Oracle, Where Shall I Submit My Precious Papers?

IST Faculty Brown Bag
Sep. 22, 2006

Dongwon Lee

Credits

- Students
  - Ergin Elmacioglu (CSE, Penn State)
  - Su Yan (IST, Penn State)
  - Ziming Zhuang (IST, Penn State)
- Colleagues
  - Lee Giles (Penn State)
  - Min-Yen Kan (NUS, Singapore)
  - Jaewoo Kang (Korea U., Korea)
  - Divesh Srivastava (AT&T Labs – Research)
What do I do?

- Databases / Data Mining
- Digital Libraries / Info. Retrieval
- XML / Web

What projects do I do?

- IBM Eclipse, 2004 & 2006
- NSF OISE, 2006
- Microsoft SciData 2005
- Penn State eBRC, 2005
Outline

- Motivation
- Simple Study
- Results
- Summary

MIT’s Prank

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

Scientific conference falls for gibberish prank. 15-04-2005. ABC News Online

[This is the print version of story http://www.abc.net.au/news/newsstory/2005/04/158752.html]

Last Update: Friday, April 15, 2005 8:19am (AEST)

Scientific conference falls for gibberish prank

A batch of computer-generated gibberish masquerading as an academic paper has been accepted at a scientific conference in a victory for pranksters at the Massachusetts Institute of Technology (MIT)

Jeremy Silver said he and two fellow MIT graduate students questioned the standards of some academic conferences, so they wrote a computer program to generate research papers complete with numerical text, charts and diagrams.

The trio submitted two of the randomly assembled papers to the World MultiConference on Systemics, Cybernetics and Informatics (WMSCI), scheduled to be held July 10-13 in Orlando, Florida.

To their surprise, one of the papers — "Scout: A Methodology for the Topical Unification of Access — The World Multi-Conference on Systemics, Cybernetics and Informatics (SCI)
Annoyance…

Get me off Your Fucking Mailing List
David Mazières and Eddie Kohler
New York University
University of California, Los Angeles
HTTP://www.danielvng.org/

Abstract
Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list.

“Dong-Won Lee” as PC?

WMSCI 2006
Some Known Questionable Venues

  - IMCSE: International Multiconference in Computer Science and Computer Engineering
  - WMSCI or SCI: World Multiconference on Systemics, Cybernetics and Informatics
  - ICCCT: International Conference on Computing, Communications and Control Technologies
  - PISTA: Conference on Politics and Information Systems: Technologies and Applications
  - SSCCI: Symposium of Santa Caterina on Challenges in the Internet and Interdisciplinary Research
  - CITSA: International Conference on Cybernetics and Information Technologies, Systems and Applications
  - ISAS: International Conference on Information Systems Analysis and Synthesis
  - CISCI: Conferencia Iberoamericana en Sistemas, Cibernética e Informática
  - SIECI: 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
  - KCPR: International Conference on Knowledge Communication and Peer Reviewing
  - International e-Conference on Computer Science
  - ...

[http://fakeconferences.org](http://fakeconferences.org) => down from a threat

Fakes Everywhere

Microsoft HoneyMonkey

SpamAssassin

STRIDER
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
Outline

- Motivation
- Simple Study
- Results
- Summary

Research Question

Can we detect the so-called “fake venues” automatically?

- Desiderata
  - Large-number of venues per year \iff scalable
  - Automatic detection \iff no human involvement
  - False positives \gg 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 ⇨ data extraction + data cleaning
- Each CFP has only finite number of PCs ⇨ scalability
- Examine quality of PC w.r.t heuristics:
  - Citation counting, productivity, centrality, betweenness, impact, …
Data Mining Models

- Outlier detection
- Clustering
- Classification

Classification w. Decision Tree

PC has feature A?
PC has feature B?
PC has feature C?
PC has feature D?

Fake?
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 (ie, collaboration)
- Dbworld: 2,979 CFPs (free text formats)
  - 16,147 distinct PC names
- Hand-selected 20 fake venues ⇔ Q
- Laborious cleaning process for venue, PC names, and citations:
  - Entity resolution
  - Name disambiguation
  - Record linkage

Another Talk
Outline

- Motivation
- Simple Study
- Results
- Summary

# of PC

![Graph showing the distribution of the number of PCs and their fraction of conferences.](image)
### # of publication of PC

![Bar chart showing the percentage distribution of publication counts for PC categories.](Image)

### Closeness centrality of PC

![Graph showing the average closeness centrality of conferences.](Image)

\[ CC(v) = \frac{n-1}{\sum_{v \in G} d(v, w)} \]
Combining All Features

- Naïve (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 (IADC)*
  - *The 4th International Conference on Computer Science and its Applications (ICCSA)*
- Not part of original Q
PSU Prank

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

PSU Prank

- Indiana’s Inauthentic Paper Detector says:
  - P1: 28.9% => inauthentic
  - P2: 61.5% => authentic
  - P3: 38% => inauthentic
PSU Prank

- April 24 – May 1, 2006
  - P1 to ICCSA on April 24, (2) P2 to IADC on April 26, and (3) P3 to ICCSA 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

“Ethan Patel” made it!
“Richard Zhang” too!

Outline

- Motivation
- Simple Study
- Results
- Summary
## Summary

- Practical setting of outlier detection
  - Semantic outlier vs. syntactic outlier
- Developing general semantic outlier detection framework
- Applying to other practical problems
  - Eg, GM counterfeit detection
- Developing general venue ranking framework
  - AppleRank project