Near-Duplicate Detection for eRulemaking

Hui Yang, Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Stuart Shulman
Library and Information Science
School of Information Sciences
University of Pittsburgh

dg.o conference 2006

Duplicates and Near-Duplicates

Looks like Not, BUT, YES!

Looks like Yes, But, NO!
Duplicates and Near-Duplicates in eRulemaking

• U.S. regulatory agencies must solicit, consider, and respond to public comments.
• Special interest groups make form letters available for generating comments via email and the Web
  – GetActive, http://www.getactive.org
• Modifying a form letter is very easy

Protect our Kids from Mercury

At the request of his friends in the energy industry, the Bush administration has directed the EPA to weaken a proposal that would have required significant reductions in mercury pollution from power plants. Exposure to mercury causes neurological and developmental problems in babies and small children. That's why we believe:"The EPA must protect our children by making all power plants install controls to step mercury pollution by 2008."

Speak Out

You can submit comments to the EPA from President Bush's friends in the energy industry. The EPA must act to protect our children from mercury pollution. For further resources on mercury, go to http://www.epa.gov.

1. Speak Out

Send your comment to the EPA. Enter your name, city, state, and zip code. You can copy and paste your comments into the form. We’ll deliver these messages to the EPA.

Individual Information

Form Letter

Personal Notes

YOUR COMMENT TO THE EPA

TO: Environmental Protection Agency
               CC: Your Representative and Senator
               FROM: (Your Name and Email)
               SUBJECT: Protect our Kids from Mercury

Dear Environmental Protection Agency,

I support Mercury2.org's mission to end mercury pollution by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution control technologies.

(Your personal note)

Send statement
Duplicates and Near-Duplicates in eRulemaking

• Some popular regulations attract hundreds of thousands of comments
• Very labor-intensive to sort through manually
• Goal:
  – Achieve highly effective near-duplicate detection by incorporating additional knowledge;
  – Organize duplicates for browsing.

What is a Duplicate in eRulemaking? (Text Documents)
Duplicate and Near-Duplicates

• Exact Copies of a form letter are easy to detect
• Non-Exact Copies are modified form letters are harder to process
  – They are similar, but not identical
  • “near duplicates”

Duplicate - Exact

| The EPA should require power plants to cut mercury pollution by 90% by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution-control technology. |
|---|---|
| The EPA should require power plants to cut mercury pollution by 90% by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution-control technology. |
The EPA should require power plants to cut mercury pollution by 90% by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution-control technology.

I urge the EPA to require controls at all power plants to stop mercury pollution. The health of our air and water, and especially our children is by far more important than any political agenda.

The EPA should require power plants to cut mercury pollution by 90% by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution-control technology.

I am writing to urge you to take prompt action to clean up mercury and other toxic air pollution from power plants. EPA’s current proposals allow far ore mercury pollution than what the Clean Air Act allows, while at the same time fail to address over sixty other hazardous air pollutants like dioxin.

I am writing to urge you to take prompt action to clean up mercury and other toxic air pollution from the power plants. EPA’s proposals permit far more mercury pollution than what the Clean Air Act allows, while at the same time fail to address over sixty other hazardous air pollutants like dioxin.
As someone who cares about protecting the health of children and our environment, I am deeply concerned about the mercury contamination of our lakes and streams. Mercury descends from polluted air into water and then works its way up the food chain. It is especially dangerous to people and wildlife that consume large amounts of fish.

I urge you to reconsider your approach and require power plants to reduce their emissions of mercury to the greatest extent possible. This is what the federal law requires, and also what the people and wildlife of this country deserve.

Thank you for your consideration.

As someone who cares about protecting our wildlife and wild places, I am deeply upset about the mercury contamination of our lakes and streams. Mercury descends from polluted air into water and then works its way up the food chain. It is especially dangerous to people and wildlife that consume large amounts of fish.

Specifically, I am concerned that EPA is proposing a cap-and-trade system to manage mercury emissions. Under such a system, not all plants would have to reduce their harmful emissions of mercury and some could even increase! This approach is unacceptable for dealing with such a toxic pollutant – which is precisely why the Clean Air Act does not allow it. I urge you to reconsider your approach and require power plants to reduce their emissions of mercury to the greatest extent possible. This is what the federal law requires, and also what the people and wildlife of this country deserve.

Thank you for your consideration.

Dear Environmental Protection Agency,

The EPA should require power plants to cut mercury pollution by 90% by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution-control technology.

I urge the EPA to require controls at all power plants to stop mercury pollution. The health of our air and water, and especially our children is by far more important than any political agenda.

Dear Environmental Protection Agency,

I urge the EPA to require controls at all power plants to stop mercury pollution. The health of our air and water, and especially our children is by far more important than any political agenda.
The EPA should require power plants to cut mercury pollution by 90% by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution-control technology.

American citizens need to stand up for their rights. Which means the freedom to pursue life, liberty, health, and happiness. Everybody has the right to wake up each morning and breathe the freshest air that this green earth can provide us, not what some government organization says that we need to put up with because they want their standards so lax. This is a democracy, by the people, for the people, not what Bush decides because it suits his mood. It concerns me to see so many people that I care about every day trying so hard to live with the mental and neurological problems that they have acquired, or were born with due to mercury poisoning.

The EPA should require power plants to cut mercury pollution by 90% by 2008. These reductions are consistent with national standards for other pollutants and achievable through available pollution-control technology.
Related Work

• Duplicate Detection Using Fingerprints
  – Hashing functions [SHA1][Rabin]
  – Fingerprint granularity [Shivakumar et al.’95]
    [Hoad & Zobel ’03]
  – Fingerprint size [Broder et al. ’97]
  – Substring selection strategy
    • position-based [Brin et al. ’95]
    • hash-value-based [Broder et al. ’97]
    • anchor-based [Hoad & Zobel ’03]
    • frequency-based [Chowdhury et al. ’02]

• Duplicate Detection Using Full-Text [Metzler et al. ’05]

Our Detection Strategy

• Group Near-duplicates based on
  – Text similarity
  – Editing patterns
  – Metadata
Document Clustering

- Put similar documents together
- How is text similarity defined?
  - Similar Vocabulary
  - Similar Word Frequencies

\[
dist(d_a, d_b) = \min(KL(p_a \parallel p_b), KL(p_b \parallel p_a))
\]

- If two documents similarity is above a threshold, put them into same cluster

Incorporating Instance-level Constraints in Clustering

- Key Block are very common
- Typical text similarity doesn’t work
  - Different words, different frequencies
- Solution: Add instance-level constraints
  - Example: must-link, cannot-link, family-link
  - These provide hints to the clustering algorithm about how to group documents
Must-links

- Two instances **must** be in the same cluster
- Created when
  - complete containment of the reference copy (*key block*),
  - word overlap > 95% (*minor change*).

Cannot-links

- Two instances **cannot** be in the same cluster
- Created when two documents
  - cite different docket identification numbers
    - People submitted comments to wrong place
Family-links

• Two instances are likely to be in the same cluster
• Created when two documents have
  – the same email relayer,
  – the same docket identification number,
  – similar file sizes, or
  – the same footer block.

How to Incorporate Instance-level Constraints?

• When forming clusters,
  – if two documents have a must-link, they must be put into same group, even if their text similarity is low
  – if two documents have a cannot-link, they cannot be put into same group, even if their text similarity is high
  – if two documents have a family-link, increase their text similarity score, so that their chance of being in the same group increases.
Evaluation

Evaluation Methodology

• We created three 1,000 email subsets
  – Two from the EPA’s Mercury dataset
docket: (USEPA-OAR-2002-0056)
  – One from DOT’s SUV dataset
docket: (USDOT-2003-16128)
• Assessors manually organized
documents into near-duplicate clusters
• Compare human-human agreement
to human-computer agreement
Experimental Setup

• **Sample Name:** NTF
• # of Docs: 1000
• # of Docs (duplicates removed): 275
• # of Known form letters: 28
• # of Assessors: 2
• Assessor 1: UCSUR13
• Assessor 2: UCSUR16

---

Experimental Setup

• **Sample Name:** NTF2
• # of Docs: 1000
• # of Docs (duplicates removed): 270
• # of Known form letters: 26
• # of Assessors: 2
• Assessor 1: UCSUR8
• Assessor 2: UCSUR9
Experimental Setup

- **Sample Name:** DOT
- **# of Docs:** 1000
- **# of Docs (duplicates removed):** 270
- **# of Known form letters:** 4
- **# of Assessors:** 2
- **Assessor 1:** SUPER (Stuart)
- **Assessor 2:** G (Grace)

Experimental Results

- Comparing with human-human intercoder agreement (measured in AC1)

<table>
<thead>
<tr>
<th></th>
<th>Macro Average (Averaged by Cluster)</th>
<th>Micro Average (Averaged by Document)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NTF</td>
<td>NTF2</td>
</tr>
<tr>
<td>Coder A / Coder B</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>Coder A / DURIAN</td>
<td>0.92</td>
<td>0.80</td>
</tr>
<tr>
<td>Coder B / DURIAN</td>
<td>0.90</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Experimental Results
- Comparing with other duplicate detection Algorithms (measured in F1)

<table>
<thead>
<tr>
<th></th>
<th>NTF</th>
<th>NTF2</th>
<th>DOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>DSC</td>
<td>0.81</td>
<td>0.80</td>
<td>0.70</td>
</tr>
<tr>
<td>I-Match</td>
<td>0.69</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>DURIAN</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Impact of Instance-level Constraints

- Number of Constraints vs. F1.
Impact of Instance-level Constraints

- Number of Constraints vs. F1.

**Graph:**
- Baseline
- Must
- Cannot
- Family
- Must+Cannot
- All

**Axes:**
- F1 on the y-axis
- DOT on the x-axis

**Data Points:**
- Various numbers indicating the relationship between the number of constraints and F1 score.
**Conclusion**

- Near-duplicate detection on large public comment datasets is practical
  - Automatic metadata extraction
  - Feature-based document retrieval
  - Instance-based constrained clustering
- Efficient
- Easily applied to other datasets
Please come to our demo (or ask us for one)

Questions?