中国科学技术大学
University of Science and Technology of China

NOKIA
Connecting People

Towards Annotating Media Contents through Social Diffusion Analysis

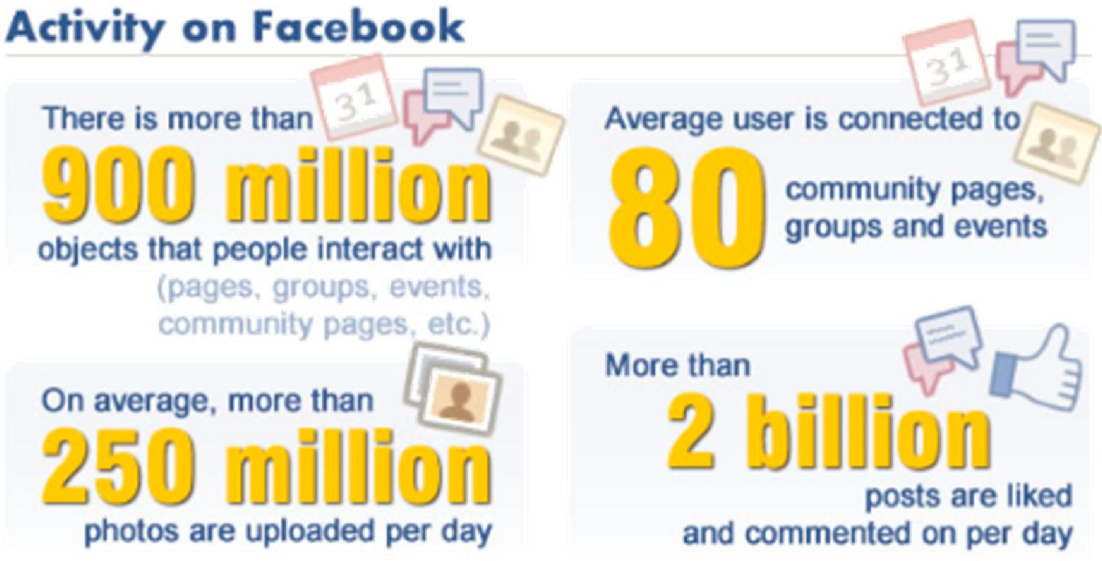
Tong Xu[1], Dong Liu[2], Enhong Chen[1], Happia Cao[2], Jilei Tian[2]

[1]University of Science and Technology of China,

[2]Nokia Research Center

Background

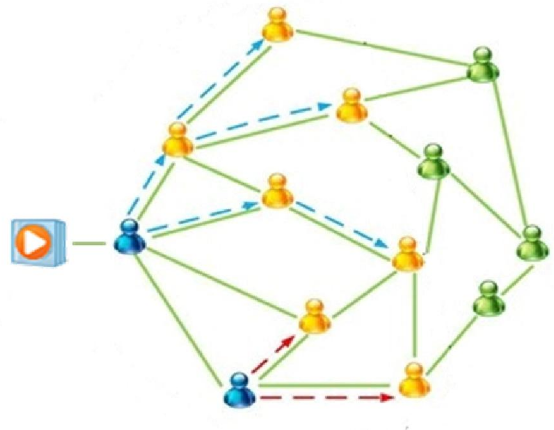
- Boom of media contents, especially the original "*user generated content*" by "*grassroot*" authors.
- Significant challenges to support efficient management and retrieval.



Compiled by Website-Monitoring.com

Background Cont.

- Thus, ***system-level*** annotations are urgently required to generalize media contents.
  - E.g., Nationality, Author, Genre, Topic …
- However, how to annotate?
  - At the same time, we realize that media contents are frequently ***shared*** in online social networks, which result in a lot of ***diffusion records***.



Common Interest (CI) based Diffusion

- Social diffusions might reflect ***common interests***.
  - E.g., big fan of love story may share new Titanic.
  - Prior arts point out that common interests influence both social connection and information flow.
  - Thus, common interests act as *intermediate*.

- Basic Assumption

  Diffusion Records of media contents usually reflect the ***common interests*** *between sharers, as well as the* ***property*** *of shared media.*

CI-based Diffusion — Preliminary

- System-level Annotation Set
  - *A pre-defined annotation set* ***T****, a subset* $\boldsymbol{T_c^*} \subset \boldsymbol{T}$ *will be selected to annotate new media .*
- Common Interest (CI) Factor
  - *A normalized* $|\boldsymbol{T}|$*-dimensional vector* ***c*** *within each pair of sharers.*
- CI-based Diffusion Graph
  - *The CI vectors act as weights in the CI-based diffusion graph, which is denoted as* ***G=<V,E,C>.***
  - ***C*** *presents the set of CI factors (vectors).*

CI-based Diffusion — Diffusion Analysis

- How to analyze diffusion process with ***graph structure?***
  - Traditional social diffusion (influence) model
  - Activating Probability (AP) - Probability of Successful Diffusion
- *CI-based Diffusion Model*
  - To calculate AP based on the correlation between property of media contents and CI factors.
  - Defined as ***Corr(T, C)***, formulated as follow:

$$w_{sr}^{i} = 1 - \prod_{z=1}^{|T|} (1 - c_{sr}^{z} \cdot t_{i}^{z}).$$

CI-based Diffusion — Maximal Likelihood

- At the same time, we extract the ***CI-based Diffusion Graph G_i*** from the diffusion records. With proper annotations, we could "***reproduce***" ***G_i*** through the social diffusion models with ***maximal likelihood***, which is formulated as follow:

$$(T_i^*, C^*) = \arg \max_{(T_i, C)} P(G_i | T_i, C),$$

Optimization Task — Two Targets

- We solve the maximal likelihood problem with ***two optimization targets***, which separately corresponds to the ***training and test*** stages in a typical supervised learning problem. To be specific:
- ***Training Stage***:

  Given labeled samples to learn the common interest factors.

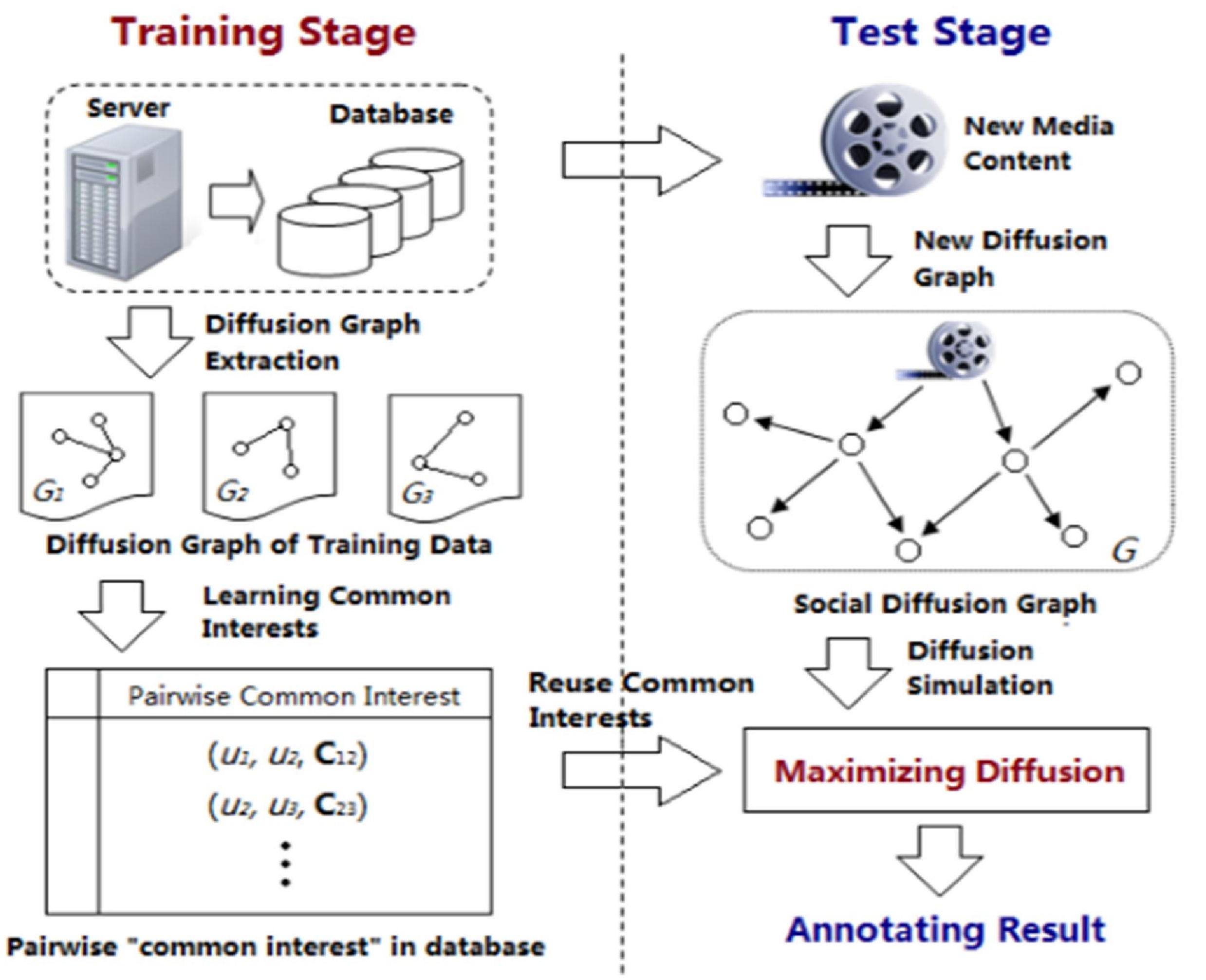
$$C^* = \arg\max_{C} \sum_{i \in I_a} D(G_i, T_i, C|_{E_i}).$$

- ***Test Stage***:

  Given learned CI factors to annotate new media contents.

$$T_i^* = \arg\max_{T_i} D(G_i, T_i, C|_{E_i}), \forall i \in I_u.$$

Optimization Task — Framework



Optimization Task — Training Stage

- The global optimization is a tough task.
  - Millions of edges.
  - Some edges even reappear in majority of samples.
- Trivial Solution
  - Higher AP leads to higher expectation.
  - Maximal Probability, i.e., $\boldsymbol{w_{sr}} = \mathbf{1}$ for all the edges.
- The optimization target could be summarized as:

$$\min \sum_{s,r,i:e_{sr}\in E_i, i\in I_a} \left(\prod_{z=1}^{|T|} (1 - c_{sr}^z \cdot t_z^i)\right),$$

$$s.t. \sum_{z=1}^{|T|} c_{sr}^z = 1.$$

Optimization Task — Test Stage

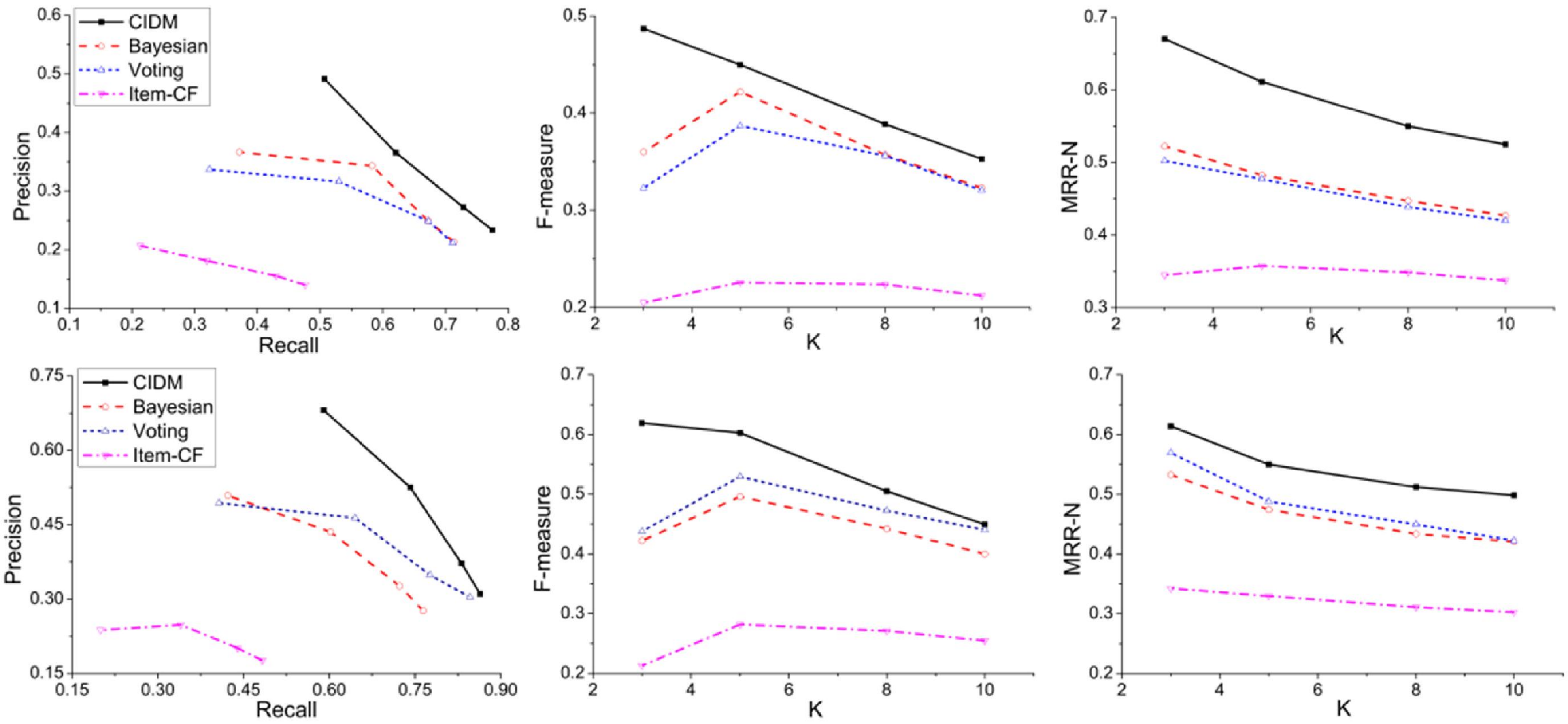
- Maximize the diffusion of test sample with adaptive CI-based diffusion model.
- Candidate is added through *greedy algorithm*.
  - In each round, the annotation with the maximum incremental diffusion will be selected. This process will repeat until enough annotations are selected.
  - Early ones might be more significant for the annotating task.

Experimental Results

- To verify the effectiveness, we perform extensive experiments on two real-world data sets that are extracted from ***Douban.com***.
- The ***voting results*** of ***individual*** viewers and ***pairwise*** sharers are introduced as baselines.
- Standard 5-fold experiments are conducted to sufficiently measure the performance.

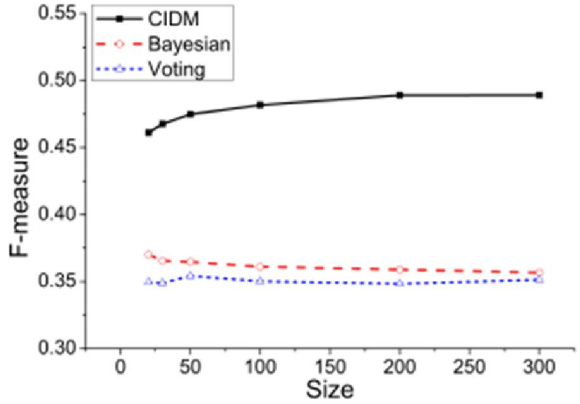
Term	Details	Douban Movie	Douban Book
Item	Total Num.	89,667	475,820
	Selected Num.	2,500	2,500
	Avg. Shared Frequency	2309.79	740.92
User	Total Num.	42,947	42,231
	Avg. Friend Num.	81.91	80.27
	Avg. Share Frequency	134.46	43.86

Experimental Results — Overall Performance

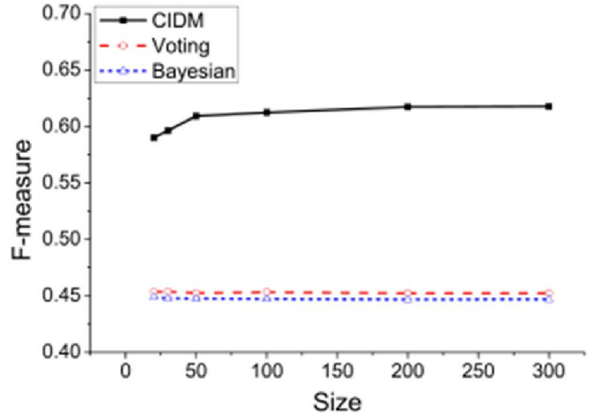


- Our approach consistently outperforms the baselines with a significant margin, especially when *K* is smaller, which indicates that unpopular annotations are successfully mined.

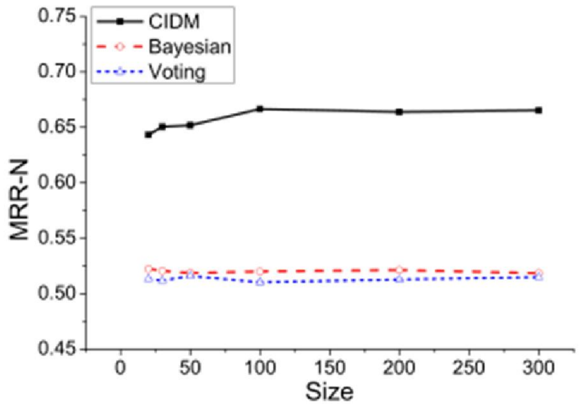
Experimental Results — Cold Start Problem

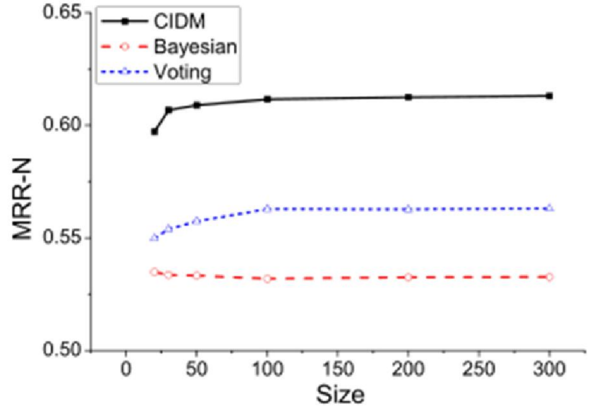


(a)



(a)





- Early viewers may contain more clear preference.
- Early viewers may be more than willing to share.

Conclusion

- A novel framework to annotate through CI-based diffusion analysis.
- ***Graph structure*** plays an important role.
- Administrators of social media should pay more attention to interest-based group, and also the detection of latent community.

Thanks!

tongxu@mail.ustc.edu.cn