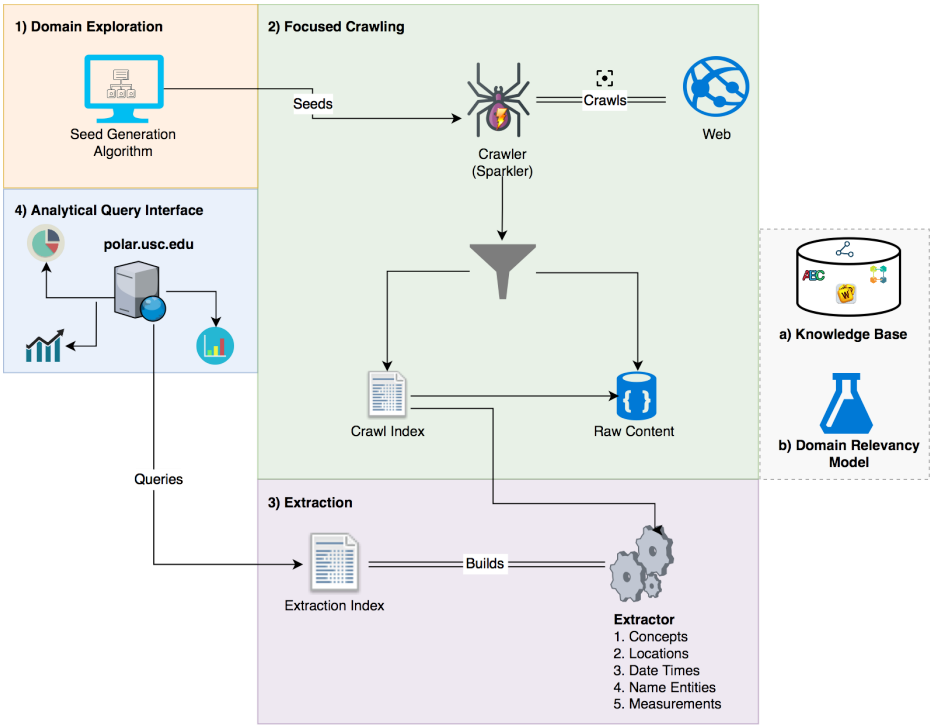
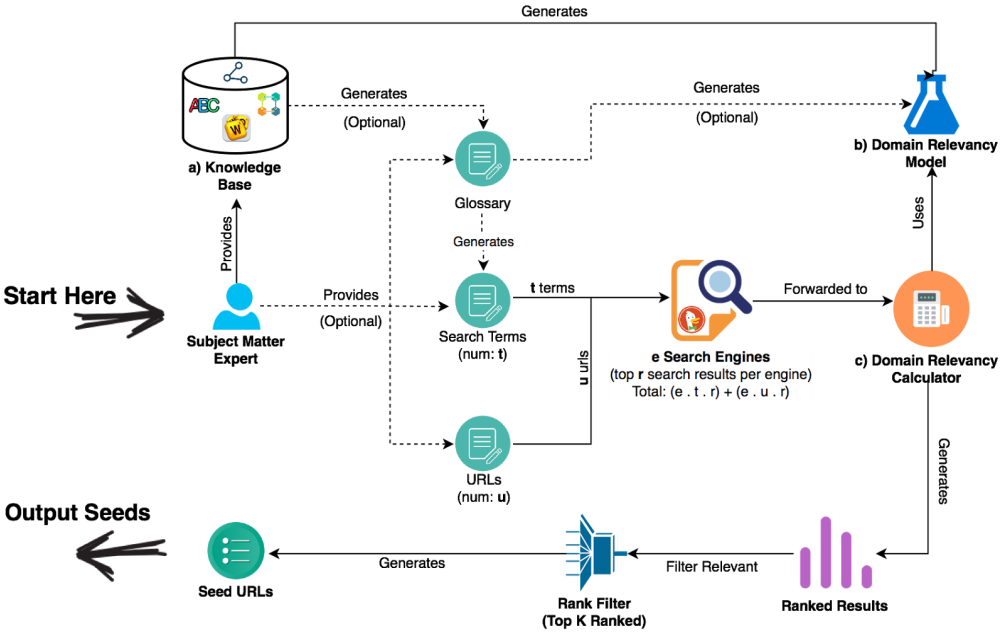
Polar Domain Discovery with Sparkler

The scientific web is vast and ever growing. It encompasses millions of textual, scientific and multimedia documents describing research in a multitude of scientific streams. Most of these documents are hidden behind forms which require user action to retrieve and thus can't be directly accessed by content crawlers. These documents are hosted on web servers across the world most often on outdated hardware and network infrastructure. Hence it is difficult and time-consuming to aggregate documents from the scientific web, more so ones those relevant to a specific domain. This makes it hard to generate any meaningful domain-specific insights. We present an automated domain discovery system (Figure 1) using *Sparkler*, an open-source, extensible, horizontally scalable crawler which facilitates high throughput and focused crawling of documents pertinent to the polar domain. With this set of highly domain relevant documents, we show that it is possible to answer analytical questions about the polar domain. Our domain discovery algorithm leverages prior domain knowledge to reach out to commercial/scientific search engines to generate seed URLs. Subject matter experts then annotate these seed URLs manually on a scale from highly relevant to irrelevant. We leverage this annotated dataset to train a machine learning model which predicts the 'domain relevance' of a given document. 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.



Domain Discovery System (Figure i)



Seed Generation Algorithm (Figure ii)

<u>Components of Domain Discovery System</u> (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.

<u>Components of Seed Generation Algorithm</u> (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.

Figure 1