



# Efficient Web-Based Linkage of Short to Long Forms

**Yee Fan Tan<sup>1</sup>, Ergin Elmacioglu<sup>2</sup>,  
Min-Yen Kan<sup>1</sup>, and Dongwon Lee<sup>2</sup>**

<sup>1</sup>National University of Singapore, Singapore

<sup>2</sup>The Pennsylvania State University, USA

{tanyeefa,kanmy}@comp.nus.edu.sg

{ergin,dongwon}@psu.edu

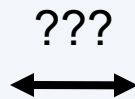
# Linkage of Short to Long Forms

- [7] S. Brin. Extracting patterns and relations from the World Wide Web. In WebDB, 1998.
- [8] M. J. Cafarella and O. Etzioni. A search engine for natural language applications. In WWW, 2005.
- [9] J. P. Callan and M. Connell. Query-based sampling of text databases. ACM TOIS, 19(2):97–130, 2001.
- [10] J. P. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In SIGMOD, 1999.
- [11] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In SIGIR, 1995.

SYMBOL	LAST	CHANGE (%)
MSFT	28.32	+0.01 (+0.04%)
GOOG	585.80	+2.80 (+0.48%)
ORCL	22.84	+0.05 (+0.22%)
SAP	54.60	+0.00 (+0.00)
SAPGF	55.05	+2.27 (+4.29%)
YHOO	26.76	-0.31 (-1.15%)
AMZN	81.62	+1.27 (+1.58%)

Short forms (*SF*)

ACSAC  
KDD  
KDID  
UbiComp  
WebDB



Long forms (*LF*)

Annual Computer Security Applications Conference  
Asia-Pacific Computer Systems Architecture Conference  
Knowledge Discovery and Data Mining  
Knowledge Discovery in Inductive Databases  
Ubiquitous Computing  
International Workshop on Web and Databases

- **For each short form, what long forms does it correspond to?**
  - We use a search engine to perform linkage

# Framework: Web-Based (Record) Linkage

- For each short form  $sf$ 
  - For each long form  $lf$ 
    - Obtain information for  $sf$  and  $lf$  using search engine
    - Compute  $score_{sf}(lf)$  using obtained information
- Rank the long forms according to  $score_{sf}(lf)$

## Search engine evidence

- Each call returns 10 results
- Contains page title, snippet, URL
- Can also download web pages
- How to postprocess?

Our focus!

## Design of queries

1. “ $sf \wedge lf$ ”  ×
  - e.g., Oh and Isahara (2008)
  - not scalable
2. “ $sf$ ” or “ $lf$ ”, or both  ✓
  - e.g., Cimiano et al. (2005)
  - linear number of queries

We only consider “ $sf$ ” or “ $lf$ ” queries

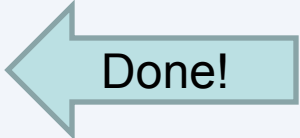
## Related Work: Short/Long Form Extraction

- **Extraction of short and long form pairs from full text articles e.g., Schwartz and Hearst (2003)**

**Scenario template creation (STC)** is the problem of generating a common semantic representation from a set of input articles. [...] Here, we use the term **salient aspect (SA)** to refer to any of such slot fillers that people would regard as important to describe a particular scenario. Figure 1 shows such a manually-built scenario template in which details about important actions, actors, time and locations are coded as slots.

STC is an important task that has tangible benefits for many downstream applications. In the **Message Understanding Conference (MUC)**, manually-generated STs were provided to guide **Information Extraction (IE)**. An ST can also be viewed as regularizing a set of similar articles as a set of attribute/value tuples, enabling multi-document summarization from filled templates.

# Contributions

- **Unify related threads of research in framework** 
- **Method**
  - First to exploit the Web for linking short form to long forms
  - Proposed an effective count-based method for this task
- **Efficiency**
  - Adaptively combine a query probing approach with our count-based method
- **Evaluation**
  - Three datasets on three different domains



# Our Count-based Methods

## 1. $count(sf \rightarrow lf)$

Example:  $sf$  = "HGP",  $lf$  = "Human Genome Project"

**Human Genome Project** - Wikipedia, the free encyclopedia  
"Genomes: 15 Years Later A Perspective by Charles DeLisi, HGP Pioneer". Human Genome News 11: 3-4. Retrieved 2005-02-03. White House Press Release. ...

**More on the sequencing of the human genome**  
The international **Human Genome Project** (HGP) and Celera Genomics ... Uses of the HGP genome assemblies in Celera genome assemblies and impact on assembly. ...

**More on the sequencing of the human genome -- Waterston et al. 100 ...**  
Approximately 60% of the underlying sequence data and 100% of the mapping data used in Celera's analysis came from the HGP, and the HGP genome assembly ...

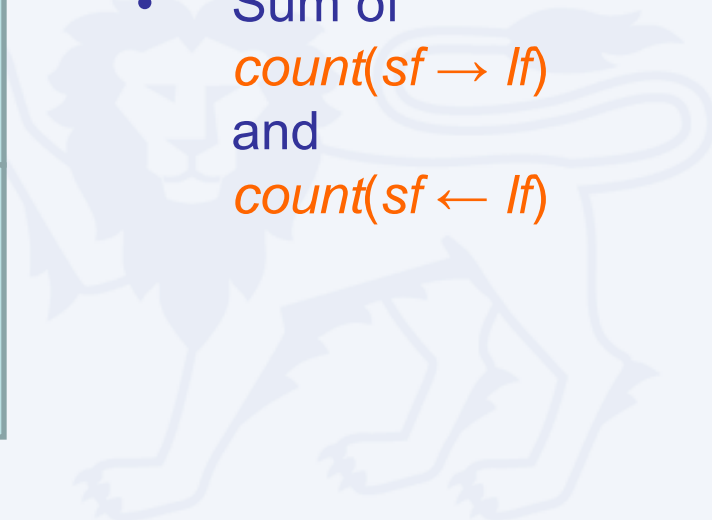
$$count(sf \rightarrow lf) = 2$$

## 2. $count(sf \leftarrow lf)$

- Interchange the roles of  $sf$  and  $lf$  in  $count(sf \rightarrow lf)$

## 3. $count(sf \leftrightarrow lf)$

- Sum of  $count(sf \rightarrow lf)$  and  $count(sf \leftarrow lf)$



# Comparison with Other Methods

## DBLP Computer Science Conferences and Workshops

KDD	Knowledge Discovery and Data Mining
KDID	Knowledge Discovery in Inductive Databases
UbiComp	Ubiquitous Computing
WebDB	International Workshop on Web and Databases

## NASDAQ Composite

AAPL	Apple Inc.
CSCO	Cisco Systems, Inc.
DELL	Dell Inc.
MSFT	Microsoft Corporation

## Human Genome Acronym List

MALDI	matrix-assisted laser desorption ionization
Mb	megabase
MHC	major histocompatibility complex
MIT	Massachusetts Institute of Technology

Search engine: Google

## Schwartz and Hearst (2003)

- Non-web based method that scans full text articles for “*lf (sf)*” patterns

## Inverse Host Frequency (Tan et al., 2006)

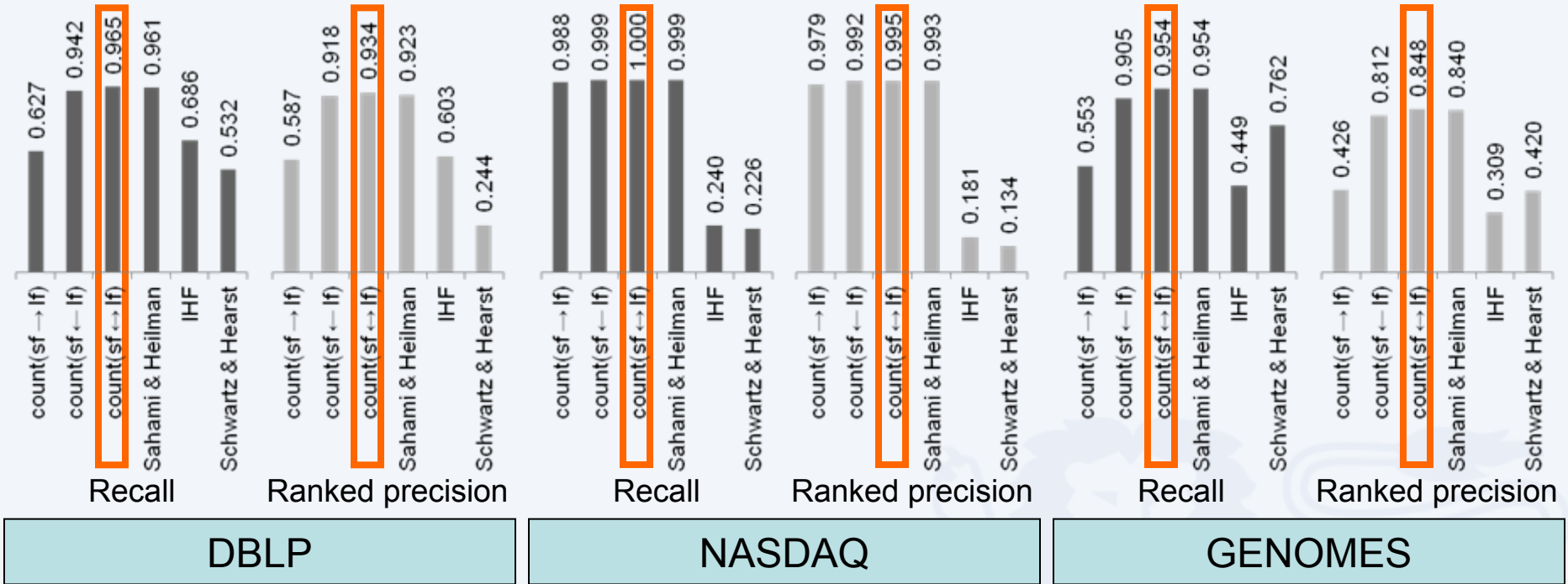
- Similarity between *sf* and *lf* vectors of URL hostnames

## Sahami and Heilman (2006)

- Variation of tf-idf cosine similarity on snippets or web pages

# Results (10 snippets)

*count(sf ↔ lf)* on snippets consistently produced best results!



- *count(sf ← lf)* much better than *count(sf → lf)*
- *count(sf ↔ lf)* is slightly better than Sahami and Heilman
  - *count(sf ↔ lf)* is simpler and faster for same number of queries
- Schwartz and Hearst, and IHF are not as competitive

# Adaptive Combination with Query Probing

- How to reduce the number of **time-consuming search engine queries**?

## Adaptive combination

- For each short form
  - Let **faster but weaker method** ( $M_w$ ) rank the long forms
  - If heuristic determines that  $M_w$  gives a poor ranking
    - Have **stronger but slower method** ( $M_s$ ) rank the long forms

- $M_w$ : Query probing
- $M_s$ : Count-based method

- Query probing

- Joint Conference on **Digital Libraries**
- European Conference on **Digital Libraries**
- **Digital Libraries**



- JCDL
- ECDL
- DL

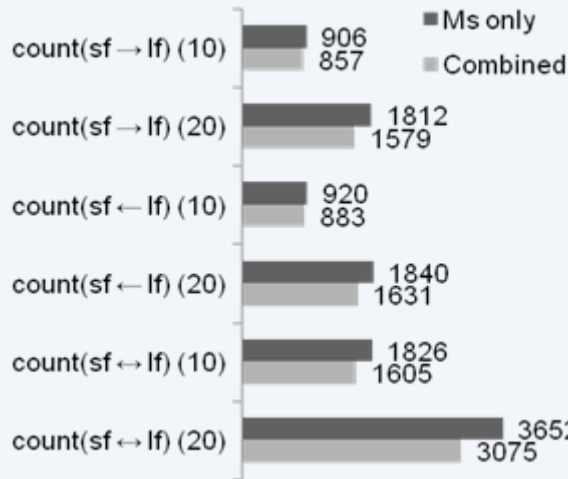
## Query probing

- Query frequently occurring  $n$ -grams from long form list
- $score_p(sf, lf) =$  number of results containing both  $sf$  and  $lf$

## Heuristic

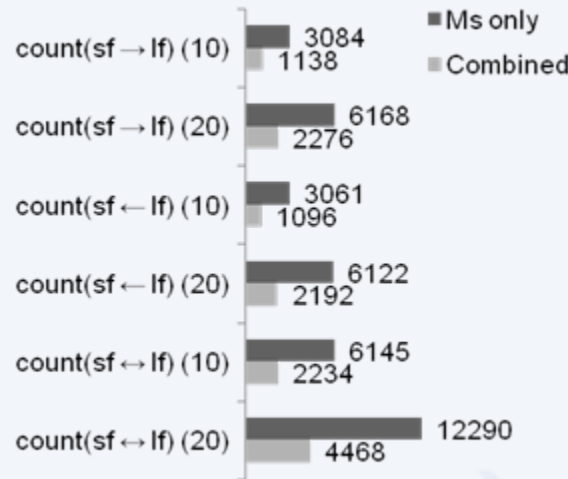
- Any  $lf$  with  $score_p(sf, lf) > 0$ ?

# Results



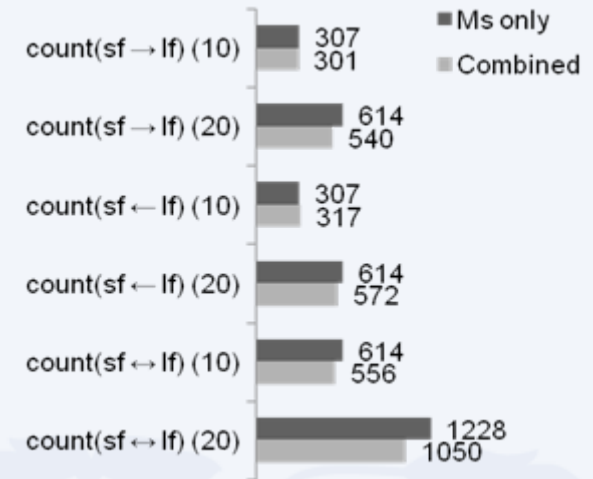
Number of search engine calls

**DBLP**

 Recall: -0.033 to +0.016  
 Ranked precision: -0.066 to +0.021


Number of search engine calls

**NASDAQ**

 Recall: -0.023 to -0.004  
 Ranked precision: -0.044 to -0.014


Number of search engine calls

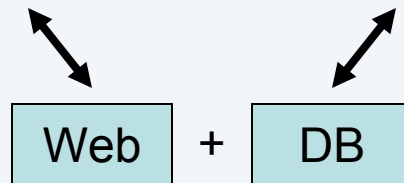
**GENOMES**

 Recall: -0.036 to +0.009  
 Ranked precision: -0.059 to +0.004

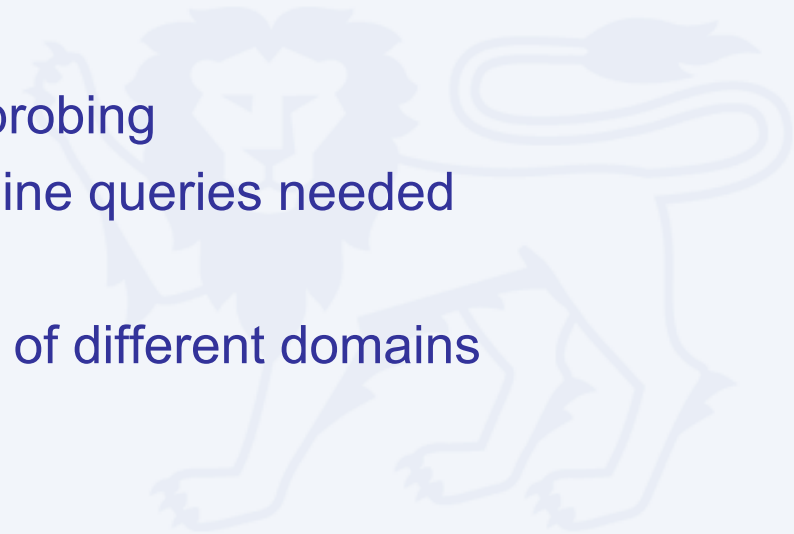
**Reduced number of search engine queries while maintaining linkage performance!**

# Conclusion

- Efficient **Web-Based Linkage** of Short to Long Forms



- **Effective**
  - Count-based method
- **Efficient**
  - Adaptive combination with query probing
  - Reduce the number of search engine queries needed
- **Evaluation**
  - Shown effective on three datasets of different domains



# Thank You

