An Introduction to Social Mining

Vladimir Gorovoy* and Yana Volkovich†

†@yvolkovich
Barcelona Media, Information, Technology & Society Group
Barcelona, Spain

* @vgorovoy
Yandex, Yandex.Uslugi
Saint Petersburg, Russia

August, 15-19 2011
Outline

1. About the course
2. Introduction
3. Opinion mining
4. Practical task
An Introduction to Social Media

About us

Vladimir Gorovoy
- Head of Yandex.Uslugi, Yandex
- Dipl. Eng. Degree in Mathematics and Computer Science from the Saint Petersburg State University

Yana Volkovich
- Research Scientist in Information, Technology and Society Group, Barcelona Media Innovation Center
- Ph.D. in Applied Mathematics from the University of Twente
- Dipl. Eng. Degree in Mathematics and Computer Science from the Saint Petersburg State University
day 1  An introduction to Social Media; Social Market; Practical task announcement;
day 2  Yandex.Market;
day 3  Social graph mining; Recommended deadline for the practical task;
day 4  Twitter, Foursquare, etc.; Results for the practical task;
day 5  New research directions; Presentations by the practical task winners.
Introduction

What is Social Media?
“Social media is like teen sex. Everyone wants to do it. No one actually knows how.”
(Avinash Kaushik, Google’s analytics evangelist)
Social Media is not only about Social Networks
**Social Media** is a media for social interaction using highly accessible and scalable communication techniques.

**Social Media** is the use of web-based and mobile technologies to turn communications into interactive dialog.
“In contrast to one-to-many communication structure of traditional mass-media, social media allows the emergence of many-to-many communication, and gives a rise to mass self-communication” [Castells, 2009]
What are the goals /purposes of Social Media?
Introduction
Social Media: examples

social communication emails, mobiles, forums, chats;
social networking facebook, google+;
social blogging/microblogging twitter, livejournal, blogger;
social sharing flickr, vimeo, youtube;
social news digg, slashdot, cnn ireport;
social bookmarking delicious, citeulike;
social knowledge, wikis wikipedia, tripadvisor;
social shopping groupon, amazon, ebay;
social apps & games foursquare, farmville;
etc.
too much data!
What could we do?

to ask right questions
Introduction
Questions

Why Social Media?
What is Social Media for?

Social Media Policies & Strategies

What is going on?
What can we learn?

Social Media

How to exploit Social Media?

Social Media Practices

Social Media Mining
Opinion Mining
Opinion mining

History

- 1993: cartoon by Peter Steiner published by The New Yorker on July 5, 1993
2011:

"On Twitter, nobody knows you're a computer program."
Opinion mining
Introduction

- people search for and are affected by online opinions;
- Consumer reviews are significantly more (12 times) trusted than descriptions that come from manufacturers. (eMarketer, Feb. 2010)
- 90% of consumers online trust recommendations from people they know; 70% trust opinions of unknown users. (Econsultancy, Jul. 2009)
People express their opinions via

- voting;
- pressing *like* or *+1*;
- rating;
- commenting;
- sharing;
- etc.
Opinion mining

Introduction (cont.)

- People evaluate/reflect on
  - items;
  - real events;
  - other people;
  - items created by others.
Opinion mining
Examples

50 MILLION
SOCIAL REVIEWS
HAVE BEEN CREATED
IN TRIPADVISOR FOR
495 000
RATED HOTELS
CouchSurfing is a hospitality exchange network and website with 3 million members in 246 countries and territories;
Opinion mining
CouchSurfing (cont.)

- survey study by [Adamic et al., 2011].
- different level of participation:
  - some prefer to host as it allows them to meet people without leaving home.
  - some use the site mainly for travel. (One interviewed participant had been couchsurfing nonstop for a year.)
Discomfort in leaving negative references: Negative ratings are seldom given publicly in part because the individual being rated can reciprocate. [Adamic et al., 2011].

Textual references (and their number) are far more important.

<table>
<thead>
<tr>
<th></th>
<th>Very Important</th>
<th>Important</th>
<th>Neutral</th>
<th>Unimportant</th>
<th>Very Unimportant</th>
</tr>
</thead>
<tbody>
<tr>
<td>how many vouches they have</td>
<td>8.6%</td>
<td>29.3%</td>
<td>37.1%</td>
<td>15.8%</td>
<td>6.9%</td>
</tr>
<tr>
<td>whether they are verified</td>
<td>9.1%</td>
<td>23.0%</td>
<td>36.4%</td>
<td>18.9%</td>
<td>11.6%</td>
</tr>
<tr>
<td>number of references received</td>
<td>23.0%</td>
<td>57.0%</td>
<td>13.7%</td>
<td>4.4%</td>
<td>1.1%</td>
</tr>
<tr>
<td>text of references received</td>
<td>47.6%</td>
<td>40.8%</td>
<td>8.2%</td>
<td>2.1%</td>
<td>0.8%</td>
</tr>
<tr>
<td>number of friends</td>
<td>4.4%</td>
<td>23.6%</td>
<td>37.7%</td>
<td>25.1%</td>
<td>8.0%</td>
</tr>
<tr>
<td>friends' friendship level</td>
<td>3.2%</td>
<td>19.0%</td>
<td>35.6%</td>
<td>25.0%</td>
<td>13.5%</td>
</tr>
</tbody>
</table>
opinion: What does A think about this item? [Pang and Lee, 2008]
meta-opinion: What do other users think about A’s opinion about this item? [Danescu et al., 2009]
Opinion mining
Rating reviews

- Amazon.com for Meta-Opinion Analysis (Danescu et al. [2009])

![Amazon.com book](image)
Opinion mining
Rating reviews: Question

- A product has a average star rating of ⭐⭐⭐.
- Aim is to write a helpful review for the product.
- Which would be your star rating choice if you can only alter the star rating of the review?
Opinion mining
Rating reviews: Social Psychology Hypotheses

Social Psychology Hypotheses:

- **Conformity** star rating is closer to the average star rating for the product;
- **Brilliant but cruel** star rating is below to the average star rating for the product;
- **Individual bias** star rating reflects the evaluators’ personal opinion about the product.
Conforming reviews are more helpful.
signed deviation = star rating - average star rating;

positive reviews are more helpful (Brilliant-but-cruel is not working).
Signed deviations vs. helpfulness ratio, in the Japanese (left) and U.S. (right) data. The curve for Japan has a pronounced lean towards the left.
Opinion mining
How is it in Russia?

- Yandex.Market
Yandex.Market is the most successful site for reviews in RuNet by the number of reviews and by reviews’ quality.

Link: bit.ly/russir2011
Practical task
Yandex Market snapshot

Yandex market

Цифровые фотоаппараты / Canon

Описание Характеристики Отзывы Обзоры Обсуждение Аксессуары

Средняя цена:

average rating
★★★★☆ 629 оценок

Поделиться
Добавить в Список покупок
Добавить к сравнению

— продвинутая зеркальная фотокамера
— поддержка сменных объективов с байонетом
— объектив в комплекте
— матрица 19 мегапикселов (22.3 x 14.9 мм)
— съемка видео разрешением до 1920x1080
— поворотный экран 3"
— влагозащищенный корпус
все характеристики

Мнение покупателей
★★★★☆ 629 оценок

user's rating
Iakov
24 апреля

advantages
disadvantages

usefulness (yes/no)
Yandex reviews usefulness:

\[
\frac{\text{useful} + 1}{\text{numvotes} + 2} - \frac{1}{2 \times (\text{numvotes} + 2)}
\]

- \text{useful} is the number of votes that rate review as useful;
- \text{numvotes} is the number of all votes that rate reviews usefulness;

The main point: \textit{usefulness} = \textit{share of useful - error}.

\textbf{Error} is a half of confidence interval;
Practical task

Learning set

Yandex.Market data set:

file:reviews.xml  Reviews and usefulness for items (digital cameras):
file:modeldata.csv  Average rating and usefulness of items;
file:categorydata.csv  Average rating and average usefulness of the product items for the selected category (digital cameras);
file:userdata.csv  Average usefulness of reviews and the number of the accepted reviews done by an author;
**Practical task**

Files (1)

file:reviews.xml  Reviews and usefulness for items (digital cameras);
- **ID** review id;
- **MODEL ID** item id;
- **AUTHOR D** author id;
- **CR TIME** writing time of the review;
- **RATING** rating of the model by the author of the review (from 1 to 5 (best));
- **TEXT** text of the review;
- **PRO** text about advantages of the model;
- **CONTRA** text about disadvantages of the model;
- **RANK** evaluation by other users of the review usefulness (from 0 to 1 (best)).
Practical task

Files (2)

file:modeldata.csv  Average rating and usefulness of items;

- **MODEL ID** item id;
- **AVG RANK** average usefulness of items’ reviews;
- **RATING** average item rating (from 1 to 5 (best));
file:categorydata.csv  Average rating and average usefulness of the product items for the selected category (digital cameras);

- CATEGORY AVG RATING  average rating of the product items for the selected category;
- CATEGORY AVG RANK  average usefulness of the product items for the selected category.
file:userdata.csv  Average usefulness of reviews and the number of the accepted reviews done by an author;

- **AUTHOR ID** author id;
- **NUM REVIEWS** the number of accepted reviews done by the author;
- **AVG RANK** average usefulness of the reviews done by the author (from 0 to 1 (best));
task 1  Given
- text and rating of the item’s review,
- average rating of the item,
- average usefulness for the item and for the category,
- average usefulness of the user,
- number of reviews from the user,

  to predict

- usefulness of the review by other users.
task 2  Given

- text of the item’s review,
- average rating of the item,
- average usefulness of the user’s reviews,
- number of reviews from the user,

to predict

- the user’s rating of the item.
Practical task

Links

- **Weka**: www.cs.waikato.ac.nz/ml/weka/
- **LingPipe**: http://alias-i.com/lingpipe/demos/tutorial/logistic-regression/read-me.html
- **Shark**: http://shark-project.sourceforge.net/Tutorials.html
- **Shogun**: http://www.shogun-toolbox.org/
Practical task

Gameplan

- deadline: evening Wed, 17th
- results: Thr, 18th
- winners present their ideas (10 min): Fri, 19th
Practical task

Questions

