Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story

Chirag Shah
CRADLE Talk
SILS, UNC Chapel Hill
September 21, 2007
Flow of this story

• What are we trying to do?
• Why are we doing it?
• How did we do it?
• What came out of it?
• How to make sense of it?
• Where do we go from here?
ContextMiner - the big picture
YouTube spread

- YouTube’s market share: 42.94% (Source: Hitwise.com)
- Average visit duration on YouTube: 13:20 mins (Source: Xeep.net)
- As of August, 2006: (Source: Wall Street Journal)
  - Videos > 6,000,000 (growth: 20% every month)
  - Storage > 45 TB
  - Views > 1,730,000,000
  - Time spent watching YouTube videos = 9,035 years
- Time taken for the spread among 50 million users
  (Source: Dean José-Marie Griffith’s talk on 09/17/2007)
  - Radio: 17 years
  - TV: 13 years
  - Internet: 5 years
  - MySpace: 3 years
  - YouTube: 1 year
Video harvesting from YouTube

Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story
Chirag Shah
SILS, UNC Chapel Hill
CRADLE Talk
September 21, 2007
Video harvesting from YouTube for election 2008
Video harvesting from YouTube for election 2008

- Number of seed queries: 56 (now 57)
  - 6 general queries and the rest are the names

- Crawl everyday (almost)

- Get top 100 results for each query

- Collect more than 20 attributes (including all the comments)

- Download flash videos

- Use YouTube APIs, screen scrapping, and other tools
ContextMiner’s journey - from Carolina to California
Overview of the collection

- Total videos
- Unique videos

Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story
Chirag Shah
SILS, UNC Chapel Hill
CRADLE Talk
September 21, 2007
Overview of the collection (as of 09/17/2007)

- Crawls = 100
- Unique videos > 10,000
- Video files > 100 GB
- Total honors > 200
- Total views > 125,000,000
- Total ratings > 800,000
- Total comments ~ 800,000
JCDL 2007 (Vancouver, BC, Canada)


YOU KNOW WHAT I LIKE ABOUT SUMMER DAYS?

THEY'RE JUST MADE FOR DOING THINGS

...EVEN IF IT'S NOTHING.

ESPECIALLY IF IT'S NOTHING.
Tipping Point

Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story

Chirag Shah
SILS, UNC Chapel Hill

CRADLE Talk
September 21, 2007
Looking for the “movers and shakers”

- **Connectors**
  - Know many people
  - Many acquaintances; may not be very strong

- **Mavens**
  - Know a lot about something
  - Special powers because of their in-depth knowledge

- **Salesmen**
  - Have skills to persuade people
  - May not have a lot of knowledge, but can convince someone
Election “connectors” on YouTube
People who posted at least one comment on many videos

Number of videos on which at least one comment was posted
Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story

Election “mavens” on YouTube

People who commented a lot in the entire collection

Total number of comments

Number of people (log scale)
Election “mavens” on YouTube

People who commented a lot for any given video
Election “salesmen” on YouTube

People who posted a lot of videos

[Bar chart showing the number of authors (log scale) and the number of videos they posted. The chart illustrates a distribution with a few authors posting a large number of videos, and many authors posting a small number of videos.]
SIGIR 2007 (Amsterdam, The Netherlands)

Detecting events from the collection

$\phi$ may not be a good representation, it's not a linear function.

We want to know how people's participation change with changing number of visitors and this may not be linear. For instance, double number of visits may not lead to doubling number of comments.

We need to learn this function.

Evaluation: show that this learned function succeeds in predicting right thing when simple $\phi$ fails.

Local changes in $\phi$ can help us understand the change in level of participation.

$\Delta \phi = 0$ No change

$\Delta \phi > 0$ More participation

$\Delta \phi < 0$ Less participation

Model: $(o, \phi)$

$\Delta o = o_i - o_{i-1}$

$\Delta o_1 = o_2 - o_1$

$\Delta o_2 = o_3 - o_2$

$\Delta o_3 = o_4 - o_3$

$\phi = \frac{\text{participants}}{\text{visitors}}$

$\phi_1 = \frac{\text{comments}}{\text{views}}$

$\phi_2 = \frac{\text{ratings}}{\text{views}}$

$\phi = \arctan\left( \frac{y_2 - y_1}{x_2 - x_1} \right)$

$\Delta o = 0$ No change

$\Delta o > 0$ More interests

$\Delta o < 0$ Less interests

Can we measure the strength/accuracy of this measure?
Model for detecting changes for a video

\[ M = (\Theta, \Phi) \]

Changes in different parameters

Level of participation

\[ \Theta = (\theta_1, \theta_2, \theta_3) \]

\[ \theta_i = \tan^{-1}\left( \frac{y^i_2 - y^i_1}{x^i_2 - x^i_1} \right) \]

\[ \Phi = (\phi_1, \phi_2) \]

\[ \phi_1 = \frac{\text{#comments}}{\text{#views}} \]

\[ \phi_2 = \frac{\text{#ratings}}{\text{#views}} \]
View counts for a video

Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story

Chirag Shah
SILS, UNC Chapel Hill

CRADLE Talk
September 21, 2007
Comments and ratings counts for a video

Crawls

Number

Comments
Ratings

Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story
Chirag Shah
SILS, UNC Chapel Hill
CRADLE Talk
September 21, 2007
Theta values

Theta values

Crawls

1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49

Theta1
Theta2
Theta3
Detecting events from the collection

Hunting for Hip, Hipsters, and Happenings on YouTube - A ContextMiner Story

Chirag Shah
SILS, UNC Chapel Hill

CRADLE Talk
September 21, 2007
Detecting events from the collection

Normalized # Comments

Crawl date
Detecting events from the collection

Normalized # Ratings

Crawl date

Normalized # Ratings

7/31/07 8/2/07 8/4/07 8/6/07 8/8/07 8/10/07 8/12/07 8/14/07 8/16/07 8/18/07 8/20/07 8/22/07 8/24/07 8/26/07 8/28/07 8/30/07 9/1/07 9/3/07 9/5/07 9/7/07 9/9/07 9/11/07 9/13/07 9/15/07 9/17/07
Interesting issues to investigate

• How to detect significant changes or events?

• How faithfully the changes in online activities reflect the real-life events?

• Who is responsible for these significant changes in online world?

• Can we detect opinions and sentiments of people by analyzing their online behavior and participation?
ContextMiner - the big picture
Becoming a part of this story

- Access the ContextMiner website
- Access the data using Developer APIs
- Build your own crawler using TubeKit

A YouTube Crawling Toolkit
Ideas, ideas, ideas...

Idea: Create a relevance/retrieval function using title, descr, and the model for the video.

Relevance $\propto$ Text similarity (title, descr, tags)
$
\propto$ Access
$
\propto$ participation $\rightarrow$ popularity

We don't have click through data

Text matching

How to evaluate?

Ideas measure "quality performance." Watch a video's rank changing and compare it with its popularity and passive participation values.

What's the point?
Moral of the story

Research areas
Policy issues
Technical challenges
Identifying and capturing context
Collection visualization
Understanding online user behavior and participation
Event detection
User interface
Retrieval performance

Tools
ContextMiner
DiscoverInfo
DIToolkit
ContextMiner APIs
TubeKit
FEX

Websites
Author’s homepage: http://www.unc.edu/~chirags
VidArch homepage: http://www.ils.unc.edu/vidarch/
ContextMiner: http://idl63.ils.unc.edu/chirag/ContextMiner/
DiscoverInfo: http://idl.ils.unc.edu/~chirag/DiscoverInfo/
DIToolkit: http://idl.ils.unc.edu/~chirag/DIToolkit
ContextMiner APIs: http://idl63.ils.unc.edu/chirag/ContextMiner/developer.php
TubeKit: http://idl.ils.unc.edu/~chirag/TubeKit/
Hunting for Hip, Hipsters, and Happenings on YouTube - *ContextMiner* Story

Chirag Shah
SILS, UNC Chapel Hill

CRADLE Talk
September 21, 2007

**Thanks!**