

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

Previous Saturday, February 23, 2002 10:00 AM PST Next

Top 20 Topics

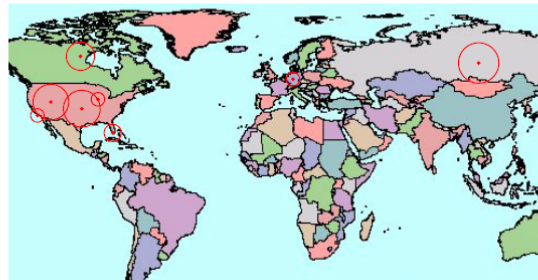
- [1] [Gold Medal](#) (37)
- [2] [George W. Bush](#) (25)
- [3] [Winter Olympics](#) (28)
- [12] [Terror Attacks](#) (20)
- NEW** [Sarah Hughes](#) (19)
- [4] [Figure Skating](#) (18)
- [10] [State Department](#) (16)
- [11] [IOC](#) (15)
- [9] [Enron Corp](#) (15)
- [13] [Former President Clin...](#) (13)
- NEW** [Irina Slutskaya](#) (13)
- NEW** [NHL](#) (12)
- [6] [FBI](#) (12)
- NEW** [Russian President Vla...](#) (10)
- NEW** [Soviet Union](#) (12)
- [14] [Anti-terrorism](#) (10)
- [7] [Daniel Pearl](#) (11)
- NEW** [Secretary of State Co...](#) (10)
- [18] [US forces](#) (9)
- NEW** [DNA](#) (8)

Top 20 Places

- [1] [UNITED STATES](#) (122)
- [2] [NEW YORK](#) (50)
- [5] [RUSSIA](#) (31)
- [6] [CANADA](#) (31)

Saturday, February 23, 2002 10:00 AM PST

Sarah Hughes



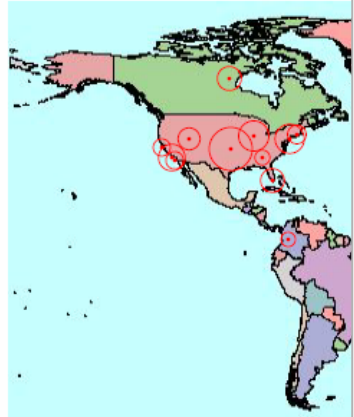
Document List (19 articles)

- [Medalists to See More Gold in Marketing Deals](#)
- [Person of the week: Sarah Hughes](#)
- [Russia cries Olympic foul](#)
- [2002 Winter Olympics](#)
- [Latest Olympics coverage](#)
- [Russian Protests Chill the Games](#)
- [Sarah Hughes: Sweet 16, and good as gold](#)
- [Cries of bad judging chill Olympics](#)
- [ISU denies Russian protest on women's gold](#)
- [O'Connor/Games' spirit soiled](#)
- [Pictures of the week](#)
- [Russia, South Korea Threaten Boycott](#)
- [Russians Bear Ill Will Over Perceived Injustices](#)
- [Russians Cry Foul](#)
- [Sarah Hughes On • Winning Gold Medal](#)
- [Skating on Thin Ice](#)
- [Team USA Defeats Russia](#)
- [U.S. Beats Russia, Canada Next](#)
- [Winter Games Insider](#)

Documents subdivided by places

- Articles referencing Sarah Hughes and RUSSIA (16)**
 - [Russians Cry Foul](#)
 - [Cries of bad judging chill Olympics](#)
 - [Russian Protests Chill the Games](#)
 - [Russians Bear Ill Will Over Perceived Injustices](#)
 - [Team USA Defeats Russia](#)
 - [U.S. Beats Russia, Canada Next](#)
 - [2002 Winter Olympics](#)
 - [Latest Olympics coverage](#)
 - [O'Connor/Games' spirit soiled](#)
 - [Person of the week: Sarah Hughes](#)
 - [Russia cries Olympic foul](#)
 - [Skating on Thin Ice](#)
 - [Winter Games Insider](#)
 - [ISU denies Russian protest on women's gold](#)
 - [Russia, South Korea Threaten Boycott](#)
 - [Sarah Hughes: Sweet 16, and good as gold](#)
- Articles referencing Sarah Hughes and UNITED STATES (13)**
 - [Russian Protests Chill the Games](#)
 - [U.S. Beats Russia, Canada Next](#)
 - [Team USA Defeats Russia](#)
 - [2002 Winter Olympics](#)
 - [Cries of bad judging chill Olympics](#)
 - [Latest Olympics coverage](#)

Legend: Same rank (grey square), Moving up (red arrow), Moving down (blue arrow), New today (yellow square)



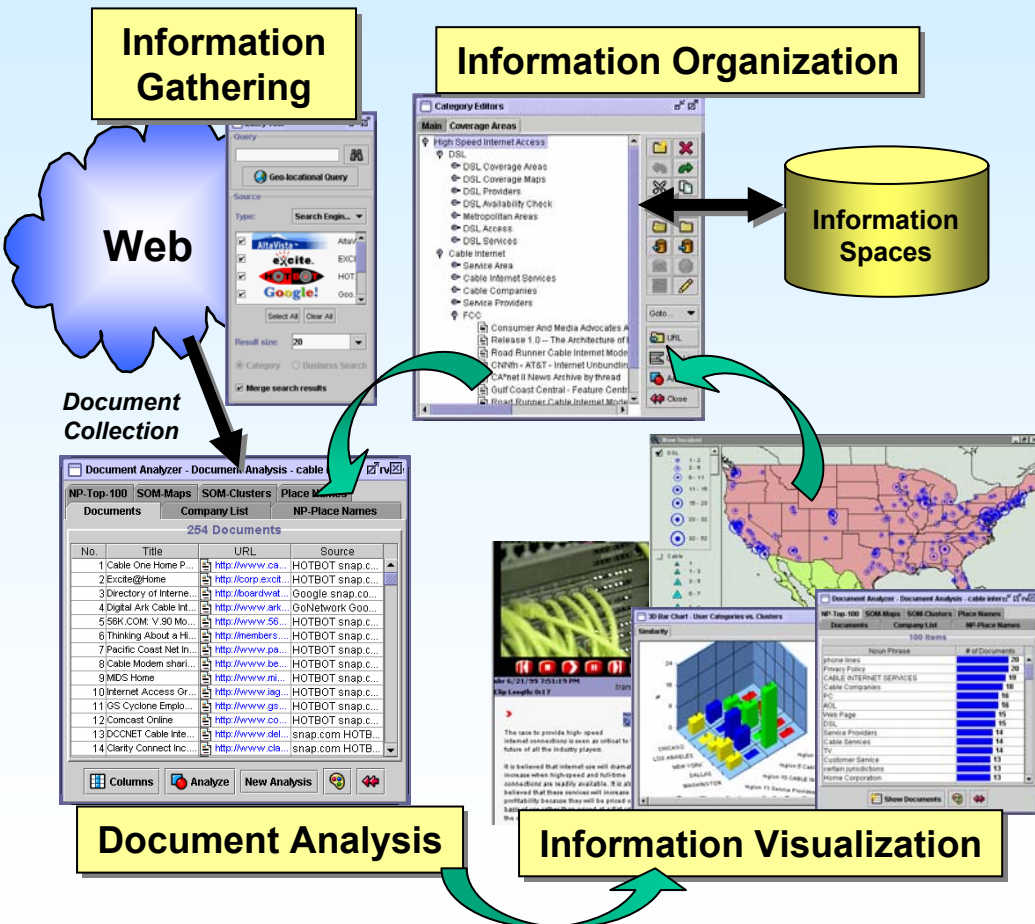
- Helps users identify the “hot topics,” and the most frequently referenced places
- Allows users to keep track of the hot topics and places

- Shows a breakdown of a topic by places and a breakdown of a place by topics

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



10 News Sites

Extracted Articles



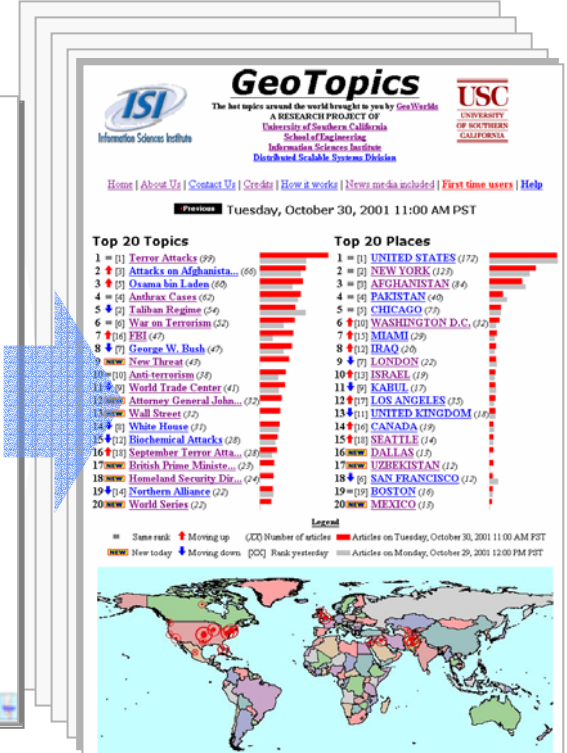
400+ News Articles

Document Analyses



92 Analysis Steps

News Compilation Results



- 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

Problems Raised by Dynamic Sources

Variable in Content

Variable in Performance

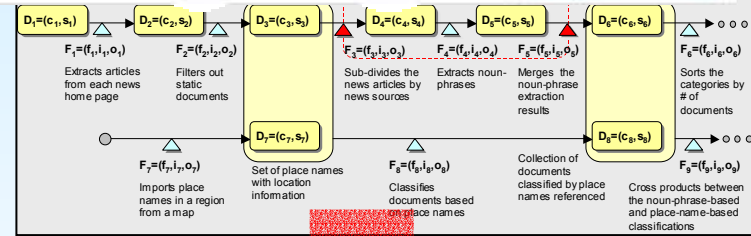
- **Composability** to string together complex information retrieval and analyses services, and generate a package application
- **Reusability** to reuse the application for recurring or similar analyses
- **Reconfigurability** to run the application on a different environment or to adapt it for dynamic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

Additional demand: Build and run it quickly!

Approach: Multi-level, Lifecycle Service Coordination Mechanism

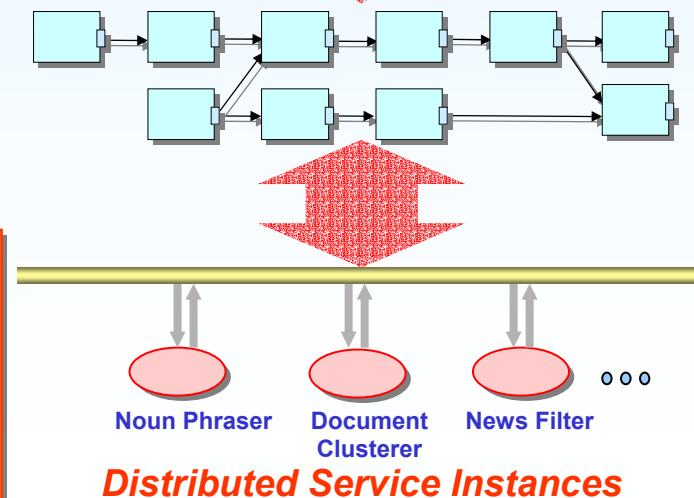
Allows non-programmers to stick to the level of detail they can handle and easily build an information management application

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

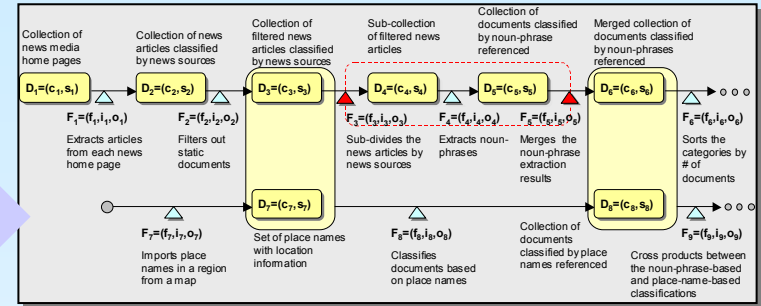
Allows efficient adaptation to different system environments and dynamic reconfiguration base on dynamic content and service conditions without affecting the high-level design



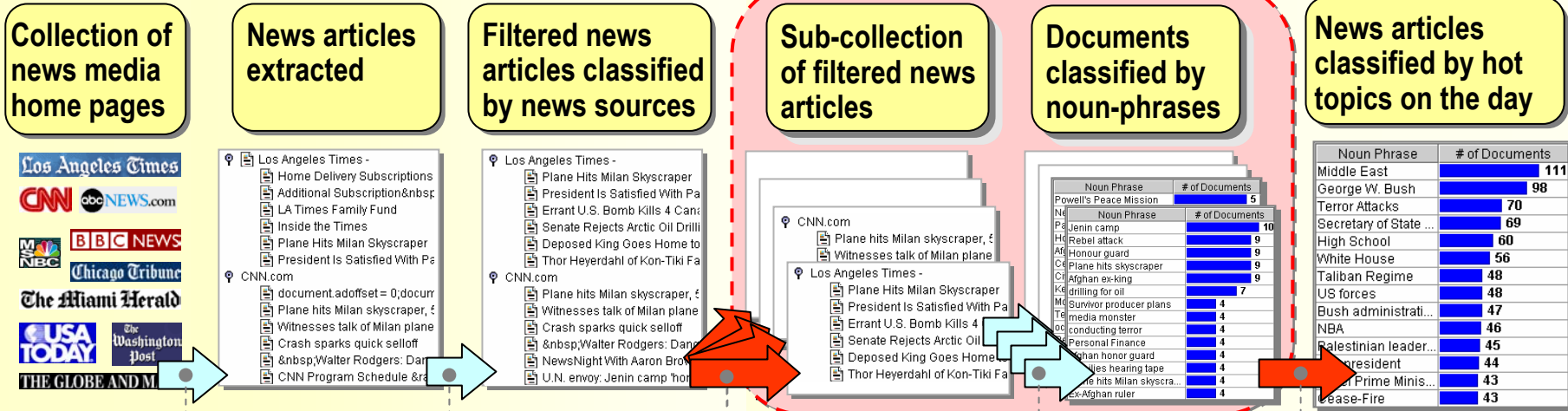
Active Document Collection Templates

e.g., News Topic Extraction

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



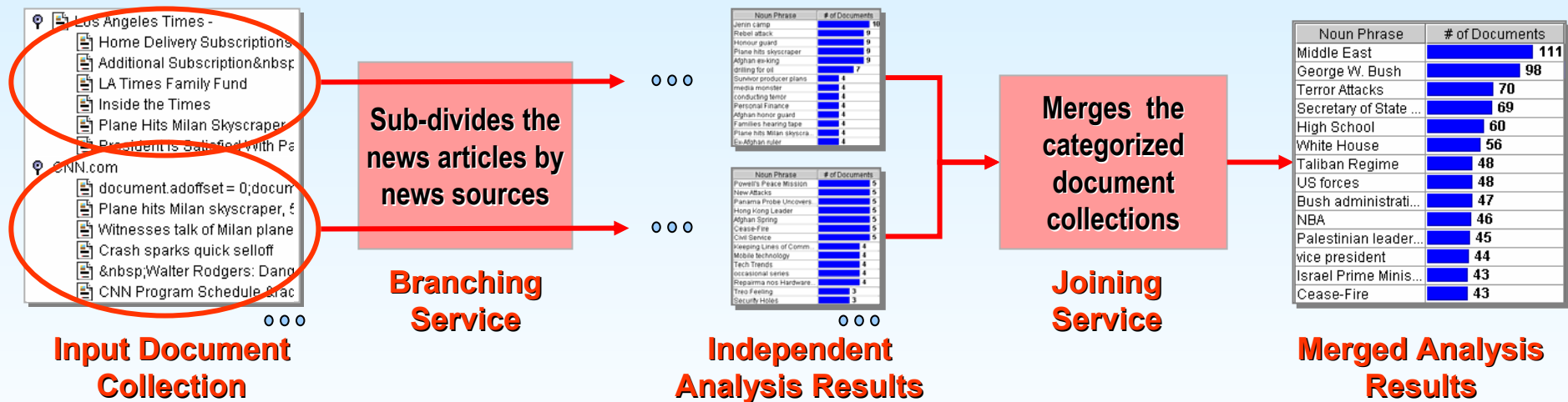
Document Collections



Transformation Actions

Dynamic Parallelism Representation

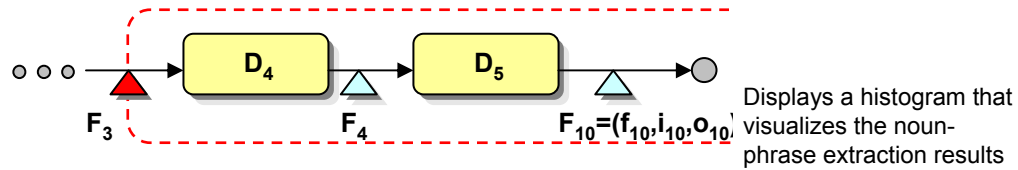
- **Service-based parallelism representation** using “branching” and “joining” services
- **Context-sensitive concurrency** representation: spawning multiple analyses based on the content and/or structure of a document collection



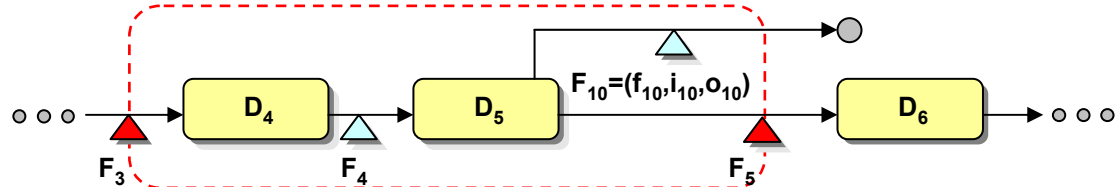
- **Better Quality of Results (Better Set of Noun-Phrases)**
- **Better Performance in Analyzing Information**
- **Easy Composition and Comprehension of an Application**

Dynamic Parallelism Representation (Branch Types)

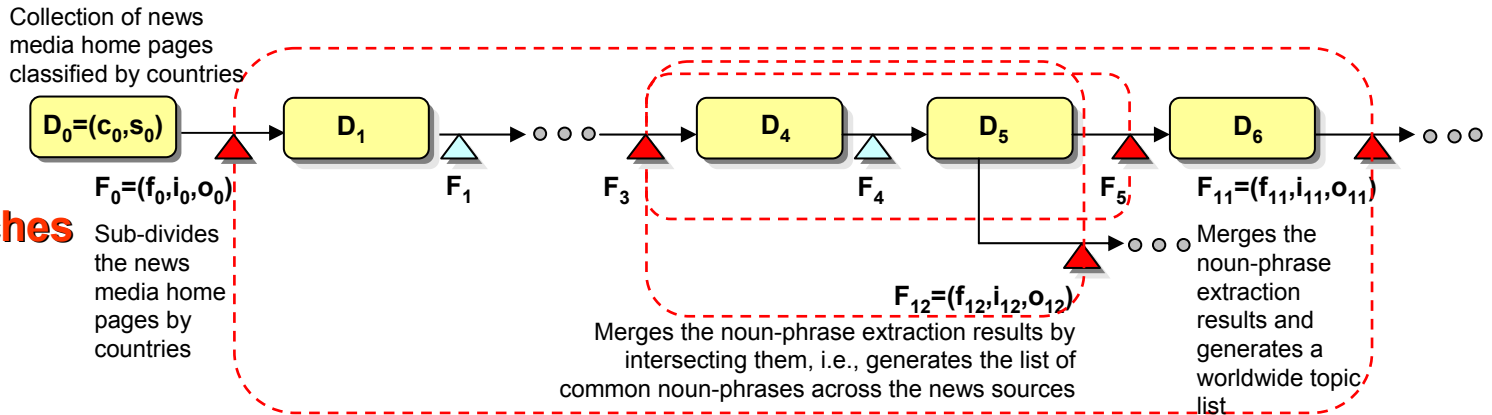
Open Branch



Partially Open Branch



Nested Branches

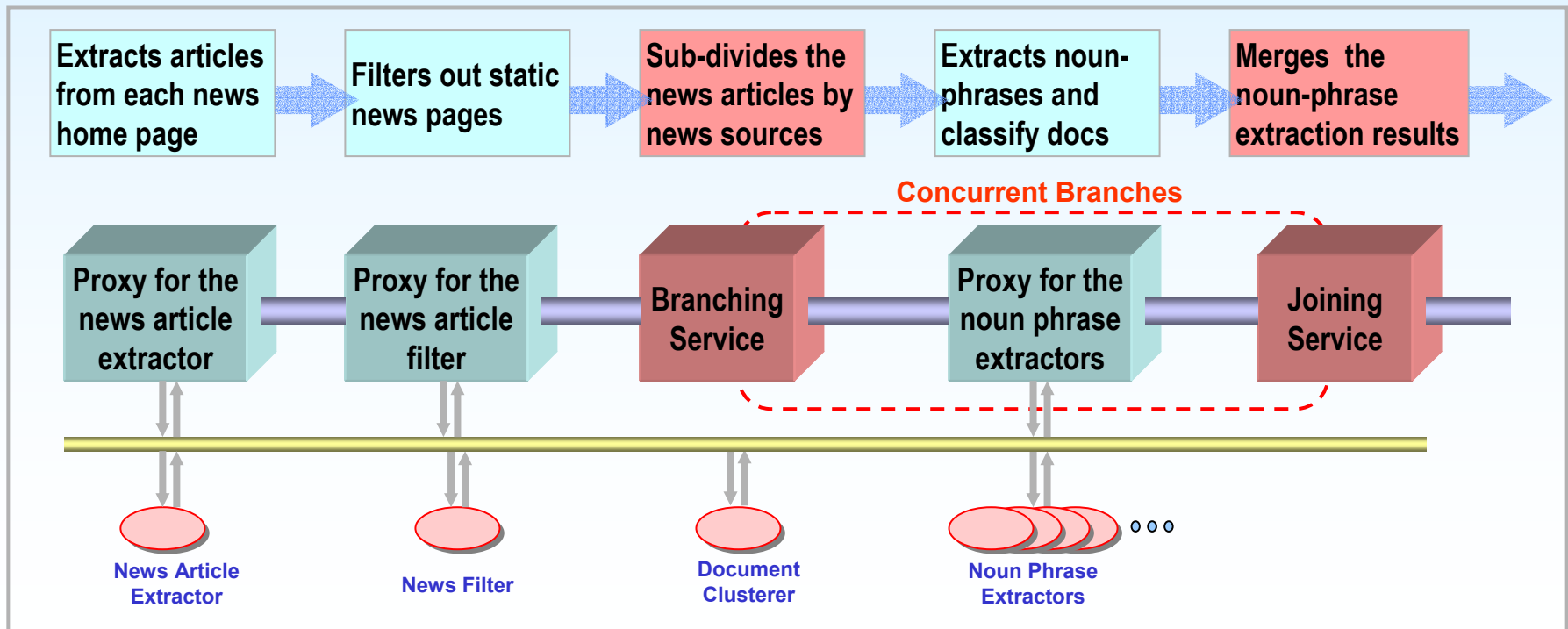


Proxy-based Run-time Service Coordination

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

Automatically figures out and inserts missing components (input and converter services)

Shows service-oriented view of the application in terms of a data flow diagram

The screenshot displays the Application Scripting Tool interface. At the top, a menu bar includes 'Tools' and 'Help'. Below it, a 'Scripts' window shows a data flow diagram with components like 'Collection Edit', 'Analyzer Fanout', 'Analyzer Index Filter', 'Brancher Branching', 'Converter Flattening Cate', 'Analyzer Noun Phraser', 'Converter Format Conve', 'Join Joining', and 'Analyzer Entropy Ranki'. A toolbar below the diagram includes 'Add', 'Alternative', 'Match', 'Save', 'Load', 'Run', and 'Initialize'. On the left, a sidebar contains 'Scripts / Results', 'Custom Categories', 'Search / Analysis', and 'Region of Interest'. The main area is divided into 'Analytic', 'Visual', 'Information Source', 'Branching', and 'Joining' tabs. The 'Analytic' tab is active, showing a list of services including 'KeywordExtraction', 'NounPhraseExtraction', 'CompanyNameExtraction', 'CategoriesToKeywordsConversion', 'ExtractKeywordsFromCategories', 'DocumentClassification', 'DocumentSummarization', 'LanguageIdentification', 'CategoryManipulation', 'CategoryFanOut', 'CategoryComparison', and 'SimilarityDividabilityComparison'. The 'NounPhraseExtraction' service is selected. To the right, a 'Service Schema' window shows details for 'NounPhraseExtraction', including its URI, short name, description, and input/output data tables.

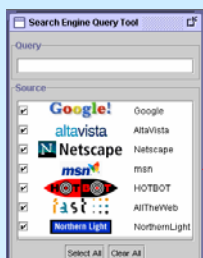
Input Data:			Output Data:		
Name	Content	Structure	Name	Content	Structure
Input\$1	DocumentC	Il	Output\$1	NounPhrases	

Guides users to select and combine semantically interoperable services

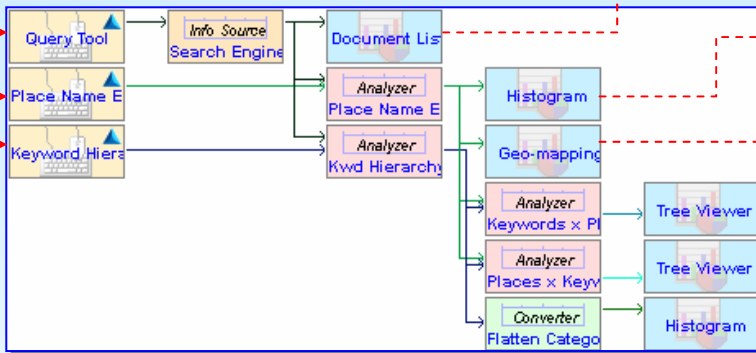
Converts and stores the application description to an XML file

Application Execution and Reconfiguration

Inputs



Application Script



Indonesia-related Information Analysis

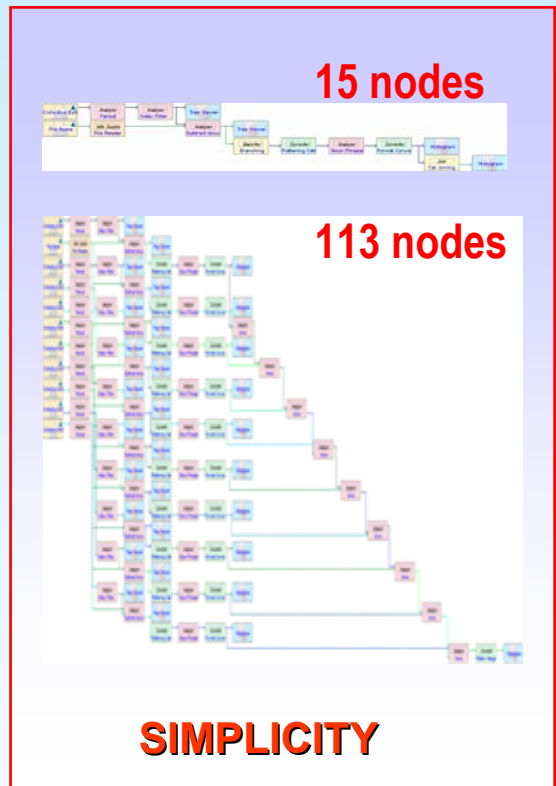
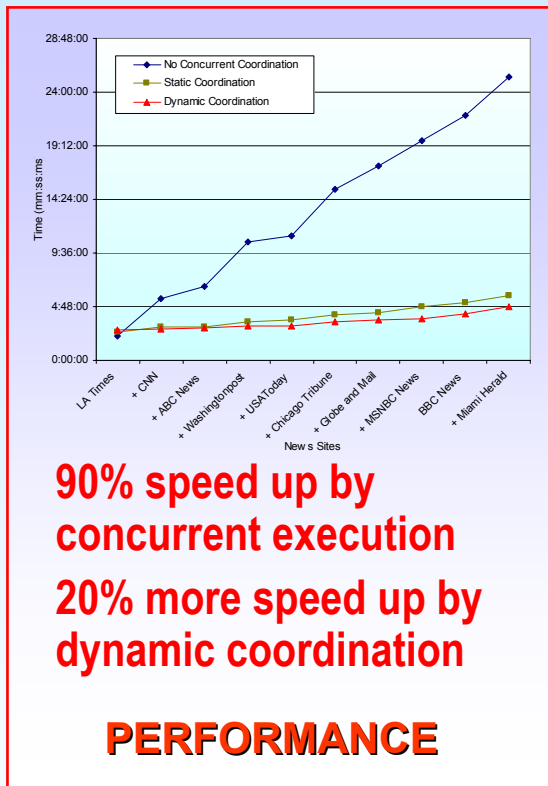
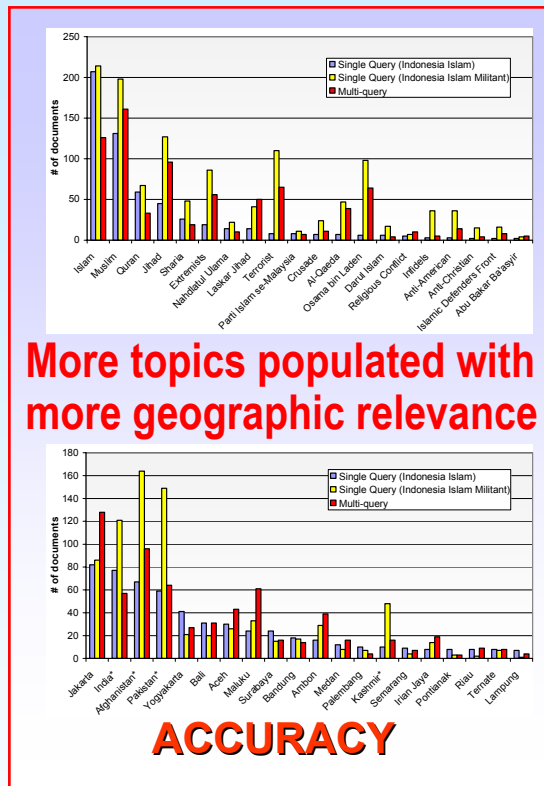
- Guides users with displaying the input services one by one to enter input data
- Shows the progress and exceptional conditions of the services during run-time
- Allows users reconfigure the application by substituting services with semantically compatible, alternative ones during design and run times

Results




Indonesia Information Portal

Quality Improvements Enabled by the Approach



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

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