



New Efforts to Promote and Enhance Content Discovery

Sustainable Scholarship 2009

September 22, 2009

Ron Snyder

Technology and Research Manager

ATR Initiatives Supporting Content Discovery

- Data for Research Explore tool
 - Also includes external APIs for search and resource retrieval
- Theme-based content classification
- Recommender systems
- Text mining and soft linking
- Tool integration using *augmented browsing* techniques
- Application of collective intelligence techniques



Demonstrations

- Data for Research Exploration Tool
 - Enhanced faceted search capabilities
 - Use of extracted keyterms (tag clouds)
- Linking
 - Pamphlets <-> Journal Articles
 - Errata
- Augmented Browsing
- Recommender system based on LDA-generated topics and auto-extracted keywords

The JSTOR Corpus

- 5.3M journal articles online
 - 2.6M research articles
 - 1.8M review articles
- 26K pamphlets
- +33M pages of OCR'd text
- ~15 billion words
- Multidisciplinary
 - Content is organized into 50 disciplines
- High-quality bibliographic and structural metadata
 - Including +40M captured references

Data for Research (DfR) Service

A self-serve tool for obtaining research data from the JSTOR archive

- Provided by a web-interface enabling researchers to identify content of interest in the JSTOR archive and to retrieve associated datasets for research purposes

A researcher-oriented exploration tool complementing the search and browse capabilities offered by the JSTOR main site

- Exposes additional fields for enhanced searching and results filtering
- Provides data visualizations for viewing aggregate and document-level data
- Links to JSTOR main site are provided for documents in search results

Theme-Based Content Classification

The objective: Develop a means for selecting content based on a theme, or topic

- We currently approximate this using a static journal/discipline categorization model
- Topic models are a better solution
 - Provide topics which *naturally* describe these documents
 - Provide a breakdown of documents into “gists”, or topics
 - An intuitive, well-studied area in natural language processing
- Generated using Latent Dirichlet Allocation (LDA) using the JSTOR discipline classifications as the base categories
 - Blei & Gerrish, Princeton
 - http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- Theme-based content selection and filtering will be available on the DfR Explorer in October, watch for it!

Text mining and soft linking

Text mining

- Current
 - Keyword extraction using word relevancy
 - Pattern matching and heuristics
- Future
 - Key *phrase* extraction
 - Concept extraction (people, places, events, time periods)

Soft Linking - linking documents based on similarity or inferred relationships

- Document signatures using keywords and weights have proven to be useful for document similarity matching

Recommender Systems

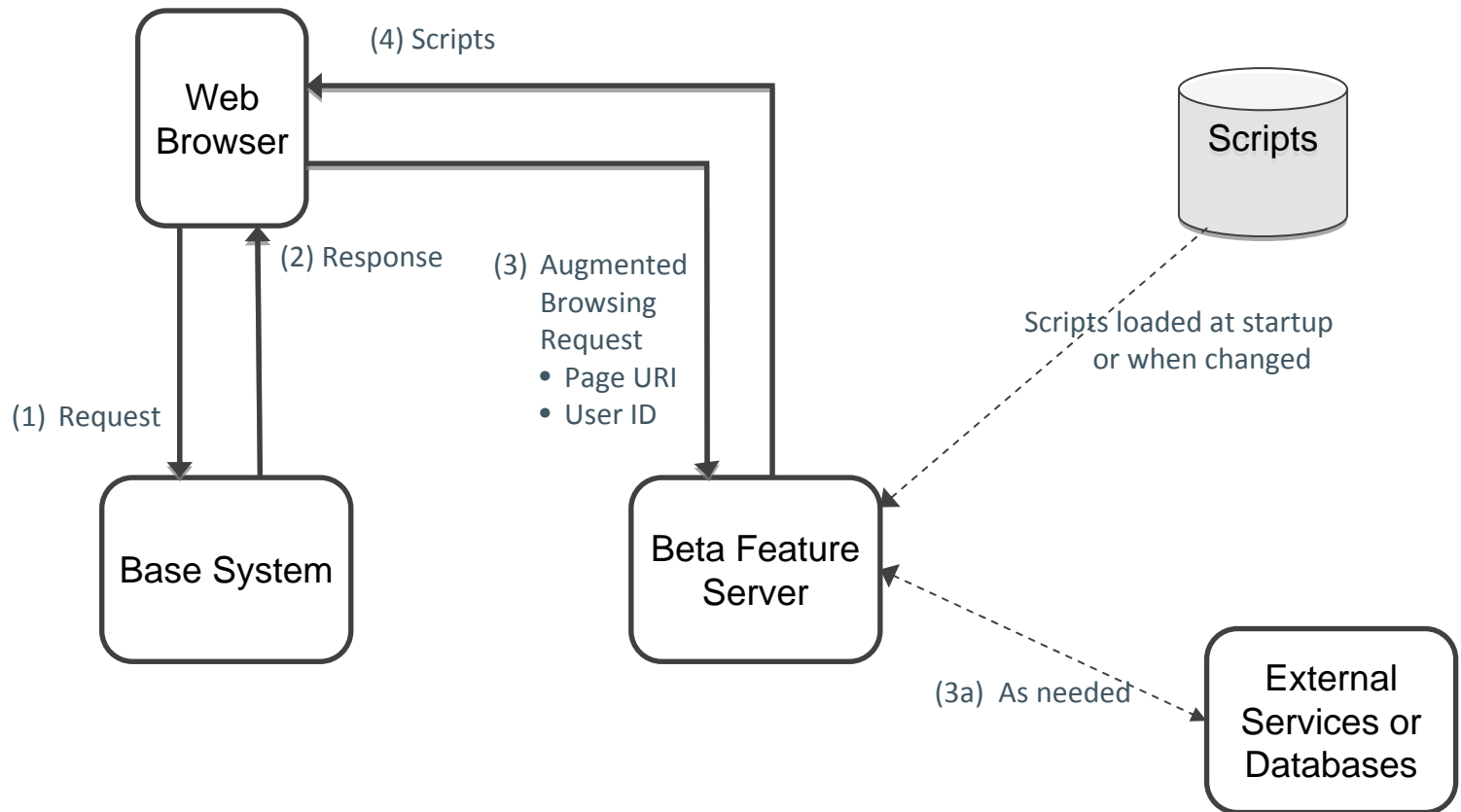
A first generation system based on document content

- Recommends similar documents using:
 - Keyword signatures from auto-extracted terms (using TF*IDF)
 - LDA topic models
- Will be incorporated into a “More Like This” feature in the DfR Explore Tool in Q4
- Also under investigation for use in a PSS recommender widget

A next generation system will be pursued in 2010 using usage-based models

- Collaborative filtering

Augmented Browsing Interaction Model



Using *Collective Intelligence* to enhance content discovery

- Collective intelligence is *the intelligence that's extracted from the collective set of interactions and contributions made by your users**
- This intelligence can be used to determine what's valuable in your application or content for a user or group

JSTOR has a large user base and a wealth of historical usage data that can be used to enhance content discovery

- Mining this data for content discovery applications will be a focus area for the ATR group in 2010

* Alag, Satnam (2009) *Collective Intelligence in Action*. Manning



Demonstrations

Journal article <-> pamphlet

- > <http://aa2vps102.jstor.org/stable/2761648?seq=11>
- > <http://aa2vps102.jstor.org/stable/2138969?seq=42>

Errata links

- > <http://aa2vps102.jstor.org/stable/3455431>

Related articles

- > <http://aa2vps102.jstor.org/stable/4477827>

PSS recommendations

- > <http://jstor-pss.appspot.com/pss/4477827>

Data for Research Explorer

- > <http://dfr.jstor.org>

World Cat linking

- > <http://aa2vps102.jstor.org/stable/30007021>