



Contribution ID: 564

Type: Poster

Improving Phenix search experience with Solr/Lucene and Nutch

Thursday 24 May 2012 13:30 (4h 45m)

During its 20 years of R&D, construction and operation the Phenix experiment at RHIC has accumulated large amounts of proprietary collaboration data that is hosted on many servers around the world and is not open for commercial search engines for indexing and searching. The legacy search infrastructure did not scale well with the fast growing Phenix document base and produced results inadequate in both precision and recall.

After considering the possible alternatives that would provide an aggregated, fast, full text search of a variety of formats (text, pdf, ppt, etc) we decided to use Nutch as a web crawler and Solr/Lucene as a search engine. Nutch support of crawling multiple domains helps Phenix aggregate collaboration data from many participating institutions. The ability of Nutch to parse large variety of file formats greatly increases the search domain.

Phenix search got a substantial boost in precision by using the Solr support for faceted navigation - indexing data under custom categories and then combining text search with a progressive narrowing of choices in available categories. Most users in Phenix know how the data is structured fairly well and can limit the search to a specific area right away, for example searching only in the mail archives or published papers.

To present XML-based Solr search results in a user-friendly manner we decided to use Drupal as a web interface to

We will report on Phenix experience searching with Solr/Lucene, Nutch and Drupal.

Authors: MORRISON, Dave (Brookhaven National Laboratory); SOURIKOVA, Irina (Brookhaven National Laboratory)

Presenter: SOURIKOVA, Irina (Brookhaven National Laboratory)

Session Classification: Poster Session

Track Classification: Collaborative tools (track 6)