collection-level analysis tools for books online

Cathy Marshall
MSR-SVC

Books Online Workshop
30 October 2008
Napa, California
the short and wonderous life of eBooks, generation 1: the excitement of multimedia

At the electrifying conclusion of his hour-long work

PLAY CONCLUSION OF THE NINTH

the audience broke out in tumultuous applause. Oblivious to everything around him, Beethoven continued to beat time after the music had stopped.

His friends on stage had to motion him to stop. They then turned him gently around so that he could at least see the ovation he had earned.

Circa 1989-1991
and the excitement of being on a computer: e.g. the ability to do in place dictionary use
the short and wonderous life of eBooks, generation 2: mobility, replacing the pBook [sic]
Backlash!

ZIPPY
I heard that th' e-book trend never really took off -- sales of th' things are tanking!

E-BOOKS WILL NEVER REPLACE P-BOOKS!

E-BOOKS ARE SPINELESS!

"PAPER TRAIL"
PAPER IS FRIENDLY!
PAPER IS CUDDLY!
YOU CAN TAKE PAPER TO BED!
YOU CAN HAVE A RELATIONSHIP WITH PAPER!

--WATCH IT, GRIFFY! YOU CAN ALSO GO TOO FAR WITH PAPER! THEN, IT'LL SLAP YOU IN TH' SNOUT!

WHY DO PUNDITS ALWAYS PREDICT THAT TH' NEWEST TECHNOLOGY WILL SUPPLANT TH' CURRENT TECHNOLOGY?! WHAT USUALLY HAPPENS IS THEY COEXIST, SIDE BY SIDE -- FOREVER!
The second wave of eBooks provoked a profound sense of anxiety—anxiety that will require *at least* 50 years of therapy.

“I’m never gonna read a book on the computer. It’s too hard to curl up with a monitor.”
the short and wonderous life of eBooks, generation 3: recapturing lost innocence

Circa 2006-2008
One last point about the page v. pixel debate

“Given the historic price-performance trajectories for storage, in a few years at least some high-end appliances will house hundreds, if not thousands, of books simultaneously, and certainly laptops with software book readers will house thousands or tends of thousands of books at once. *Think of portable personal digital libraries, not portable electronic books,* as the future role of these appliances.”

Idea 1: harness the wisdom of crowds
Many ways to implement a 'wisdom of crowds' approach

<table>
<thead>
<tr>
<th>Explicit judgments</th>
<th>Whole text</th>
<th>Within-text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recommending a book; rating a movie</td>
<td>Quoting a passage</td>
</tr>
<tr>
<td>Implicit actions</td>
<td>Sharing a newspaper article; linking to a web page</td>
<td>Annotating a selection</td>
</tr>
</tbody>
</table>
Example: aggregating (student) annotations
First, recall that readers’ personal annotations are not always meaningful

“So of this I’m starting to skim more. As it talks about things that are less relevant. I can’t remember – here’s another highlight, so I definitely read this far. Uh. I think I read to the end of the chapter. Yeah, there’s another highlight.”

A student flipping through e-book pages
Observations of annotations in the aggregate

- consensus is significantly more common than predicted by a strict probabilistic calculation of overlap

- annotators converge on important text that is different than the text that the authors and publisher designate as important
Annotation: collective effects

We have proposed in [15] a clustering process called the time-constrained clustering which takes into account both visual characteristics and temporal locality of shots. Let $C_i$ be the $i$th cluster, and $w$, $x$, and $y$ be elements in a cluster, each of which is a video shot, we impose the clustering criteria as follows. For all $x \in C_i$

1) $\max_{w \in C_i} d(x, w) \leq \delta$

2) $\max_{y \in C_i} d(x, y) \leq T$

3) $d(x, w) > \delta$ or $d_i(x, w) > T$,
   for all $w \in C_j$, $j \neq i$.  

reader 1

We have proposed in [15] a clustering process called the time-constrained clustering which takes into account both visual characteristics and temporal locality of shots. Let $C_i$ be the $i$th cluster, and $w$, $x$, and $y$ be elements in a cluster, each of which is a video shot, we impose the clustering criteria as follows. For all $x \in C_i$

1) $\max_{w \in C_i} d(x, w) \leq \delta$

2) $\max_{y \in C_i} d(x, y) \leq T$

3) $d(x, w) > \delta$ or $d_i(x, w) > T$,
   for all $w \in C_j$, $j \neq i$.  

reader 2

reader 3

We have proposed in [15] a clustering process called the time-constrained clustering which takes into account both visual characteristics and temporal locality of shots. Let $C_i$ be the $i$th cluster, and $w$, $x$, and $y$ be elements in a cluster, each of which is a video shot, we impose the clustering criteria as follows. For all $x \in C_i$

1) $\max_{w \in C_i} d(x, w) \leq \delta$

2) $\max_{y \in C_i} d(x, y) \leq T$

3) $d(x, w) > \delta$ or $d_i(x, w) > T$,
   for all $w \in C_j$, $j \neq i$.  

reader 4

We have proposed in [15] a clustering process called the time-constrained clustering which takes into account both visual characteristics and temporal locality of shots. Let $C_i$ be the $i$th cluster, and $w$, $x$, and $y$ be elements in a cluster, each of which is a video shot, we impose the clustering criteria as follows. For all $x \in C_i$

1) $\max_{w \in C_i} d(x, w) \leq \delta$

2) $\max_{y \in C_i} d(x, y) \leq T$

3) $d(x, w) > \delta$ or $d_i(x, w) > T$,
   for all $w \in C_j$, $j \neq i$.  

reader 5

reader 6
visualizing collective intelligence
Idea 2: make gathering easier

Keep gathering snouts as low as practical to pick up downed ears
Gathering has interesting requirements
let's look at an example...

“I was searching one of the things that [the professor] had suggested ... Things to look for in the critical reading. And just one of the offhand things that she tossed off was, ‘do they say anything about universals versus particulars?’... And that caught my interest. ... First I searched for ‘particular’ and he uses the word particular a lot. So, you know, like a ‘particular theory’, a ‘particular couch,’ whatever. And not in the sense that I’m looking for...”

--an English Lit grad student
Break it down

• limit the scope of a search
e.g. just Shakespeare’s plays within a larger corpus of English literature

• define a thesaurus-like tree of search terms
e.g. references to specific types of birds

• gather together multiple passages in a space and use tools to work with separate search results in that space
e.g. organize ‘bird passages’ along different dimensions, say mythological v. natural
Idea 3: textual discovery tools
***I am not a text mining person***
So let's keep it dead-simple...
Moby Dick, chapter 1, Loomings (via wordle)
Moby Dick, chapter 41, Moby Dick
(again, via wordle)
conclusions: the high order bits

- Books online can be seen as *datasets* rather than traditional collections

- Books online give us the opportunity to take advantage of *collective* effects

- Don’t forget analytic/*gathering* tools for humans too, of course!