A Unified Microblog User Similarity Model for Online Friend Recommendation

Shi Feng, Le Zhang, Daling Wang, Yifei Zhang

Northeastern University

NLPCC 2014

Outline

- Motivation
- The Characteristics of Friend Relationship in Microblogs
- Friend Recommendation by Combining Multiple Measures
- Experiments
- Conclusion and Future Work

Outline

- Motivation
- The Characteristics of Friend Relationship in Microblogs
- Friend Recommendation by Combining Multiple Measures
- Experiments
- Conclusion and Future Work

Motivation

- Real-life friends from school-mates, colleagues, neighbors
- Extend real-life social relations into online virtual social networks
- Microblogging services
  - Weibo, Twitter
  - 536 million registered users
  - More than 100 million tweets are generated per day

Motivation

- Friend relationship in microblog is different
  - The users can follow someone without his or her permissions (more casual)
  - The users may add a friend link to someone because of similar hobbies, tags, locations or hot topics

- Major contributions
  - Find critical features for friend recommendation
  - Propose a similarity model by linearly combining multiple measures
  - Validate the effectiveness of the proposed method on a real-world microblog dataset

Outline

- Motivation
- The Characteristics of Friend Relationship in Microblogs
- Friend Recommendation by Combining Multiple Measures
- Experiments
- Conclusion and Future Work

Crawled Dataset

- Who is the target user's potential good friend in microblogs? It can be determined by many features because the microblog is full of personal and social relation information

Dataset Features	NO. of Features	Percentage of the Whole Dataset
Independent Users	1,459,303	--
Friend Links	3,853,864	--
Users that All Friends in the Dataset	9,646	--
Users with Tags	1,017,443	69.7%
Users with Location Information	1,292,942	88.6%
Users with Check-in Information	457,520	31.4%
Users with Hot topics	537,741	36.8%

Where the user is from

Where the user have been

High coverage

Statistics of the Crawled Dataset

- The average similarity between users of friends and strangers (based on cosine similarity)

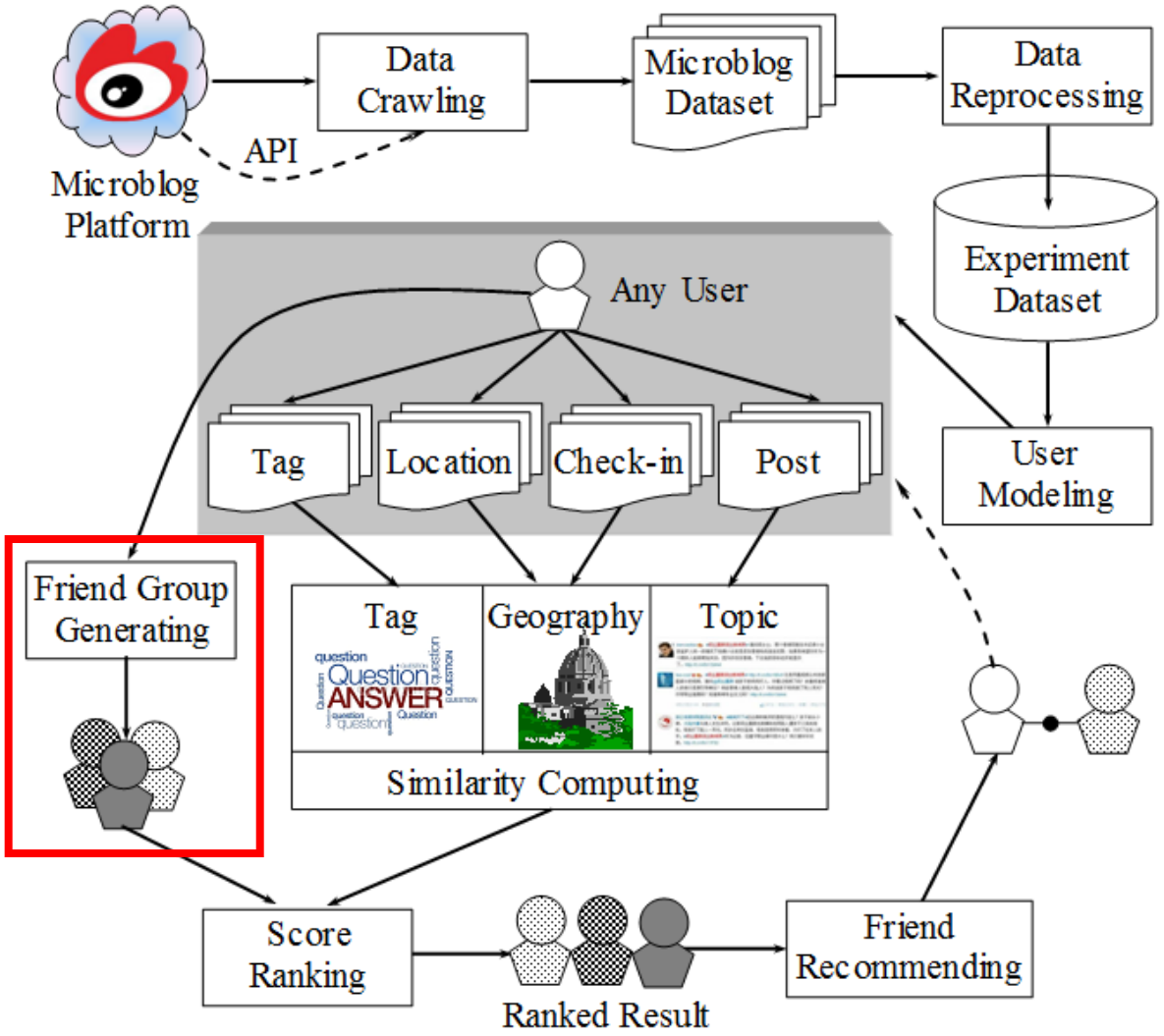
	Tag	Location	Check-in	Hot topic
Friends	4.4×10^{-3}	0.082	0.037	3.0×10^{-4}
Strangers	1.6×10^{-3}	0.038	0.022	2.3×10^{-4}

- These features are good indicators for friend recommendation

Outline

- Motivation
- The Characteristics of Friend Relationship in Microblogs
- Friend Recommendation by Combining Multiple Measures
- Experiments
- Conclusion and Future Work

Overall Framework of the Proposed Model



Candidate Friend Set Generation

- Select users that are friends' friends

$$f_r(u) = \bigcup_{i=1}^{n} f(u_i) - f(u)$$

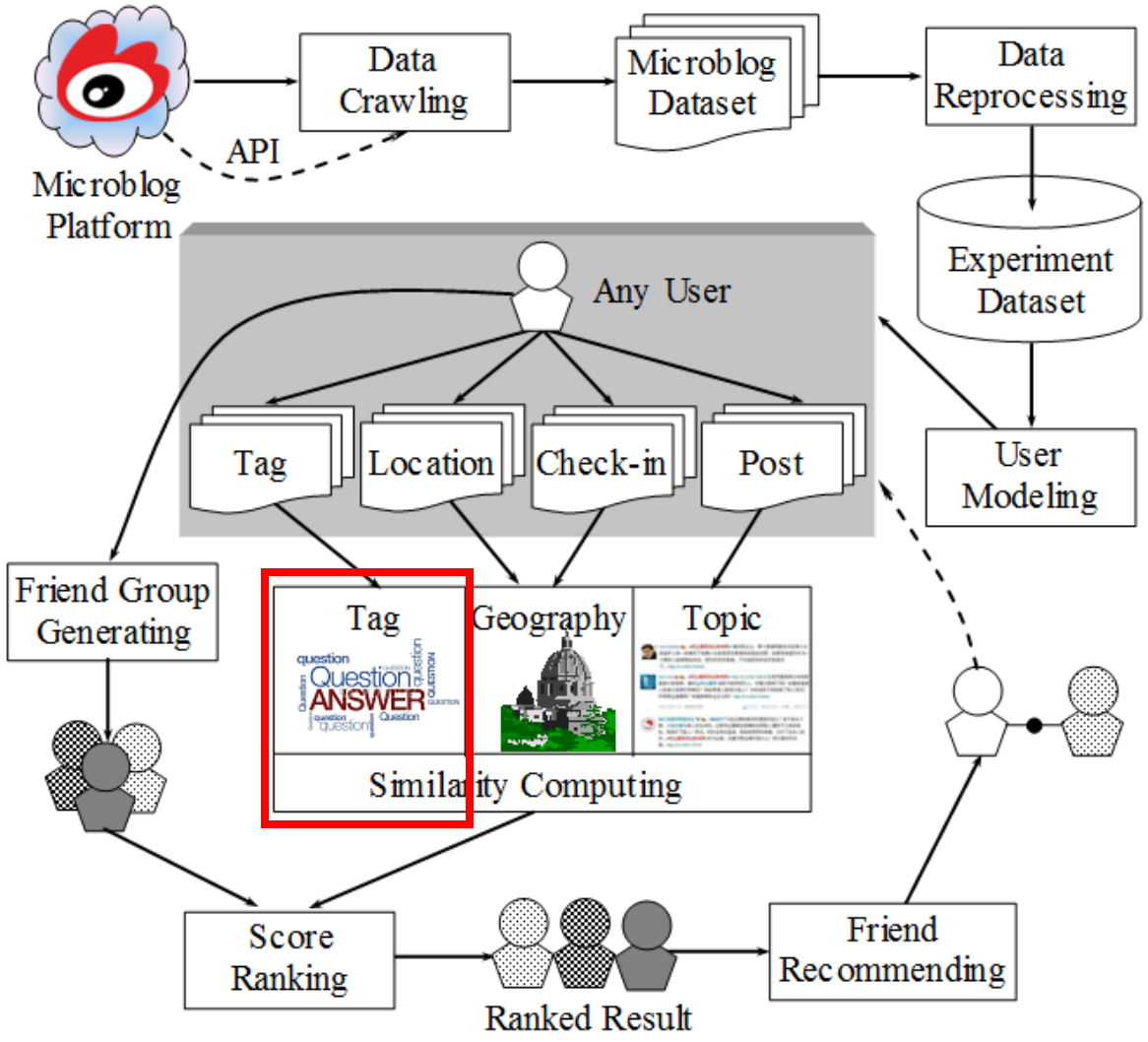
  - Rank the users by their common friends
  - Extract the top k users in $f_r(u)$
- Select the most popular k users in microblogs
  - Assumption: The celebrities are usually good candidate friends
- Combine these two set together to form the final candidate friend set $f_c(u)$ that has $2k$ users

Overall Framework of the Proposed Model

社会计算 信息检索 自然语言处理

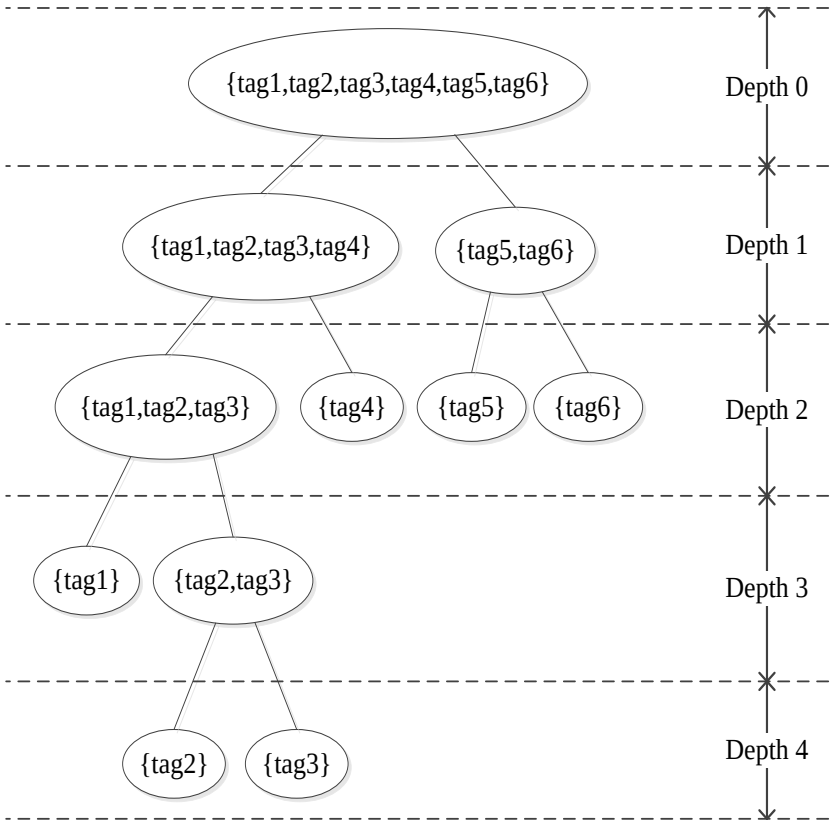
百度 ACL NLP

自然语言处理 搜索



User Tag Similarity

- Challenges
  - User tag vectors are very sparse
  - Many OOV words for WordNet
- Build tag tree
  - hierarchical clustering
  - Recalculate the similarity based on the tree



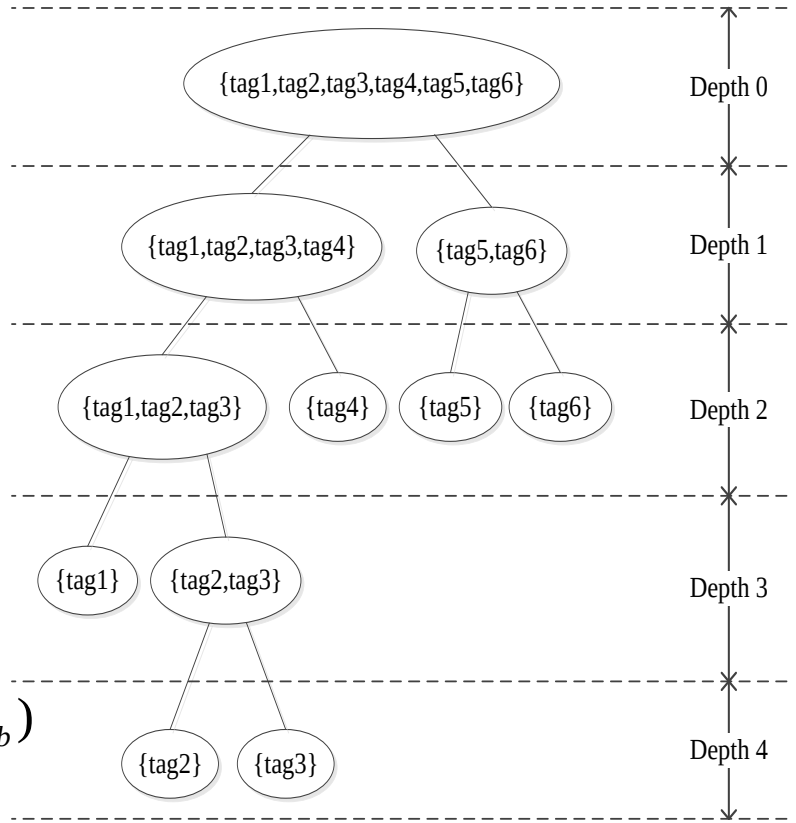
User Tag Similarity

- Similarity between two tags

$$sim_{tt}(t_a, t_b) = \frac{1}{\sum_{k,k+1 \in SP(a,b)} 2/(d(k)+d(k+1))}$$

- Similarity between two users

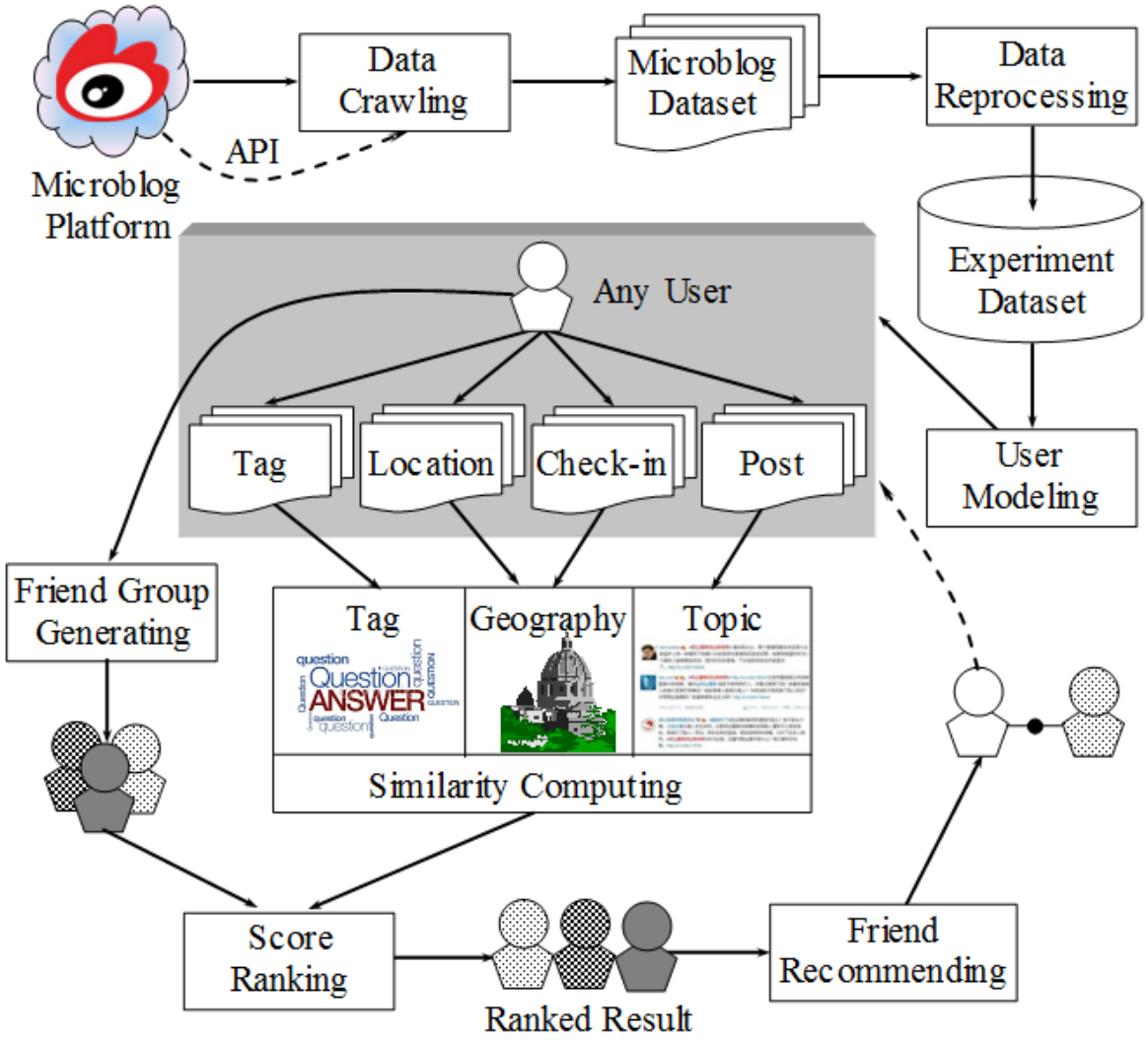
$$sim_{ts}(u_i, u_j) = \frac{1}{|T(u_i)||T(u_j)|} \sum_{t_a \in T(u_i)} \sum_{t_b \in T(u_j)} sim_{tt}(t_a, t_b)$$



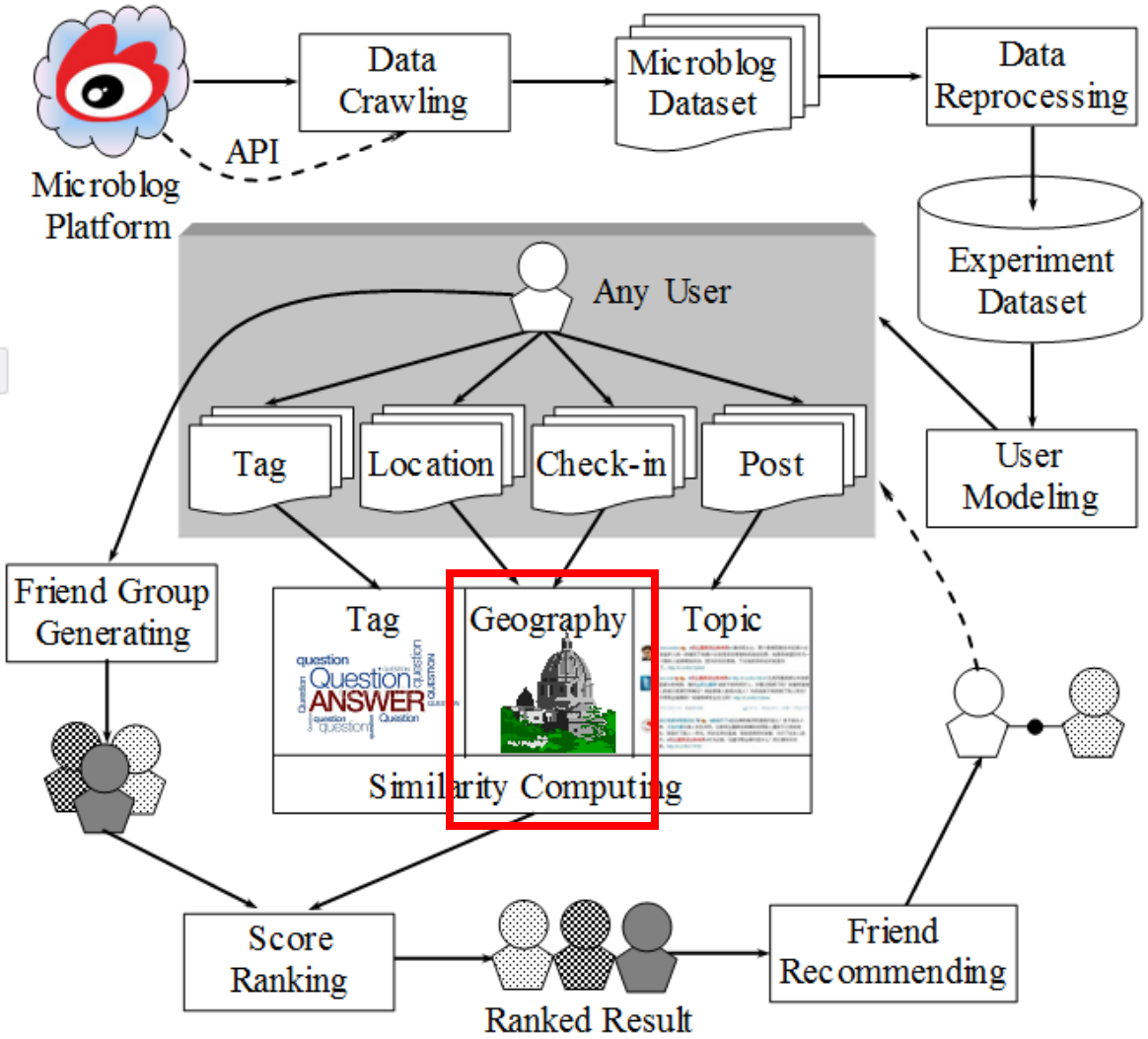
Overall Framework of the Proposed Model

所在地：黑龙江 哈尔滨

所在地：黑龙江 哈尔滨



Overall Framework of the Proposed Model

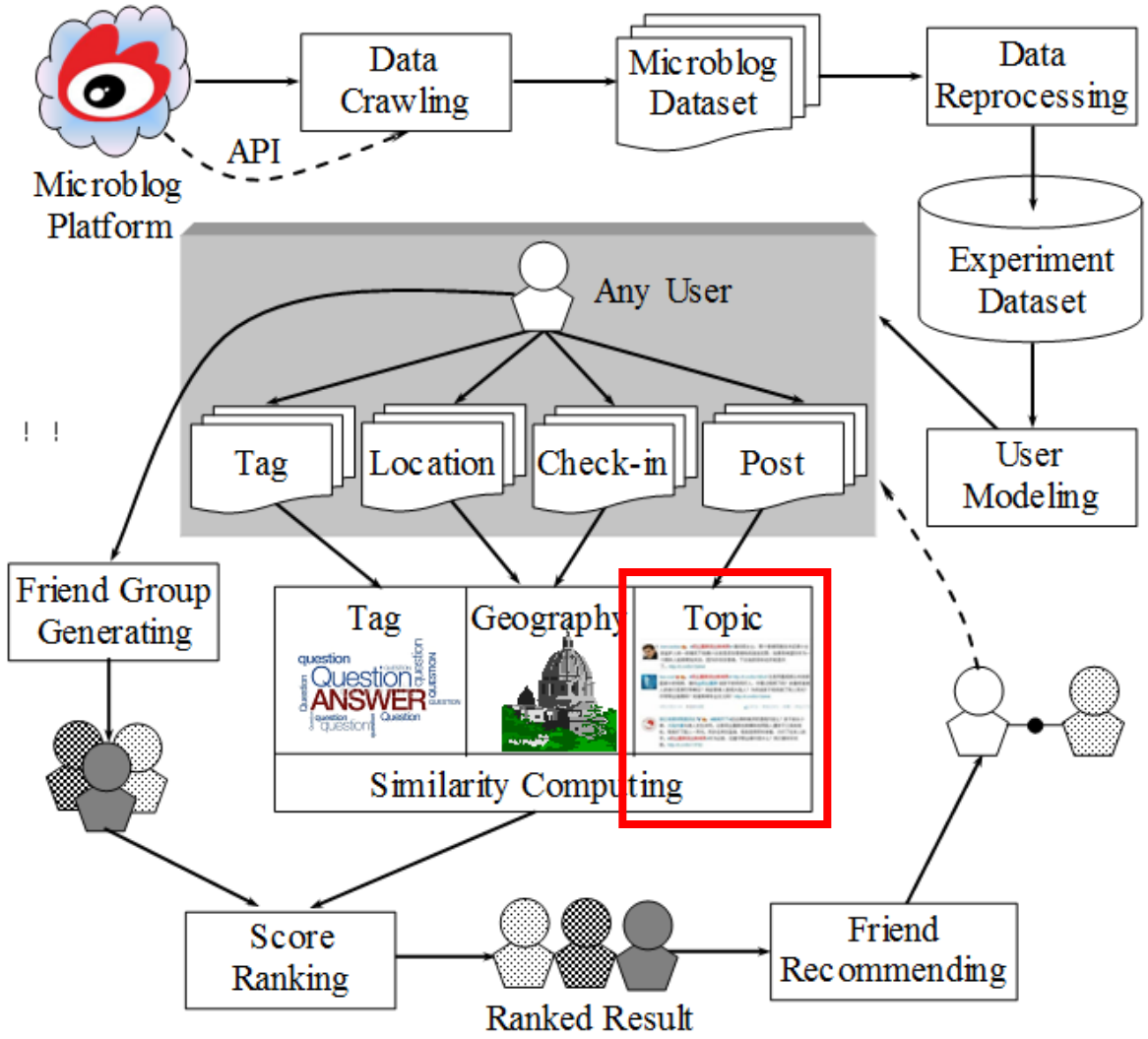


User Geography Similarity

- Location Similarity
  - $sim_{ct}(u_i,u_j)=1$, if the two users have the same location
  - $sim_{ct}(u_i,u_j)=0$, if the two user do not have same location
- Check-in Similarity
  - Divide the check-in information into 12 categories
  - Represent check-in using a vector with 12 dimensions
  - $chk(u)=\{cp_1,cp_2,\ldots,cp_{12}\}$
  - $simchk(u_i,u_j)=cos(chk(u_i),chk(u_j))$
- Geography similarity

$$sim_{loc}(u_i,u_j)=\gamma\cdot sim_{ct}(u_i,u_j)+(1-\gamma)\cdot sim_{chk}(u_i,u_j)$$

Overall Framework of the Proposed Model



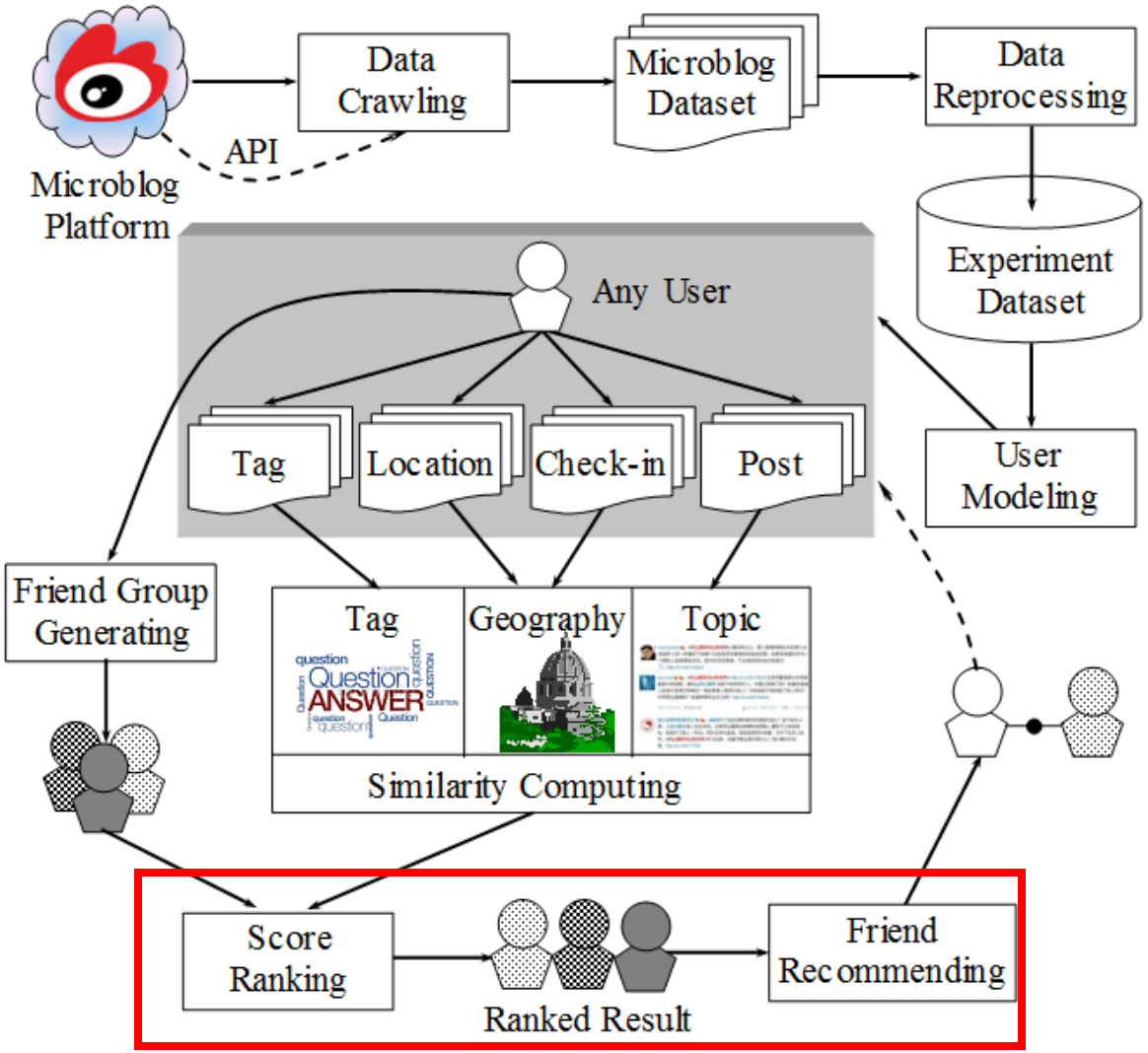
User Hot Topic Similarity

- The hot topic discussion that user takes part in could reflect his/her interests and hobbies

$$sim_{tp}(u_i, u_j) = Jaccard(TP(u_i), TP(u_j)) = \frac{|TP(u_i) \cap TP(u_j)|}{|TP(u_i) \cup TP(u_j)|}$$

- If two users have discussed more hot topics in common, they will have bigger similarity

Overall Framework of the Proposed Model



Unified User Similarity

- Tag, Geography, Hot topic information

$$sim(u,u_i) = \alpha \cdot sim_{ts}(u,u_i) + \beta \cdot sim_{loc}(u,u_i) + (1-\alpha-\beta) \cdot sim_{tp}(u,u_i)$$

- Rank the users in candidate friend set by $sim(u,u_i)$

Outline

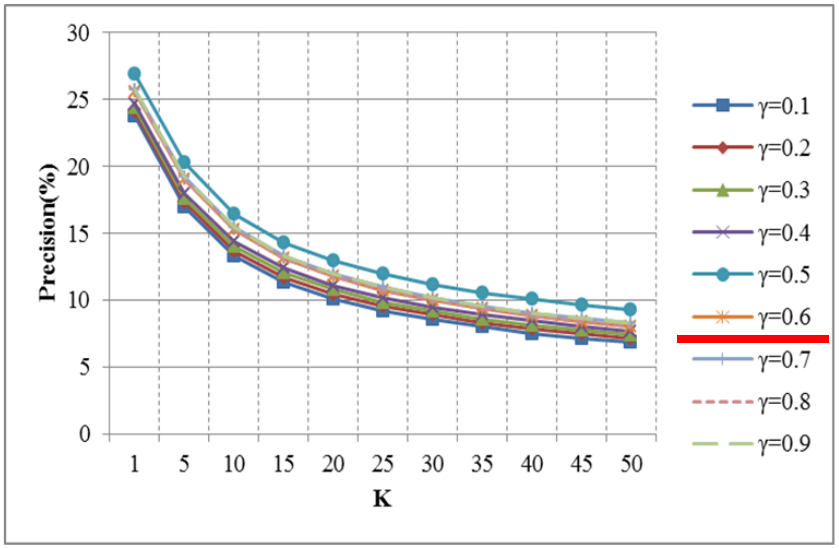
- Motivation
- The Characteristics of Friend Relationship in Microblogs
- Friend Recommendation by Combining Multiple Measures
- Experiments
- Conclusion and Future Work

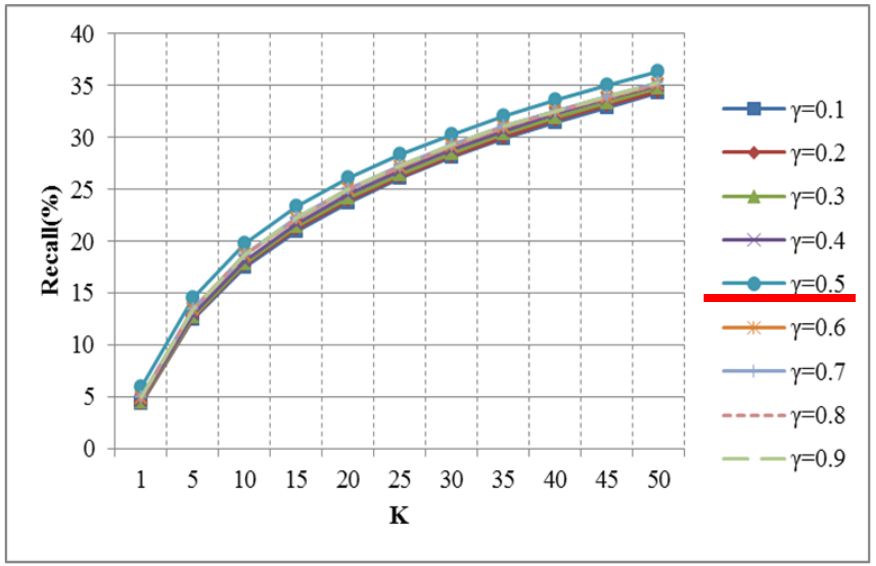
Experiment Setup

- We conduct the 5-fold cross validation on the crawled dataset
- We randomly partition user's current friends and non-friends into 5 groups respectively
- We randomly put one group of friends and one group of non-friends together to form a subset of the crawled data
- For each run, four of the five subsets are used for training and the remaining one subset is used for testing
- Precision, Recall and F-Measure are used for evaluation

Parameter Tuning for Location Similarity

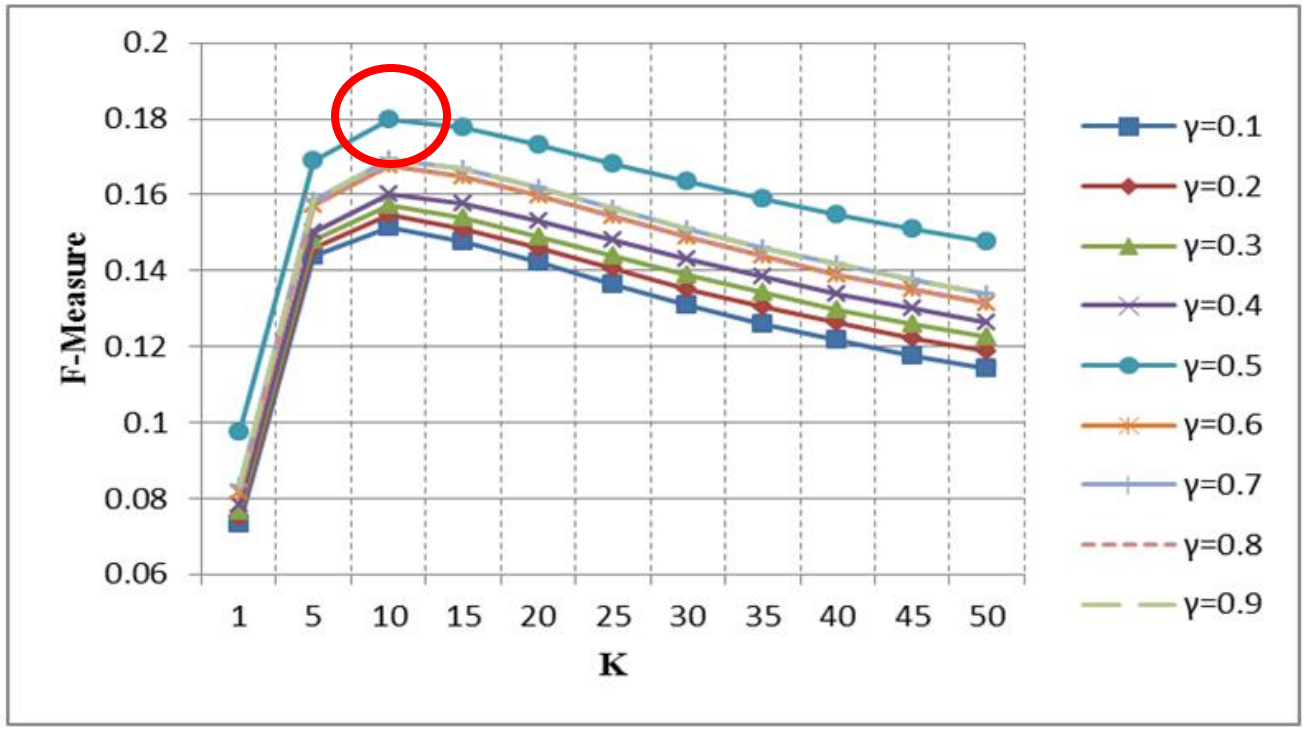
$$sim_{loc}(u_i, u_j) = \gamma \cdot sim_{ct}(u_i, u_j) + (1-\gamma) \cdot sim_{chk}(u_i, u_j)$$





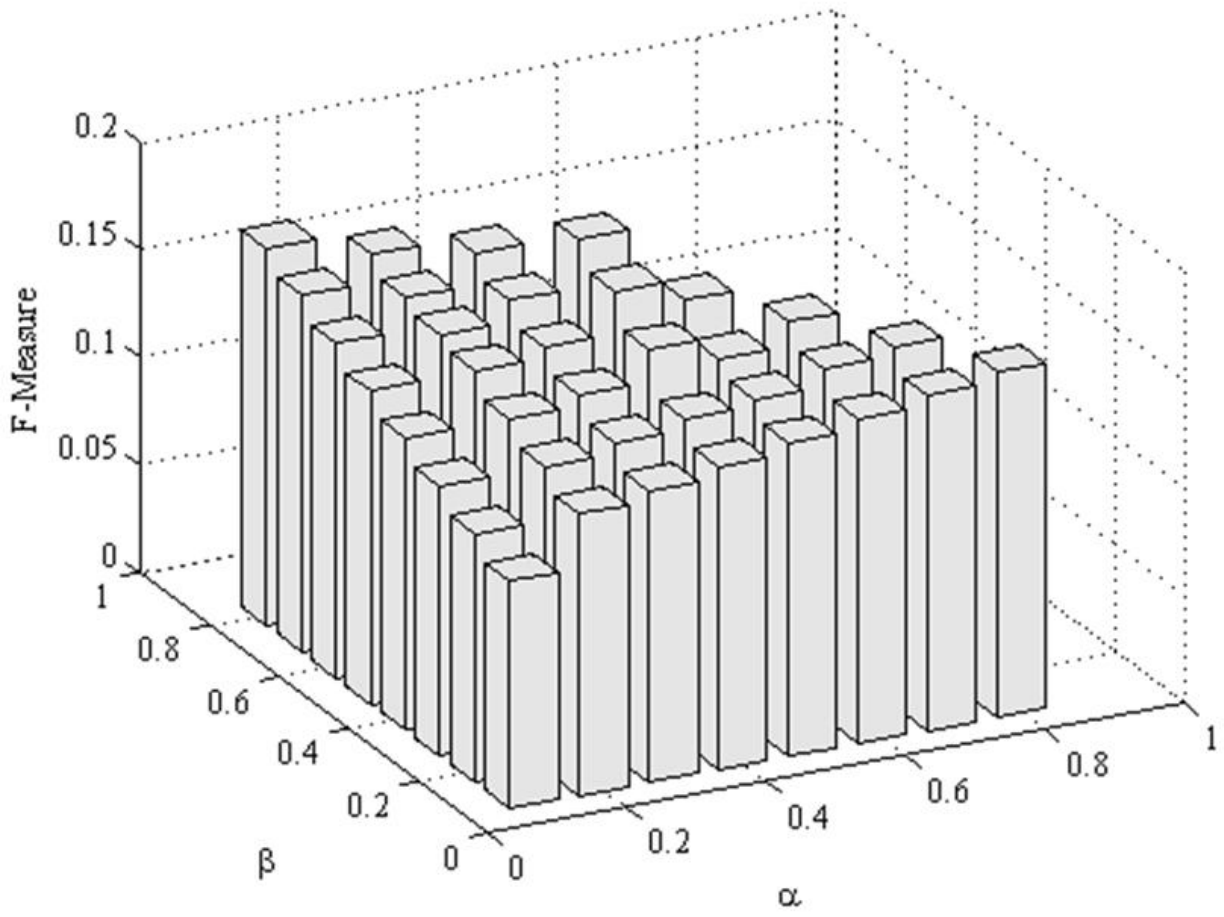
Parameter Tuning for Location Similarity

$$sim_{loc}(u_i, u_j) = \gamma \cdot sim_{ct}(u_i, u_j) + (1-\gamma) \cdot sim_{chk}(u_i, u_j)$$

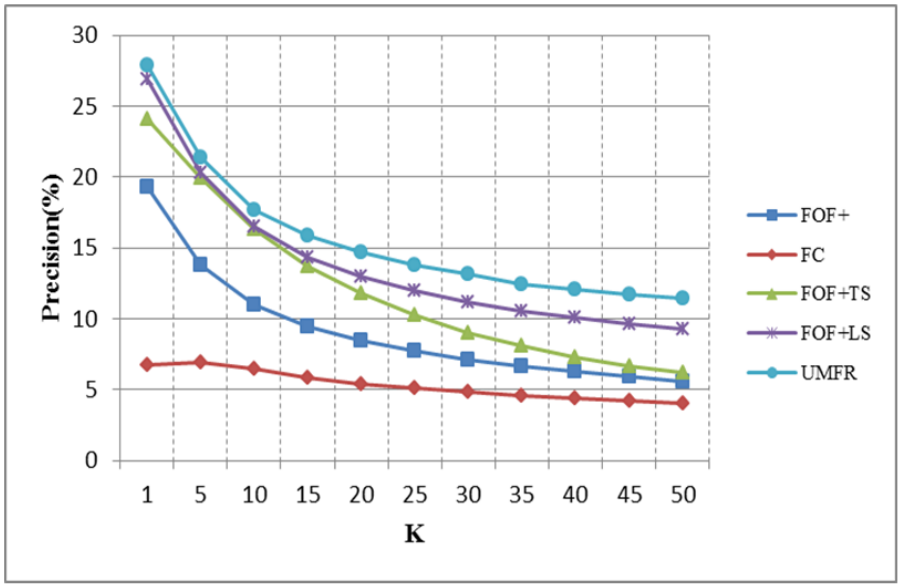


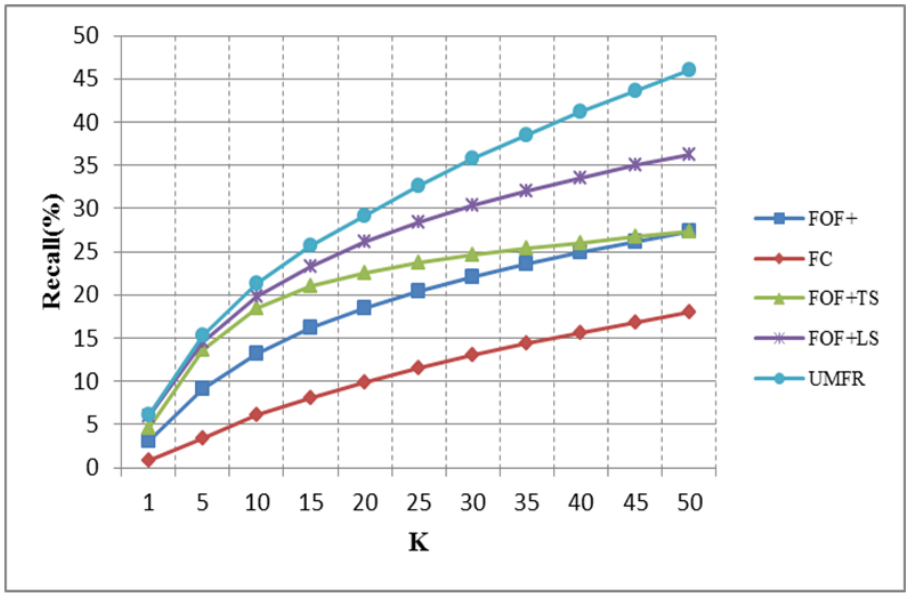
Parameter Tuning for Friend Recommendation

$$sim(u,u_i) = \alpha \cdot sim_{ts}(u,u_i) + \beta \cdot sim_{loc}(u,u_i) + (1-\alpha-\beta) \cdot sim_{tp}(u,u_i)$$

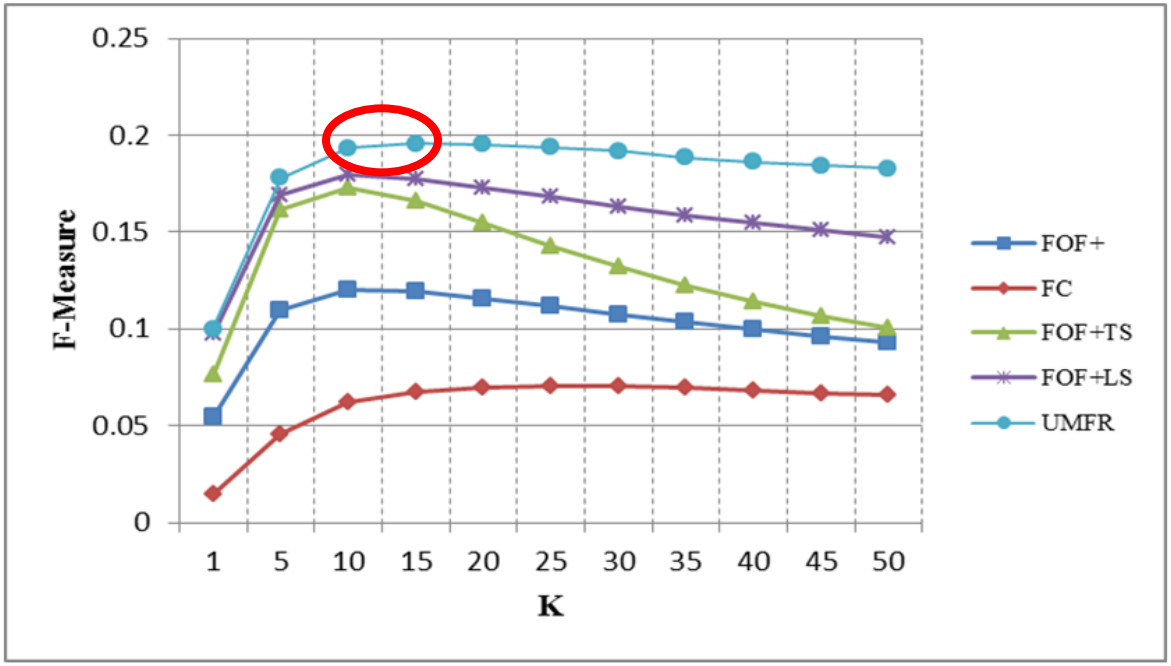


Recommendation Results





Recommendation Results



Outline

- Motivation
- Related Work
- Learning Sentiment Lexicon from Massive Microblog Data
- Sentiment Lexicon Optimization
- Experiments
- Conclusion and Future Work

Conclusion and Future Work

- The friend relationship in microblogs are quite different from other traditional social media
  - More casual/Unidirectional friendship
  - Tags, Locations, Check-ins, Hot topics are good indicators for friend recommendation
- A linearly combination model of multiple measurements are proposed for calculating similarity in microblogs
- Future work
  - More measures, such as time factors
  - New similarity measurement

Thank you for your attention!