Dynamic Coordination of Information Management Services for Processing Dynamic Web Content

In-Young Ko, Ke-Thia Yao, and Robert Neches

\{IKo, KYao, RNeches\}@isi.edu

Information Sciences Institute
University of Southern California

4676 Admiralty Way
Marina del Rey, CA 90292, U.S.A.

WWW2002, Honolulu, HI
May 9, 2002
Example Process of Dynamic Web Content

GeoTopics Portal
Daily News Reportage Analysis
www.isi.edu/geoworlds/geotopics

- Helps users identify the “hot topics,” and the most frequently referenced places
- Allows users to keep track of the hot topics and places
- Shows a breakdown of a topic by places and a breakdown of a place by topics

Top 20 Topics
1. Gold Medal
2. George W. Bush
3. Winter Olympics
4. Thomas H. Kean
5. Sarah Hughes
6. Figure skating
7. State Department
8. IOC
9. Baron Corp
10. Former President Clin.
11. Irina Slutskaya
12. NHL
13. FBI
14. Russian President Vla...
15. Soviet Union
16. Anti-terrorism
17. Daniel Pearl
18. Secretary of State Cont.
19. US Forces
20. DNA

Top 20 Places
1. UNITED STATES
2. NEW YORK
3. RUSSIA
4. CANADA

Document List (19 articles)
- Medina to See More Gold in Marketing Deals
- Person of the week: Sarah Hughes
- Russian curling team in curling season
- Russian Bear B U.S. Curling Season
- Russian Bear B World Curling
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World
- Russian Bear B World

Documents subdivided by places
- Articles referencing Sarah Hughes and RNAI (16)
  - Russian Curling
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World

- Articles referencing Sarah Hughes and RNAI (16)
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World
  - Russian Bear B World

www.isi.edu/geoworlds/geotopics

May 9, 2002
WWW2002
USC ISI’s GeoWorlds
Web-based Geo-spatial Information Management System

http://www.isi.edu/geoworlds

- Provides a toolset of Web information management
- Provides a library of information management services as software components
- GeoTopics is assembled from the component library
GeoTopics: Daily News Reportage Analysis

www.isi.edu/geoworlds/geotopics

- News Sources
- Extracted Articles
- Document Analyses
- News Compilation Results

10 News Sites
400+ News Articles
92 Analysis Steps

- Need complex analyses to filter, classify and correlate the articles
- Need to regularly perform the same set of analyses to update the portal site to maintain the latest information
- Need to handle dynamic situations such as format changes, instability of the sources
Problems and Challenges

Capabilities Required to Process Dynamic Sources

- **Composability** to string together complex information retrieval and analyses services, and generate a package application
- **Reusability** to reuse the application for recurring or similar analyses
- **Reconfigurability** to run the application on a different environment or to adapt it for dynamic situations
- **Context-sensitivity** to improve quality and performance

Additional demand: **Build and run it quickly!**
Approach: Multi-level, Lifecycle Service Coordination Mechanism

Allow non-programmers to stick to the level of detail they can handle and easily build an information management application.

Application-level, Design-time Coordination:
Specify what types of document collections are required and how they can be transformed.

Proxy-level, Run-time Coordination:
Automatically create run-time agents (proxies) that invoke, synchronize and monitor the distributed information management services to perform the task.

Allows efficient adaptation to different system environments and dynamic reconfiguration base on dynamic content and service conditions without affecting the high-level design.
Active Document Collection Templates
  e.g., News Topic Extraction

- Represent transformation actions between document collections
- Represent concurrent analysis processes

---

Transformation Actions

- Extracts articles from each news home page
- Filters out static news pages
- Sub-divides the news articles by news sources
- Extracts noun-phrases and classify articles
- Merges the noun-phrase extraction results

---

Parallel Branches

- Collection of news media home pages
- News articles extracted
- Filtered news articles classified by news sources
- Sub-collection of filtered news articles
- Documents classified by noun-phrases
- News articles classified by hot topics on the day

---

Collection of news media home pages
- Los Angeles Times
- CNN
- BBC News
- Chicago Tribune
- The Miami Herald
- USA Today
- The Globe and Mail

News articles extracted
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Filtered news articles classified by news sources
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Sub-collection of filtered news articles
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Documents classified by noun-phrases
- Noun Phrase
  - # of Documents
  - Middle East: 111
  - Terror Attack: 79
  - Secretary of State: 69
  - High School: 68
  - White House: 56
  - Taliban Regime: 48
  - US Forces: 40
  - Bush administration: 37
  - NBA: 46
  - 9/11 anniversary: 44
  - Prime Time: 43

News articles classified by hot topics on the day
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

---

Merges the categorization results

---

Represent transformation actions between document collections
Represent concurrent analysis processes

---

Transformation Actions

- Extracts articles from each news home page
- Filters out static news pages
- Sub-divides the news articles by news sources
- Extracts noun-phrases and classify articles
- Merges the noun-phrase extraction results

---

Parallel Branches

- Collection of news media home pages
- News articles extracted
- Filtered news articles classified by news sources
- Sub-collection of filtered news articles
- Documents classified by noun-phrases
- News articles classified by hot topics on the day

---

Collection of news media home pages
- Los Angeles Times
- CNN
- BBC News
- Chicago Tribune
- The Miami Herald
- USA Today
- The Globe and Mail

News articles extracted
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Filtered news articles classified by news sources
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Sub-collection of filtered news articles
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Documents classified by noun-phrases
- Noun Phrase
  - # of Documents
  - Middle East: 111
  - Terror Attack: 79
  - Secretary of State: 69
  - High School: 68
  - White House: 56
  - Taliban Regime: 48
  - US Forces: 40
  - Bush administration: 37
  - NBA: 46
  - 9/11 anniversary: 44
  - Prime Time: 43

News articles classified by hot topics on the day
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

---

Merges the categorization results

---

Represent transformation actions between document collections
Represent concurrent analysis processes

---

Transformation Actions

- Extracts articles from each news home page
- Filters out static news pages
- Sub-divides the news articles by news sources
- Extracts noun-phrases and classify articles
- Merges the noun-phrase extraction results

---

Parallel Branches

- Collection of news media home pages
- News articles extracted
- Filtered news articles classified by news sources
- Sub-collection of filtered news articles
- Documents classified by noun-phrases
- News articles classified by hot topics on the day

---

Collection of news media home pages
- Los Angeles Times
- CNN
- BBC News
- Chicago Tribune
- The Miami Herald
- USA Today
- The Globe and Mail

News articles extracted
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Filtered news articles classified by news sources
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Sub-collection of filtered news articles
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Documents classified by noun-phrases
- Noun Phrase
  - # of Documents
  - Middle East: 111
  - Terror Attack: 79
  - Secretary of State: 69
  - High School: 68
  - White House: 56
  - Taliban Regime: 48
  - US Forces: 40
  - Bush administration: 37
  - NBA: 46
  - 9/11 anniversary: 44
  - Prime Time: 43

News articles classified by hot topics on the day
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

---

Merges the categorization results

---

Represent transformation actions between document collections
Represent concurrent analysis processes

---

Transformation Actions

- Extracts articles from each news home page
- Filters out static news pages
- Sub-divides the news articles by news sources
- Extracts noun-phrases and classify articles
- Merges the noun-phrase extraction results

---

Parallel Branches

- Collection of news media home pages
- News articles extracted
- Filtered news articles classified by news sources
- Sub-collection of filtered news articles
- Documents classified by noun-phrases
- News articles classified by hot topics on the day

---

Collection of news media home pages
- Los Angeles Times
- CNN
- BBC News
- Chicago Tribune
- The Miami Herald
- USA Today
- The Globe and Mail

News articles extracted
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Filtered news articles classified by news sources
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Sub-collection of filtered news articles
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Documents classified by noun-phrases
- Noun Phrase
  - # of Documents
  - Middle East: 111
  - Terror Attack: 79
  - Secretary of State: 69
  - High School: 68
  - White House: 56
  - Taliban Regime: 48
  - US Forces: 40
  - Bush administration: 37
  - NBA: 46
  - 9/11 anniversary: 44
  - Prime Time: 43

News articles classified by hot topics on the day
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

---

Merges the categorization results

---

Represent transformation actions between document collections
Represent concurrent analysis processes

---

Transformation Actions

- Extracts articles from each news home page
- Filters out static news pages
- Sub-divides the news articles by news sources
- Extracts noun-phrases and classify articles
- Merges the noun-phrase extraction results

---

Parallel Branches

- Collection of news media home pages
- News articles extracted
- Filtered news articles classified by news sources
- Sub-collection of filtered news articles
- Documents classified by noun-phrases
- News articles classified by hot topics on the day

---

Collection of news media home pages
- Los Angeles Times
- CNN
- BBC News
- Chicago Tribune
- The Miami Herald
- USA Today
- The Globe and Mail

News articles extracted
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Filtered news articles classified by news sources
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Sub-collection of filtered news articles
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

Documents classified by noun-phrases
- Noun Phrase
  - # of Documents
  - Middle East: 111
  - Terror Attack: 79
  - Secretary of State: 69
  - High School: 68
  - White House: 56
  - Taliban Regime: 48
  - US Forces: 40
  - Bush administration: 37
  - NBA: 46
  - 9/11 anniversary: 44
  - Prime Time: 43

News articles classified by hot topics on the day
- Los Angeles Times
- CNN.com
- Chronicle
- NY Times
- USA Today

---

Merges the categorization results

---

Represent transformation actions between document collections
Represent concurrent analysis processes
Dynamic Parallelism Representation

- Service-based parallelism representation using “branching” and “joining” services
- Context-sensitive concurrency representation: spawning multiple analyses based on the content and/or structure of a document collection

- Merges the categorized document collections
- Input Document Collection
- Branching Service
- Independent Analysis Results
- Joining Service
- Merged Analysis Results

- Better Quality of Results (Better Set of Noun-Phrases)
- Better Performance in Analyzing Information
- Easy Composition and Comprehension of an Application
Dynamic Parallelism Representation

(Branch Types)

Open Branch

Partially Open Branch

Nested Branches

Displays a histogram that visualizes the noun-phrase extraction results.

Collection of news media home pages classified by countries.

F₀=(f₀, i₀, o₀)

Merges the noun-phrase extraction results by intersecting them, i.e., generates the list of common noun-phrases across the news sources.

F₁₀=(f₁₀, i₁₀, o₁₀)

Merges the news media home pages by countries.

F₆=(c₀, s₀)

F₁=(f₁, i₁, o₁)

F₃=(f₃, i₃, o₃)

F₄=(f₄, i₄, o₄)

F₅=(f₅, i₅, o₅)

F₁¹=(f₁¹, i₁¹, o₁¹)

F₁₂=(f₁₂, i₁₂, o₁₂)

F₂=(f₂, i₂, o₂)

F₃=(f₃, i₃, o₃)

F₄=(f₄, i₄, o₄)

F₅=(f₅, i₅, o₅)

F₆=(f₆, i₆, o₆)
Proxy-based Run-time Service Coordination

Application instantiation by allocating service instances & creating client-side proxies that perform:

- Dynamic data binding during run time
- Data-driven service invocation and synchronization
- Token-based concurrency control
- Dynamic spawning and merging of service branches by the branching and joining service proxies

- **Proxies hide the system dependent information from the high-level application**
- **Proxies show the centralized view of the distributed services**
Application Scripting Tool

- Converts and stores the application description to an XML file.
- Shows service-oriented view of the application in terms of a data flow diagram.
- Guides users to select and combine semantically interoperable services.
- Automatically figures out and inserts missing components (input and converter services).
- Converts and stores the application description to an XML file.
Application Execution and Reconfiguration

Guides users with displaying the input services one by one to enter input data
Shows the progress and exceptional conditions of the services during run-time
Allows users reconfigure the application by substituting services with semantically compatible, alternative ones during design and run times
Quality Improvements Enabled by the Approach

More topics populated with more geographic relevance

ACCURACY

PERFORMANCE

S IMPLICITY

GeoTopics Example:
- **Rapid Development**: took less than a week to develop GeoTopics which is composed of 92 services

Indonesia-related Topic Analysis Example:
- **High Reuse**: GeoTopics could be adapted to retrieve Indonesia related information from Web search engines and analyze them for specialized topics less than an hour

90% speed up by concurrent execution
20% more speed up by dynamic coordination
Conclusion

Features

- High-level Application Development by using Active Document Collection Templates
- Data Flow Based Service Coordination Mechanism
- Service Based Parallelism Representation and control

Benefits

- Allow non-programmer users to efficiently develop information management applications
- Support context-sensitive parallelism control
- Improve performance and quality of information management processes
- Allow applications to be adapted and reused for different system environments and tasks