Constructing Effective and Efficient Topic-Specific Authority Networks For Expert Finding in Social Media

Reyyan Yeniterzi & Jamie Callan

SoMeRA 2014

Social Media for Expert Search

2

- 72% of the companies use internal social media to find experts within the organization and improve collaboration
  - McKinsey Global Institute survey with >4200 companies
- 56% of the companies use social media for recruiting
  - SHRM 2011 survey on 'Social Networking Websites and Staffing'

Expert Retrieval Background

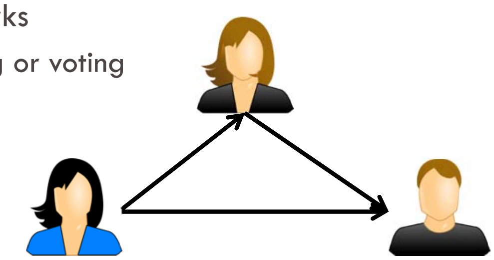
3

- Expert Finding Task
  - TREC Enterprise Track 2005-2008
  - W3C and CSIRO Collections
- State-of-the-art Approaches
  - Profile-based Models [Balog, 2006]
  - Document-based Models [Balog, 2006; Macdonald, 2006]
  - Graph-based Models [Serdyukov, 2008]
  - Learning-based Models [Fang, 2010]

Expert Retrieval in Social Media

4

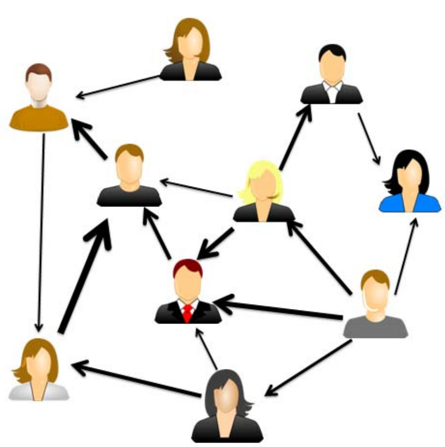
- Is writing topic-specific content enough for being considered an expert ?
- One also needs to have topic-specific influence over other users
  - authority estimation
  - user authority networks
    - reading, commenting or voting



Outline

- Authority-based approaches
  - PageRank [Brin and Page, 1998]
  - Topic-Sensitive PageRank [Haveliwala, 2002]
  - HITS [Kleinberg, 1999]
- Topic-Candidate Graphs
- Experiments
  - Finding topic-specific expert bloggers
- Conclusion

PageRank (PR) [Brin and Page, 1998]



$$PR(u) = \frac{1-d}{|U|} + d \sum_{i \in IL_u} \frac{PR(i)}{OL(i)}$$

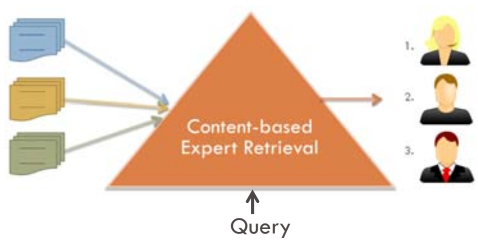
- Graph
  - topic-independent
    - all users
    - all user activities over all documents

Topic-Sensitive PageRank (TSPR)

[Haveliwala, 2002]

7

- the PageRank graph
- TSPR Approach
  - PageRank approach +
  - Teleportation is possible only to users that are associated with topic-relevant content

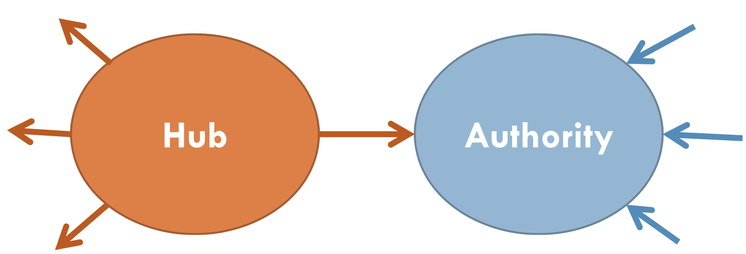


Hyperlink-Induced Topic Search (HITS)

[Kleinberg, 1999]

8

- Hub: Sum of authority scores of outgoing edges
- Authority: Sum of hub scores of incoming edges

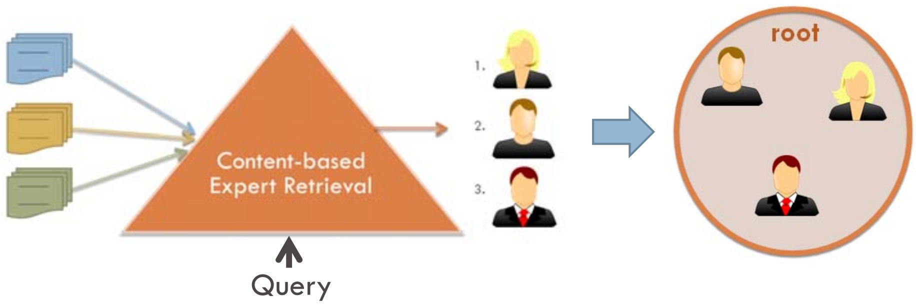


- Applied to more topic-specific authority networks
  - to focus the computational effort on relevant nodes

Constructing HITS Graph

9

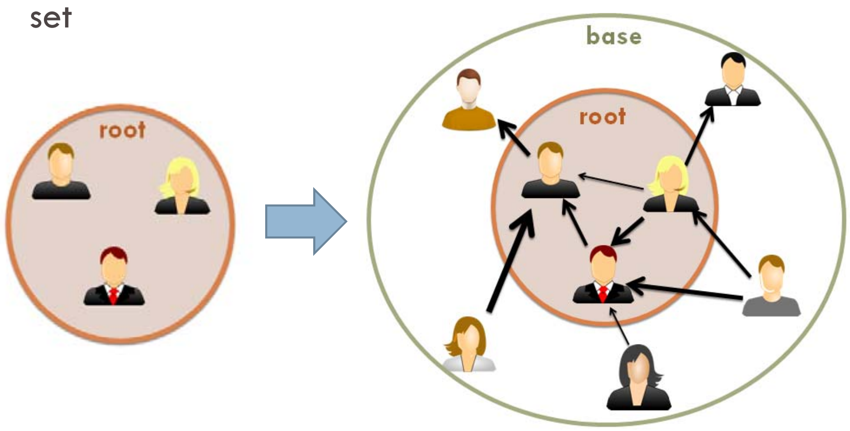
- Step 1: Retrieve an initial list of expert candidates, which is called the root set



Constructing HITS Graph

10

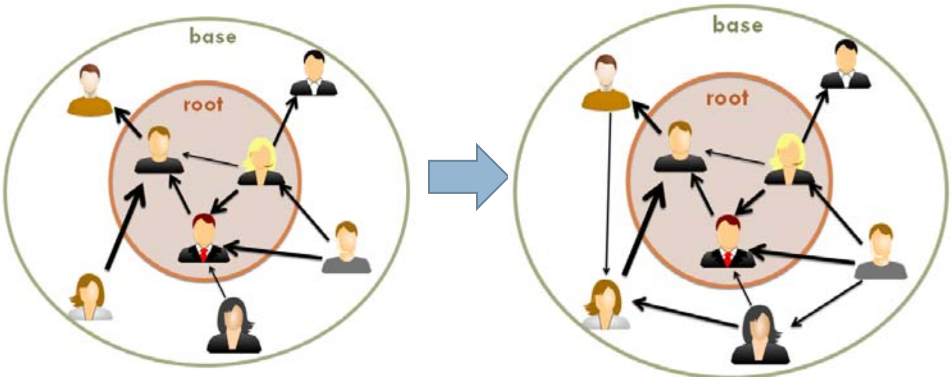
- Step 2 : Expand root set into base set, which consists of users who are connected to/from users in the root set



Constructing HITS Graph

11

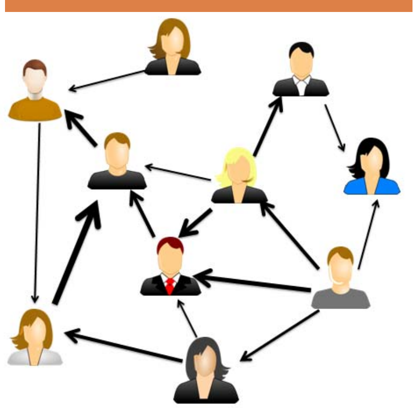
- Step 3 : Use all users in base set as nodes and all existing interactions among them as edges

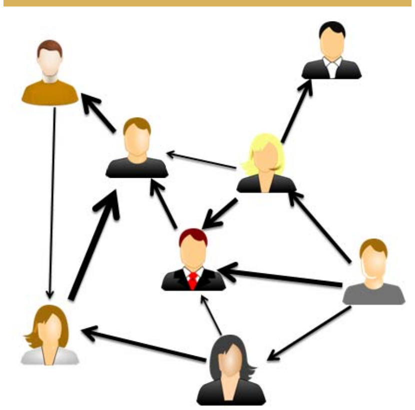


Graph Properties: Nodes & Edges

12

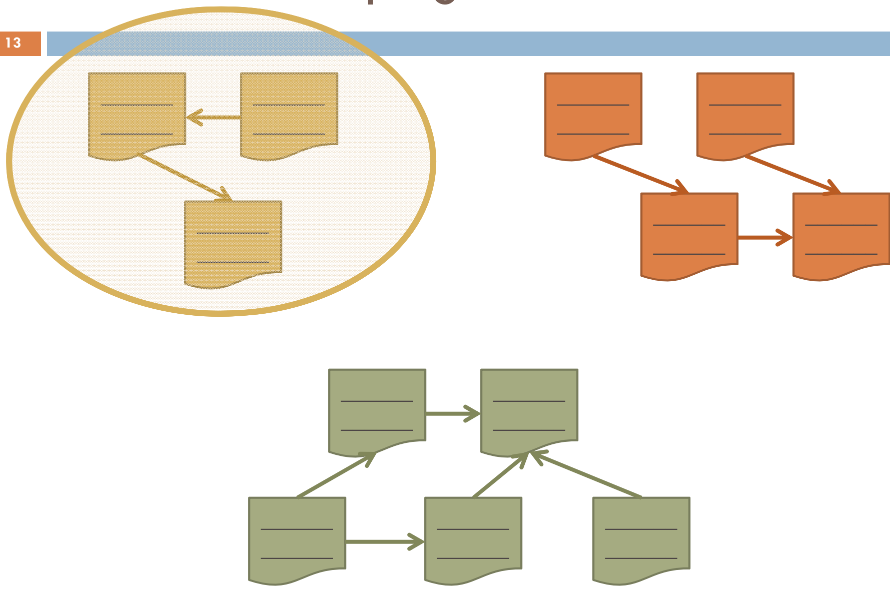
PageRank Graph	HITS Graph





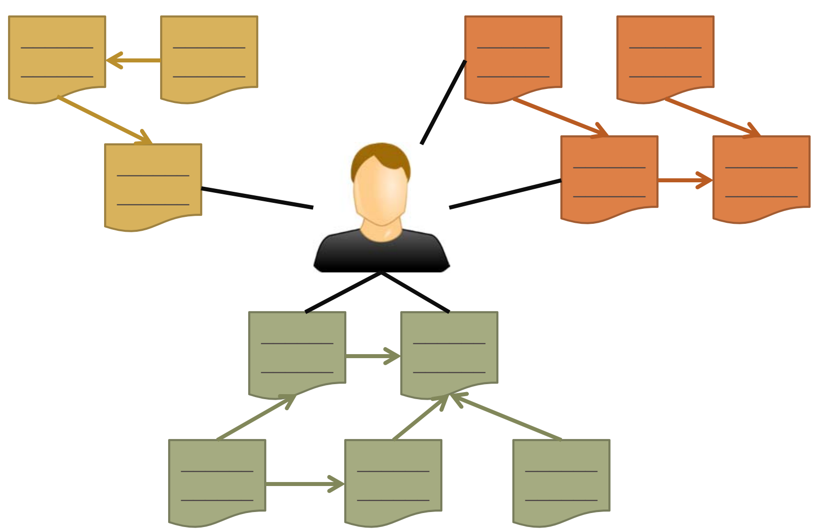
HITS on web pages

13



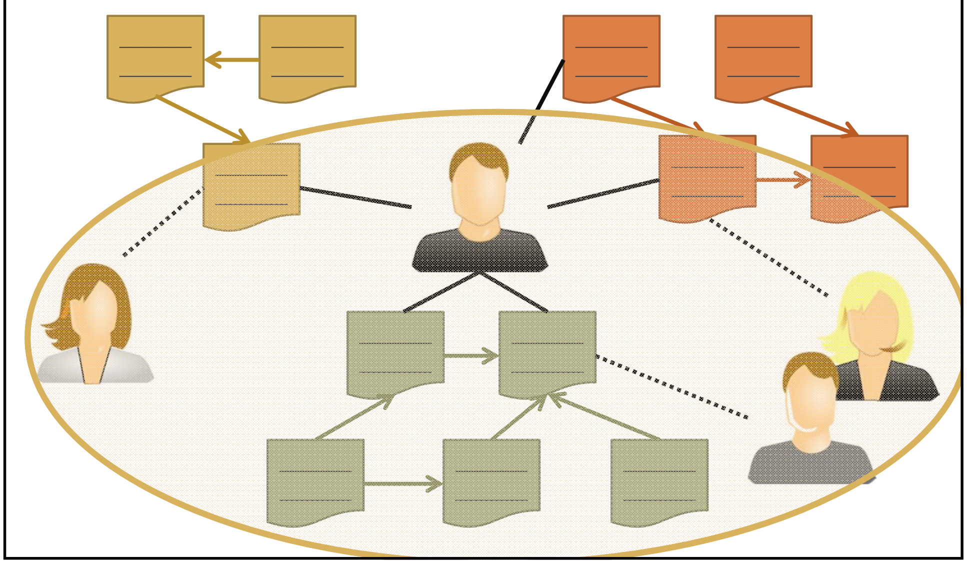
HITS on users

14



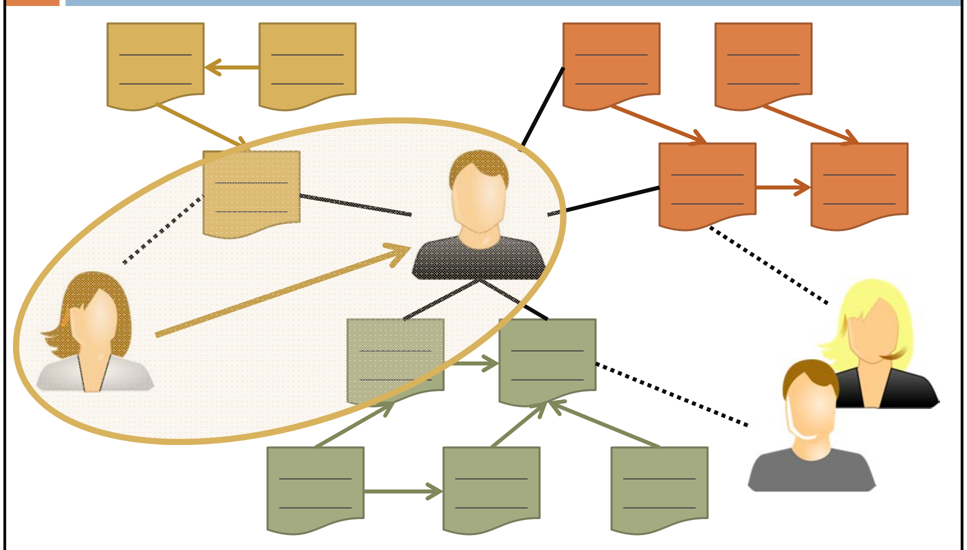
HITS on users

15



Topic-Candidate (TC) graphs

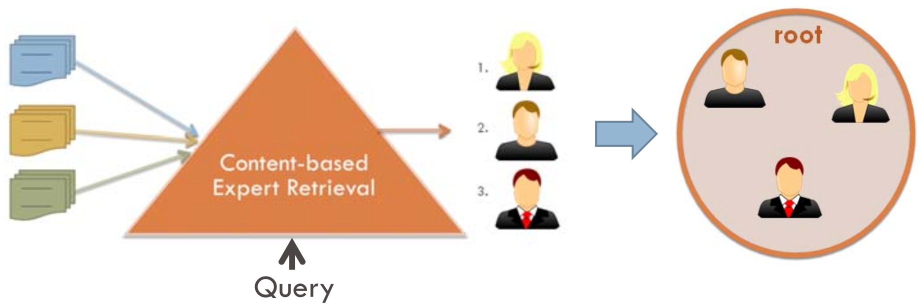
16



Constructing Topic-Candidate Graph

17

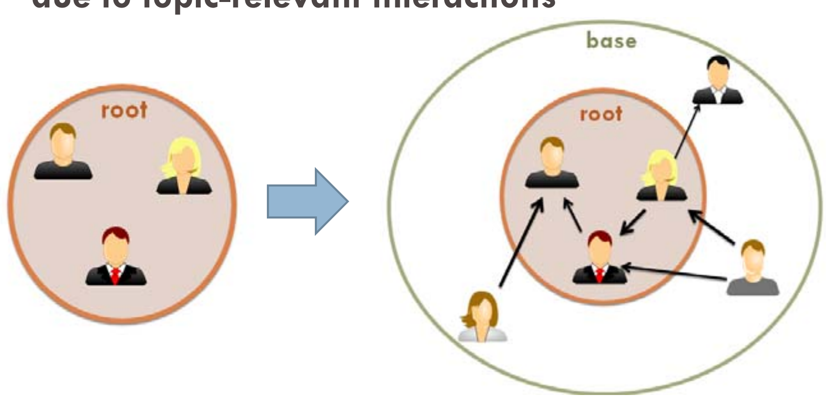
- Step 1: Retrieve an initial list of expert candidates, which is called the root set



Constructing Topic-Candidate Graph

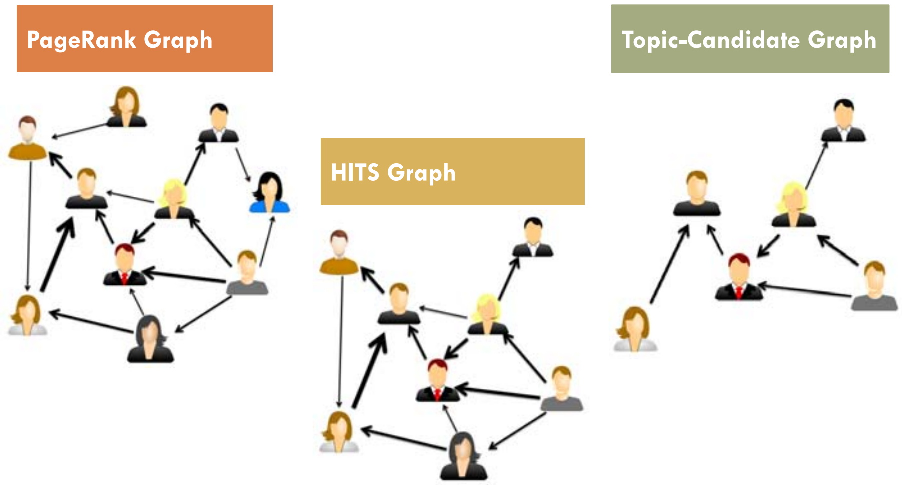
18

- Step 2 : Expand root set into base set, which consists of users who are connected to/from users in root set **due to topic-relevant interactions**



Comparison of Graphs

19



Experiments

- Finding topic-specific expert bloggers
  - Reading and commenting activity as authority signals

Dataset

21

- Intra-organizational blog collection from a large multinational IT firm

# Posts	165,414
# Comments	783,356
# Employees	>100,000
# Posters	20,354
# Commenters	42,169
# Readers	92,360

- Access logs
  - cover 44 of the 56 months of the collection

Evaluation Data

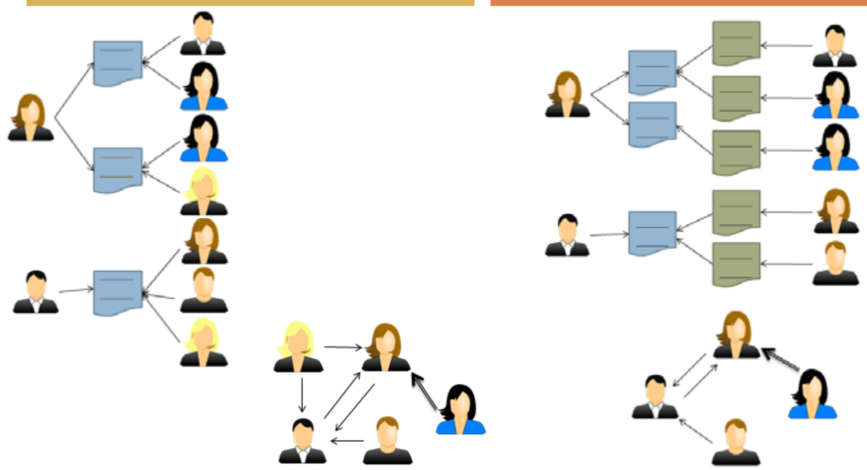
22

- 40 work related topics
  - Selected from the access logs of company search engine
  - Created by the company employees
- Candidate Pools
  - Top 10 candidates retrieved from content-based approaches
- Assessments – (The collection is not public)
  - Performed by author Yeniterzi
  - 4-point scale
    - not an expert, some expertise, an expert, very expert

Authority Networks

23

Reading | **Commenting**



Content-based Experiments

24

	NDCG @1	NDCG @3	NDCG @10
Profile [Balog, 2006]	.7000	.6689	.6494
Votes [MacDonald, 2006]	.3667	.4090	.4140
ReciprocalRank [MacDonald, 2006]	**.7083**	**.7003**	**.7281**
CombSUM [MacDonald, 2006]	.6417	.6334	.6168
CombMNZ [MacDonald, 2006]	.5333	.5295	.5124
IRW [Serdyukov, 2008]	.5167	.5189	.5159

Authority-based Re-ranking

25

$$final = \ content^{\alpha}\ reading^{\beta}\ commenting^{\gamma}$$

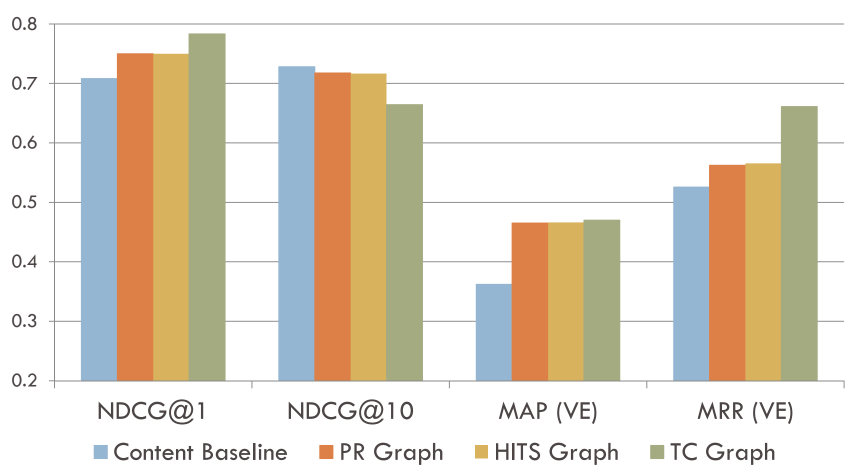
where

$$\alpha + \ \beta + \gamma = 1$$

- Parameter optimization
  - 5-fold cross validation

PageRank on Three Types of Graph

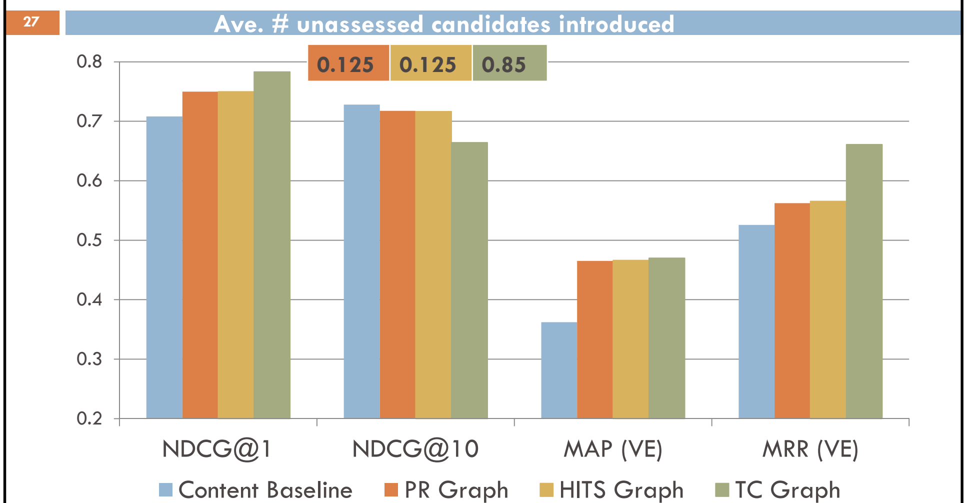
26



MRR (VE) improvement is statistically significant with $p< 0.05$
MAP (VE) improvement is statistically significant with $p< 0.10$

PageRank on Three Types of Graph

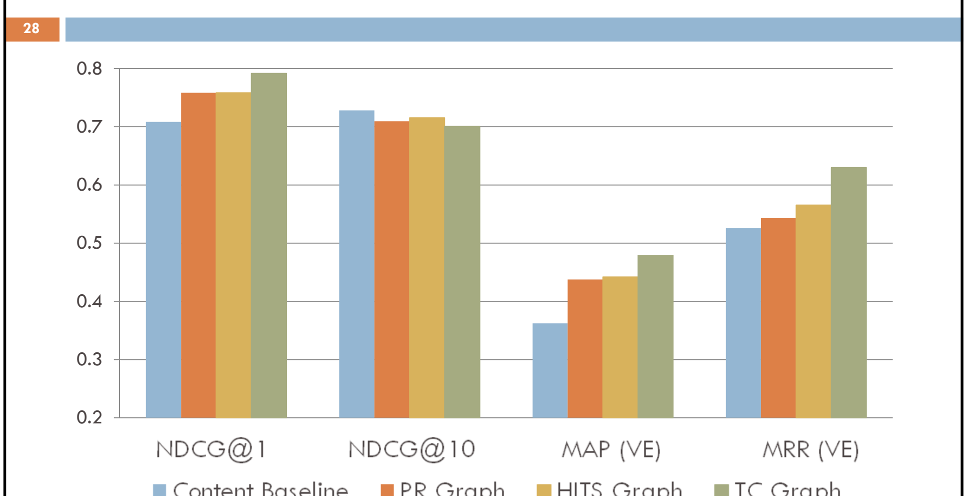
27



MRR (VE) improvement is statistically significant with $p < 0.05$
MAP (VE) improvement is statistically significant with $p < 0.10$

TSPR on Three Types of Graph

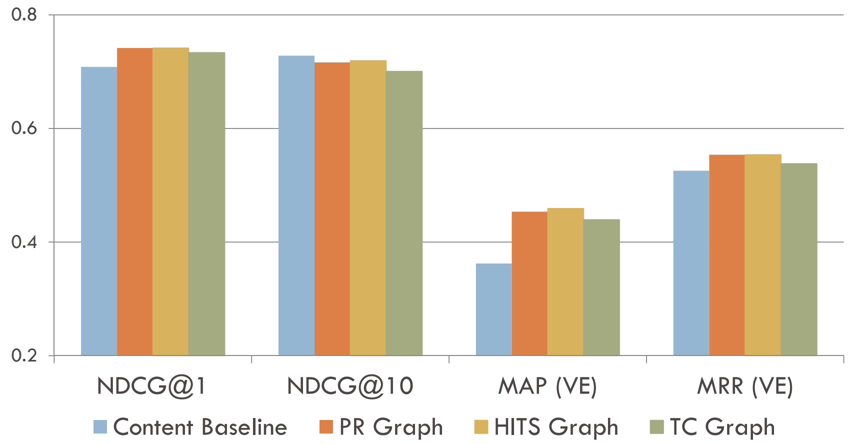
28



MRR (VE) improvement is statistically significant with $p < 0.05$

HITS on Three Types of Graph

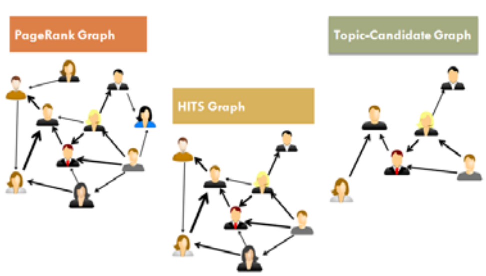
29



Graph Size and Running Time Analysis

30

Graph	Average # Nodes		Average # Edges	
	R	C	R	C
PR	92K	43K	1,631K	214K
HITS	57K	14K	1,480K	138K
TC	**7K**	**1K**	**9K**	**2K**



Approach	Graph	Approximate Running Times (in sec)	
		R	C
PR	PR	1,203	85
	HITS	1,116	49
	TC	**4**	**1**
TSPR	PR	1,222	93
	HITS	1,248	65
	TC	**2**	**0.4**
HITS	PR	478	73
	HITS	344	26
	TC	**3**	**0.5**

Conclusion

31

- Topic-Candidate graphs
- Statistically significant improvements @ MRR ($p<0.05$) with PageRank and TSPR approaches
  - Effectiveness
    - 4% @ NDCG@1
    - 8% @ MAP(VE)
    - 17% @ MRR(VE)
  - Efficiency
    - Reading: 20 min to 2 sec
    - Commenting: 1 min to 0.4 sec

Thank you