Information Access Challenges in the Blogspace

Gilad Mishne
Overview

- Blogs 101

- Properties of the blogspace
  - From an information access point of view

- Application test cases
  - http://moodviews.com
  - Profile-based product and ad matching

- Challenges and ongoing efforts
Web logs

- In a nutshell: online diaries

- More formally, web pages which
  - Include periodic, time-stamped entries ('posts')
  - Are ordered in reverse timeline
  - Are published through a CMS
  - Are authored by a single person or a small community
  - Contain commentary about the blogger's life, current events, web-related issues, ...
  - Regularly updated
  - Allow visitor feedback

- Blogspace/blogosphere:
  - The totality of blogs
Blog examples

My Journey: Lordi celebratory gig (Helsinki)

chain saw in his hand! That was really monstrous and horrifyingly electric. They also sang their another master piece "Would you love the monster man?". If they were looking for an answer then the crowd said it all. They were at the stage till 9:30 pm after which other bands took over.

Finland is one rock crazy country having internationally reputed bands like HIM, Rasmus and the now Eurovision winner 'Lordi'. Earlier this week there was huge outrage against a local tabloid when it published the picture of Lordi without his monster mask. The band has asked the media to publish only their monster pictures and address them always by their band names.

The winner of the Eurovision has to host the next edition, so next year the event will be held in Finland.

(Original copyright of image applied)

posted by greensatya @ 9:40 AM

17 Comments:

- At 11:16 AM, killer is I said...
  :doc... I can't believe that u didn't go to see such event. U weren't part of such crowd so, i die to see such events. That must have have been great, 80,000 ppl... thats awesome, its more like woodstock except it was for 1 day only. But looks really awesome, i have heard HIM, thats gud band. And u didn't took any pics of show?? if yes, then post them.

- At 11:34 AM, Anonymous said...
  where are you in the pictures?
  curiosity.
Blog examples
The blogspace as a corpus

- Differences from other web corpora
  - Content
  - Structure
  - Timeline
  - Growth

![Graph showing growth of weblogs from March 2003 to January 2006 by Technorati. The graph indicates a doubling in size every 5 months, with a consistent doubling over the last 36 months. The graph shows 26.6 million blogs tracked.]
The blogspace vs. the web

<table>
<thead>
<tr>
<th></th>
<th>Blogs</th>
<th>Web pages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content type</strong></td>
<td>personal, diary-like, commentary, observations,</td>
<td>Anything, really (but in most cases - low on sentiment content)</td>
</tr>
<tr>
<td></td>
<td>sentiments, moods</td>
<td></td>
</tr>
<tr>
<td><strong>Content format</strong></td>
<td>Template</td>
<td>Custom</td>
</tr>
<tr>
<td><strong>Links</strong></td>
<td>Dense; links change frequently; different types</td>
<td>Links usually static</td>
</tr>
<tr>
<td><strong>Timeline</strong></td>
<td>Frequently updated; reverse-chrono-logically-sorted</td>
<td>Usually static; time does not play major role</td>
</tr>
<tr>
<td><strong>Represents</strong></td>
<td>A person’s life</td>
<td>Information</td>
</tr>
</tbody>
</table>
Structure of the blogspace

- The World Wide Web:
  - A scale-free network

- The blogspace:
  - A scale-free (scale-weak?)
  - and small-world network

- What does this mean?
  - There are a few influential blogs, and a “long tail”
  - Information diffuses quickly
  - There are lots of communities
  - A good model of a social network (better than the web)
    - Not surprising – this is a network of people
Overview

✓ Blogs 101

✓ Properties of the blogspace
  ✓ From an information access point of view

▪ Application test cases
  ▪ http://moodviews.com
  ▪ Profile-based product and ad matching

▪ Challenges and ongoing activities
Timeline-oriented sentiments

- Many bloggers indicate their mood at the time of writing

<table>
<thead>
<tr>
<th>Date</th>
<th>Rate of change</th>
<th>Absolute counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>05/05/2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/10/2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>08/15/2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>09/21/2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/29/2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/05/2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>01/11/2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02/17/2016</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questions:
- Can we predict the mood using the bloggers’ text?
- Can we explain changes in the “global mood”?
- ...
Mood level prediction:

- Language model comparison identifies “distinctive mood-related words”
- “Mood recipe” calculated for each mood

\[
\text{Happy} = 23.345 + 0.0318 \times \text{total-posts} + (-2.4026 \times \text{count(always)} + (-114.9877 \times \text{count(day)} + 16.2727 \times \text{count(excited)} + 55.3942 \times \text{countfinally} + 129.2576 \times \text{count(happy)} + 223.8079 \times \text{count(home)} + (-246.8737 \times \text{count(know)} + 506.9564 \times \text{count(lol)} + 5.7815 \times \text{count(thoughtful)} + (-88.1313 \times \text{count(will be)}
\]

Prediction accuracy: 0.83 correlation

- Up to 0.95 on some moods
http://www.moodviews.com

- Explaining irregular moods:
  - Identify “spike”
  - Language model during spike compared to “expected” model
  - Distinctive words used as a query to an index of events
Moodsignals in action

Changes in "excited" over 07/10/2005 - 07/20/2005

Peak (1) 07/15/2005 11h - 07/16/2005 11h

Overused terms during the peak period:

- harry
- potter
- chat

Harry Potter and the Half-Blood Prince released

From Wikinews, the free news source you can write!

July 16, 2005

Harry Potter and the Half-Blood Prince by J.K. Rowling was released today to hundreds of special midnight parties at bookstores. The book is expected to sell tens of thousands of copies in the first hour, and the waves of fans just keep on coming. As midnight spreads across the Atlantic and the North American continent, more and more stores will start selling.

The Royal Mail UK postal service delivered four hundred tons of the pre-ordered books this morning.
Blogger profiling

- Lots of work on collecting knowledge from the blogspace as a collective
  - Sales predictions, political reflections, …

- What can be mined from *individual* blogs?

- Intuition: a blog can be used to create a profile of the human behind it

- Textual profiles of blogs can be derived using
  - Keyword extraction
  - Summarization
  - your-favorite-method-here
Blogger profiles in action 1

- Task: match blogger with products
  - In particular, books she is likely to appreciate

1. Create textual profile of blog
   - Use corpus-comparison techniques

2. Identify typical products related to profile
   - Use a large DB of products - Amazon

- Evaluation:
  - Compare the derived categories to those present in an explicit list of “desired products” by the blogger
  - Performs 121% better than simply trying to locate references to products in the blog
Blogger profiles in action 2

- Task: contextual advertising in blog posts
  - More difficult than contextual advertising on other web pages, since blog posts can be very un-topical
Blogger profiles in action 2

- Task: contextual advertising in blog posts
  - More difficult than contextual advertising on other web pages, since blog posts can be very un-topical

1. Extract textual profile of blog post components
   - Post, blog, community, comments, time, tags, similar posts

2. Combine models; distill single profile of post; match with ads by similarity

- Evaluation:
  - Used a corpus of blogs/ad used by the largest Dutch blogging platform
  - Human assessors
  - Outperform state-of-the-art in contextual advertising by 20%
Main inf. access challenges

- In one sentence
  - Identify how the intrinsic properties of the blogspace can be used to access the information in it
    - Subjective, personal content
    - Temporal profile
    - Social-network structure

- In particular
  - Modeling bloggers and communities
  - Catering for different search requirements
  - Data quality
  - <the-next-big-thing>
Questions?