POLAR DOMAIN DISCOVERY WITH SPARKLER

Chris a. Mattmann[2] , Karanjeet Singh[1], Nithin Krishna Ottilingam[1], Siri Jodha S. Khalsa[3], Ruth Duerr[4]

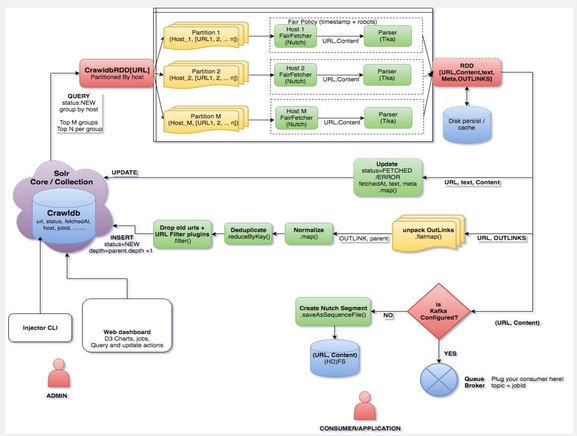
University of Southern California, NASA Jet Propulsion Laboratory, University of Colorado, The Ronin Institute

Motivation

- The scientific web is the collection of Scientific Data Repositories (SDR) available on the internet. This web-sample study[1] of 100 SDRs in the internet indicates that they are often multidisciplinary and encompass data in variety of formats including multimedia, text, statistical, GIS etc.
- Conventional scrapers and crawlers do not have domain knowledge, lack context and are agnostic to data content. They do not have the ability to selectively crawl parts of SDRs which are specific to a particular domain of interest. They do not possess the ability to understand the contents of document and predicting if it's of interest or not.
- Most scientific literature is SDRs exists outside the realm of the surface web [2]. There exist multiple data access and availability challenges while downloading data from SDRs [3], including data exchange issues and data download delays due to lack of data replication.
- Conventional scrapers and crawlers are agnostic to data access policies, they are single threaded bound by data exchange times.
- Aggregating domain specific documents from the scientific web is time-consuming and difficult. This makes it hard to generate any meaningful domain-specific insights.

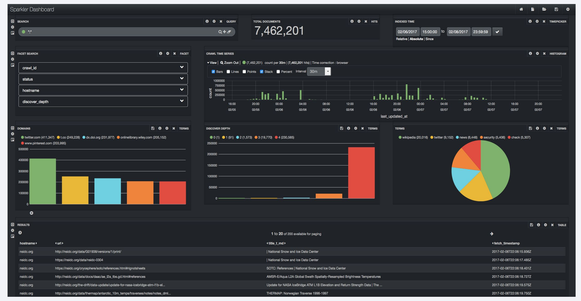
Sparkler

- *Sparkler*, an open-source, extensible, horizontally scalable crawler which facilitates high throughput and focused crawling of documents pertinent to the polar domain.
- We extend Sparkler with this model to focus crawling, to polar specific documents. Sparkler avoids disruption of service by 1) partitioning URLs by hostname such that every node gets a different host to crawl and by 2) inserting delays between subsequent requests.
- With an NSF-funded supercomputer Wrangler, we scaled our domain discovery pipeline to crawl about 200k polar specific documents from the scientific web, within a day.



Sparkler Dashboard

A near real time crawl analytics dashboard to search through the fetched content and metadata. Assist to guide the subsequent crawl cycles.



Domain Relevance Model (DRM)

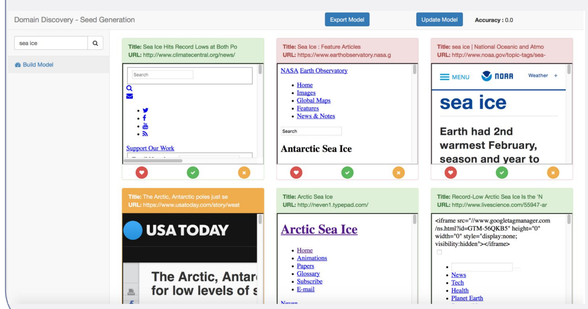
We aim to provide context and domain knowledge to the crawler by building a machine learning model (DRM) which is capable of predicting the relevance of a given document to the domain.

We build a SVM[4] (Support Vector Machine) classifier which contains a vocabulary of domain ontologies and english stop word filtered text extracted from a web page as input. The classifier predicts if the web page is relevant to the domain or not.

We experimented two SVM models - (a) Linear kernel with SGD (Stochastic Gradient Descent) and (b) RBF (Radial Basis Function) kernel. While SGD classifier uses stochastic approximation to solve optimization problems, RBF leverages a popular gaussian based kernelized learning algorithm for SVM classification. We use Stratified 10-Fold cross validation to evaluate the results.

ML Algorithm	Model	Labeled Examples	Training Accuracy
SVM	Linear with SGD	336	68.54%
	Radial Basis Function	336	**83.83%**

Our user interface[6] allows subject matter experts to search for documents on the web and incrementally train the DRM.



Polar Deep Insights

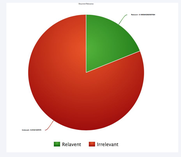


Figure: Ratio of relevant to irrelevant crawled documents v

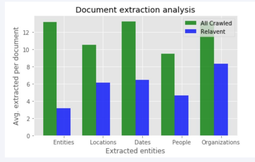


Figure: Ratio of relevant to irrelevant crawled documents

The results we preset are on 50k documents crawled from top polar and earth science data repositories and websites listed below.

ncdc.noaa.gov, permafrosttunnel.crrel.usace.army.mil, icdc.cen.uni-hamburg.de, neo.sci.gsfc.nasa.gov, nichols.edu, toolkit.climate.gov, psc.apl.uw.edu, nsidc.org, antarcticglaciers.org, ospo.noaa.gov, pmel.noaa.gov, ccin.ca, epa.gov, cpom.org.uk, climatecentral.org, www2.usgs.gov, neven1.typepad.com, arctic.noaa.gov, antarcticsun.usap.gov, arctic.gov, esa.int and climate.gov

Our domain relevance model found only 9578 documents of the 50k relevant.

Insight Generation

- We extract dates, times, named entities, locations, terms of interest (etc) from documents and study spatial, temporal and conceptual trends to generate insights.
- Extracting these entities from documents is computationally expensive.
- The DRM ensures that we are able to retrieve atleast as much meaningful insights from a subset of relevant documents as compared to a broad set of documents.
- Though fewer documents are parsed and fewer entities per document on average are extracted, we observe lesser noisy insights.

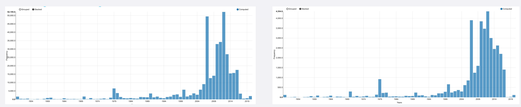


Figure: Temporal 'sea-ice' trend remains same before and after filtering irrelevant documents. Significant events are at 1970, 1979, 2000, 2007 and 2012.

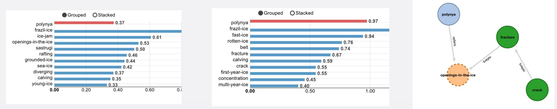


Figure: Correlation trends for 'polynya' is more meaningful. Returns terms related to 'polynya' and thus produces richer inferences.

Future Work

- Scale up our pipeline for continuous crawling, extraction and visualization.
- Investigate more factors contributing to the Domain Relevancy. For eg: Count of Inlinks/Outlinks.
- Couple insights with it's source (scraped web pages) for evidence.
- Include more polar specific ontologies to gain specific insights about the cryosphere.
- Continue improving visual interface.

Architecture

Components of Domain Discovery System (Figure i)

1) Domain Exploration : Using the KB, DRM and commercial/scientific search engines to generate URLs specific to the DOI. Detailed in the figure ii.

2) Focused Crawling : Using DRM to adjudicate crawler decisions on expanding relevant document subtrees.

3) Extraction : Parsing crawled documents to extract entities of interest and build an extraction index.

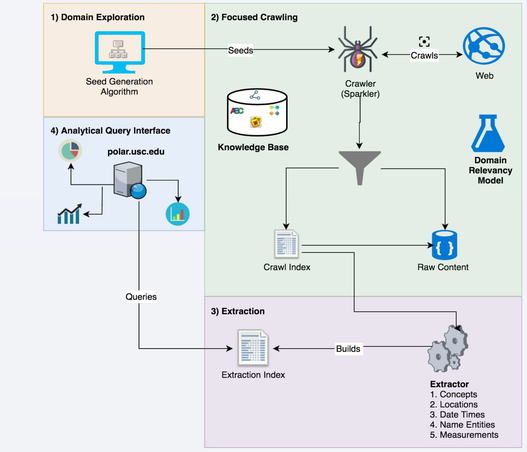
4) Analytical query interface : A dynamic visual interface to query and aggregate entities of interest from the extraction index and generate insights.

Components of Seed Generation Algorithm (Figure ii)

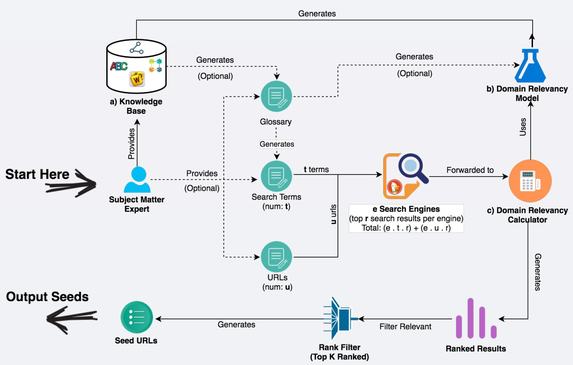
a) Knowledge Base (KB) : The glossary of terms relevant to the domain of interest(DOI) and optionally their relationship with each other.

b) Domain Relevancy Model (DRM) : Machine learning model built on top of the KB and a Subject Matter Expert(SME) annotated dataset, describing the relevancy of URLs to the DOI.

c) Domain Relevancy Calculator (DRC) : Uses DRM to compute confidence values for relevancy.



Domain Discovery System



Seed Generation Algorithm (Figure ii)

References

Polar Deep Insights Earth Cube : http://bit.ly/2rEuWvn

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com