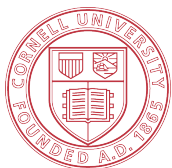


Query-Focused Opinion Summarization for User-Generated Content



Cornell University



IBM Research

Lu Wang¹, Hema Raghavan², Claire Cardie¹, and Vittorio Castelli³

¹Department of Computer Science
Cornell University

²LinkedIn

³IBM T. J. Watson Research Center

Introduction

YAHOO!
Answers

twitter

CNN

WORDPRESS

facebook.

AP

The New York Times

Introduction

- Question: Is Big Data Spreading Inequality? (from New York Times opinion page)



The Dangers of High-Tech Profiling

SEETA PEÑA GANGADHARAN, OPEN TECHNOLOGY INSTITUTE

Big data systems have been used to target minorities. But these systems can be used to help them.



A Way Toward Greater Equality

CHRISTOPHER WOLF, FUTURE OF PRIVACY FORUM

Big data can also advance the interests of minorities and actually fight discrimination.



Implement 'Technological Due Process'

DANIELLE KEATS CITRON, AUTHOR, "HATE CRIMES IN CYBERSPACE"

Oversight of scoring algorithms would go a long way to ensure their fairness and accuracy for both government and private systems.



It Can Be Used for Good in the Community

MAURICE MITCHELL, COMMUNITY ORGANIZER

Prescriptions for our most pressing social issues emerge from the patterns found in the bonanza of collected data points.



Losing Out on Jobs

OLON BAROCAS, PRINCETON AND ANDREW SELBST, PUBLIC CITIZEN

When companies use patterns in large datasets to hire employees, they may unknowingly rely on previous poor decisions.



Extending Credit Through Data

JAKE ROSENBERG, LENDUP

Meaningful data such as on-time rent and bill payments, or even payday loan repayments, do not make it into traditional credit bureau data files.

Introduction

- Question: What is the long term effect of piracy on the music and film industry?
(from Yahoo! Answers)
 - Best answer: Rising costs for movies and music. ... If they sell less, they need to raise the price to make up for what they lost. The other thing will be music and movies with less quality.
 - Answer 1: Its bad... really bad. (Just watch this movie and you will find out ... Piracy causes rappers to appear on your computer).
 - Answer 2: By removing the profitability of music & film companies, piracy takes away their motivation to produce new music & movies. If they can't protect their copyrights, they can't continue to do business. ...
 - Answer 3: It is forcing them to rework their business model, which is a good thing. In short, I don't think the music industry in particular will ever enjoy the huge profits of the 90's. ...
 - Answer 4: Please-People in those businesses make millions of dollars as it is!! I don't think piracy hurts them at all!!!

Introduction

- Question: What is the long term effect of piracy on the music and film industry?
(from Yahoo! Answers)
 - Best answer: Rising costs for movies and music. ... If they sell less, they need to raise the price to make up for what they lost. The other thing will be music and movies with less quality.
 - Answer 1: Its bad... really bad. (Just watch this movie and you will find out ... Piracy causes rappers to appear on your computer).
 - Answer 2: By removing the profitability of music & film companies, piracy takes away their motivation to produce new music & movies. If they can't protect their copyrights, they can't continue to do business. ...
 - Answer 3: It is forcing them to rework their business model, which is a good thing. In short, I don't think the music industry in particular will ever enjoy the huge profits of the 90's. ...
 - Answer 4: Please-People in those businesses make millions of dollars as it is!! I don't think piracy hurts them at all!!!

The Problem

- We present a submodular function-based framework for generating opinion summary that encapsulates different perspectives for a given opinion question and a set of relevant answers or documents.

Datasets

- Community Question Answering (QA)
 - Yahoo! Answers

- Blogs
 - TAC 2008 opinion summarization track

Related Work

- Reviews
 - Hu and Liu, (2004) and Lerman et al., (2009): product reviews summarization
- News and editorials
 - Stoyanov and Cardie (2006): opinion expression identification
 - Paul et al., (2010): contrastive viewpoints summarization
- User generated content
 - Liu et al. (2008): construct question taxonomies for community QA
 - Tomasoni and Huang (2010): add coverage and quality constraints

Related Work

- Submodular functions
 - Lin and Bilmes (2010) introduce the submodular functions for multi-document summarization.
 - Lin and Bilmes (2011), Sipos et al. (2012): unsupervised and supervised learning with submodular functions for newswire summarization.
 - Dasgupta et al., 2013: news and comments summarization.

Outline

- The Submodularity-Based Framework
 - Definition
 - Dispersion functions
 - Submodular functions
 - Summary generation
- Experimental Setup
 - Datasets
 - Evaluation metrics
- Results
 - Main results
 - Discussion: the choice of text similarity metric and dispersion function

Outline

- The Submodularity-Based Framework
 - Definition
 - Dispersion functions
 - Submodular functions
 - Summary generation
- Experimental Setup
 - Datasets
 - Evaluation metrics
- Results
 - Main results
 - Discussion: the choice of text similarity metric and dispersion function

Submodular Functions

- **Input:** a set of sentences $V = \{s_1, \dots, s_n\}$
- **Goal:** select a subset $S \subseteq V$ that maximizes an objective function $f : 2^V \rightarrow R$, under a certain budget.

- **Definition:**

A function $f : 2^V \rightarrow R$ is submodular iff for all $s \in V$ and every $S \subseteq S' \subseteq V$, it satisfies

$$f(S \cup \{s\}) - f(S) \geq f(S' \cup \{s\}) - f(S')$$

Submodular Functions

- However, there are limitations...
 - Redundancy measured with pairwise dissimilarities between sentences is not submodular.
 - Existing work on redundancy handling:
 - Implicitly encoded in the objective function, or
 - Combined with a reward function for diversity
 - Our solution:
 - Adding dispersion functions (Dasgupta et al., 2013) to enforce lexical, semantic, or topic dissimilarity.

Outline

- The Submodularity-Based Framework
 - Definition
 - Dispersion functions
 - Submodular functions
 - Summary generation
- Experimental Setup
 - Datasets
 - Evaluation metrics
- Results
 - Main results
 - Discussion: the choice of text similarity metric and dispersion function

Submodular Functions with Dispersion

- Dispersion functions

- Summation of distance:
$$h_{sum} = \sum_{u \neq v} d(u, v)$$

- Minimum of distance:
$$h_{min} = \min_{u \neq v} d(u, v)$$

- $d(\cdot, \cdot)$ is a distance function defined based on the dissimilarity between pairwise sentences.

Submodular Functions with Dispersion

- Text dissimilarity metrics
 - Lexical dissimilarity

$$d_{lex}(u, v) = 1 - TFIDF(u, v)$$

- Semantic dissimilarity

$$d_{sem}(u, v) = 1 - \sum_{rel_i \in u, rel_j \in v} WordNet(rel_i, rel_j)$$

- Topic dissimilarity

$$d_{topic}(u, v) = JSDivergence(P_u, P_v)$$

Submodular Functions

- **Part One:** Relevance function
 - Produce preference ordering based on the query
- **Part Two:** Coverage functions
 - Topic
 - Polarity
 - Authorship
 - Content

Submodular Functions

- **Part One:** Relevance function
 - A statistical ranker (ListNet, Cao et al., 2007) is trained to generate ranks for the answers or sentences.

- $$r(S) = \sum_{i=0}^{|S|-1} \sqrt{\text{rank}_i^{-1}}$$

Submodular Functions

- **Part Two: Coverage functions**
 - Clustering-based coverage (Topic, Authorship, and Polarity)
 - \mathbf{T} is a partition of the sentences according to a topic distribution
 - *Topic coverage*: $t(S) = \sum_{T \in \mathbf{T}} \sqrt{S \cap T}$
 - Similarly for *Authorship* and *Polarity*
 - Saturation-based coverage
 - *Content coverage*: $c(S) = \sum_{v \in V} \min(\text{coverage}(v, S), \theta \cdot \text{coverage}(v, V))$

Full Objective Function

- Given the submodular functions (relevance + coverage), and the dispersion functions (diversity), our objective is represented as:

$$F(S) = r(S) + \alpha \cdot t(S) + \beta \cdot a(S) + \gamma \cdot p(S) + \eta \cdot c(S) + \delta \cdot h(S)$$



Summary Generation

- Greedy algorithm: for each iteration, we select the sentence that maximizes the objective function.

Outline

- The Submodularity-Based Framework
 - Definition
 - Dispersion functions
 - Submodular functions
 - Summary generation
- Experimental Setup
 - Datasets
 - Evaluation metrics
- Results
 - Main results
 - Discussion: the choice of text similarity metric and dispersion function

Datasets

- Community Question Answering (QA)
 - Yahoo! Answers: about 3.9 million questions in total
 - To get the opinion questions, we run a opinion question classifier
 - 130,609 questions
 - 80% is for training the ranker; 20% is for testing.
 - We do not have human constructed gold-standard summary.

Datasets

- Blogs
 - TAC 2008 opinion summarization track
 - 25 queries
 - We have human-constructed summary in the form of snippets.
 - E.g., given a question “Why do people like Starbucks better than Dunkin Donuts?”, gold-standard snippets include “Starbucks makes a great espresso; Dunkin Donuts espresso stinks.”

Evaluation

- Automatic evaluation
 - Community QA
 - Jensen-Shannon (JS) divergence (Louis and Nenkova, 2013)
 - Blogs
 - ROUGE scores (Lin and Hovy, 2003) and JS divergence
- Human evaluation
 - Community QA
 - Amazon Mechanical Turk
 - Blogs
 - Pyramid (Dang, 2008)

Outline

- The Submodularity-Based Framework
 - Definition
 - Dispersion functions
 - Submodular functions
 - Summary generation
- Experimental Setup
 - Datasets
 - Evaluation metrics
- **Results**
 - **Main results**
 - Discussion: the choice of text similarity metric and dispersion function

Results on Community QA

- Comparisons:
 - Best answer voted by the users
 - Lin and Bilmes (2011): combines content coverage with diversity reward function.
 - Dasgupta et al. (2013): adds semantic similarity based dispersion function.

Results on Community QA

- Automatic Evaluation – JS divergence
 - Average length of best answers is 102.7 words

	Length=100	Length=200
Best Answer	0.3858	--
Lin and Bilmes (2011)	0.3398	0.2008
Lin and Bilmes (2011) + query	0.3379	0.1988
Dasgupta et al. (2013)	0.3316	0.1939
Our system	0.3017	0.1758

Results on Community QA

- Automatic Evaluation – JS divergence
 - Average length of best answers is 102.7 words

	Length=100	Length=200
Best Answer	0.3858	--
Lin and Bilmes (2011)	0.3398	0.2008
Lin and Bilmes (2011) + query	0.3379	0.1988
Dasgupta et al. (2013)	0.3316	0.1939
 Our system	0.3017	0.1758

Results on Community QA

- Automatic Evaluation – JS divergence
- Test on different component of the objective function

	Length=100	Length=200
Rel(levance)	0.3424	0.2053

Results on Community QA

- Automatic Evaluation – JS divergence
- Test on different component of the objective function

	Length=100	Length=200
Rel(levance)	0.3424	0.2053
Rel+Aut(hor)	0.3375	0.2040

Results on Community QA

- Automatic Evaluation – JS divergence
- Test on different component of the objective function

	Length=100	Length=200
Rel(levance)	0.3424	0.2053
Rel+Aut(hor)	0.3375	0.2040
Rel+Aut+TM (Topic Model)	0.3366	0.2033

Results on Community QA

- Automatic Evaluation – JS divergence
- Test on different component of the objective function

	Length=100	Length=200
Rel(levance)	0.3424	0.2053
Rel+Aut(hor)	0.3375	0.2040
Rel+Aut+TM (Topic Model)	0.3366	0.2033
Rel+Aut+TM+Pol(arity)	0.3309	0.1983

Results on Community QA

- Automatic Evaluation – JS divergence
- Test on different component of the objective function

	Length=100	Length=200
Rel(evance)	0.3424	0.2053
Rel+Aut(hor)	0.3375	0.2040
Rel+Aut+TM (Topic Model)	0.3366	0.2033
Rel+Aut+TM+Pol(arity)	0.3309	0.1983
Rel+Aut+TM+Pol+Cont(ent)	0.3102	0.1851

Results on Community QA

- Automatic Evaluation – JS divergence
- Test on different component of the objective function

	Length=100	Length=200
Rel(evance)	0.3424	0.2053
Rel+Aut(hor)	0.3375	0.2040
Rel+Aut+TM (Topic Model)	0.3366	0.2033
Rel+Aut+TM+Pol(arity)	0.3309	0.1983
Rel+Aut+TM+Pol+Cont(ent)	0.3102	0.1851
Rel+Aut+TM+Pol+Cont+Disp(ersion)	0.3017	0.1758

Results on Community QA

- Human Evaluation
 - 100 questions on Amazon Mechanical Turk, and each question was evaluated by four Turkers
 - Turkers are asked to provide two rankings based on
 - Overall quality (informativeness and non-redundancy, other desiderata can be considered)
 - Information diversity
 - Turkers are encouraged, but not required, to write down the reasons for their rankings.

Results on Community QA

- Human Evaluation
 - Summaries from different systems are in random order
 - Best answer (~100 words)
 - Dasgupta et al. (2013) (100 words)
 - Our system (100 words and 200 words)
 - One noisy summary (irrelevant to the question)
 - Two duplicate questions are added to test intra-annotator agreement.
- We reject HITs if:
 - the noisy summary is ranked higher than other summaries, or
 - inconsistent ranking are submitted for both of the duplicate questions


Results on Community QA

- Human Evaluation
 - Inter-annotator agreement (Fleiss' kappa):
 - Overall quality: 0.28 (fair agreement)
 - Information diversity: 0.43 (moderate agreement)

	Length	Overall Quality		Information Diversity	
		Ranked as Best Summary	Average Rank	Ranked as Best Summary	Average Rank
Best answer	102.7	31.9%	2.69	9.6%	3.29
Dasgupta et al. (2013)	100	11.0%	2.83	5.0%	2.94
Our system	100	12.5%	2.50	6.7%	2.43
Our system	200	44.6%	1.98	78.7%	1.34

Results on Community QA

- Human Evaluation
 - Inter-annotator agreement (Fleiss' kappa):
 - Overall quality: 0.28 (fair agreement)
 - Information diversity: 0.43 (moderate agreement)

	Length	Overall Quality		Information Diversity	
		Ranked as Best Summary	Average Rank	Ranked as Best Summary	Average Rank
Best answer	102.7	31.9%	2.69	9.6%	3.29
Dasgupta et al. (2013)	100	11.0%	2.83	5.0%	2.94
Our system	100	12.5%	2.50	6.7%	2.43
 Our system	200	44.6%	1.98	78.7%	1.34

Sample Summary

- Question: What is the long term effect of piracy on the music and film industry?
- Our system summary:
 - Rising costs for movies and music. The other thing will be music and movies with less quality.
 - Now, with piracy, there isn't the willingness to take chances. American Idol is the result of this. The real problem here is that the mainstream music will become even tighter. Record labels will not won't to go far from what is currently like by the majority.
 - I hate when people who have billions of dollars whine about not having more money. But it's also like the person put the effort into it and they aren't getting paid. I don't see anything wrong with burning a mix cd or a cd for a friend.
 - It is forcing them to rework their business model, which is a good thing.
 - By removing the profitability of music & film companies, piracy takes away their motivation to produce new.

Results on Blogs

- Comparisons:
 - Best system in TAC 2008
 - Ranker (ListNet) trained on Yahoo! Answers
 - Lin and Bilmes (2011)
 - Dasgupta et al. (2013)

Results on Blogs

- Automatic evaluation – ROUGE and JS divergence

	ROUGE-2	JS Divergence
Best system in TAC'08	0.2923	0.3286
Ranker (ListNet)	0.3200	0.2293
Lin and Bilmes (2011)	0.2732	0.2330
Lin and Bilmes (2011) + query	0.2732	0.2349
Dasgupta et al. (2013)	0.2618	0.2370
Our system	0.3234	0.2258

Results on Blogs

- Automatic evaluation – ROUGE and JS divergence

	ROUGE-2	JS Divergence
Best system in TAC'08	0.2923	0.3286
Ranker (ListNet)	0.3200	0.2293
Lin and Bilmes (2011)	0.2732	0.2330
Lin and Bilmes (2011) + query	0.2732	0.2349
Dasgupta et al. (2013)	0.2618	0.2370
Our system	0.3234	0.2258




Results on Blogs

- Human Evaluation:
 - Inter-annotator agreement (Cohen's kappa): 0.68 (substantial)

	Pyramid F-score
Best system in TAC'08	0.2225
Lin and Bilmes (2011)	0.2790
Our system	0.3620

Results on Blogs

- Human Evaluation:
 - Inter-annotator agreement (Cohen's kappa): 0.68 (substantial)

	Pyramid F-score
Best system in TAC'08	0.2225
Lin and Bilmes (2011)	0.2790
 Our system	0.3620

The Choice of Text Similarity and Dispersion Function

- Which text similarity metric performs better?
- Which dispersion function performs better?

Yahoo! Answers		
	Dispersion (SUM)	Dispersion (MIN)
Semantic similarity	0.3143	0.3129
Topical similarity	0.3101	0.3106
Lexical similarity	0.3017	0.3071



Measured in JS divergence

The Choice of Text Similarity and Dispersion Function

- Which text similarity metric performs better?
- Which dispersion function performs better?

Blogs		
	Dispersion (SUM)	Dispersion (MIN)
Semantic similarity	0.2216	0.2772
Topical similarity	0.2128	0.3234
Lexical similarity	0.2167	0.3117

Measured in ROUGE-2 scores

Conclusion

- We have presented a submodular function-based opinion summarization framework.
- Our approach outperforms state-of-the-art methods that are also based on submodularity in community QA and blogs opinion summarization.
- We have shown that our framework is able to statistically learn sentence relevance and encouraging the summary to cover diverse topics.
- We also study the effect of different text similarity metrics on submodularity-based summarization.

Thank you!