Interest Analysis using Semantic PageRank and Social Interaction Content

Chung-chi Huang and Lun-wei Ku

Introduction: Motivation (1/2)

- Both content providers and consumers
  - E.g., movie reviews and etc.
- There exists keyword extraction tools to digest information
- Need more
  - Highlighting the words that interest us/catch our eyes

Introduction: Motivation (2/2)

- Keywords != words of interest
  - Interesting words!=keywords
- Keywords: from authors’ perspectives
  - I.e., the statistics of the article content alone
- Words of interest: need to combine readers’ perspectives

Introduction: Purpose (1/2)

- In this paper
  - Predict topic words catching readers’ eyes after article reading
- In prediction
  - Social interaction data of great importance
    - Reader information not public
  - PageRank algorithm used to help
    - Consider semantic features

Introduction: Purpose (2/2)

- These interesting words can be used
  - As social tags
  - In article recommendation
  - In sentiment analysis

Introduction: Example Web Post

The article:

府城.西市場*(traditional market)*謝宅*(the old house)* 歡迎喜愛旅行與體驗生活*(life style)*的好朋友來玩；1905年淺草商場，台南人稱大菜市； 古老的布料行集散地，與迪化街齊名。雖沒落，但 ...
昔日華麗市場*(traditional market)*仍保一絲光采。一群同樣熱愛台南*(the historical city)*老房子*(the old house)*的夥伴，近10個月的懷胎，完成了　這個夢想的空間。陡峭的樓梯，奇妙的格局 ...
□此契約屬於房屋不動產契約，支付的爲房租費用*(rental fees)*，...手繪私房地圖*(exclusive map)*...
讓大家簡單而直接的去體驗與感受屬於原本純粹簡單的美好生活*(life style)* 一棟四十多年的老房子*(the old house)*坐落在台南市*(the historical city)*紛擾喧鬧的市場*(traditional market)*中經歷過近十個月不斷的反覆討論與修正 ... 從此來台南*(the historical city)*晃盪的旅人們可以住在一個像家的地方...
早起喝碗牛肉湯*(bouillon)*吃菜粽帶個營養三明治 中午到市場*(traditional market)*去嚐個虱目魚湯
再轉進這數百年記憶的巷弄間尋找秘密的記憶 台南*(the historical city)*府城.西市場*(traditional market)*謝宅*(the old house)* 有四個樓層 ... 可以基本住四個人 ...

Its social interaction content (i.e., its response posts):

Post 1: 我想要預約12/19~12/20. 人數*(head count)*6~8個左右. 請問:1.還有空房間嗎? 2.費用*(rental fees)*是多少?
Post 2: 我們人數*(head count)*有6人，是一群喜愛老房子*(the old house)*的學生，希望能親身體驗謝宅*(the old house)*的故事。想進一步了解相關資訊與費用*(rental fees)*。
...

Scores of interest preferences for words (w.r.t. the topic of the article):

謝宅*(the old house)*: 0.25, 台南*(the historical city)*: 0.15, 生活*(life style)*: 0.09,
市場*(traditional market)*: 0.05, ... 費用*(rental fees)*: 0.0002, ...

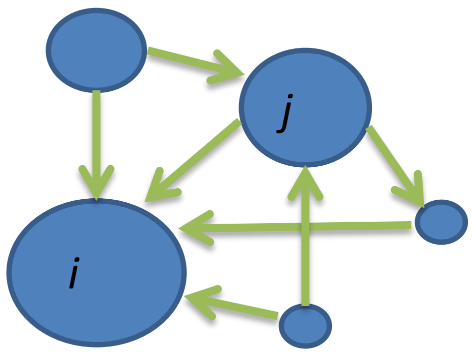
Top-ranked predicted words of interest for future readers:

1. 謝宅*(the old house)* 2. 費用*(rental fees)* 3. 台南*(the historical city)* 4. 市場*(traditional market)* ...

- Keyword extractors find frequent words
- Feedback covers topics of less-frequent/single-occurrence article words
- Combine article with feedback
  - Single-appearance word given more attention

Method: PageRank on Web Pages

- PageRank introduced to find important web pages
  - Nodes: web pages
  - Edges: incoming and outgoing links
  - PageRank iterates to find the probability of a random walker landing on any web page

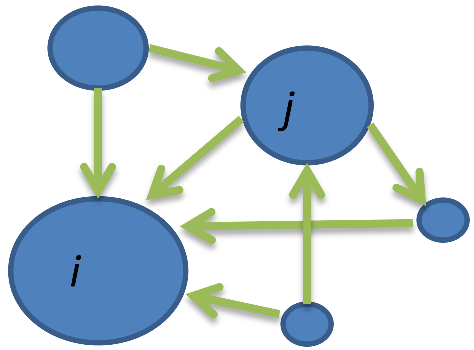


$$\mathrm{PR}(i) = \frac{1-d}{N} + d \times \sum_{j:j\to i} \frac{w(j,i)}{\sum_{k:j\to k} w(j,k)} \mathrm{PR}(j)$$

Method: PageRank in Our Paper (1/5)

- Nodes: words in sentences
- Words within window size have edges
  - Directed from words to words that follow
- Iteration formula
  - $\mathrm{PR}(i) = (1 - d) \times \mathrm{IntPref}(i) +$

$$d \times \sum_{j:j\to i} \frac{w(j,i)}{\sum_{k:j\to k} w(j,k)} \mathrm{PR}(j)$$

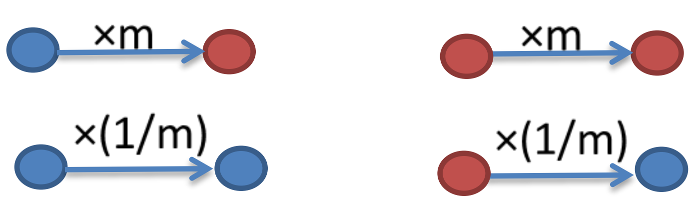


Method: PageRank in Our Paper (2/5)

- Semantic features of word nodes used
  - (1) word group:
    - Intuition: content words ( ) likely to be interests than function words ( )
    - a) slightly content word centered model

    - b) moderately content word centered model

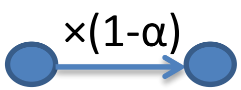
    - c) aggressively content word centered model



Method: PageRank in Our Paper (3/5)

- Semantic features of word nodes used
  - (2) content source of a word pair:
    - Word pairs from articles

    - Word pairs from reader feedback



    - Both authors’ and readers’ voice are heard

Method: PageRank in Our Paper (4/5)

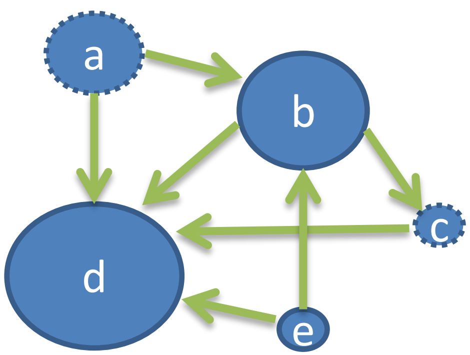
- Semantic features of word nodes used
  - (3) words' degrees of reference:
    - Intuition: highly referenced words among authors and readers likely to be interests
    - A node weighted by 1+DR(the node)
    - DR(the node) defined as

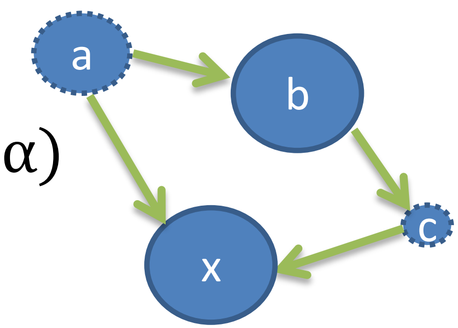
      num(reader response with the node) / num(reader response)
    - Article counted as "a reader response"

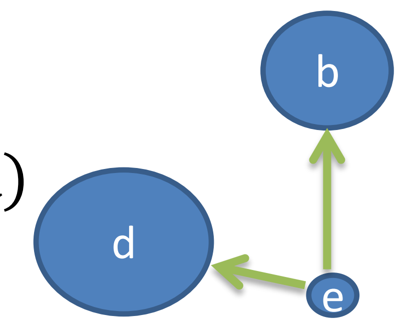
Method: PageRank in Our Paper (5/5)

- Incorporate semantic features into PageRank

$$
\begin{aligned}
\mathrm{PR}(i) = {} & (1-d) \times \mathrm{IntPref}(i) + \\
& d \times \Big\{ \alpha \times \sum\nolimits_{j:j \to i} \frac{w(j,i)}{\sum_{k:j \to k} w(j,k)} \mathrm{PR}(j) \times \big(1 + \mathrm{DR}(i)\big) + \\
& (1-\alpha) \times \sum\nolimits_{j:j \to i} \frac{w(j,i)}{\sum_{k:j \to k} w(j,k)} \mathrm{PR}(j) \times \big(1 + \mathrm{DR}(i)\big) \Big\}
\end{aligned}
$$

- α 

$+(1-\alpha)$ 

$+(1-\alpha)$ 

Method: Interest Preference Model

- Estimate topical interest preference score
- 1. Tfidf(w)
- 2. $\Pr(w|t) = \mathrm{freq}(w,t)/\mathrm{freq}(*,t)$
- 3. $\Pr(t|w) = \mathrm{freq}(w,t)/\mathrm{freq}(w,*)$
- 4. $\mathrm{entropy}(w) = -\sum_{t'} \Pr(t'|w) \times \log(\Pr(t'|w))$
- 5. $\text{Pr-Entropy}(w|t) = \Pr(w|t)/2^{\mathrm{entropy}(w)}$
- 6. $\text{Pr-Entropy}(t|w) = \Pr(t|w)/2^{\mathrm{entropy}(w)}$
- While PageRank uses local info, these use global

Method: Informativity of Reader Feedback

- Not all interaction content responds to the article
  - Check informativity of readers' response sentence and select informative ones
- 1) coverage:
  - Compute ngram coverages
    - To ensure the topic cohesion
  - BLEU: coverages weighted and favor longer ngrams
- 2) focus:
  - The percentage of words certain in topics
    - To have more focused topic

Experiments: Data Sets

- 6,600 articles collected from www.wretch.cc
  - Along with their feedback
- Most of the blog posts in Chinese
  - CKIP segmenter used for segmentation
- 30 articles for testing (avg 17.6 responses)

Experiments: Gold Standards

- Two judges annotated interested words
- To evaluate our system on majority readers
  - Judges related to the responding readers and found their interests in their feedback
  - Only ½ replies responded with reader interest info and they covered one/two topic words in the articles

Evaluation (1/4)

- Top-N nDCG, P, MRR used for evaluation
- Content-word weighting mechanisms

	nDCG	P	MRR
w/o	.778	.397	.728
agr@m=2	.765	.390	.719
agr@m=4	.754	.370	.707
mod@m=2	.782	.390	.747
mod@m=4	.765	.390	.719
slg@m=2	**.792**	**.397**	**.741**
slg@m=4	**.792**	**.397**	**.741**

  – Slightly performed the best; aggressive is too much

Evaluation (2/4)

- Different window sizes

	WS=2	WS=3	WS=6	WS=10
nDCG	.765	**.792**	.774	.733
P	.410	.397	.343	.350
MRR	.736	**.741**	.741	.686

- In blogosphere words bond in proximity
  - In contrast to large window size in news articles

Evaluation (3/4)

- Estimation strategies for IntPref w/o reader feedback

@N=5	nDCG	P	MRR
entropy	.677	.287	.659
tfidf	.719	.313	.676
PR+tf	.657	.310	.632
PR+Pr(w|tp)	.631	.290	.583
PR+Pr(tp|w)	.673	.317	.639
PR+PrEntropy(w|tp)	.636	.283	.584
PR+PrEntropy(tp|w)	**.773**	**.337**	**.725**
PR+tfidf	**.792**	**.397**	**.741**

@N=3	nDCG	P	MRR
entropy	.667	.356	.644
tfidf	.651	.389	.638
PR+tf	.655	.350	.617
PR+Pr(w|tp)	.562	.328	.539
PR+Pr(tp|w)	.659	.350	.622
PR+PrEntropy(w|tp)	.562	.328	.539
PR+PrEntropy(tp|w)	**.757**	**.428**	**.717**
PR+tfidf	**.767**	**.506**	**.728**

- Entropy, tfidf beats PR+tf
- *PR+tfidf* achieves the best performance
- Entropy helps especially when better estimation is used

Evaluation (4/4)

- We trained tfidf and PR+tfidf with social interaction content

@N=5	# sentences in FB used	judges' interest	general readers' interest		
		nDCG	hit rate	nDCG	MRR
tfidf+FB$_{none}$ (=tfidf)	0	.719	.10	.087	.075
tfidf+FB$_{all}$	1314 (=100%)	.699	.10	.079	.072
PR+tfidf+FB$_{none}$ (=PR+tfidf)	0	.792	.19	.137	.122
PR+tfidf+FB$_{Coverage}$	393 (=30%)	**.803**	**.34**	.221	.182
PR+tfidf+FB$_{Focus}$	476 (=36%)	**.766**	**.28**	.164	.139
PR+tfidf+FB$_{Coverage+Focus}$	321 (=24%)	**.808**	**.33**	.210	.177

- Using all reader feedback is no better than using none
- *Coverage* and *Focus* select useful data and contribute to interest analysis
  - *Coverage* boosts hit rate relatively by 240% and 79%
- The combination filters out ¾ reader sentences
  - ¼ of the social data still help

Future Work

- Word omission happens in blogosphere especially in reader responses
  - Recover these words
- Connection between reader sentiment and reader interest
  - Sentiment analysis on interaction content help interest analysis?
  - Interest analysis help on-topic sentiment detection?

Conclusion

- Propose a work that predicts reader interest using
  - Semantic PageRank
  - Social data
- They are simple but helpful
  - Semantic features e.g., parts-of-speech and degrees of reference
  - Selection of informative reader responses
  - Topical interest preference model