Dynamic Coordination of Information Management Services for Processing Dynamic Web Content

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

{lKo, KYao, RNeches}@isi.edu

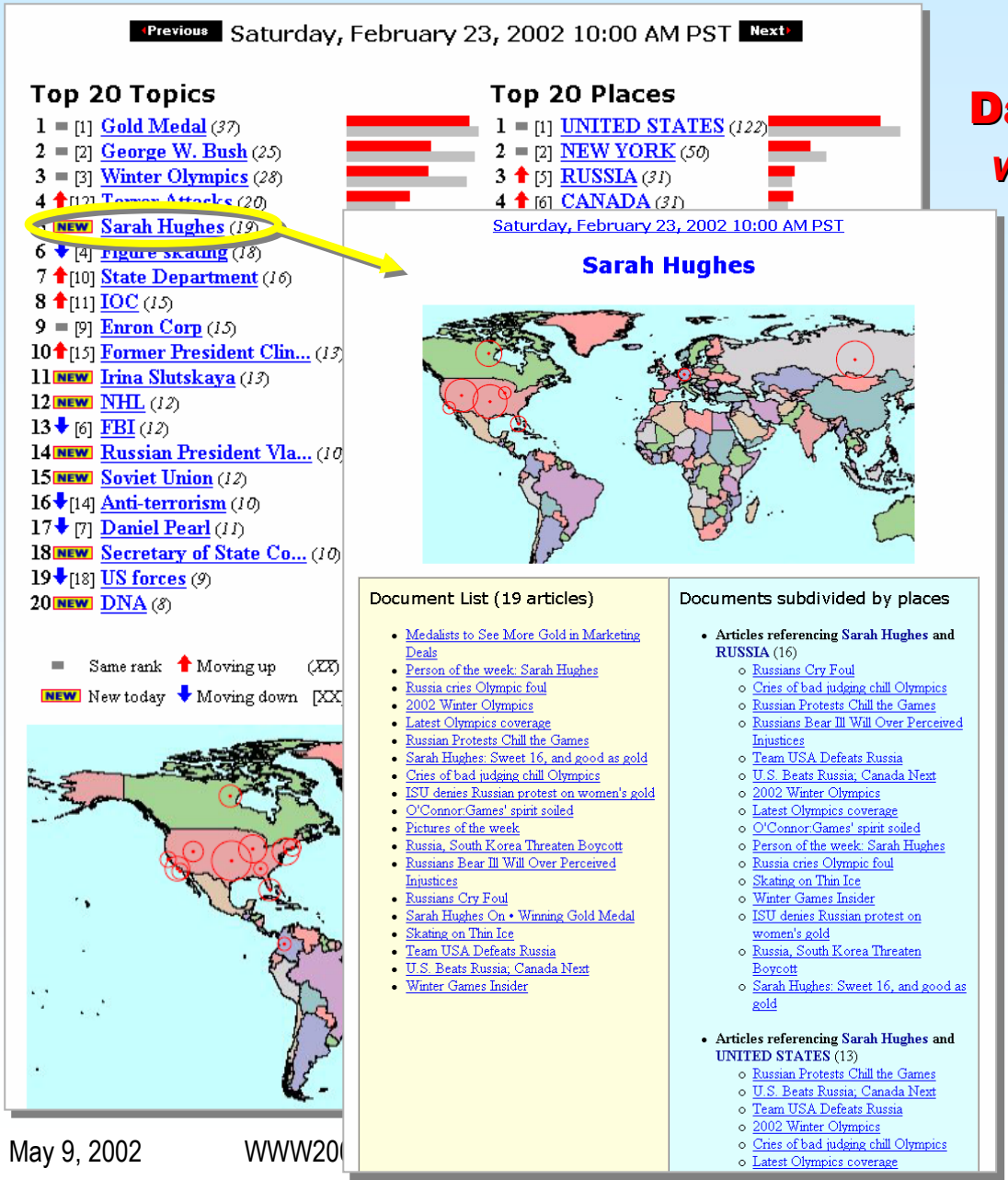
Information Sciences Institute
University of Southern California

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

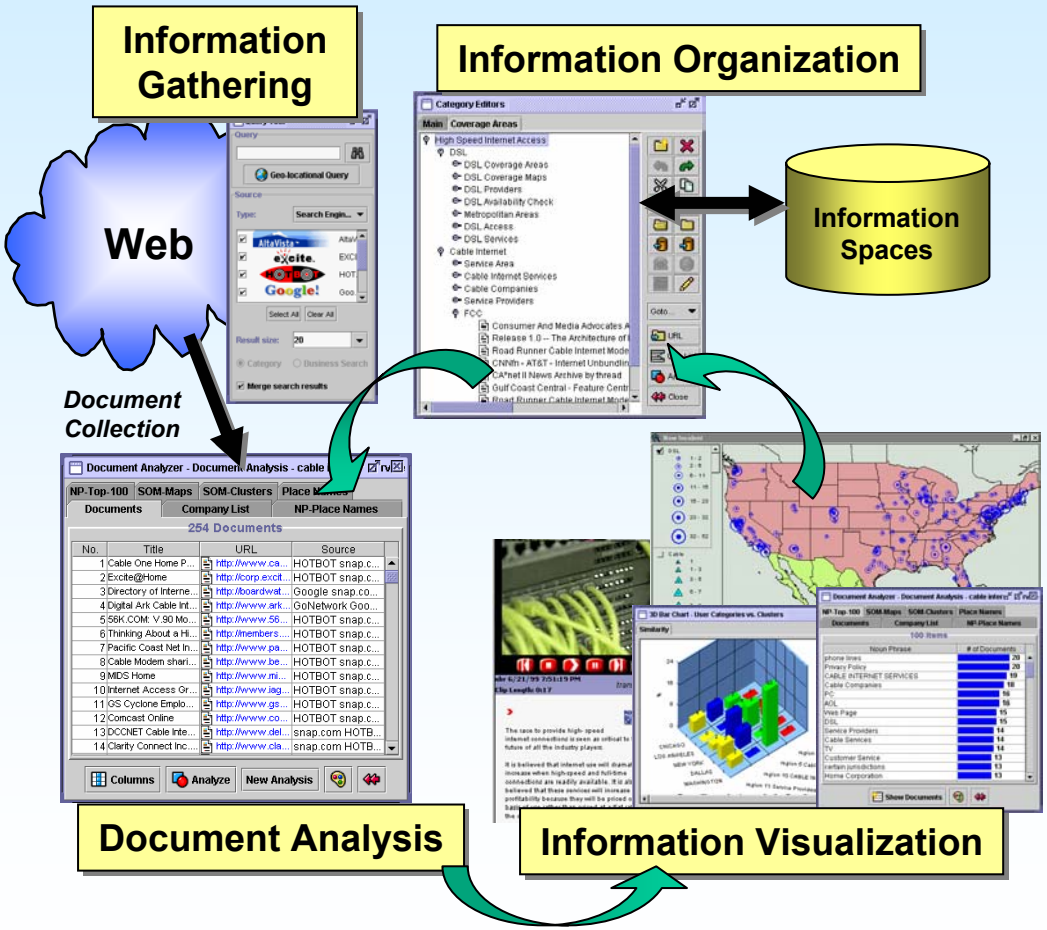
- Helps users identify the "hot topics," and the most frequently referenced places
- Allows users to keep track of the hot topics and places

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

USC ISI's GeoWorlds

Web-based Geo-spatial Information Management System

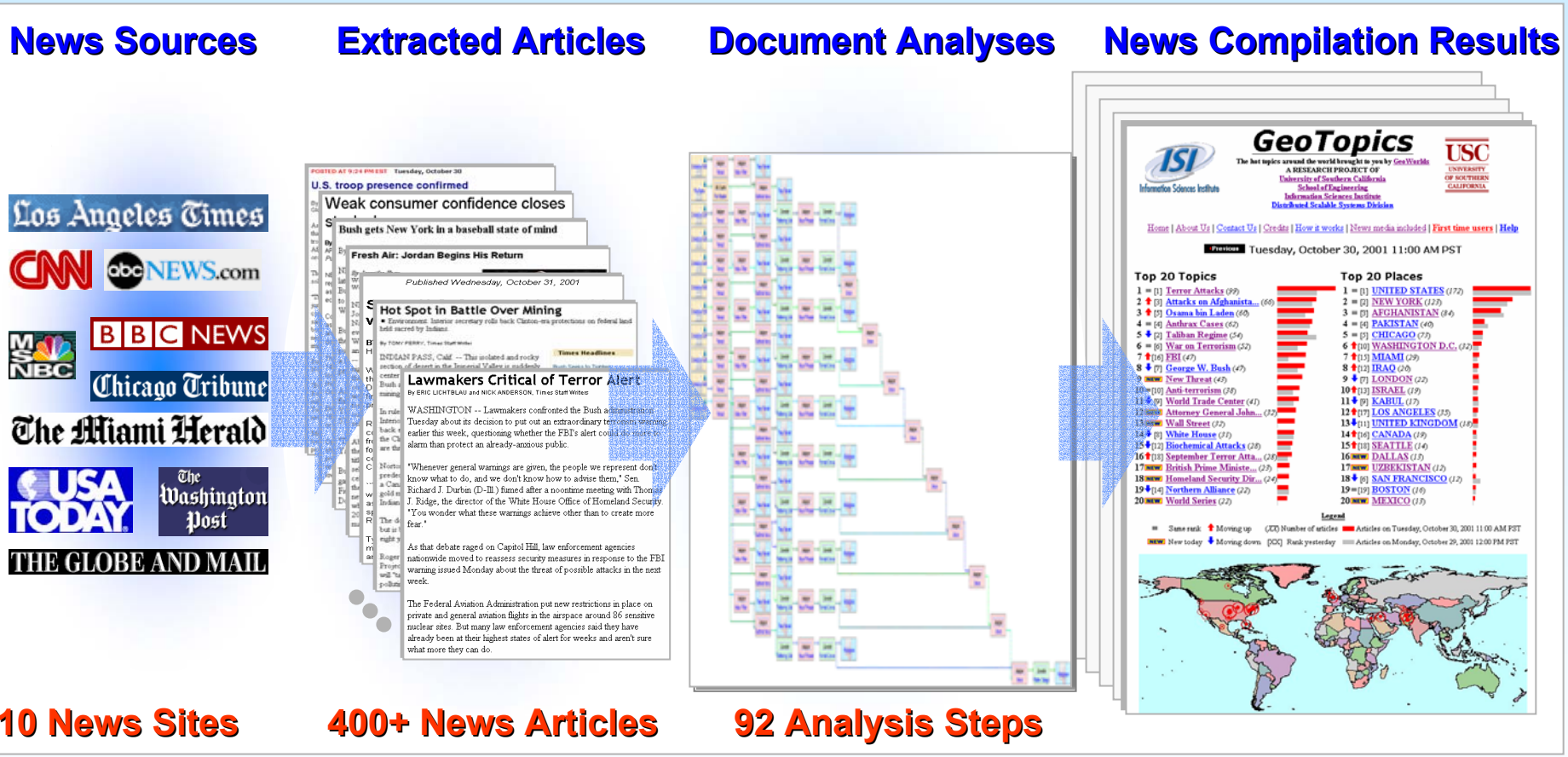
http://www.isi.edu/geoworlds



- 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



- 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

Problems Raised by Dynamic Sources

- **Variable in Content**
- **Variable in Performance**

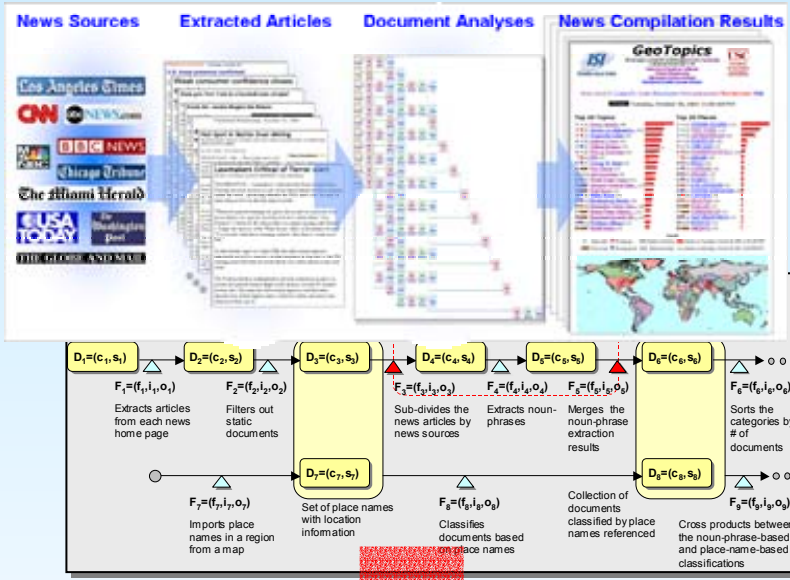
Capabilities Required to Process Dynamic Sources

- **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!

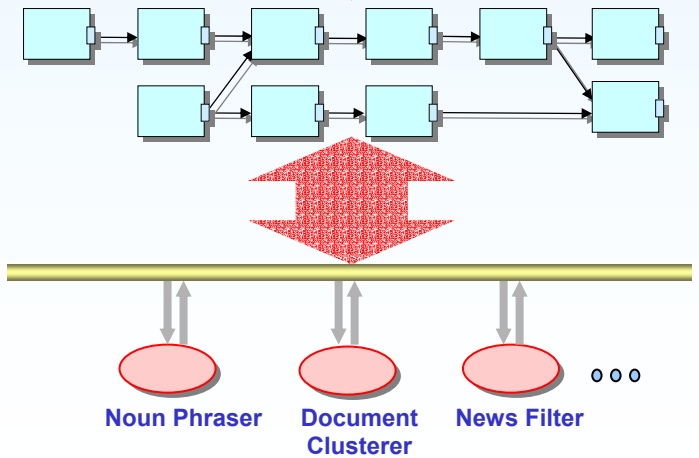
Approach: Multi-level, Lifecycle Service Coordination Mechanism

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 invocate, synchronize and monitor the distributed information management services to perform the task

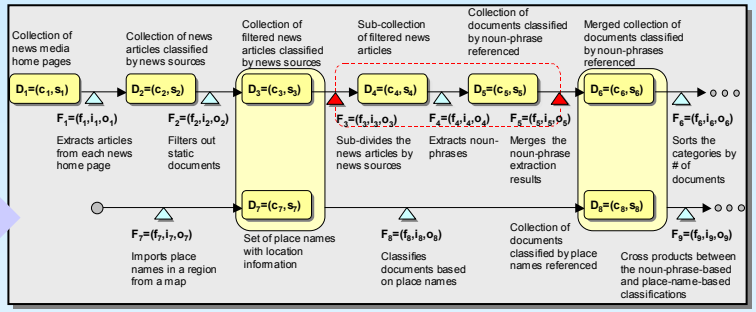


Allows efficient adaptation to different system environments and dynamic reconfiguration base on dynamic content and service conditions without affecting the high-level design

Active Document Collection Templates

e.g., News Topic Extraction

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



Parallel Branches

Document Collections

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

Noun Phrase	# of Documents
Middle East	111
George W. Bush	98
Terror Attacks	70
Secretary of State ...	69
High School	60
White House	56
Taliban Regime	48
US forces	48
Bush administrati...	47
NBA	46
Palestinian leader...	45
...resident	44
...Prime Minis...	43
...ease-Fire	43

Extracts articles from each news home page

Filters out static news pages

Sub-divides the news articles by news sources

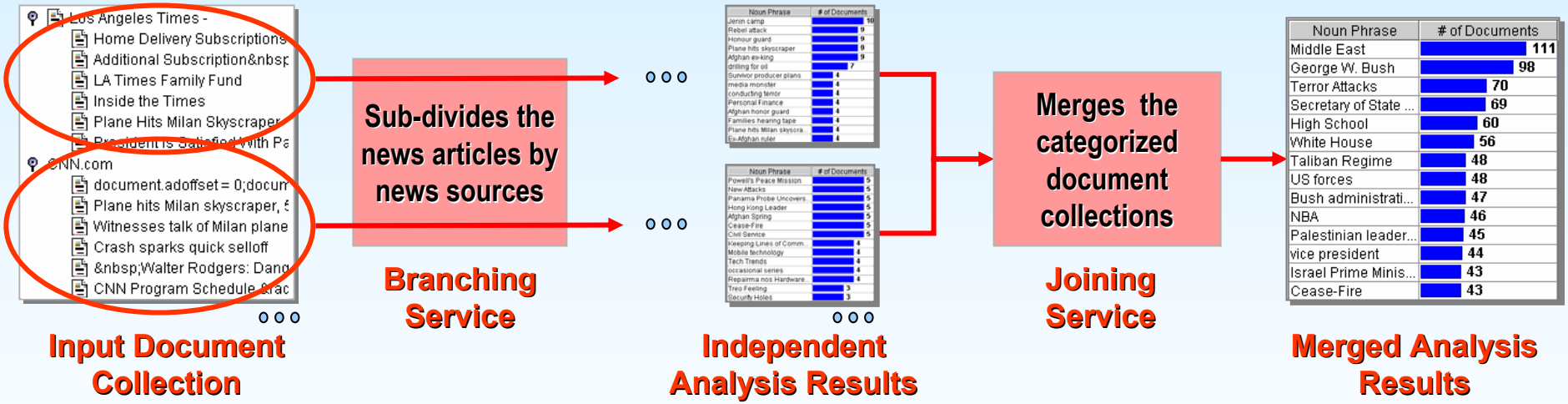
Extracts noun-phrases and classify articles

Merges the noun-phrase extraction results

Transformation Actions

Dynamic Parallelism Representation

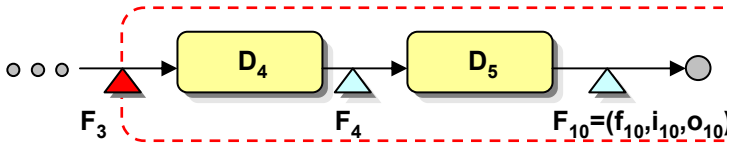
- **Service-based parallelism representation** using "branching" and "joining" services
- **Context-sensitive concurrency representation:** spawning multiple analyses based on the content and/or structure of a document collection



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

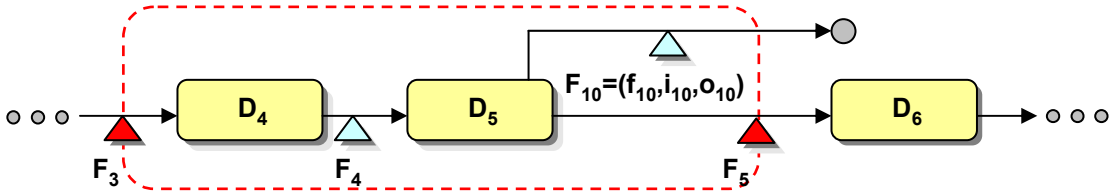
Dynamic Parallelism Representation
(Branch Types)

Open Branch



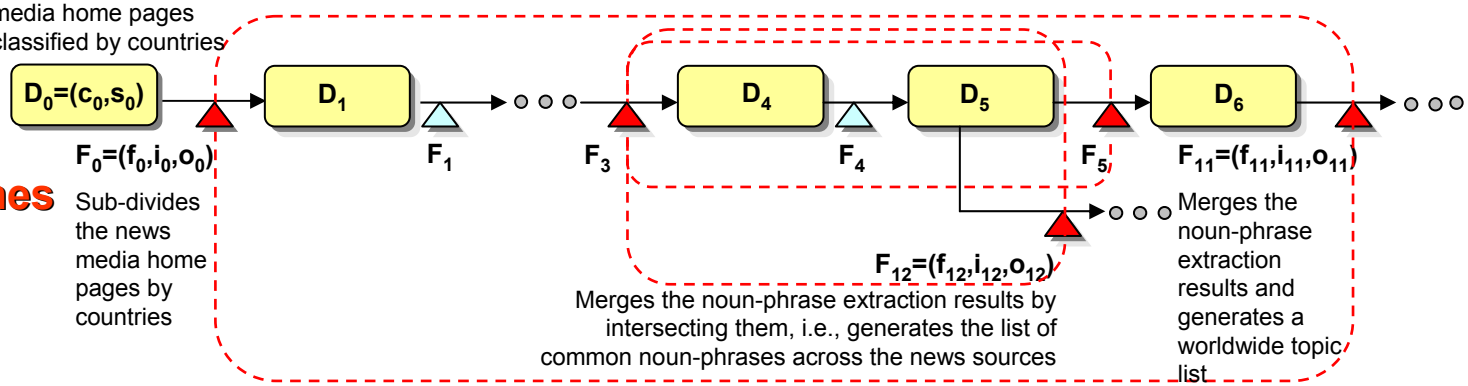
Displays a histogram that visualizes the noun-phrase extraction results

Partially Open Branch



Nested Branches

Collection of news media home pages classified by countries



Sub-divides the news media home pages by countries

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

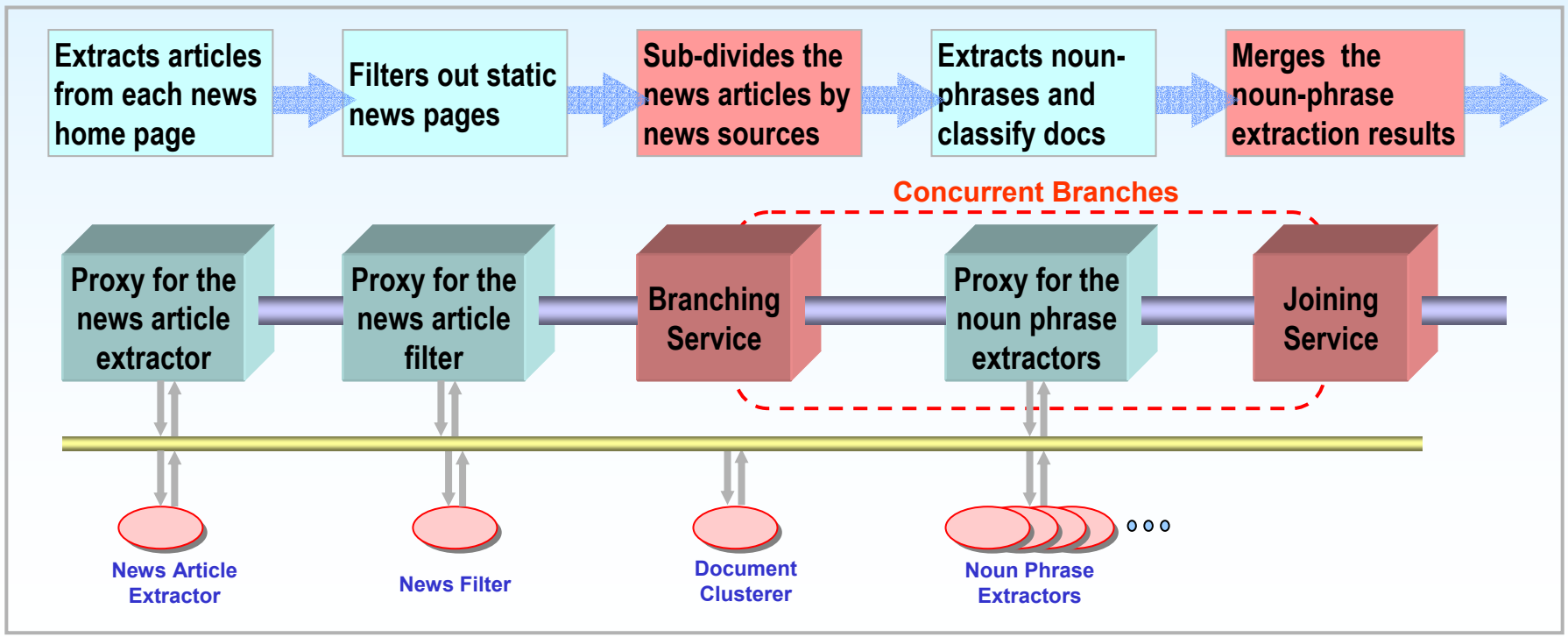
Merges the noun-phrase extraction results and generates a worldwide topic list

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

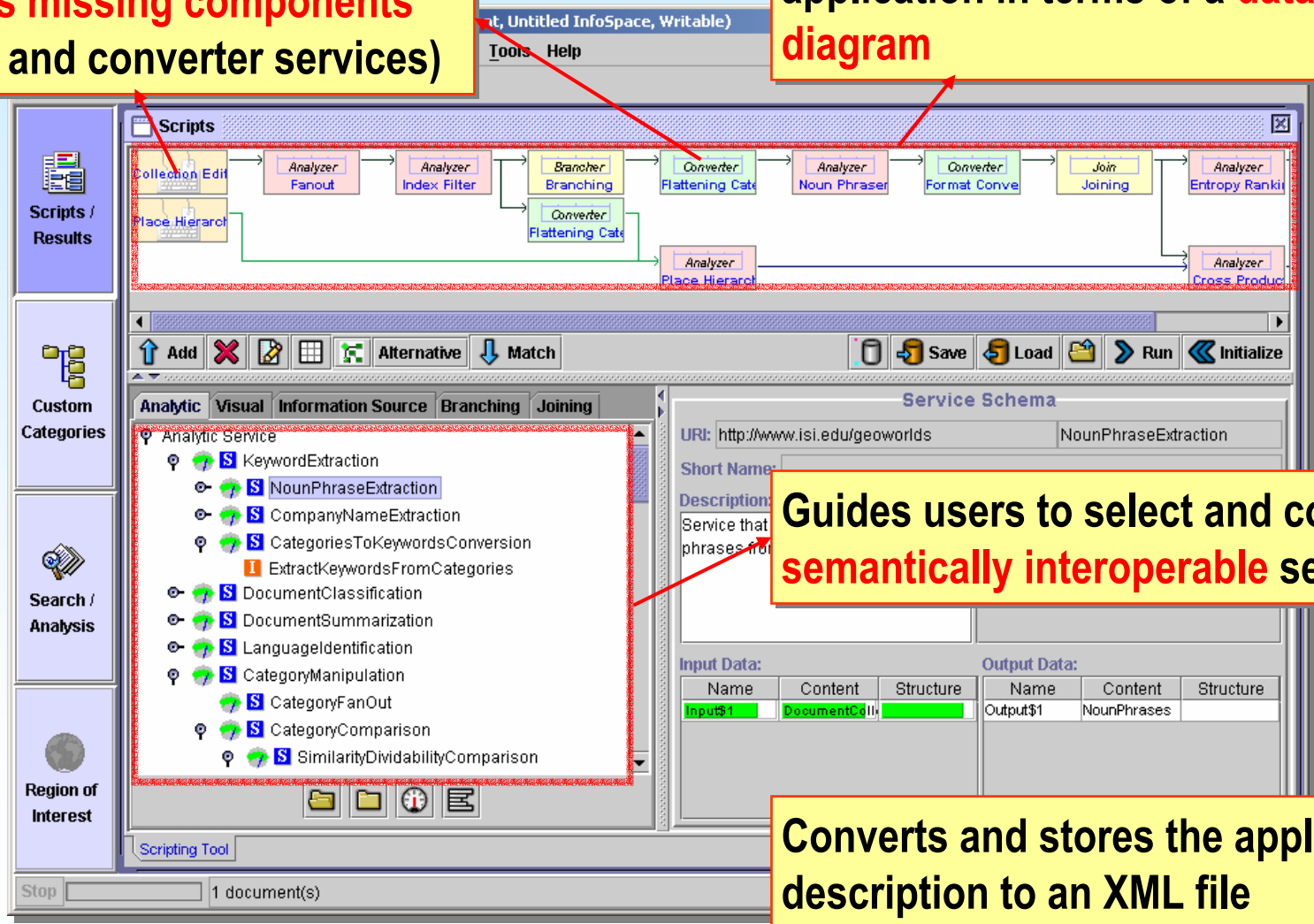
- Proxies hide the system dependent information from the high-level application
- Proxies show the centralized view of the distributed services



Application Scripting Tool

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

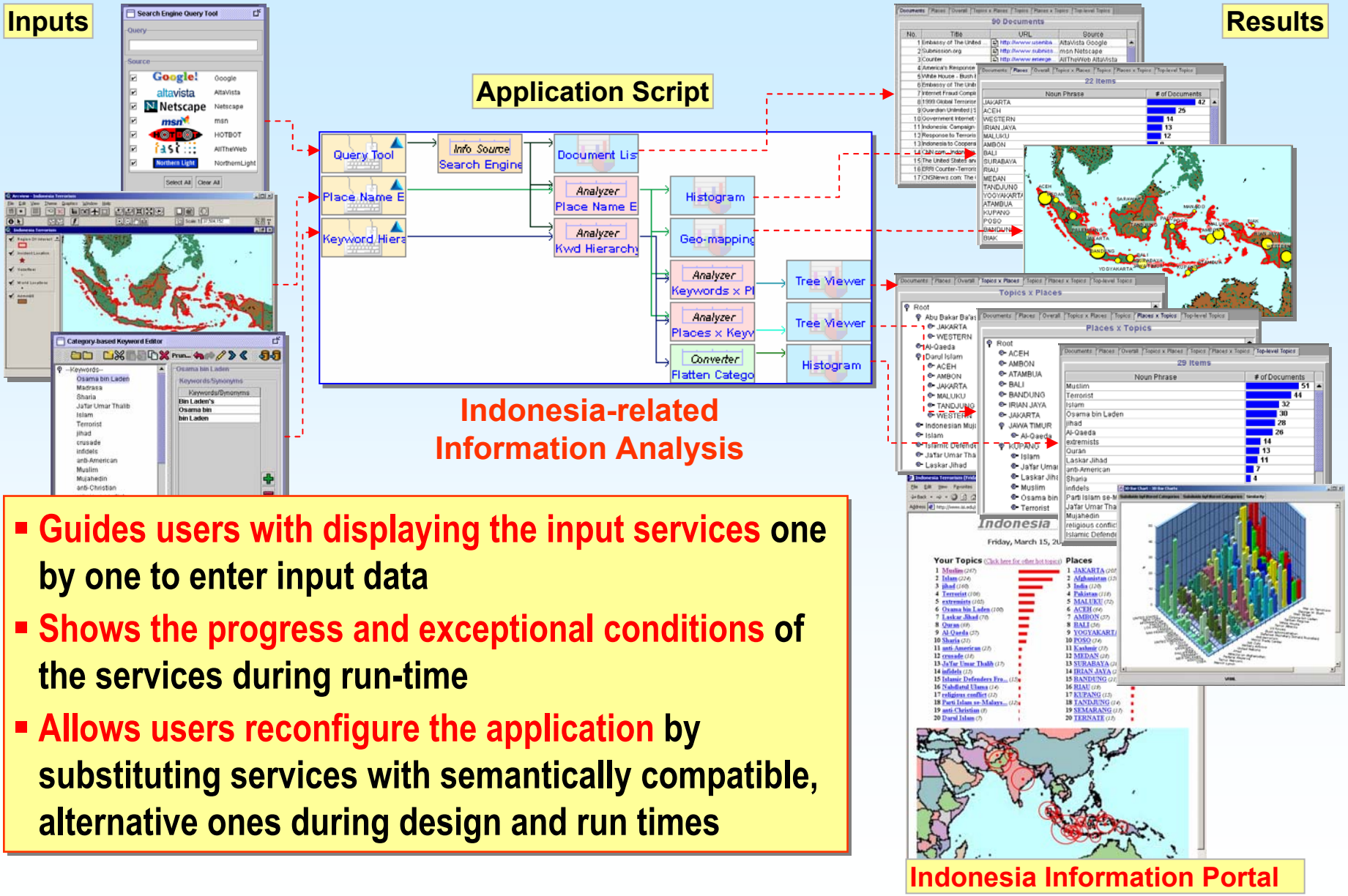
Shows service-oriented view of the application in terms of a data flow diagram



Guides users to select and combine semantically interoperable services

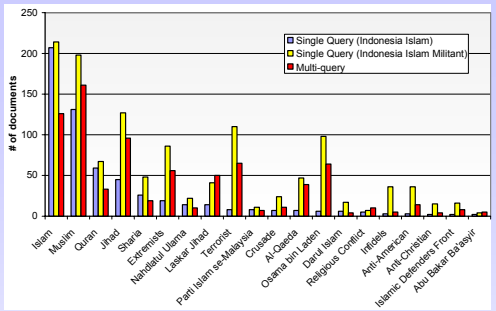
Converts and stores the application description to an XML file

Application Execution and Reconfiguration

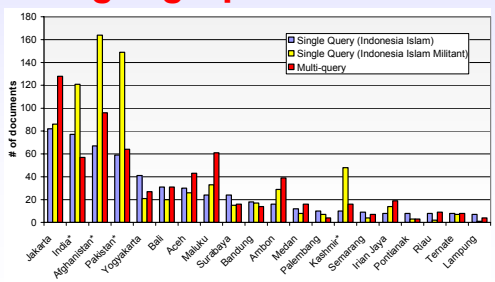


- **Guides users with displaying the input services** one by one to enter input data
- **Shows the progress and exceptional conditions** of the services during run-time
- **Allows users reconfigure the application** by substituting services with semantically compatible, alternative ones during design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

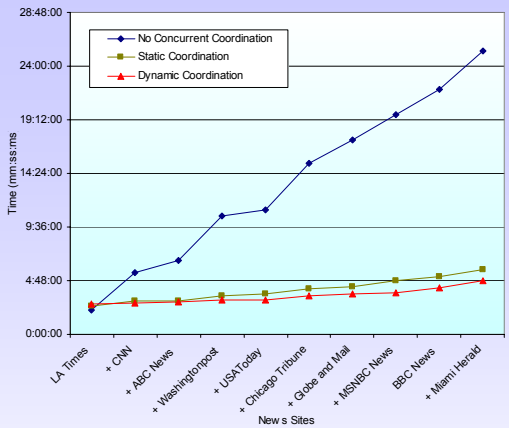
Quality Improvements Enabled by the Approach



More topics populated with more geographic relevance



ACCURACY



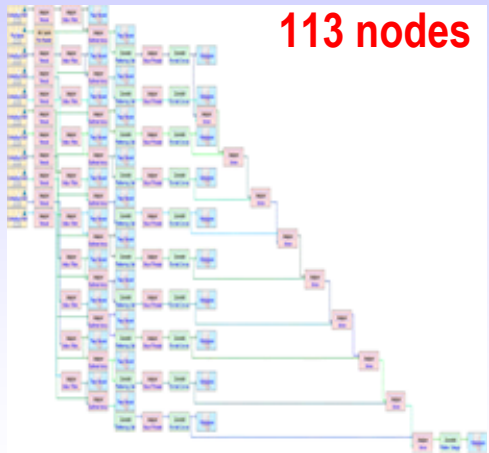
90% speed up by concurrent execution

20% more speed up by dynamic coordination

PERFORMANCE

15 nodes

113 nodes



SIMPLICITY

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