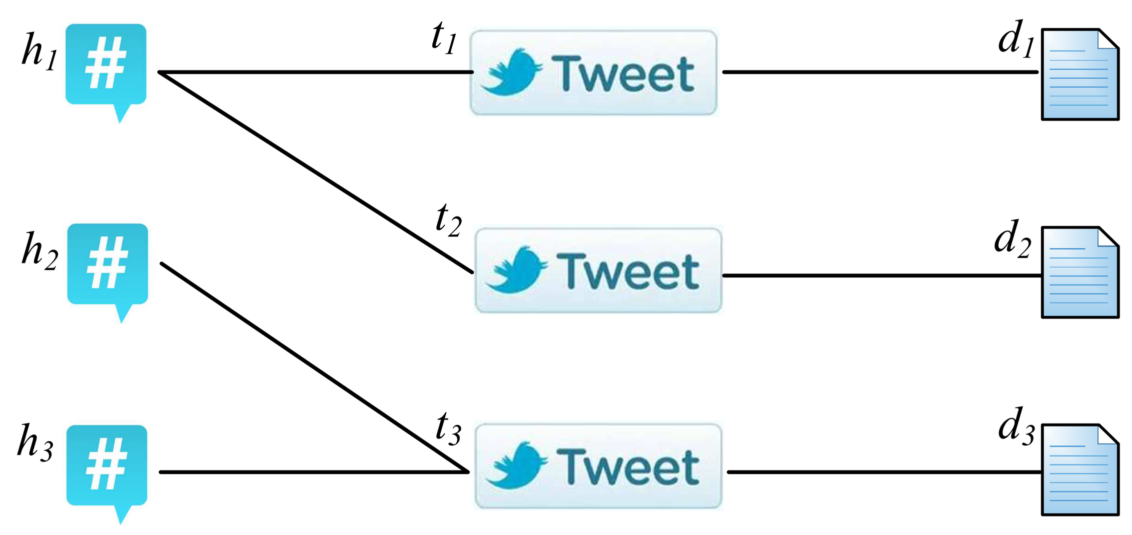
Hashtag Recommendation for Hyperlinked Tweets

Surendra Sedhai **Aixin Sun**
School of Computer Engineering, Nanyang Technological University, Singapore
surendra001@e.ntu.edu.sg axsun@ntu.edu.sg

NANYANG TECHNOLOGICAL UNIVERSITY

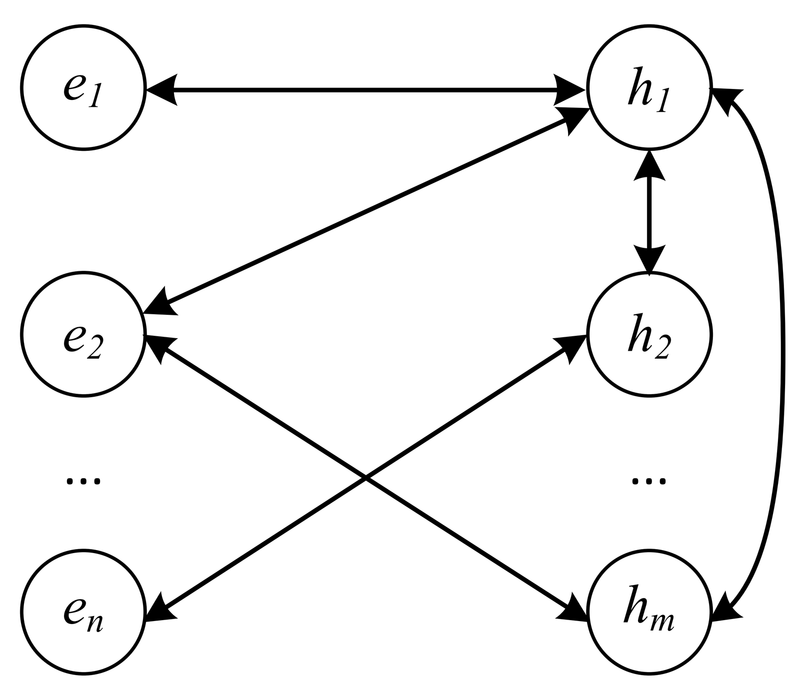
Introduction



- **Hyperlinked tweet**: a tweet containing one or more hyperlinks to external documents.
- **Hashtag recommendation for hyperlinked tweets?**
  - → Presence of hyperlink in a tweet is a strong indication of tweet being more informative.
  - → Functions of hashtags for providing right context to interpret the tweets, tweet categorization, and tweet promotion, can be extended to the linked documents.
- **Recommendation in two phases**
  - → Candidate hashtag selection
  - → Recommendation by learning to rank

Candidate Hashtag Selection

- **Candidate hashtag selection**: selecting a subset of hashtags from all existing hashtags that have been used to annotate any of the observed tweets with or without hyperlinks.
- **Selected through five schemes**:
  - → Top 20 most voted hasthags from the top 50 **most similar tweets**.
  - → Top 20 most voted hasthags from the top 50 **most similar webpages**.
  - → Top 20 most used hashtags for tweets from the **domain of the hyperlink**.
  - → Top 20 highly ranked hashtags based on **named entities** by Random Walk with Restart (RWR) model.
  - → Top 20 highly scored hashtags based on **named entities** by Language Translation (LT) model.
- **Entity-hashtag graph and RWR**



  - → $P(h_j|e_i)$, $P(e_i|h_j)$: the number of times a hashtag h_j is used to annotate a tweet linking to a document containing a named entity e_i, divided by the frequencies of e_i and h_j.
  - → $P(h_k|h_j)$: asymmetric hashtag co-occurrence
- **Language Translation model**: named entities and hashtags as descriptions of the same content in two different languages: $Score(h_j) = \sum_{e_i \in N_e} P(h_j|e_i)$, where N_e is the named entities in the linked webpage of the tweet.

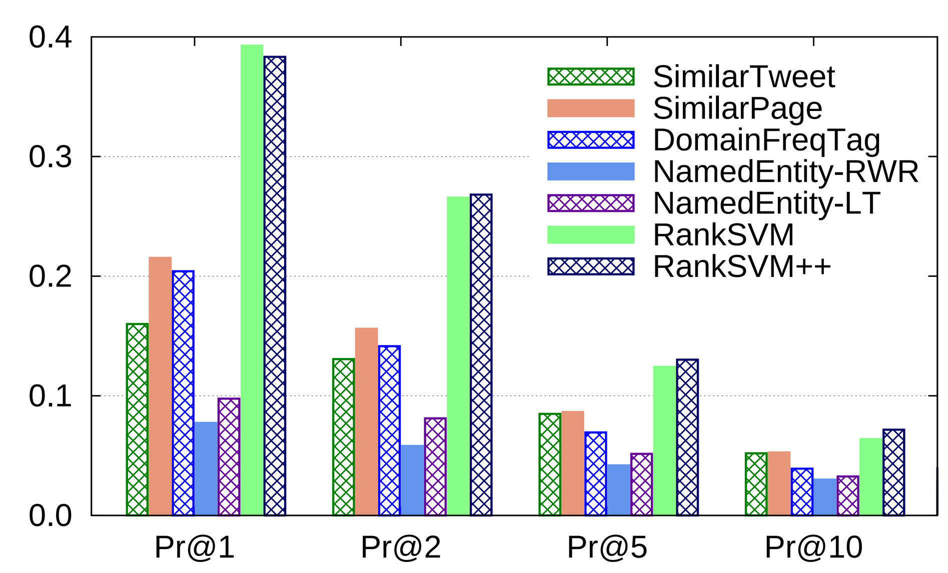
Recommendation by Learning To Rank

- **Pairwise Learning to Rank**:
  - → Learning: Let h_i^+ be a positive candidate hashtag and h^- be a negative candidate hashtag; then the pair $\langle h^+, h^- \rangle$ is a positive instance and $\langle h^-, h^+ \rangle$ is a negative instance in learning the model.
  - → Recommendation: Let H_c be the set of candidate hashtags. The recommendation score of candidate hashtag h_i: $f(h_i) = \sum_{h_j \in H_c, h_i \neq h_j} I(h_i, h_j)$, where $I(h_i, h_j) = 1$ if $\langle h_i, h_j \rangle$ is classified as positive and 0 otherwise.
- **Two sets of features**:
  - → Five binary features: set to 1 if the hashtag is selected by each of the 5 selection schemes.
  - → Four binary features: Wikipedia entry? Top-level category in Yahoo! hierarchy? Popular domain? Hashtag matches webpage domain?

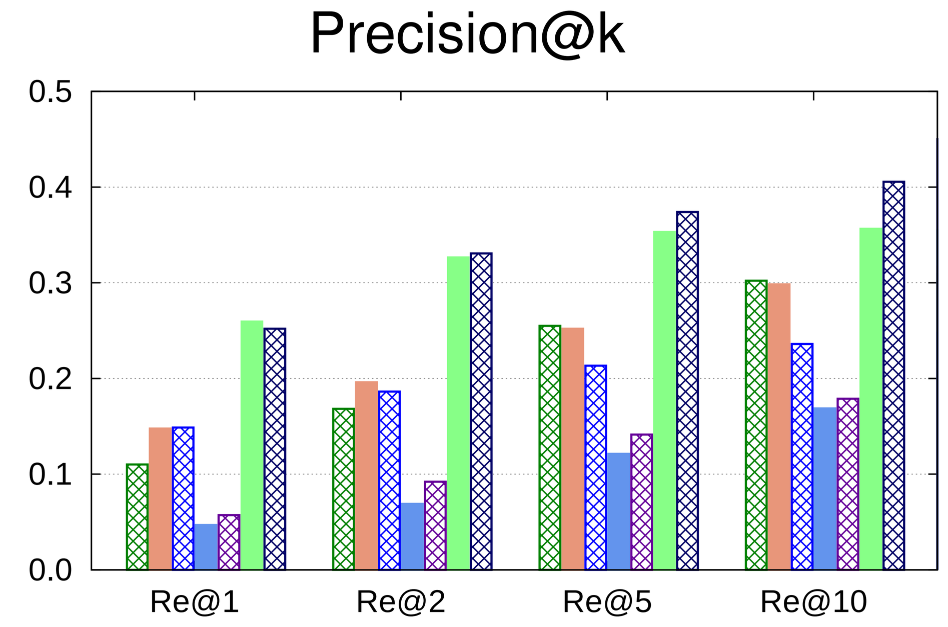
Dataset

- **Data collection:** Two months (May 1 to Jun 30, 2013) of sampled tweets using Twitter streaming API guided by hashtags.org: 24 million tweets published by 11.9 million users, containing 6.9 million links with 3.4 million distinct URLs; 1.37 million downloaded pages are in English.
- **Training and Testing** 15,000 randomly selected hyperlinked tweets from the first 40 days for training. 7,000 hyperlinked tweets from the remaining 20 days for testing.

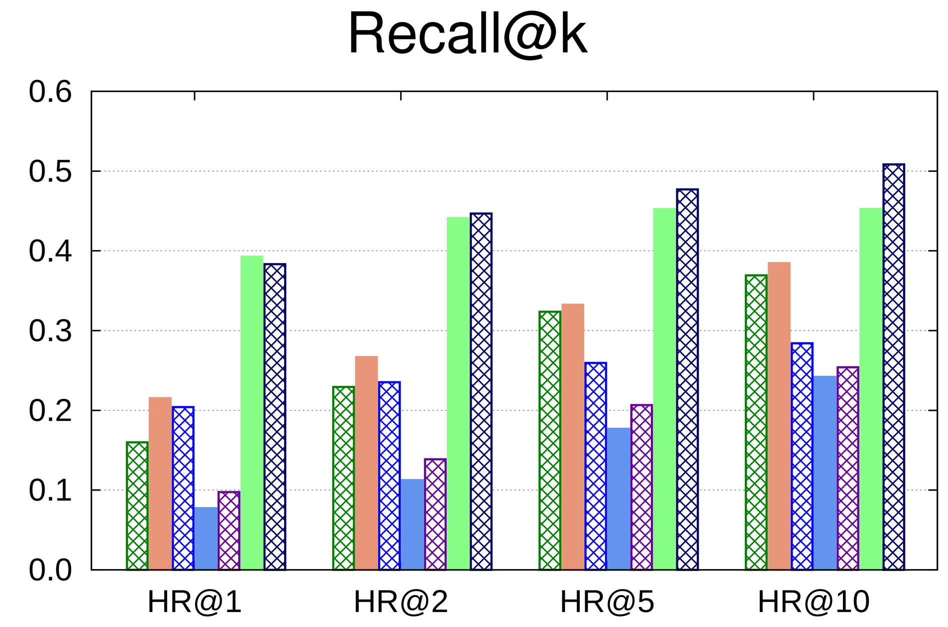
Result



Precision@k



Recall@k



HitRate@k