Searching and Navigating Petabyte-Scale Files Systems Based on Facets

Jonathan Koren, Yi Zhang, Sasha Ames,
Andrew Leung, Carlos Maltzahn, Ethan Miller
Information Retrieval and Knowledge Management Lab
Storage Systems Research Center
University of California, Santa Cruz

Outline of Talk

- Introduction and Motivation
  - Challenges in Petabyte Scale file systems
- Faceted Metadata
  - What it is, Where it comes from
- Programatic Interface
  - ViewFS and QUASAR
- Search/Browse User Interface
  - Faceted Search Interfaces, Personalization and Collaboration
- Conclusion

Need for a New Approach

- Monolithic hierarchy in traditional file system could be disorienting
  - Assumes users are familiar with the layout of file repository (e.g. naming conventions)
  - Multiple reasonable locations to place a file
- Keyword-based search often fails
  - Need to know how the files are described
  - Bad for exploration
  - Does not support expert users
  - Relevance ranking is hard

Faceted Search as Navigation

- Faceted search can help
  - Avoid explicitly organizing the files
  - Convert search from an interrogation to a browsing scenario
- Search becomes the primary interaction method with a file system
  - For user: faceted search is easy navigation
  - Our goal: make search a first-class function

Faceted Search

- Information Retrieval technique popular for large data repositories
  - libraries and e-commerce sites
- Faceted metadata
  - key-value pairs (keys == facets)
  - Facets group values in semantically meaningful ways
- Each facet creates a parallel categorization scheme
- Users mix and match facet-value pairs to find their files
  - multiple valid "paths" to a file

Acquiring Faceted Metadata

- Explicit Metadata
  - Leverage the easily parseable existing metadata
  - Example: ID3Tags
- Automatically Generated Metadata
  - Extract metadata from parseable file contents
    - Example: "Call me at 555-1212" -> <phone=5551212>
  - Metadata can propagate to related unannotated files
    - Examples: Soules et al.'s Connections and Provenance
- User Annotations
  - Manually provided (e.g. tags)
  - Example: Graffiti [Maltzahn 2007]

Storing Faceted Metadata

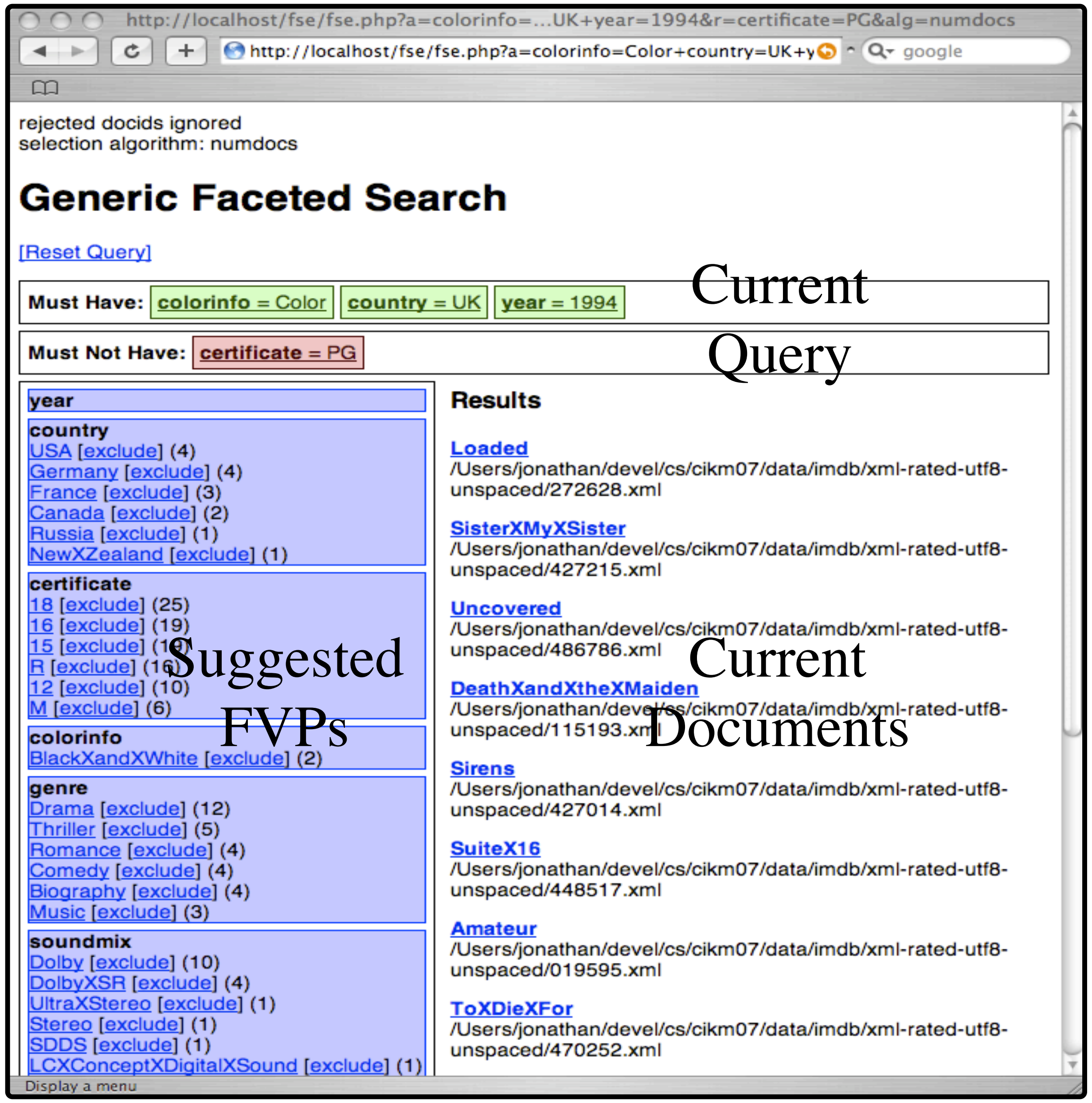
- Many file system search tools store metadata and the index separately from the file store
- Problems with Separate Stores
  - Require frequent reindexing of the store
  - Require notification method to keep the store and index synced
  - Not POSIX compliant
- Proposed new file system: ViewFS

ViewFS

- Stores metadata within the file system
  - Tight couple between the index and the store
- Modifies to POSIX interface to support both keyword and structured queries
- Queries can be used as file and directory names
  - Newly created QUASAR query language
  - Backwards compatible to existing POSIX paths
  - Designed for faceted metadata
- Virtual directories become ubiquitous
  - Current query is analogous to CWD

Search/Browse User Interface

- User creates a complex query through a point-and-click interface
- User is presented with:
  - Ranked list of matching files
  - Current query
  - Suggestions for query refinement
- User refines the query or selects the file



http://localhost/fse/fse.php?a=colorinfo=...UK+year=1994&r=certificate=PG&alg=numdocs
http://localhost/fse/fse.php?a=colorinfo=Color+country=UK+y
google
Display a menu
Current Query
Suggested FVPs
Current Documents

rejected docids ignored
selection algorithm: numdocs

Generic Faceted Search

[Reset Query]

Must Have: colorinfo = Color | country = UK | year = 1994

Must Not Have: certificate = PG

year

country
USA [exclude] (4)
Germany [exclude] (4)
France [exclude] (3)
Canada [exclude] (2)
Russia [exclude] (1)
NewXZealand [exclude] (1)

certificate
18 [exclude] (25)
16 [exclude] (19)
15 [exclude] (19)
R [exclude] (16)
12 [exclude] (10)
M [exclude] (6)

colorinfo
BlackXandXWhite [exclude] (2)

genre
Drama [exclude] (12)
Thriller [exclude] (5)
Romance [exclude] (4)
Comedy [exclude] (4)
Biography [exclude] (4)
Music [exclude] (3)

soundmix
Dolby [exclude] (10)
DolbyXSR [exclude] (4)
UltraXStereo [exclude] (1)
Stereo [exclude] (1)
SDDS [exclude] (1)
LCXConceptXDigitalXSound [exclude] (1)

Results

Loaded
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/272628.xml

SisterXMyXSister
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/427215.xml

Uncovered
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/486786.xml

DeathXandXtheXMaiden
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/115193.xml

Sirens
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/427014.xml

SuiteX16
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/448517.xml

Amateur
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/019595.xml

ToXDieXFor
/Users/jonathan/devel/cs/cikm07/data/imdb/xml-rated-utf8-unspaced/470252.xml

Challenges for Faceted Search Interfaces

- Diverse File Types
  - Large variety of file types, each with different facets
- Facet overloading problem
  - Too much metadata to present to users
- Ranking files is hard
  - Web search has explicit relationships among web pages
  - Files on a disk have few links that are useful for search

Adaptive Personalization

- Many parts of the shared file system are irrelevant to a particular user, so don't display them.
- Personalization
  - Explicit and implicit feed on query results
  - Contents of files, user access patterns
- Collaborative Recommendations
  - Compares users to each other
  - Good when you have many users

Handling Diverse File Types

- Present the facets that both prevalent in the currently selected files, and have a suitable values
  - Presents the major features of the search space
  - As search narrows, facets unique to that segment of the search space become available for query refinement
  - Used in mobile faceted search
- Meta-facets
  - Some facets are semantically similar
  - Cluster facets that have similar values together

Handling Facet Overload

- Present facets relevant to a specific user under a specific context
  - A user's interest is focused on only a small segment of the entire file system
  - System observes which files the user is most interested in
  - The facets in these documents are considered relevant to the user
- This information is shared among all the users through collaborative and content-based recommendations

Handling Ranking Challenge

- Modern IR ranking techniques leverage information about the relationship among documents
  - Anchor text in hyperlinks, Site reputation, link flux, etc.
- Files system typically do not have this information
- Using implicit user feedback
  - Connections [Soules 2005] and Provenance [Shah 2007]
- We propose to learn user models from implicit and explicit feedback

Conclusion

- Search should become a first-class function
- Faceted search allows both browsing and navigation
- Potential programatic interface for supporting faceted search (ViewFS, QUASAR)
- Outlined some problems with applying faceted search to file systems
  - Personalization and collaboration is an attractive method to overcoming some of these

Fin