



NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006

## Research Hypothesis

**Hypothesis: Using external resources as in URL would help disambiguate author names with mixed citation**

- **Many factors to consider:**
  - Which external resources to use: URL, web page contents, affiliation, etc
  - How to use: both internal and external? How to mix?
  - How to apply external resources? Weighting?
- **Preliminary study focuses on the case using URL and simple weighting**

Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 7

NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006

## External Resources

- Lay people doing this task with unfamiliar publications may use a search engine, using paper title as query
- Our method tries to approximate this
- For each citation  $c$  in  $C$ 
  - Query search engine with title of  $c$  as phrase search to obtain a set of relevant URLs
  - Represent  $c$  by a feature vector of relevant URLs and weighting scheme
- **Apply hierarchical agglomerative clustering (HAC) on  $C$  to derive  $k$  clusters**
  - Cosine similarity
  - Tested with [single link](#), [complete link](#) and [group average](#)

Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 8

NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006

## Weighting: Inverse Host Frequency (IHF)

- **Observation**
  - Not all URLs are equally useful
  - e.g., aggregator services
- **Desired weighting scheme**
  - Low weights to aggregator web sites
  - High weights to personal and group publication pages
- **Inverse Host Frequency (IHF)**
  - Similar to Inverse Document Frequency (IDF) in information retrieval
- Consider citations of top 100 authors in DBLP (by number of citations)
- For each such citation, query search engine with its title to obtain URLs, truncate them to their hostnames
- If a hostname  $h$  has frequency  $f(h)$ , then its IHF is
 
$$\text{IHF}(h) = \log_2 \frac{\max_h f(h) + 1}{f(h) + 1} + 1$$

Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 9

NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006

## Weighting: Inverse Host Frequency (IHF)

- **We notice that using hostnames alone may be problematic**
  - Especially when a host has multiple hostnames or is represented by an IP address with dissimilar distributions
  - e.g. [www.informatik.uni-trier.de](http://www.informatik.uni-trier.de), [ftp.informatik.uni-trier.de](http://ftp.informatik.uni-trier.de) and 136.199.54.185 are the same host
- **Therefore, we also experimented with**
  - Domain (e.g. uni-trier.de)
  - Resolving hostnames to IP addresses

Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 10

NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006

## Evaluation

- **Dataset**
  - Manually-disambiguated dataset of 24 ambiguous names in computer science domain
  - Each ambiguous name represented 2 unique authors ( $k = 2$ ) except for one where it represented 3
  - Each name is attributed to 30 citations on average
  - Proportion of largest class ranges from 50% to 97%
- **Search engine**
  - Google (<http://www.google.com/>)

Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 11

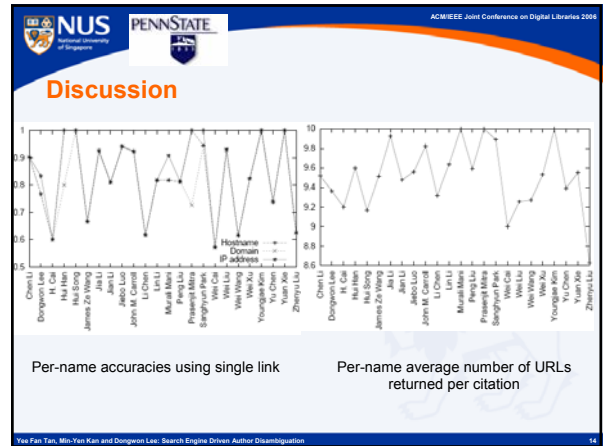
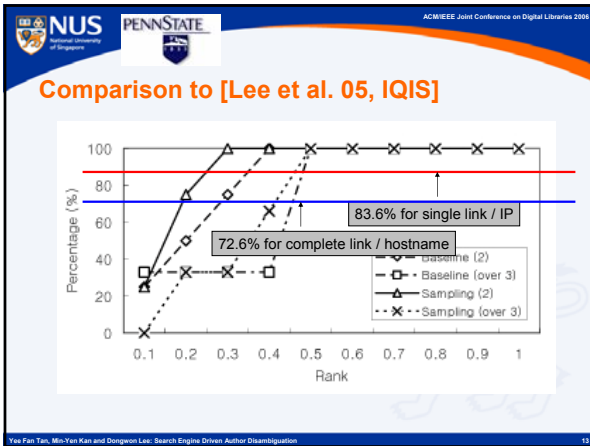
NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006

## Evaluation

Method	Hostname	Domain	IP address
Single link	0.827	0.801	0.836
Complete link	0.726	0.798	0.734
Group average	0.806	0.810	0.812

- **Single link performs best**
  - Good for clustering citations from different publication pages together (some pages list only selected publications)
  - Some authors have disparate research areas, not well represented by a centroid vector
- **Resolving hostnames to IP addresses give best accuracy**

Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 12



- NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006
- ### Discussion
- **Apparent correlation between accuracy and average number of URLs returned per citation**
    - Author names with few URLs tend to fare poorly since results are mainly aggregator web sites
  - **We do not observe any apparent relation between accuracy and number of citations for an author name**
    - Our algorithm is scalable for large number of citations
  - **Analysis of returned URLs is very fast, execution time is dominated by search engine querying**
    - Querying may already be done while spidering, so our algorithm is time-efficient
- Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 15

- NUS PENNSTATE ACM/IEEE Joint Conference on Digital Libraries 2006
- ### Conclusion
- **Summary**
    - We focused on using URLs returned from searching citation titles
    - Respectable average accuracy of 0.836 using IP addresses with single link HAC clustering
  - **Future work**
    - Explore other sources of information, such as the publication venues of the citations as well as utilizing the actual contents of the web pages
    - Combine knowledge gained externally and internally to obtain improved performance
- Yee Fan Tan, Min-Yen Kan and Dongwon Lee: Search Engine Driven Author Disambiguation 16