

Content Extraction & Duplicate Analysis & Recognition (CEDAR)

Susan Lubar

781-271-2860 • slubar@mitre.org

Army MOIE



Problem

- **Analysts review open source information from many different Web sites to produce reports.**
- **Approximately 30% of documents on the Web are duplicates or near-duplicates.**
- **Duplicates result in wasted labor and missed information during analysis.**

Background

North Korea said to have fueled missile for test launch
Satellite photos show accelerating preparations

By Carol Giacomo, Reuters | June 19, 2006

WASHINGTON — North Korea is believed to have completed fueling a missile capable of reaching Alaska, raising the probability of an imminent test launch, US officials said yesterday.

The United States plans to join Japan in a sharp response if the test goes ahead.

Washington has warned Pyongyang against the launch in a message passed to North Korean diplomats at the United Nations but there was no response, American officials told Reuters.

The officials, speaking on condition of anonymity, said Pyongyang could still decide to scrap the launch, but that was unlikely given the complexity of siphoning fuel back out of a missile prepared for launch.

The test is expected to involve a Taepodong-2 missile with an estimated range of 2,175 to 2,670 miles.

ADVERTISMENT: Get Globe home delivery at 50% off and receive a DUNKIN' DONUTS® Card for up to \$40!

ARTICLE TOOLS: PRINT FRIENDLY, EMAIL TO A FRIEND, WORLD WIDE WEB FEED, REPRINTS & LICENSING

MORE: Check World stories, Latest world news

LATEST WORLD NEWS:

- Green groups call for buffer against whaling jobs
- UN opens new chapter in struggle for human rights
- Norway and Iceland PMs applaud pro-whaling vote
- Mexico forgets Zapata's demands at election
- Luxury brands take baby steps in India



North Korea said to have fueled missile for test launch

Satellite photos show accelerating preparations

By Carol Giacomo, Reuters | June 19, 2006

WASHINGTON — North Korea is believed to have completed fueling a missile capable of reaching Alaska, raising the probability of an imminent test launch, US officials said yesterday.

The United States plans to join Japan in a sharp response if the test goes ahead.

Washington has warned Pyongyang against the launch in a message passed to North Korean diplomats at the United Nations but there was no response, American officials told Reuters.

The officials, speaking on condition of anonymity, said Pyongyang could still decide to scrap the launch, but that was unlikely given the complexity of siphoning fuel back out of a missile prepared for launch.

The test is expected to involve a Taepodong-2 missile with an estimated range of 2,175 to 2,670 miles.

North Korea said to have fueled missile for test launch
Satellite photos show accelerating preparations

By Carol Giacomo, Reuters | June 19, 2006

WASHINGTON — North Korea is believed to have completed fueling a missile capable of reaching Alaska, raising the probability of an imminent test launch, US officials said yesterday.

The United States plans to join Japan in a sharp response if the test goes ahead.

Washington has warned Pyongyang against the launch in a message passed to North Korean diplomats at the United Nations but there was no response, American officials told Reuters.

The officials, speaking on condition of anonymity, said Pyongyang could still decide to scrap the launch, but that was unlikely given the complexity of siphoning fuel back out of a missile prepared for launch.

The test is expected to involve a Taepodong-2 missile with an estimated range of 2,175 to 2,670 miles.

ADVERTISMENT: nytimes.com World Cup 06 Complete coverage of the world's most popular sporting event

ARTICLE TOOLS: PRINT FRIENDLY, EMAIL TO A FRIEND, WORLD WIDE WEB FEED, REPRINTS & LICENSING

MORE: Check World stories, Latest world news

LATEST WORLD NEWS:

- Green groups call for buffer against whaling jobs
- UN opens new chapter in struggle for human rights
- Norway and Iceland PMs applaud pro-whaling vote
- Mexico forgets Zapata's demands at election
- Luxury brands take baby steps in India



North Korea said to have fueled missile for test launch

Satellite photos show accelerating preparations

By Carol Giacomo, Reuters | June 19, 2006

WASHINGTON — North Korea is believed to have completed fueling a missile capable of reaching Alaska, raising the probability of an imminent test launch, US officials said yesterday.

The United States plans to join Japan in a sharp response if the test goes ahead.

Washington has warned Pyongyang against the launch in a message passed to North Korean diplomats at the United Nations but there was no response, American officials told Reuters.

The officials, speaking on condition of anonymity, said Pyongyang could still decide to scrap the launch, but that was unlikely given the complexity of siphoning fuel back out of a missile prepared for launch.

The test is expected to involve a Taepodong-2 missile with an estimated range of 2,175 to 2,670 miles.

Key

- - Core Content
- - Extraneous Content

Core content is identical even though extraneous content is not

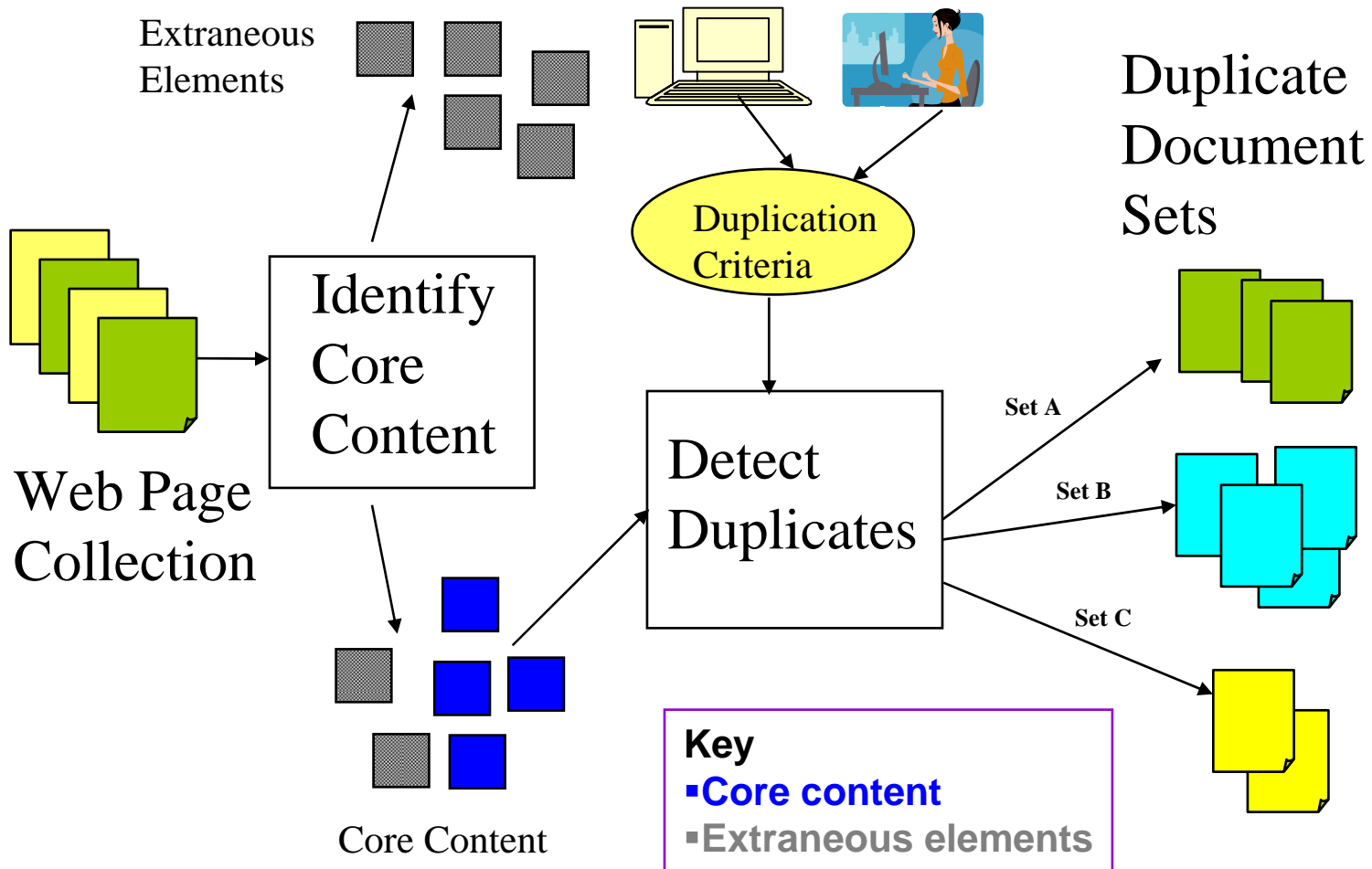
Objective

- **Produce a system to identify duplicate core content within Web pages**
 - **Use heuristics along with a statistical classifier to identify core content in Web pages**
 - **Establish criteria to define what should be considered near-duplicates in Web pages**
 - **Automatically identify duplicates and near-duplicates based on core content extracted**

Activities

- **Create Gold Standard Data Set**
 - **1621 documents collected from 28 sources**
 - **Core content identified**
 - **Duplicates and near-duplicates identified**
 - **Documents and duplicate information manually reviewed**
- **Automatically identify core content in Web pages**
- **Process core content of documents to determine duplicates**

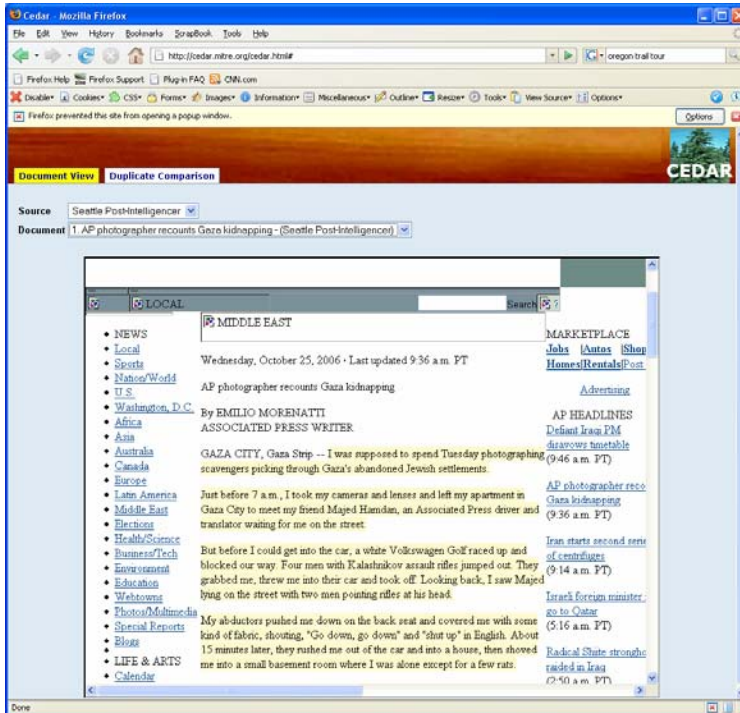
Highlight



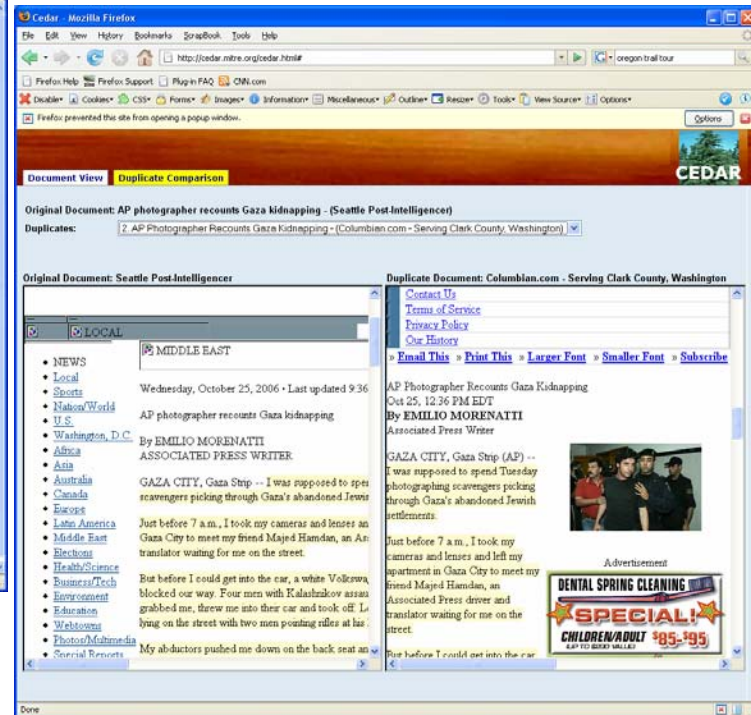
Grouping duplicates simplifies the review of open source information

Demonstration

Content Extraction View



Duplicates View



Viewing core content and duplicates identified by CEDAR

MITRE

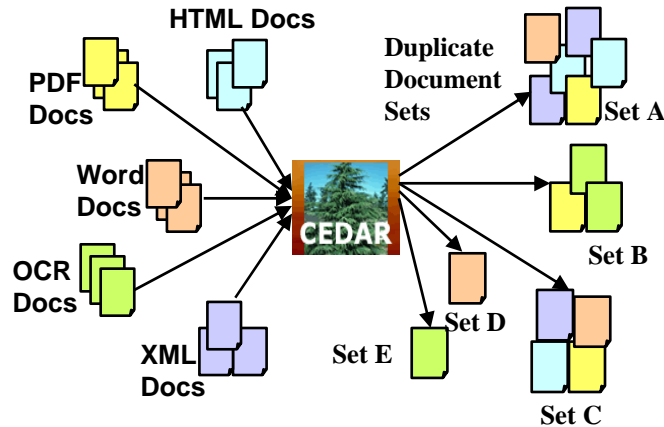
© 2007, The MITRE Corporation

Impacts

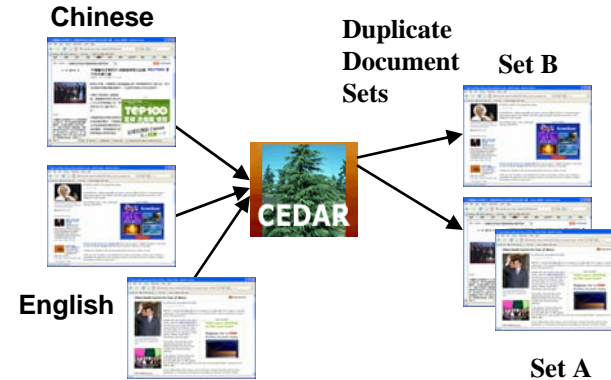
- **Improve quality of analyses**
 - **Timeliness: Less time spent sifting through duplicates**
 - **Accuracy: More complete review of source material**
 - **Relevance: Easier to ensure documents are relevant to analysis**

Future Plans

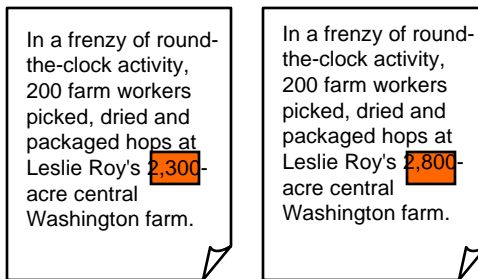
Find Duplicates Across Multiple Document Types



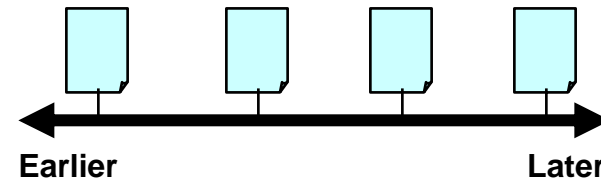
Find Duplicates Across Multiple Languages



Identify Changes in Names, Locations, Numbers



Order Duplicates by Publication Time



Ideas for extending CEDAR's capabilities