Taxonomy-based Query-dependent Schemes for Profile Similarity Measurement

Suppawong Tuarob, Prasenjit Mitra, C. Lee Giles

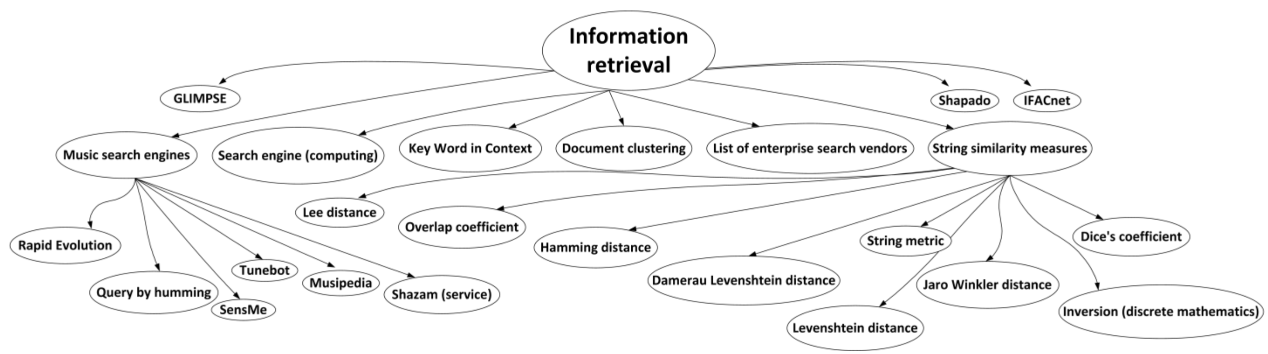
Computer Science and Engineering, Information Sciences and Technology
The Pennsylvania State University

Contributions

- We propose 10 query dependent schemes for computing similarity between 2 profiles
- We obtain resources such as the topic taxonomy from Wikipedia, Authors' profiles from ArnetMiner, and author and paper databases from Citeseer^{X}.
- We provide anecdotal results that show great promises on the proposed schemes.

Definition: Topic Taxonomy and Topic Library

- A ***topic taxonomy*** is a hierarchy of topics, where a node is a topic and each edge represents sub-topic relationship.



- A ***topic library*** is a set of topics taken from a topic taxonomy.

Definition: User Profile

- Given a topic library ***T***.
- Profile of user ***U*** is defined by a set of weighted topics:

$$P_U = \{< t_{u1}, w_{u1} >, ..., < t_{un}, w_{un} >\}$$

- Where $\{t_{u1}, ..., t_{un}\} \subseteq T$ and $\{w_{u1}, ..., w_{un}\}$ are real numbers between 0 and 1.

Definition: Query

- Given a topic library ***T***.
- Query ***Q*** is defined by a set of weighted topics:

$$Q = \{< t_{q1}, w_{q1} >, \ldots, < t_{qk}, w_{qk} >\}$$

- Where $\{t_{q1}, \ldots, t_{qk}\} \subseteq$ ***T*** and $\{w_{q1}, \ldots, w_{qk}\}$ are real numbers between 0 and 1.

Problem Definition

- Given Profile of two users $\boldsymbol{P_A}$ and $\boldsymbol{P_B}$, and a query $\boldsymbol{Q}$
- We aim to compute:
  - **`ProfileSimilarity(Q, PA, PB)`**
  - A function that returns a real number between 0 and 1, representing the level of profile similarity.

Resources

- Topic Taxonomy from Wikipedia
- Author research interests from ArnetMiner
- Author and Paper Databases from Citeseer^X

Topic Taxonomy from Wikipedia

- Extract 758,336 topics and their sub-topics relationship from Wikipedia.
- Pre-compute a shortest path between each pair of topics for fast look-ups, producing 139,736,685 shortest path entries.

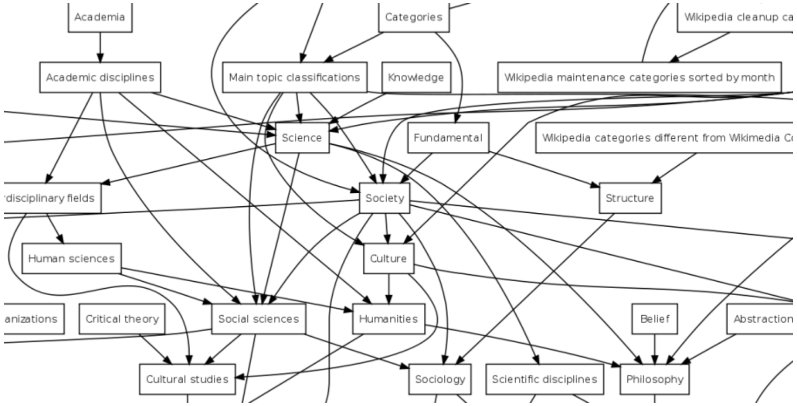
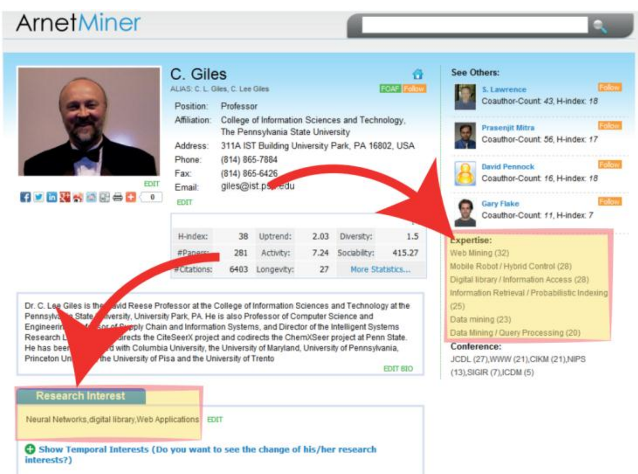


Image from: http://en.wikipedia.org/wiki/Wikipedia:Categorization

Author research interests from ArnetMiner

- Use research interests to define user profiles.
  - Extract each research interest (as a keyword) from ArnetMiner.org and map the keyword to topics using WikipediaMiner



wikipediaminer

Topic	Weight
Library_science	0.07692308
Data_mining	0.07692308
Machine_learning	0.05128205
Computational_neuroscience	0.05128205
Neural_networks	0.05128205
Archival_science	0.05128205
Digital_Humanities	0.05128205
Digital_libraries	0.05128205
Data_analysis	0.05128205
Formal_sciences	0.05128205
Software_architecture	0.02564103
Web_applications	0.02564103

C Lee Giles' Profile

Author and Paper Databases from Citeseer^X

- Citeseer^X hosts over 1.5 million scholarly documents.
- The author information (names, affiliations, lists of publications, etc.) is extracted from the documents as part of the meta-data extraction.
- We obtain a database of 307,262 authors from 1,077,513 documents.

Topic Similarity Function $TS(t_q, t_a, t_b)$

- An atomic function that computes the similarity between two topics t_a and t_b, given a query topic t_q.

$$TS(t_q, t_a, t_b) = \frac{|LCP(t_q, t_a, t_b)|}{min(|SP(t_q, t_a)|, |SP(t_q, t_b)|)}$$

- ***SP*(t_{start}, t_{end})** is a shortest path from topic t_{start} to topic t_{end} in the topic taxonomy
- **LCP(t_q, t_a, t_b)** is the longest common path between $SP(t_q, t_a)$ and $SP(t_q, t_b)$.

Profile Similarity Schemes

- We propose 10 query dependent schemes for calculating profile similarity, divided into 3 families: ***Topic Overlap*** based, ***Summation*** based, and ***Maximization*** based.

Family	Scheme Name	Acronym
Topic Overlap	User Uniform Overlap	UUO
	User Weighted Overlap	UWO
Summation	User Weighted Sum, Query Weighted	UWS-QW
	User Weighted Sum, Query Uniform	UWS-QU
	User Uniform Sum, Query Weighted	UUS-QW
	User Uniform Sum, Query Uniform	UUS-QU
Maximization	User Weighted Max, Query Weighted	UWM-QW
	User Weighted Max, Query Uniform	UWM-QU
	User Uniform Max, Query Weighted	UUM-QW
	User Uniform Max, Query Uniform	UUM-QU

Schemes: Topic Overlap Based

- Measure the topic overlapness of the two profiles.

$$\mathbf{ProfileSim}_{UUO}(Q, P_A, P_B) = \frac{1}{U_U} \cdot \sum_{\substack{<t_a, w_a> \\ \in P_A}} \sum_{\substack{<t_b, w_b> \\ \in P_B}} \begin{cases} TS(t_q, t_a, t_b) & ; if\ t_a = t_b \\ 0 & ; Otherwise \end{cases}$$

$$\mathbf{ProfileSim}_{UWO}(Q, P_A, P_B) = \frac{1}{W_U} \cdot \sum_{\substack{<t_a, w_a> \\ \in P_A}} \sum_{\substack{<t_b, w_b> \\ \in P_B}} \begin{cases} (w_a + w_b) \cdot TS(t_q, t_a, t_b) & ; if\ t_a = t_b \\ 0 & ; Otherwise \end{cases}$$

Schemes: Summation Based

- Sum over the similarity of each pair of topics between two users and takes the average.

$$\text{ProfileSim}_{UWS-QW}(Q, P_A, P_B) = \frac{1}{W_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} \frac{w_q}{W_U} \cdot \left(\sum_{\substack{<t_a, w_a> \\ \in P_A}} \sum_{\substack{<t_b, w_b> \\ \in P_B}} (w_a + w_b) \cdot TS(t_q, t_a, t_b) \right)$$

$$\text{ProfileSim}_{UWS-QU}(Q, P_A, P_B) = \frac{1}{U_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} \frac{1}{W_U} \cdot \left(\sum_{\substack{<t_a, w_a> \\ \in P_A}} \sum_{\substack{<t_b, w_b> \\ \in P_B}} (w_a + w_b) \cdot TS(t_q, t_a, t_b) \right)$$

$$\text{ProfileSim}_{UUS-QW}(Q, P_A, P_B) = \frac{1}{W_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} \frac{w_q}{U_U} \cdot \left(\sum_{\substack{<t_a, w_a> \\ \in P_A}} \sum_{\substack{<t_b, w_b> \\ \in P_B}} TS(t_q, t_a, t_b) \right)$$

$$\text{ProfileSim}_{UUS-QU}(Q, P_A, P_B) = \frac{1}{U_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} \frac{1}{U_U} \cdot \left(\sum_{\substack{<t_a, w_a> \\ \in P_A}} \sum_{\substack{<t_b, w_b> \\ \in P_B}} TS(t_q, t_a, t_b) \right)$$

Schemes: Maximization Based

- Pick the pair of topics between the two users that maximizes the similarity.

$$\text{ProfileSim}_{UWM-QW}(Q, P_A, P_B) = \frac{1}{W_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} \frac{w_q}{M_U} \cdot \left(\max_{\substack{<t_a, w_a> \in P_A, \\ <t_b, w_b> \in P_B}} \{(w_a + w_b) \cdot TS(t_q, t_a, t_b)\} \right)$$

$$\text{ProfileSim}_{UWM-QU}(Q, P_A, P_B) = \frac{1}{U_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} \frac{1}{M_U} \cdot \left(\max_{\substack{<t_a, w_a> \in P_A, \\ <t_b, w_b> \in P_B}} \{(w_a + w_b) \cdot TS(t_q, t_a, t_b)\} \right)$$

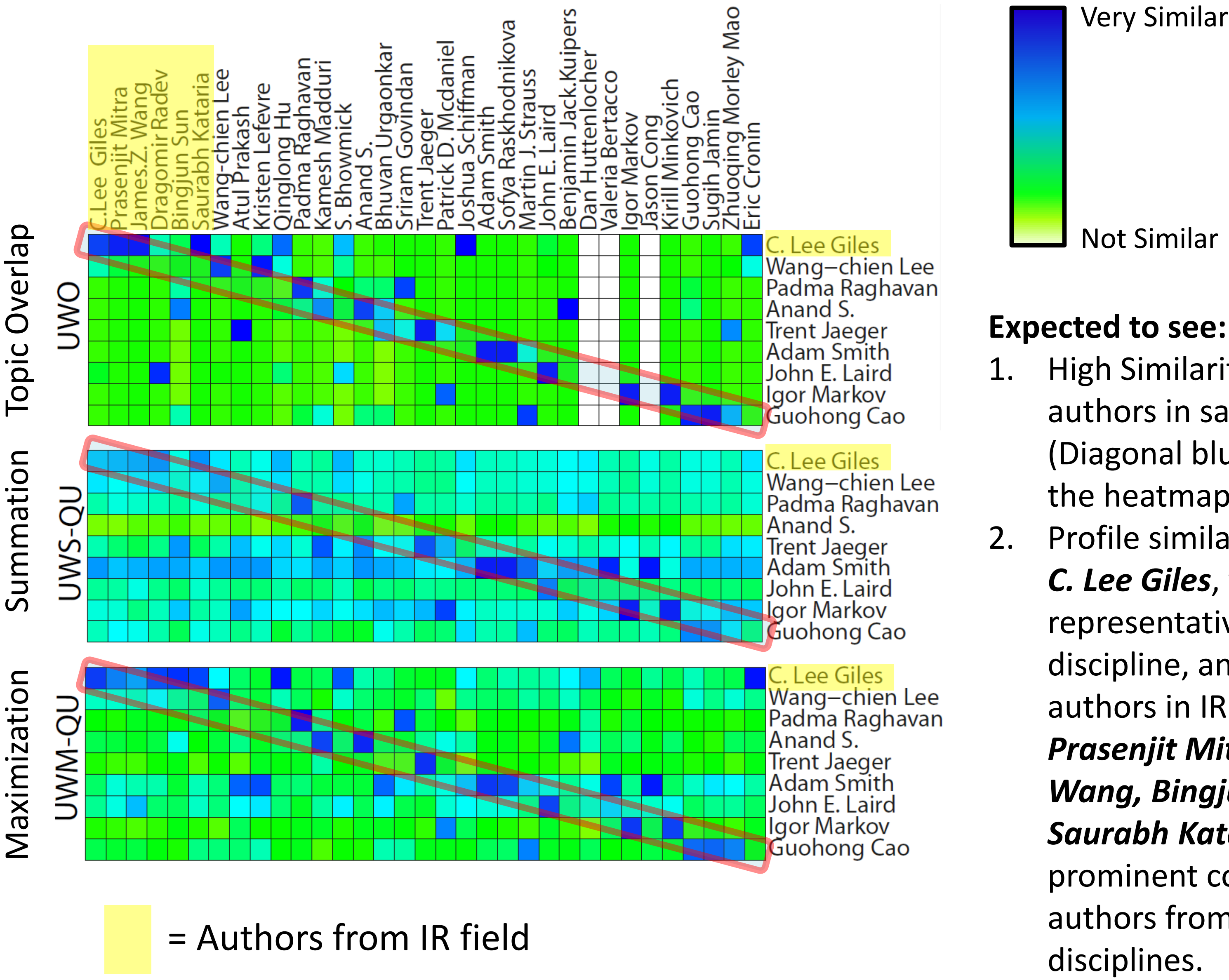
$$\text{ProfileSim}_{UUM-QW}(Q, P_A, P_B) = \frac{1}{W_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} w_q \cdot \left(\max_{\substack{<t_a, w_a> \in P_A, \\ <t_b, w_b> \in P_B}} \{TS(t_q, t_a, t_b)\} \right)$$

$$\text{ProfileSim}_{UUM-QU}(Q, P_A, P_B) = \frac{1}{U_Q} \cdot \sum_{\substack{<t_q, w_q> \\ \in Q}} \left(\max_{\substack{<t_a, w_a> \in P_A, \\ <t_b, w_b> \in P_B}} \{TS(t_q, t_a, t_b)\} \right)$$

Anecdotal Results

- 34 authors are chosen from 9 different computer science disciplines.
- Inter-similaities are compute between them using paper "***TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages***", as the query.

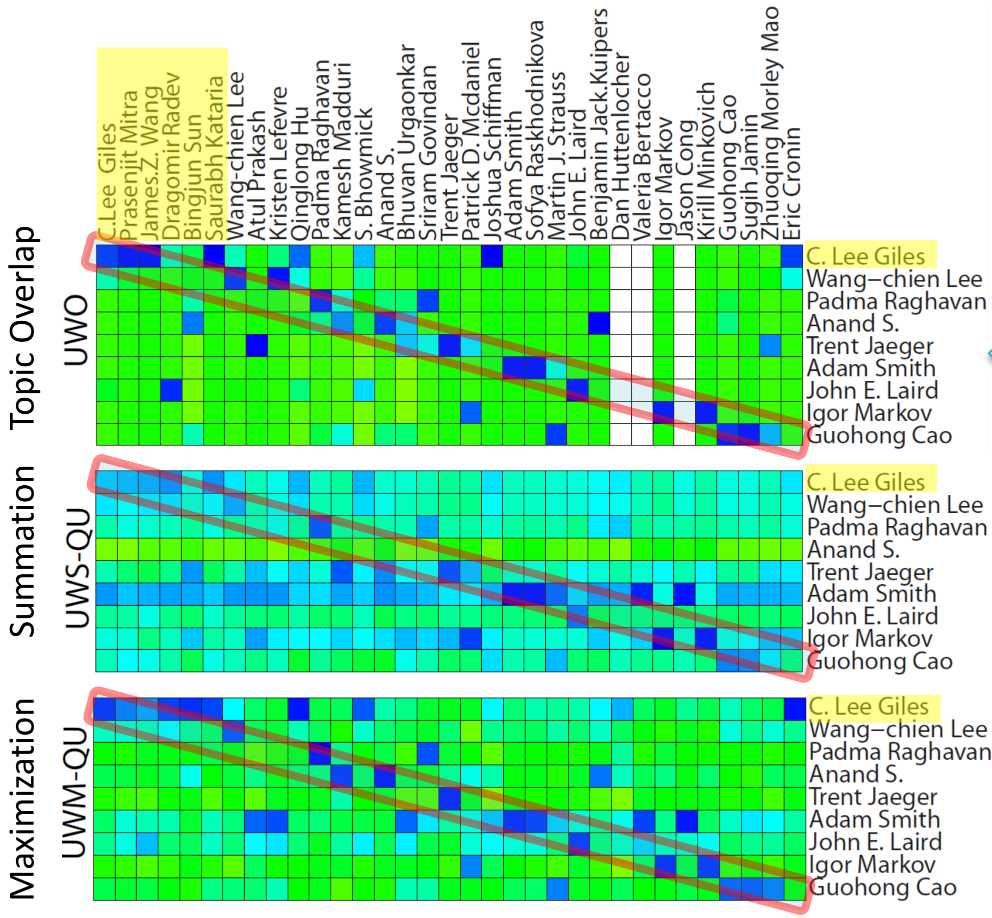
Anecdotal Results (cont.)



Expected to see:

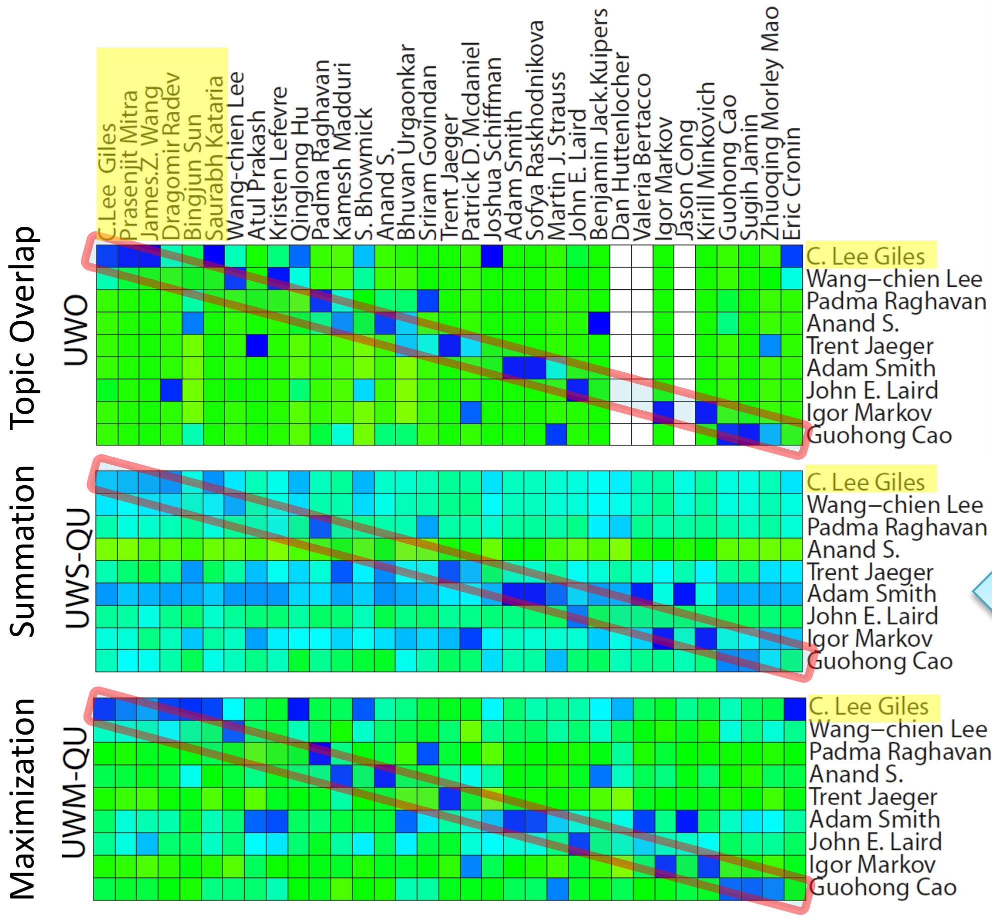
1. High Similarity among authors in same disciplines. (Diagonal blue trend across the heatmap)
2. Profile similarities between ***C. Lee Giles***, who is the representative of IR discipline, and the other authors in IR field (i.e. ***Prasenjit Mitra***, ***James Z. Wang, Bingjun Sun, and Saurabh Kataria***) are highly prominent compared to authors from other disciplines.

Anecdotal Results (cont.)



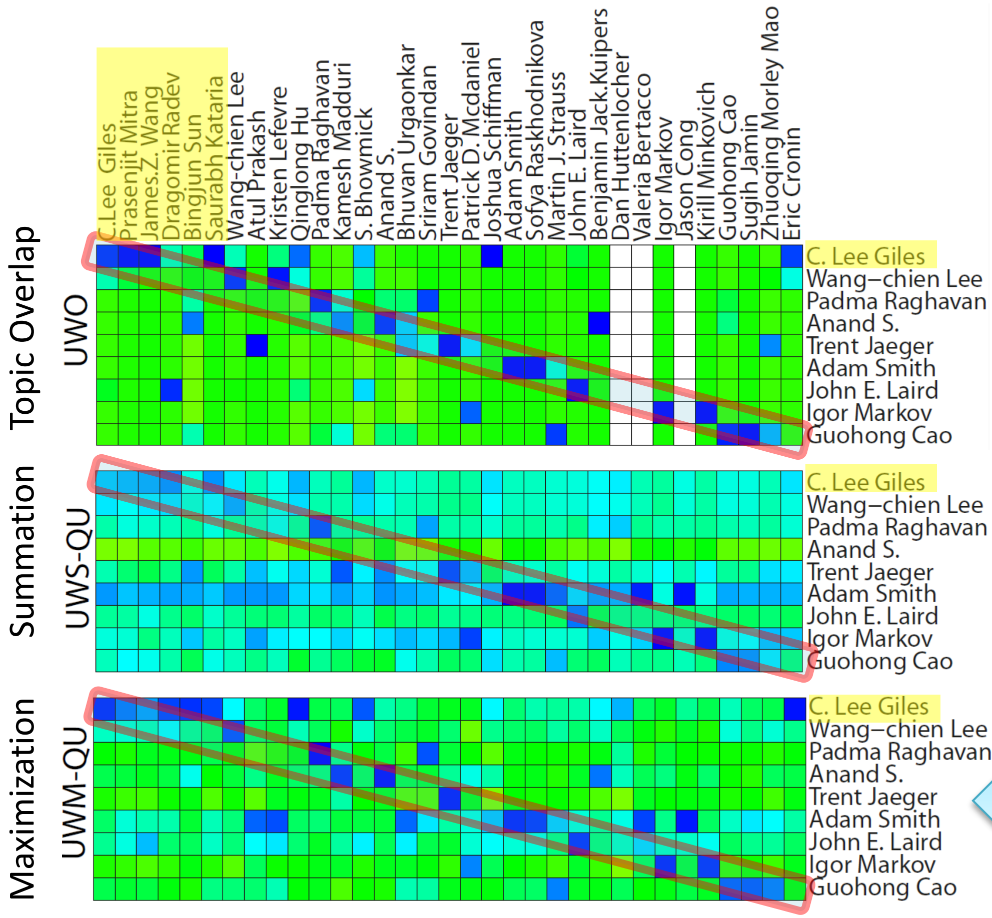
The ***topic overlap*** based schemes (UUO and UWO) give correct results. The dark blue grids tend to form a diagonal line across the heatmaps, implying high profile similarities among authors within the same research areas. However, the similarity levels are very strict–the heatmaps display only either dark blue grids or green (even white) grids. These high contrasts are expected since the topic overlap based schemes are not able to capture partial similarities.

Anecdotal Results (cont.)



The ***summation based*** schemes are able to compute partial similarities. However, these schemes do not yield accurate results. First, the profile similarities are not distinctive across the disciplines–the heatmaps show light blue grids spreading all over. Second, sometimes self-similarity levels are inferior to the similarities against others, which is not intuitive. For example, the similarities between C. Lee Giles and himself are even less than the similarities between C. Lee Giles and Bingjun Sun.

Anecdotal Results (cont.)



The ***maximization*** based schemes yield both correct and more accurate results than the other two families. Especially, the UWM-QU and UWM-QW schemes show promising diagonal blue patterns across the heatmaps. Furthermore, the profile similarities between C. Lee Giles, who is the representative of IR discipline, and the other authors in IR field (i.e. Prasenjit Mitra, James Z. Wang, Bingjun Sun, and Saurabh Kataria) are highly prominent compared to authors from other disciplines. This is expected since the query that we use is a publication from the IR field.

Conclusions

- We propose 10 schemes for profile similarity calculation divided into three families: topic overlap based, summation based, and maximization based.
- The anecdotal results show that the maximization based schemes, especially UWM-QU and UWM-QW, yield most accurate results as they are able to capture partial similarity between two topics.
- We also invest our efforts harvesting resources such as the topic taxonomy from Wikipedia, the high quality list of authors from Citeseer^{X}, and the author research interests from ArnetMiner.

References

- [1] mediawiki:org=wiki=Manual : Page table.
- [2] mediawiki:org=wiki=Manual : Categorylinks table.
- [3] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Capturing missing edges in social networks using vertex similarity. In Proceedings of the sixth international conference on Knowledge capture, K-CAP '11, pages 195{196, New York, NY, USA, 2011. ACM.
- [4] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Collabseer: a search engine for collaboration discovery. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, JCDL '11, pages 231{240, New York, NY, USA, 2011. ACM.
- [5] S. D. Gollapalli, P. Mitra, and C. L. Giles. Ranking authors in digital libraries. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, JCDL '11, pages 251{254, New York, NY, USA, 2011. ACM.
- [6] S. D. Gollapalli, P. Mitra, and C. L. Giles. Similar researcher search in academic environments. In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '12, pages 167{170, New York, NY, USA, 2012. ACM.
- [7] M. A. Hearst. TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1):33{64, 1997.
- [8] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, pages 538{543, New York, NY, USA, 2002. ACM.
- [9] S. Kataria, P. Mitra, and S. Bhatia. Utilizing context in generative bayesian models for linked corpus. In In AAAI, 2010.
- [10] J. M. Kleinberg. Hubs, authorities, and communities. ACM Computing Surveys, 31(4es):5{es, 1999.
- [11] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford Digital Library Technologies Project, 1998.
- [12] P.-N. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Journal of School Psychology, 19(1):51{56, 2005.
- [13] J. Tang and J. Zhang. ArnetMiner : Extraction and Mining of Academic Social Networks. Architecture, pages 990{998, 2008.
- [14] P. Treeratpituk and C. L. Giles. Disambiguating authors in academic publications using random forests. In Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, JCDL '09, pages 39{48, New York, NY, USA, 2009. ACM.