) *()* (undo)
(cur) (prev) 18:06, 28 Apr 2007 24.125.244.154 (talk | contribs) (7,068 bytes) *()* (undo)
(cur) (prev) 16:53, 28 Apr 2007 170.76.35.11 (talk | contribs) (7,077 bytes) *()* (undo)
(cur) (prev) 03:51, 26 Apr 2007 TJ Spyke (talk | contribs) (7,077 bytes) *(NN)* (undo)
(cur) (prev) 03:48, 26 Apr 2007 142.167.41.161 (talk | contribs) (7,345 bytes) *(/* References */ a*
(cur) (prev) 22:28, 22 Apr 2007 TJ Spyke (talk | contribs) (7,088 bytes) *(nicknames aren)* (undo)

Where we're at

1 Building history trees

2 Visualizing history trees

3 Computing on history trees

4 Improving tree building

5 Conclusion

Computationally determining acceptance

Two questions in computational analysis:

1. Was a particular revision retained or discarded?
2. At a particular time, what revision is most representative of the article's “true” state?

Our intuition: use later editing activity to determine the community's response.

Computationally determining acceptance

Difficulties:

- Cannot just check for a revert – what if the revert is reverted?
- Ultimate end state of the article is inaccessible, so cannot see if revision is in path to end state.

Goal: A notion of "accepted" revisions that

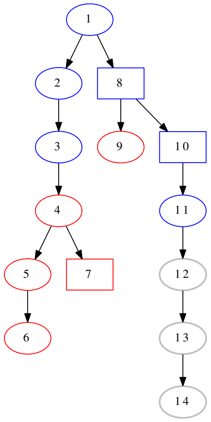
- Handles reverted reverts
- Avoids saying things about revisions overly affected by the temporal frontier
- Is well-defined

k-acceptance

k-acceptance: see whether there's a path of *k* edits leading from a revision.

Acceptance near the end of history is undefined.

A sample tree indicating *k*-acceptance for $k = 3$:



Picking *k*

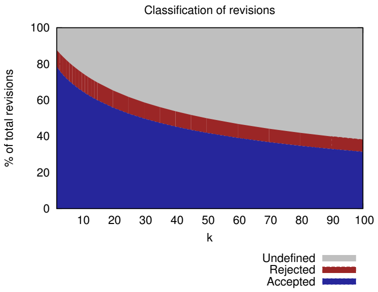
Two properties of *k*-acceptance:

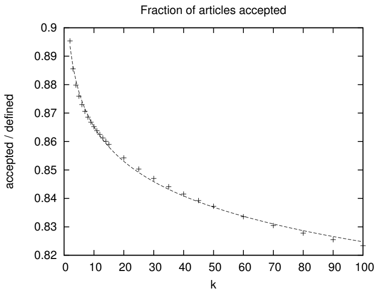
1. Increasing *k* decreases the likelihood of labeling as accepted a revision which is ultimately rejected.
2. Increasing *k* decreases the number of revisions for which *k*-acceptance can be computed.

We therefore want to select a *k* that balances the twin concerns of accuracy and defineability (and thus usefulness).

Picking *k* (cont.)

We computed acceptance over all of Wikipedia's main namespace for various values of *k*.





It seems that $10 \leq k \leq 20$ (say, $k = 15$) is a reasonable choice.

Where we're at

1. Building history trees
2. Visualizing history trees
3. Computing on history trees
4. Improving tree building
5. Conclusion

Limitations of basic tree building

There are some actions which may constitute discarded work, but which do not appear as reverts.

- Manually almost-revert a revision
- Partially revert a revision
- Undo a revision while making other changes

Framework for more advanced tree building

We attempted to go beyond reverts based on the following intuition:

> *An editor has in their mind some target state for the article and edits the article so as to bring it closer to their desired target state.*

This suggests a modification to our earlier algorithm:

- If a revision is identical to some prior revision, it is a revert node whose parent is the most recent identical revision.
- Otherwise, the revision is an edit node whose parent is the *most similar* previous revision [Sabel, 2007].

Similarity metrics

We consider two metrics for comparing revisions:

Cosine Similarity (CS)
: Standard cosine similarity between bag-of-words representations.

Adoption Coefficient (AC) [Sabel, 2007]
: Metric of retained text based on longest-common-subsequence.

In both cases, we adjust by a damping factor.

Problems with CS and AC

We have found some difficulties applying these algorithms in practice:

- $O(n^2)$ performance (and LCS itself is slow)
- Word-level tokens allow large changes to dominate small changes
- Extreme vandalism breaks the underlying assumption and produces unpredictable results
- AC does not respond well to content moves
- CS results can be difficult to understand

Further, it is unclear how some cases (such as undoing old revisions) should be handled, and this difficulty remains for any future solutions.

Where we're at

1. Building history trees
2. Visualizing history trees
3. Computing on history trees
4. Improving tree building
5. Conclusion

Conclusion and contributions

- Basic revision history structuring (steps back from Sabel 2007)
- In-line visualization method for history trees
- Computational method for determining whether a revision has been "accepted" by the editing community around a page in a well-defined manner
- Consideration of additional structuring algorithms and explanation of current defects
- There are open questions about doing a better job of detecting discarded work (and, in some cases, what "better" even means)

Questions?

Funded by NSF grants IIS 05-34420 and IIS 08-08692