Towards Breaking the Quality Curse. 
A Web-Querying Approach to 
Web People Search.

Dmitri V. Kalashnikov
Rabia Nuray-Turan
Sharad Mehrotra
Dept of Computer Science
University of California, Irvine
Searching for People on the Web

- Approx. 5-10% of internet searches look for specific people by their names [Guha&Garg’04]

- **Challenges:**
  - *Heterogeneity of representations:*
    - Same entity represented multiple ways.
  - *Ambiguity in representations:*
    - Same representation of multiple entities.

- WePS systems address both these challenges.
Search Engines v WePS

WePS system return clusters of web pages wherein pages of the same entity are clustered together

Google results for “Andrew McCallum”

WePS results for “Andrew McCallum”

**Advantages:**
*Improved Precision, Ability for user to refine the query and filter away pages of irrelevant namesakes, resolves “Famous Person Problem”*
Clustering Search Engines v WePS

Examples of clustering search engines: Clusty – http://www.clusty.com
Kartoo – http://www.kartoo.com

Search results of “Andrew McCallum” using Clusty

WePS system results of “Andrew McCallum”

Differences:
While both return clustered search results, Clustering search engines cluster similar/related pages based on broad topics (e.g., research, photos) while WePS system cluster based on disambiguating amongst various namesakes.
WePS System Architecture

*Classified based on where references are disambiguated*

- **Client side solution**
  - the client resolves references

- **Proxy Solution**
  - Proxy resolves references

- **Server side solution**
  - The server resolves references
Current WePS Approaches

- **Commercial Systems:**
  - Spock – [http://www.spock.com](http://www.spock.com)

- **Multiple research proposals**
  - [Kamha et al. 04; Artiles et. al ’07 &’05; Bekkerman et. al. ’05; Bollegala et. al. ’06; Matsou et. al. 06] etc.

- **Web people search task included in Sem-Eval workshop in 2007**
  - 16 teams participated in the task

- **WePS techniques differ from each other in:**
  - data analyzed for disambiguation & Clustering approach used

**Type of data analyzed**

- **Baseline**
  - exploit information in result pages only
  - Keywords, named entities, hyperlinks, emails, …
  - [Bekkerman et.al. 05], [Li, et. al. 05][Elmachioglu, et. al. 07]

- **exploit external information**
  - Ontologies, Encyclopedia, …
  - [Kalashnikov, et. al, 07]

- **Exploit Web content**
  - Crawling
    - [Kanani et. al. 07]
  - Search engine Statistics
Contributions of this paper

• Exploiting Co-occurrence statistics from the search engine for WePS.

• A novel classification approach based on skylines to support reference resolution.

• Outperforms existing techniques on multiple data sets
Approach Overview

Top-K Webpages

Preprocessing

1. Extract NEs
2. Filter

Preprocessed pages

Initial Clustering

1. Compute similarity using TF/IDF on NE
2. Cluster based on TF/IDF similarity

Post processing

1. Rank clusters
2. Extract cluster sketches
3. Rank pages in clusters

Final clusters

Refining Clusters using Web Queries

1. Web Query
2. Feature Extraction
3. Skyline based classifier
4. Cluster refinement
Preprocessing: Named Entity Extraction

- Locate query name on web page
- Extract $M$ neighboring names entities
  - organization
  - people Names
- Stanford Named Entity Recognizer & GATE used
Preprocessing: Filtering Named Entities

**Input:** Extracted NEs

- **Location Filter**
  - Remove location NEs by looking up a dictionary

- **Ambiguous name Filter**
  - Remove one-word person names, such as ‘John’

- **Ambiguous Org. Filter**
  - Remove common English words

- **Same last name Filter**
  - Remove the person names with the same last name.

**Output:** Cleaned NEs
Initial Clustering

- **Intuition:**
  - If two web pages “significantly” overlap in terms of associated NEs, then the two pages refer to the same individual.

- **Approach:**
  - Compute similarity between web pages based on TF/IDF over Named Entities
  - Merge all pairs of Web pages whose TF/IDF similarity exceeds a threshold
  - Threshold learnt over the training dataset.
Cluster Refinement

- Initial Clustering based on TF/IDF similarity:
  - **Accurate** – references resolved are usually correct.
  - **Conservative** – too many unresolved references

Key Observation:
- Web may contain additional information that may further resolve unresolved references.
  - Co-occurrence between contexts associated with unresolved references

Co-occurrence statistics can be learnt by querying the search engine
Using Co-occurrence to Disambiguate

What if context NEs $C_i$ and $C_j$ are very “general”
• The co-occurrence counts could be high for all namesakes!

Solution -- Normalize $|N \cap C_i \cap C_j|$ using Dice similarity

Merge Web pages $D_i$ and $D_j$ based on $|N \cdot C_i \cdot C_j|/|N|$?
Dice Similarity

\[ Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \]

• Two possible versions of dice:

\[ Dice_1(N \cdot C_i, N \cdot C_j) = \frac{2|N \cdot C_i|}{|N \cdot C_i| + |C_i \cdot C_j|} \]

\[ Dice_2(N, C_i \cdot C_j) = \frac{2|N \cdot C_i \cdot C_j|}{|N| + |C_i \cdot C_j|} \]
Dice2 Similarity -- Example

Let an $N O_1 O_2$ query be:

Q: “Andrew McCallum” AND “Yahoo” AND “Google”

Let counts be:

- “Andrew McCallum” AND “Yahoo” AND “Google” – 522
- “Andrew McCallum” AND “Yahoo” – 2470
- “Andrew McCallum” AND “Google” – 8950
- “Andrew McCallum” – 38500
- “Yahoo” AND “Google” – 73000000

Dice:

$Dice_1 = 0.05 \quad Dice_2 = 7.15e^{-6}$

- $Dice_2$ captures that Yahoo and Google are too unspecific Organization Entities to be a good context for “Andrew McCallum”
- $Dice_2$ leads to visibly better results in practice!
Web Queries to Gather Co-Occurrence Statistics

**People NEs**
- Lise Getoor
- Ben Taskar

**Organization**
- UMASS
- ACM

**People NEs**
- Charles A. Sutton
- Ben Wellner

**Organization**
- DBLP
- KDD

**Context**
- Association between context
  - (L.G. OR B.T.) AND (C.A.S. OR B.W.)

**Associationamongst Context within pages relevant to original query N**
- N AND Q1 (#97)
- N AND Q2 (#2,570)
- N AND Q3 (#720)
- N AND Q4 (#2,190K)

**8 web queries**

**Ex:** Q1 = ("Lise Getoor" OR "Ben Taskar") AND ("Charles A. Sutton" OR "Ben Wellner")
## Creating Features from Web Co-Occurrence Counts

### 8-Feature Vector

<table>
<thead>
<tr>
<th>Feature</th>
<th>Formula</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>$</td>
<td>N P_i P_j</td>
</tr>
<tr>
<td>2:</td>
<td>$\text{Dice}(N, P_i P_j)$</td>
<td>$0.000918$</td>
</tr>
<tr>
<td>3:</td>
<td>$</td>
<td>N P_i O_j</td>
</tr>
<tr>
<td>4:</td>
<td>$\text{Dice}(N, P_i O_j)$</td>
<td>$0.000561$</td>
</tr>
<tr>
<td>5:</td>
<td>$</td>
<td>N O_i P_j</td>
</tr>
<tr>
<td>6:</td>
<td>$D(N, O_i P_j)$</td>
<td>$0.00073$</td>
</tr>
<tr>
<td>7:</td>
<td>$</td>
<td>N O_i O_j</td>
</tr>
<tr>
<td>8:</td>
<td>$D(N, O_i O_j)$</td>
<td>$0.008552$</td>
</tr>
</tbody>
</table>
Role of a Classifier

- Classifier used to mark edges as “merge” (+) or “do not merge” (-) resulting in the final cluster.

- Since the features are “similarity” based, the classifier should satisfy the dominance property
Dominance Property

• Let \( \mathbf{f} \) and \( \mathbf{g} \) be 8D feature vectors
  – \( \mathbf{f} = (f_1, f_2, \ldots, f_8) \)
  – \( \mathbf{g} = (g_1, g_2, \ldots, g_8) \)

• Notation
  – (\textbf{f up dominates} \textbf{g}) if \( f_1 \leq g_1, f_2 \leq g_2, \ldots, f_8 \leq g_8 \)
  – (\textbf{f down dominates} \textbf{g}) if \( f_1 \geq g_1, f_2 \geq g_2, \ldots, f_8 \geq g_8 \)

• Dominance Property for classifier
  – Let \( \mathbf{f} \) and \( \mathbf{g} \) be the features associated with edges \((d_i, d_j)\)
    and \((d_m, d_n)\) respectively.
  – If \( \mathbf{f} \leq \mathbf{g} \) and the classifier decides to merge \((d_i, d_j)\), then it
    must also merge \((d_m, d_n)\)
Skyline-Based Classification

- A skyline classifier consists of an up-dominating classification skyline
  - All points “above” or “on” it are declared “merge” (“+”)
  - The rest are “don’t merge” (“-”)
  - Automatically takes care of dominance

Our goal is to learn a classification skyline that supports a good final clusters of search results.
Greedy Approach

- Initialize classification skyline with the point that dominates the entire space.
- Pool of points to add to CS
  - Down-dominating skyline on active “+” edges
- Maintain “current” clustering
  - Clustering induced by current CS
- Choosing a point from pool to CS that
  - **Stage1**: Minimizes loss of quality due to adding new “-” edges
  - **Stage2**: Maximizes overall quality
- Iterate until done
- **Output best skyline discovered**
Example: Adding a point to CS

Current CS = \{q,d,c\}
Point Pool = \{f,e,u,h\}
Example: Considering f

What if add f to CS?

Stage 1: Adding edge k
Fp = 0.805

Stage 2: Adding edge f
Fp = 0.860
Example: Considering e

What if add e to CS?

Stage 1: Adding edges k and v  
\[ \text{Fp} = 0.718 \]

Stage 2: Adding edge e  
\[ \text{Fp} = 0.774 \]
Example: Considering u

Stage 1: Adding edge i
Fp = 0.805

Stage 2: Adding edge u
Fp = 0.833

What if add u to CS?
Example: Considering h

What if add h to CS?

Stage 1: Adding edges i and t
Fp = 0.625
Stage 2: Adding edge h
Fp = 0.741
Example: Stage 2

After Stage 1 only f and u remain

Stage 1 (for f): $F_p = 0.805$  Stage 2 (for f): $F_p = 0.860$
Stage 1 (for u): $F_p = 0.805$  Stage 2 (for u): $F_p = 0.833$

Thus adding f to CS!
Why “Skyline-based” classifier?

• Specialized classifier
  – Designed specifically for the problem at hand
  – Is aware of the clustering underneath

• Fine-tunes itself to a given clustering quality metric
  – Such as $F_1$, $F_P$, $F_B$

• Takes into account the dominance in 8D data
  – Explained shortly

• Takes into account the transitivity of merges
  – Classifying one edge as “+” might lead to classifying multiple edges as “+”
  – For instance when two large clusters are merged
Post Processing

• **Page Ranking**
  - Use the original ranking from search engine

• **Cluster Ranking**
  - each cluster we simply select the highest ranked page and use that as the order of the cluster.

• **Cluster Sketches**
  - coalesce all pages in the cluster into a single page
  - removing the stop words we compute the
  - TF/IDF of the remaining words
  - Select set of terms above a certain threshold (or top N terms)

---

**Post processing**

1. Rank clusters
2. Extract cluster sketches
3. Rank pages in clusters

---

**Refining Clusters using Web Queries**

1. Web Query
2. Feature Extraction
3. Skyline based classifier
4. Cluster refinement
Experiments

• Datasets
  - WWW’05
    • By Bekkerman and McCallum
    • 12 different Person Names all related to each other. Each Person name has different number of namesakes.
  - SIGIR’05
    • By Artiles et al
    • 9 different Person Names, each with different number of namesakes.
  - WEPS
    • For Semeval Workshop on WEPS task (by Artiles et al)
    • Training dataset (49 Different Person Names)
    • Test Dataset (30 Different Person Names)
Quality Measures

- $F_P$: Harmonic mean of purity and inverse purity

\[
Purity = \sum_{A_i \in A} \frac{|A_i|}{|R|} \max_{S_j \in S} \frac{|A_i \cap S_j|}{|A_i|}.
\]

\[
InversePurity = \sum_{S_j \in S} \frac{|S_j|}{|R|} \max_{A_i \in A} \frac{|A_i \cap S_j|}{|S_j|}.
\]

- $F_B$: Harmonic mean of B-cubed precision and recall.

\[
Precision = \sum_{r \in R} w_r \cdot \frac{|A_r \cap S_r|}{|A_r|}.
\]

\[
Recall = \sum_{r \in R} v_r \cdot \frac{|A_r \cap S_r|}{|S_r|}.
\]
Experiments

• Baselines
  – Named Entity based Clustering (NE)
    • TF/IDF merging scheme that uses the NEs in the Web Pages
  – Web-based Co-Occurrence Clustering (WebDice)
    • Computes the similarity between Web pages by summing up the individual dice similarities of People-People and People-Organization similarities of Web pages.
Experiments on WWW & SIGIR Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>SkyLine</th>
<th>WebDice</th>
<th>NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam Cheyer</td>
<td>92.8/85.8</td>
<td>95.7/91.5</td>
<td>86.5/73.9</td>
</tr>
<tr>
<td>Andrew McCallum</td>
<td>95.0/89.5</td>
<td>95.6/93.1</td>
<td>92.5/88.2</td>
</tr>
<tr>
<td>Andrew Ng</td>
<td>88.4/84.1</td>
<td>89.1/84.2</td>
<td>82.4/74.6</td>
</tr>
<tr>
<td>Ann Hill</td>
<td>91.2/89.6</td>
<td>76.3/73.2</td>
<td>87.8/85.0</td>
</tr>
<tr>
<td>Bill Mark</td>
<td><strong>89.2/84.2</strong></td>
<td>86.4/81.1</td>
<td>77.7/69.7</td>
</tr>
<tr>
<td>Brenda Clark</td>
<td>96.6/94.4</td>
<td><strong>97.2/95.5</strong></td>
<td>93.1/89.7</td>
</tr>
<tr>
<td>Christine King</td>
<td>88.1/82.8</td>
<td><strong>90.2/86.2</strong></td>
<td>86.9/91.0</td>
</tr>
<tr>
<td>David Israel</td>
<td><strong>86.9/80.7</strong></td>
<td>79.5/72.4</td>
<td>80.3/73.3</td>
</tr>
<tr>
<td>David Mulford</td>
<td><strong>85.9/78.8</strong></td>
<td>83.9/74.9</td>
<td>77.6/65.9</td>
</tr>
<tr>
<td>Fernando Pereira</td>
<td>82.3/71.9</td>
<td><strong>86.3/77.9</strong></td>
<td>83.3/73.5</td>
</tr>
<tr>
<td>Helen Miller</td>
<td><strong>92.1/91.0</strong></td>
<td>88.7/87.2</td>
<td>90.7/88.5</td>
</tr>
<tr>
<td>Leslie Kaelbling</td>
<td>97.7/95.4</td>
<td><strong>98.3/96.6</strong></td>
<td>86.6/74.0</td>
</tr>
<tr>
<td>Lisa Harris</td>
<td>83.7/79.6</td>
<td><strong>84.5/79.8</strong></td>
<td>84.4/79.8</td>
</tr>
<tr>
<td>Lynn Voss</td>
<td><strong>84.7/79.5</strong></td>
<td>72.0/61.6</td>
<td>66.3/60.5</td>
</tr>
<tr>
<td>Mary Johnson</td>
<td><strong>89.0/88.8</strong></td>
<td>83.5/83.2</td>
<td>79.1/77.3</td>
</tr>
<tr>
<td>Nancy Thompson</td>
<td><strong>93.9/91.8</strong></td>
<td>87.0/80.8</td>
<td>77.7/74.5</td>
</tr>
<tr>
<td>Samuel Baker</td>
<td><strong>89.8/88.6</strong></td>
<td>71.9/67.0</td>
<td>69.5/69.0</td>
</tr>
<tr>
<td>Sarah Wilson</td>
<td><strong>90.3/87.6</strong></td>
<td>84.8/81.2</td>
<td>71.9/68.7</td>
</tr>
<tr>
<td>Steve Hardt</td>
<td>76.6/61.0</td>
<td><strong>84.0/72.1</strong></td>
<td>43.9/30.6</td>
</tr>
<tr>
<td>Tom Mitchell</td>
<td><strong>89.1/86.2</strong></td>
<td><strong>89.8/87.5</strong></td>
<td>86.9/83.5</td>
</tr>
<tr>
<td>William Cohen</td>
<td>87.9/79.2</td>
<td><strong>89.9/82.9</strong></td>
<td>86.1/76.7</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>89.1/84.4</strong></td>
<td>86.4/81.4</td>
<td>80.5/74.2</td>
</tr>
</tbody>
</table>
Experiments on the WEPS Dataset

<table>
<thead>
<tr>
<th>Systems</th>
<th>SkyLine F_P/B</th>
<th>WebDice F_P/B</th>
<th>NE F_P/B</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>85.7/78.8</td>
<td>85.5/78.5</td>
<td>79.4/71.5</td>
</tr>
<tr>
<td>test</td>
<td>84.8/78.7</td>
<td>73.5/63.2</td>
<td>78.0/70.2</td>
</tr>
</tbody>
</table>
Effect of Initial Clustering

![Graph showing the effect of initial clustering on different algorithms. The graph plots Fp against Threshold, with lines representing Skyline, Dice, DTC, and SVM. Each line has a different marker and line style, indicating distinct performance trends.]
Conclusion and Future Work

• Web co-occurrence statistics can significantly improve disambiguation quality in WePS task.
• Skyline classifier is effective in making merge decisions.
• However, current approach has limitations:
  – large number of search engine queries to collect statistics → will work only as a server side solution given limitations today.
  – Based on the assumption that context of different entities is very different. May not work as well on the difficult case when contexts of different entities significantly overlap.