Social Context Summarization

Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, Juanzi Li

Tsinghua University
IBM, China Research Lab

July 25, 2011, SIGIR 2011, Beijing
Outline

- Motivation
- Prior Work
- Social Context Summarization
- Solving via Dual Wing Factor Graph Model
- Experimental Result and Case Study
- Conclusion and Future Work
Motivation

• Document summarization
Motivation

• Context and social context
Motivation

• How do (social) contexts help summarization?

Textual information
  • Integrate context information to help estimate the informativeness of sentences

Social influence
  • If a user is an opinion leader, his/her comments should be more important than others.

Information propagation
  • If a comment has been forwarded or replied by many other users, the comment should be more important than others.
Challenges

How to formally **define** the social context?

- To incorporate users, user generated contents, social networks, and implicit networks (such as the forward/reply network).

How to **formalize** the problem in a principled framework?

- To capture the dependency between the social context information and the quality of summaries.

How to **validate** on a real-world data set?

- The social context contains inevitable noise.
Outline

• Motivation

• Prior Work

• Social Context Summarization

• Solving via Dual Wing Factor Graph Model

• Experimental Result and Case Study

• Conclusion and Future Work
Prior Work

- Document summarization w/o context

**Document**

**Summary**

**Sentence Scoring**

Linguistic features
- *rhetorical structure, lexical chains*

Statistical features
- *term significance, sentences similarity*

**Sentence Selection**

Unsupervised approaches

- Binary classification: *Naïve Bayes* [Kupiec:95], *Maximum Entropy* [Osborne:02], *Genetic Algorithm* [Yeh:05]

Supervised approaches

- Sequence labeling: *HMM* [Conroy:01], *CRF* [Shen:07]
Prior Work

• Binary classification vs. sequence labeling
  – Example: document $d$ with sentence set $S_d = \{s_1, s_2, s_3, s_4\}$ associated with $Y_d = \{y_1, y_2, y_3, y_4\}$.

\[
p(x) = \prod_{i=1}^{4} f_i
\]

\[
p(x) = \prod_{i=1}^{4} f_i \prod_{i=1}^{3} g_{i,i+1}
\]
Prior Work

• Binary classification vs. sequence labeling
  – Example: document $d$ with sentence set $S_d = \{s_1, s_2, s_3, s_4\}$ associated with $Y_d = \{y_1, y_2, y_3, y_4\}$.

Factor graph representation

**Nodes**
- represent variables or factors

**Edges**
- associate factors with variables

$$p(x) = \prod_{i=1}^{4} f_i$$

$$p(x) = \prod_{i=1}^{4} f_i \prod_{i=1}^{3} g_{i,i+1}$$
Prior Work

• Document summarization with context
  – Contents from external documents [Wan:08], cited articles [Mei:08], documents selected by users [Mani:98]
  – Text surrounding the hyperlink [Amitay:00, Delort:03]
  – Search engine click-through data [Sun:05]
  – Commented sentence selection [Hu:08] or opinionated text [Ganesan:10, Paul:10]
Prior Work

LIMITATION

Lacks of a unified approach
✓ To explore the non-textual information from social context
✓ To take full advantage of conventional techniques in summarization

– Commented sentence selection [Hu:08] or opinionated text [Ganesan:10, Paul:10]
Outline

• Motivation
• Prior Work
• Social Context Summarization
• Solving via Dual Wing Factor Graph Model
• Experimental Result and Case Study
• Conclusion and Future Work
Social Context Summarization

- Social Context and SCAN (in Twitter scenario)

Web document $d$

- Associated user set $U_d = \{u_i\}$
- Associated tweet set $M_d = \{m_i\}$

User relations
- $E_{du} = \{(u_i, u_j)\}$

Tweet relations
- $E_{dm} = \{(m_i, m_j)\}$

Sentence relations
- $E_{ds} = \{(s_i, s_j)\}$

Social context of $d$
- $C_d = <M_d, U_d>$

Social Context Augmented Network (SCAN)
- $G_d = (S_d, C_d, E_d)$
- $E_d = E_{ds} \cup E_{dm} \cup E_{du}$
Social Context Summarization

- **SCAN**
  - Three-layer perspective

- **Summarization objective**
  A summary consists of

  - **Subproblem 1:** Key Sentence Extraction
    - Important tweets $M_d^*$
    - Important sentences $S_d^*$

  - **Subproblem 2:** Tweet Summarization

A summary consists of important tweets $M_d^*$ and important sentences $S_d^*$. The process involves extracting key sentences and then summarizing tweets.
QUESTION

How to capture the mutual reinforcement between the two subproblems to facilitate generating a HQ summary?
Outline

• Motivation
• Prior Work
• Social Context Summarization
• Solving via Dual Wing Factor Graph Model
• Experimental Result and Case Study
• Conclusion and Future Work
Solving via Dual Wing Factor Graph Model

• Overview
  - Supervised framework

Take advantage of conventional techniques in summarization

Explore the non-textual information from social context

- Overview
  - Supervised framework

- Supervised framework
  - Take advantage of conventional techniques in summarization
  - Explore the non-textual information from social context

Reinforce!
Solving via Dual Wing Factor Graph Model

• Modeling

Local attribute factor

Sentence dependency factor

Tweet dependency factor
Solving via Dual Wing Factor Graph Model

- **Modeling**

  - **Local attribute factor**
  - **Sentence dependency factor**
  - **Tweet dependency factor**

- **Cross-domain dependency factor**
Solving via Dual Wing Factor Graph Model

- Modeling

How to define these factors?
Solving via Dual Wing Factor Graph Model

• Modeling
  – Local attribute factor
    • Maximum entropy \( f_{i,c}(\lambda_c, y_i) = \exp(\lambda_c x_{i,c} y_i) \)
  – Dependency factor
    • Binary value \( g_{ij,c}(\mu_c, y_i, y_j) = \begin{cases} e^{\mu_c} & \text{if some condition holds} \\ 1 & \text{otherwise} \end{cases} \)

What condition?

Sentence dependency

Tweet dependency

Cross-domain dependency

\( y_i \neq 1 \) or \( y_j \neq 1 \)

\( y_i \leq y_j \)

\( y_i = y_j \)

Only for similarity > \( \theta_g \) or \( \theta_h \)
Solving via Dual Wing Factor Graph Model

- **Modeling**
  - Objective function
    \[
    \max_{\Theta} \frac{1}{Z} \prod_{i,j \in T} \prod_{c \in C} f_{i,c}(\lambda_c, y_i) \cdot g_{ij,c}(\mu_c, y_i, y_j) \cdot h_{ij}(\nu, y_i, y_j)
    \]

- **Parameter estimation**
  - L-BFGS

- **Inference algorithm**
  - Loopy belief propagation
  - Could be distributed easily!
Outline

• Motivation
• Prior Work
• Social Context Summarization
• Solving via Dual Wing Factor Graph Model
  • Experimental Result and Case Study
• Conclusion and Future Work
Experimental Result and Case Study

- Data preparation
  - 4,874,389 users
  - 404,544,462 tweets
  - 200,000 most frequent URLs
  - 12,964,166 tweets

- Distribution of frequency
- Evaluate on only HQ websites
  - CNN, BBC, MTV, ESPN, Mashable.
Experimental Result and Case Study

- **Annotation method**
  - Manual annotation on [Amazon Mechanical Turk](http://www.amazon.com/mechanicalturk)
  - 1,145 HITs (Human Intelligent Tasks) from 158 different users

- **Evaluation methods**
  - F1 and ROUGE-1,2

- **Baseline methods**
  - Supervised: SVM, logistic regression (LR), CRF, SVM+, LR+ (include neighbors’ features)
  - Unsupervised: random, Doc/ParaLead, PageRank
Experimental Result and Case Study

• Feature
  – 6 basic features for sentences
  – 5 basic features for tweets
    • author’s PageRank (influence from user network)

• Result
  
  Improved with social context
  Not improved
Experimental Result and Case Study

• Analysis on results
  – Why not significantly improved on Tweet Summarization?
  – Significantly improved on the MTV and ESPN domains
    • Fewer documents and corresponding tweets
Experimental Result and Case Study

- Analysis on results
  - Impact of thresholds $\theta_g$ and $\theta_h$

Smaller $\theta_g$
More LQ relations

Greater $\theta_g$
Less HQ relations

Smaller $\theta_h$
More LQ relations

Greater $\theta_h$
Less HQ relations
Experimental Result and Case Study

- Analysis on results
  - Factor contribution analysis

Sentence length
Sentence position in paragraph
Tweet length
Averaged TF-IDF
• Case study

High-freq terms in Doc:
- women, dead person, body,
- Liverpool Airport

High-freq term in Tweets:
- Weekend At Bernie’s

"Women try to take body on plane at Liverpool airport"
Outline

• Motivation
• Prior Work
• Social Context Summarization
• Solving via Dual Wing Factor Graph Model
• Experimental Result and Case Study
• Conclusion and Future Work
Conclusion and Future Work

• Conclusion
  – Formally defined the social context that incorporates users, user generated contents, and associated networks.
  – Proposed a Dual Wing Factor Graph Model to capture the dependency among SCANs.
  – Conducted experiments on real world dataset, and the proposed method outperforms the baseline methods.

• Future work
  – Explore friendship among user network more “elegantly”.

Thanks! QA