Opinion Mining in Web Forums

Manos Karvounis
Thodoris Georgiou
Yannis Ioannidis

University of Athens
In the beginning

- man created the Web
Let there be structure
Opinion Mining

\( (H, O, f, SO)_1 \)
\( (H, O, f, SO)_2 \)
\( (H, O, f, SO)_3 \)
\( (H, O, f, SO)_4 \)
\( (H, O, f, SO)_5 \)

\( H = \) opinion holder
\( O = \) object,
\( f = \) feature of O
\( SO = \) semantic orientation of opinion on f
Current Research

- Focused mostly on (product, movie, ...) reviews
- There are many other forms of opinions on the Web
- The current framework might be insufficient (opinion orientation, document)

Well, I finally decided to Kindle after much thought and research. *I am so glad that 3G and Wi-Fi are available Down Under. I was concerned about the time it takes to charge the Kindle but all my concerns were unfounded.* It took less than 3 hours to charge my Kindle using the USB cable. I didn't even have to use an adaptor! The other good thing is that I am able to continue reading using my iphone as I downloaded the apps. Awesome!

And now you're guilty of what you just accused me of. When you agree with what they're saying, it's ok, but when you don't it's not.
Opinion Mining on Web Forums

• What is a Web Forum
  • website that accommodates dynamic discussions
    • are separated in autonomous conversations (thread)
• Users participate in a thread by writing a post
• They can also quote a previous post and reply
What is there to extract

- Many conversations on current events and hot topics
- People agree and disagree, expressing their opinions in a peer-to-peer fashion
- A wealth of information on how people feel about what is happening around them, and why
Key Points

- Agreements and disagreements play a significant role in the discussion (opinion orientation)
  - Only if we extract these relationships can we identify the different sides of the argument
- Individual posts do not carry the complete topic information (document)
  - We need to look at the whole discussion structure
- And many more …
Extracting Topics of Debate between Users on Web Discussion Boards

Thodoris Georgiou
Manos Karvounis
Yannis Ioannidis
Main Aim and Motivation

- Discover the topic of every post in large discussions (threads)
  - Following a user-oriented approach it is important to know the exact topics every user has discussed about

- Challenges:
  - Single posts may not contain sufficient information to extract topic
    - More than 10% of the posts don’t contain any important keywords
  - Topic of the whole thread can’t be used due to the constant change of the subtopics
Observations

- The flow of the discussion always changes and new sub-topics get introduced.

- A post that replies to another tends to maintain the same topic (minor differences may exist though).

- Users mostly reply either to very recent posts or at the first post of the discussion.
The Method

- Create reply graph
  - Quotations
  - Patterns for direct replies
- Find post neighbors to create discussion contexts for every post
- Get topic keywords for every post
- Set as topic of a post the most frequent keywords of its context
- Link users with topics and extract pairs that share common interests
Reply Graph
The Method

- Create reply graph
  - Quotations
  - Reply patterns
- Find post neighbors to create discussion contexts for every post
- Get topic keywords for every post
- Set as topic of a post the most frequent keywords of its context
- Link users with topics and extract pairs that share common interests
Neighbors of Post $i$

Search depth: 2
The Method

- Create reply graph
  - Quotations
  - Reply patterns
- Find post neighbors to create discussion contexts for every post
- Get topic keywords for every post
- Set as topic of a post the most frequent keywords of its context
- Link users with topics and extract pairs that share common interests
Post: “Sadly, that is not the situation with many children. Some come with health problems, or behavioral problems, or mixed race, and there are many African-American children, and a married couple walks by them, without a second look backwards.”

Keywords: African-American children, behavioral problems, health problems
The Method

- Create reply graph
  - Quotations
  - Reply patterns
- Find post neighbors to create discussion contexts for every post
- Get topic keywords for every post
- Set as topic of a post the most frequent keywords of its context
- Link users with topics and extract pairs that share common interests
Context of Post $i$
The Method

- Create reply graph
  - Quotations
  - Reply patterns
- Find post neighbors to create discussion contexts for every post
- Get topic keywords for every post
- Set as topic of a post the most frequent keywords of its context
- Link users with topics and extract pairs that share common interests
The Results

- The set of topics being discussed in the web discussion board
  - Topics are represented as sets of keywords and phrases
- Every user is linked with the topics he has discussed about

- We can extract pairs of users that are linked with the same topic
Utilizing the Quoting System of Web Forums to Estimate User Agreement

Manos Karvounis
Thodoris Georgiou
Yannis Ioannidis
Main Aim

• Discover agreements & disagreements *without* using topic information

• Reasons for that:
  • NLP is difficult
    • “disagree” yielded results for less than 2% of the posts
  • We can use the relationships discovered to better understand the topic itself and improve recall
Observations (or assumptions)

- People tend to create support groups in a discussion
  - Reinforce their opinions
  - The general sides of an argument are often limited
- When the conversation is established, one is more likely to respond to the arguments he disagrees with than those he agrees
  - Agreement doesn’t need explicit reinforcement
  - Disagreement is constantly fueled by the inability to reach common ground
- Every group has its leader(s)!
Key Points

- Two users replying often to each other are very likely to disagree
  - Evaluation of 813 replies: only 4% expressed agreement, many were part of long reply chains
- Few “power users” have most of the posts and form the backbone of the discussion. The rest of the users comment on their opinions
  - In 3110 posts 25% of the users had 65% of the posts
The Results

• Two sets of users are created
  • Users in the same set very likely to agree
  • Users in different sets very likely to disagree
• To achieve this, we will apply coloring on a graph representation of the discussion
The Graph

User A

Originally posted by User C

Then I guess I was wrong to assume that you agree...

You certainly were! Check my last post again and read carefully what I wrote...

[Diagram of a network with nodes labeled A, B, C, D, E, F, G, H, I and connections indicated with numbers 4, 6, 9, 14, 18, and 35.]
The Algorithm
The Algorithm
The Algorithm
The Algorithm
Key Points

- Start with the most certain disagreement
- Continue moving on the paths with the most certain disagreement
- Spread the current choices to the local discussion
  - Remote parts of the graph might indicate a slightly different topic
  - The coloring choice that must prevail must be a balance between the closest one and the strongest one
A real example
Evaluation

EVALUATION RESULTS

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>86%</td>
<td>91%</td>
</tr>
</tbody>
</table>

Manual evaluation of 3000 posts in the context of 30 randomly chosen threads with various topics (politics, religion, science, society etc).
Future Directions and Impact
Putting it all together

Topic Keywords per User

2 Groups

Same Keywords & Same Group = Agreement
Same Keywords & Dif. Group = Disagreement

Many groups, with their keywords
Future Directions

• The extracted topic keywords correspond to the topics that the connected groups agree/disagree on
• This can be used to more easily identify the opinions expressed
  • They will be near the keywords
  • We don’t need to extensively search every post, since we know the agreement/disagreement relationships
A glimpse of the future