Detecting Fake Conferences and Cleaning Data Objects Therein

The Pennsylvania State University
U.S.A.

Dongwon Lee
dongwon@psu.edu

Outline

- Warm-Up
- AppleRank Project
  - Fake Conferences
- Data Linkage Project
  - Group Linkage
  - Other Linkage Techniques
- Conclusion

Penn State University

- Founded in 1855
- 23 campuses throughout PA state
- Main campus at State College, PA
- 84,000 students, 20,800 faculty
- $1.2 billion endowment
- “Nittany Lion”

Penn State ≠ U. Penn

- Two CompSci-related divisions:
  - Dept. of Computer Science & Engineering (CSE)
  - College of Info. Sciences & Technology (IST)

Penn State University

- State College, PA
  - Out of nowhere, but close to everywhere
  - West: 2.5 hours to Pittsburgh
  - East: 4 hours to New York
  - South: 3 hours to Washington DC
  - North: 3 hours to Buffalo

Penn State University

- 5 DB/IR Faculty
  - CSE:
    - Wang-Chien Lee
  - IST:
    - C. Lee Giles
    - Dongwon Lee
    - Prasenjit Mitra
    - James Wang

Penn State University

BLAST
This Talk

- Based on:
  - "Oracle, Where Shall I Submit My Papers?" ACM CACM 07
  - "Measuring Conference Quality by Mining Program Committee Characteristics", JCDL 07
  - "Group Linkage", ICDE 07

- Slides for this talk are available at:
  - http://pike.psu.edu => talks

Credits

- Students @ Penn State
  - Ergin Elmacioglu
  - Byung-Won On
  - Su Yan
  - Ziming Zhuang

- Collaborators
  - C. Lee Giles (Penn State, USA)
  - Nick Koudas (U. Toronto, Canada)
  - Divesh Srivastava (AT&T Labs, USA)

Outline

- Warm-Up
- AppleRank Project
  - Fake Conferences
- Data Linkage Project
  - Group Linkage
  - Other Linkage Techniques
- Conclusion

The AppleRank Project

- Fuzzy but interesting questions to ask:
  - How do we know if a conference X is better than a conference Y? In what aspect?
  - How can a system recommend an ideal venue to submit your paper?
  - In an interdisciplinary organization, how do we know if a scholar A in a field X is better than a scholar B in a field Y? If so, what does that mean?
  - Given your research interests, which professor is going to be your best advisor?

The AppleRank Project

- Goal: Can we rank academic entities termed as apples (eg, venues, authors, groups, papers) better?
- Better in what?
  - Accuracy (subjective vs. objective)
  - Computationally efficient?
    - Incremental computation
    - Unifying metrics?
    - Testing
    - in CiteSeer
    - Human subject
  - http://pike.psu.edu/applerank/

MIT's Prank

http://pdos.csail.mit.edu/scigen/

The World Multi-Conference on Systemics, Cybernetics and Informatics (SCI)
Annoyance…

“Dong-Won Lee” as PC?

Some Known Questionable Venues
- From http://www.inesc-id.pt/~aml/trash.html
  - IMCSE: International Multiconference in Computer Science and Computer Engineering
  - WMSCI or SCI: World Multiconference on Systemics, Cybernetics and Informatics
  - ECCIT: International Conference on Computing, Communications and Control Technologies
  - PISTA: Conference on Politics and Information Systems: Technologies and Applications
  - ESSOS: Symposium of Santa Catarina on Challenges in the Internet and Society: Security and Privacy Research
  - OTSA: International Conference on Cybernetics and Information Technologies, Systems and Applications
  - ISAS: International Conference on Information Systems Analysis and Synthesis
  - ISCBI: Conferência Iberoamericana de Educacion, Cibernética e Informática
  - SECII: Symposium Iberoamericano de Educacion, Cibernética e Informática
  - WCAC: World Congress in Applied Computing
  - Any IPSI International Conference or journal
  - Any GESTS international conference or journal
  - Any GESTS international conference or journal

Fake Venues
- According to fakeconferences.org,
  - “… fake venues are ones that are organized for the revenue, not for the advancement of science. They share a lot in common: an abundance of varying, vaguely connected topics, high frequency of conference, spam mailings, obscure organizers and sponsors, and poor peer reviewing and randomly accepting papers …”
- WMSCI has listed close to 300 research topics as relevant in its Call-For-Paper (CFP), and reportedly accepted 2,165 and 2,904 papers in 2003 and 2004, respectively

Differences in Disciplines
- Computer Science
  - Peer-reviewed conferences
  - Top conferences have 5-15% acceptance rate
  - Specialized and small conferences (attendance of 500+)
  - Often value conferences > journals
- Pure Sciences (eg, Math, Physics)
  - Pre-print at Arxiv.org
  - Rigorous reviews for journals
  - Huge flagship conference (ICM 98 attracted ~4000)
- Social Sciences
  - Often value journals > conferences
  - Conferences are mostly for gathering or short abstract based screening
  - Rigorous reviews for journals

Research Question
- Can we detect the so called “fake venues” automatically?
- Desiderata
  - Large-number of venues per year
  - Scalable
  - Automatic detection
  - Little human involvement
  - Avoiding false positives is more important than false negatives

Histogram of CFPs in dbworld
Candidate Features

- Good vs. bad venues
  - Citation counting (e.g., Impact Factor)
  - Acceptance rate
  - Reputation (e.g., society)
  - History
  - ...
- At the end, none satisfy our desiderata. Need something else…

Research Hypothesis

Qualities of venues are closely correlated with those of PC members of the venues

- PC member list can be readily available from CFP $\bowtie$ data extraction + data cleaning
- Each CFP has only finite number of PCs $\bowtie$ scalability
- Examine quality of PC w.r.t heuristics:
  - Citation counting, productivity, centrality, betweenness, impact, …

Classification w. Decision Tree

<table>
<thead>
<tr>
<th>PC has feature A?</th>
<th>PC has feature B?</th>
<th>PC has feature C?</th>
<th>PC has feature D?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

5 Classification Features

- # of PC
- # of publication of PC
- # of co-authors of PC
- Closeness centrality of PC
- Betweeness centrality of PC

Set-Up

- ACM DL: downloaded data of 1950-2004
  - 0.6M authors, 0.7M articles
  - 1.2M edges (i.e., collaboration)
- Dbworld: 2,979 CFPs (free text formats)
  - 16,147 distinct PC names
- Hand-selected 20 fake venues $\bowtie$ Q
- Laborious cleaning process for venue, PC names, and citations:
  - Entity resolution
  - Name disambiguation
  - Record linkage

# of PC

- 2nd part of the talk
Combining All Features

- Naive (C4.5)
  - Precision: 0.877
  - Recall: 0.965
- Bagging
  - Precision: 0.899
  - Recall: 0.979
- Boosting
  - Precision: 0.938
  - Recall: 0.964

More than “usual suspects”

- Classification detected two:
  - The 2nd International Advanced Database Conference X
  - The 4th International Conference on Computer Science and its Applications Y
  - Not part of original Q

PSU Prank

- Apr. 10, 2006, we generated 3 bogus papers using MIT SC1gen software:
  - P1 by Ethan Patel
  - P2 by Simon R. Hathaway
  - P3 by Richard Zhang

Indiana’s Inauthentic Paper Detector says:

- P1: 28.9% => inauthentic
- P2: 61.5% => authentic
- P3: 38% => inauthentic

PSU Prank
PSU Prank

- April 24 – May 1, 2006
  - P1 to X on April 24
  - P2 to Y on April 26
  - P3 to X on May 1
- May 15, 2006
  - P1 and P2 accepted w/o reviews
  - P3 rejected w/o reviews
  - Asked for reviews or any rationale ∆ no response so far

Summary

- Practical setting of outlier detection
  - Semantic outlier vs. syntactic outlier
- Future Work
  - Extra features (affiliations, dates, professional societies, etc)
  - Correlation with Impact Factor
  - A generic academic entity ranking framework (extend current work for journals and for research disciplines other than Computer Science)
  - Apply to other fake detection problems?
    - Eg, GM counterfeit detection
  - Developing generic academic entity ranking framework
    - AppleRank Project

Outline

- Warm-Up
- AppleRank Project
  - Fake Conferences
  - Data Linkage Project
    - Group Linkage
    - Other Linkage Techniques
- Conclusion

The Data Linkage Project

- In the AppleRank project, one needs to do laborious cleaning for venue, PC, citations:
  - Entity resolution
  - Name disambiguation
  - Record linkage
- Re-visit the traditional record linkage problem to cope with novel challenges
  - The Data Linkage Project
  - http://pike.psu.edu/linkage/

Eg. ACM DL Portal

Jeffrey D. Ullman
@ Stanford Univ.

Eg. DBLP: split names
Eg. DBLP: mixed names

Eg. World-Wide Web

Technical Landscape

Group Linkage

Group Linkage Example
Popular Group Similarity

- Jaccard
  - Intuitive, cheap to run
  - Error-prone
  
  \[ \text{sim}(g_1, g_2) = \frac{|g_1 \cap g_2|}{|g_1 \cup g_2|} \]

Q: Can we combine Jaccard and Bipartite Matching for Group Linkage?

Intuition for Better Similarity

- Two groups are similar if:
  - A large fraction of elements in the two groups form matching element pairs
  - There is high enough similarity between matching pairs of individual elements that constitute the two groups

Framework

- Threshold Version
- Task:
  - Two groups \( g_1 \) and \( g_2 \) and a user-selected threshold \( \theta \) are given
  - Measure the similarity \( \text{sim} \) between \( g_1 \) and \( g_2 \)
  - If \( \text{sim} \geq \theta \), then \( g_1 \) and \( g_2 \) are similar groups!

Group Similarity

- Two groups of elements:
  - \( g_1 = \{r_{11}, r_{12}, \ldots, r_{1m_1}\} \)
  - \( g_2 = \{r_{21}, r_{22}, \ldots, r_{2m_2}\} \)
  - The group measure \( \text{BM} \) is the normalized weight of the maximum bipartite matching \( M \) in the bipartite graph \( (N = g_1 \cup g_2, E=g_1 \times g_2) \)

\[ \text{BM}_{\rho, \Theta}(g_1, g_2) = \frac{\sum_{i,j} (\text{sim}(r_{i1}, r_{j1}))}{m_1 + m_2 - |M|} \]

such that \( \text{sim}(r_{i1}, r_{j1}) \geq \rho \)
- \( \text{BM}(g_1, g_2) \geq \Theta \)

Example (\( \rho = 0.3, \Theta = 0.9 \))

\[ \text{BM}_{0.3, 0.9}(g_1, g_2) = \frac{0.9 + 0.7}{3 + 2} - 0.53 < 0 \]

Therefore, \( g_1 \not\approx g_2 ! \)

Challenge

- Each BM group measure uses the maximum weight bipartite matching
  - Bellman-Ford: \( O(V^2E) \)
  - Hungarian: \( O(V^3) \)
- Large number of groups to match
  - \( O(NM) \)
Solution: Greedy matching

- Bipartite matching computation is expensive because of the requirement.
  - No node in the bipartite graph can have more than one edge incident on it.
- Let’s relax this constraint:
  - For each element \( e_i \) in \( g_1 \), find an element \( e_j \) in \( g_2 \) with the highest element-level similarity \( \sim \).
  - For each element \( e_j \) in \( g_2 \), find an element \( e_i \) in \( g_1 \) with the highest element-level similarity \( \sim \).

Upper/Lower Bounds

\[
BM_{\text{sim}, \rho}(g_1, g_2) = \frac{\sum_{(r_i, r_j) \in \rho} \left( \text{sim}(r_i, r_j) \right)}{m_1 + m_2}
\]

\[
UB_{\text{sim}, \rho}(g_1, g_2) = \frac{\sum_{(r_i, r_j) \in \rho} \left( \text{sim}(r_i, r_j) \right)}{m_1 + m_2 - |M|}
\]

\[
LB_{\text{sim}, \rho}(g_1, g_2) = \frac{\sum_{(r_i, r_j) \in \rho} \left( \text{sim}(r_i, r_j) \right)}{m_1 + m_2 - |S_1 \cup S_2|}
\]

Properties:
- Numerator of UB is at least as large as that of BM.
- Denominator of UB is no larger than that of BM.
- => UB is the upper-bound of BM.

MAX Heuristics

\[
MAX_{\text{sim}, \rho}(g_1, g_2) = \max_{(r_i, r_j) \in g_1 \times g_2} \text{sim}(r_i, r_j)
\]

- Two groups with high BM will share at least one pair of very similar elements.
  - Use MAX to quickly identify those.
  - No guarantee of avoiding false identification.
- We proposed 4 group similarity measures:
  - BM, UB, LB, and MAX.

Theorem & Algorithm

\[
BM_{\text{sim}, \rho}(g_1, g_2) \leq UB_{\text{sim}, \rho}(g_1, g_2)
\]

Theorem 1

- IF \( UB(g_1, g_2) < \theta \rightarrow BM(g_1, g_2) < \theta \rightarrow g_1 \neq g_2 \)
- ELSE IF \( LB(g_1, g_2) \geq \theta \rightarrow BM(g_1, g_2) \geq \theta \rightarrow g_1 \approx g_2 \)
- ELSE, compute BM\( g_1, g_2 \)

Goal: \( BM(g, g) \geq \theta \)

Differences

- Traditional Bipartite Matching and BM
  - \( O(V^2) \) or \( O(V^3) \)
  - Exact algorithm
- UB and LB
  - \( O(V) \)
  - Filtering algorithm
- MAX
  - \( O(E) \)
  - Approximation algorithm

Suggestion: Pipelined Approach
1. Use UB, LB, or MAX
2. Use BM afterward

\( UB|BM, LB|BM, MAX|BM \)
Evaluation

- Evaluated Selection version (vs. Join version)
- Use bibliography data set from ACM and DBLP digital libraries
  - Authors with his/her publication lists
- Various cases
  - Real vs. Synthetic
  - Uniform vs. Skewed
  - Jaccard vs. 4 proposals (BM, UB, LB, and MAX)
  - Hybrid as blocking method
- Main evaluation metric: AVG recall

BM vs. Jaccard

S1: Left: 300 DBLP groups
Right: 700,000 ACM groups + 1/3 or 3 dummy groups

MAX vs. UB

R2Net: Left: 100 DBLP groups on Network topics
Right: 700,000 ACM groups

ACM Dataset

R2Net: Left: 100 DBLP groups on Network topics
Right: 700,000 ACM groups

Summary

- When entities have a group of elements in them, group linkage is useful and efficient
- Directions
  - Group Linkage => Group Join
  - More efficient implementation
    - Approximate Group Linkage
  - Hierarchical Group Linkage: OLAP
  - Group => Tree, Graph
    - Quasi-Clique
  - Application to Image Retrieval
Outline
- Warm-Up
- AppleRank Project
  - Fake Conferences
- Data Linkage Project
  - Group Linkage
  - Other Linkage Techniques
- Conclusion

Data Linkage Project
- Exploring various directions:
  - Adaptability ⇔ Adaptive Linkage (JCDL 07)
  - Web ⇔ Search Engine Based Linkage
  - Scalability ⇔ Parallel Linkage
  - Privacy ⇔ Privacy-Preserving Linkage
  - Video ⇔ Video Linkage

Outline
- Warm-Up
- AppleRank Project
  - Fake Conferences
- Data Linkage Project
  - Group Linkage
  - Other Linkage Techniques
- Conclusion

Conclusion
- The AppleRank Project
- The Data Linkage Project
- We are only in a very early stage
- Many interesting yet practical problems
  - Surged interest from DB, DM, WWW, NLP, and AI communities

Thank You!