SAND2021-14483PE

Sandia National Laboratories

Exceptional service in the national interest

Intelligent Web Crawling

Sheetal Narvekar

SANDIA LABS, LIVERMORE CA

Nov 15, 2021

ENERGY NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Web Crawling

Web Crawler

A Web Crawler, also known as a Web Spider, crawls the world wide web using a systematic approach to find relevant information.

Purpose

A web crawler is used for storing, analyzing, and indexing information on a page and finding additional pages by following links.

Common Use cases

Standard uses for web crawlers include indexing content for search engines such as Google® or Bing®, Real Simple Syndication (RSS) feed generation, web site testing, and many other reasons.

Image Source: https://pxhere.com/en/photo/499507/ https://askleo.com/whats-the-difference-between-the-web-and-the-internet/ https://commons.wikimedia.org/wiki/File:Web-browser-openclipart.svg Accessed Nov. 2, 2021.

How Web Crawling Works

Seed URL

Each crawl begins with one or more URLs that constitute as the seed set.

Frontier

The frontier is the set of links that have yet to be visited by the crawler. The frontier is constantly updated as the crawler downloads pages and finds new links to visit.

Indexing

The retrieved information is organized and ranked based on its relevance.

Image Source: https://pxhere.com/en/photo/887957 https://www.freeimageslive.co.uk/free_stock_image/searchbuttonjpg https://www.flickr.com/photos/24617197@N02/31712357957 Accessed Nov. 2, 2021.

Challenges

- The sheer volume of information available via the internet and large databases leads to difficulties in locating the correct information relevant to a task.

- Results depend on the parameterization of the search engines' crawlers

- The user conducting the search needs to possess a moderate understanding of the subject matter being sought

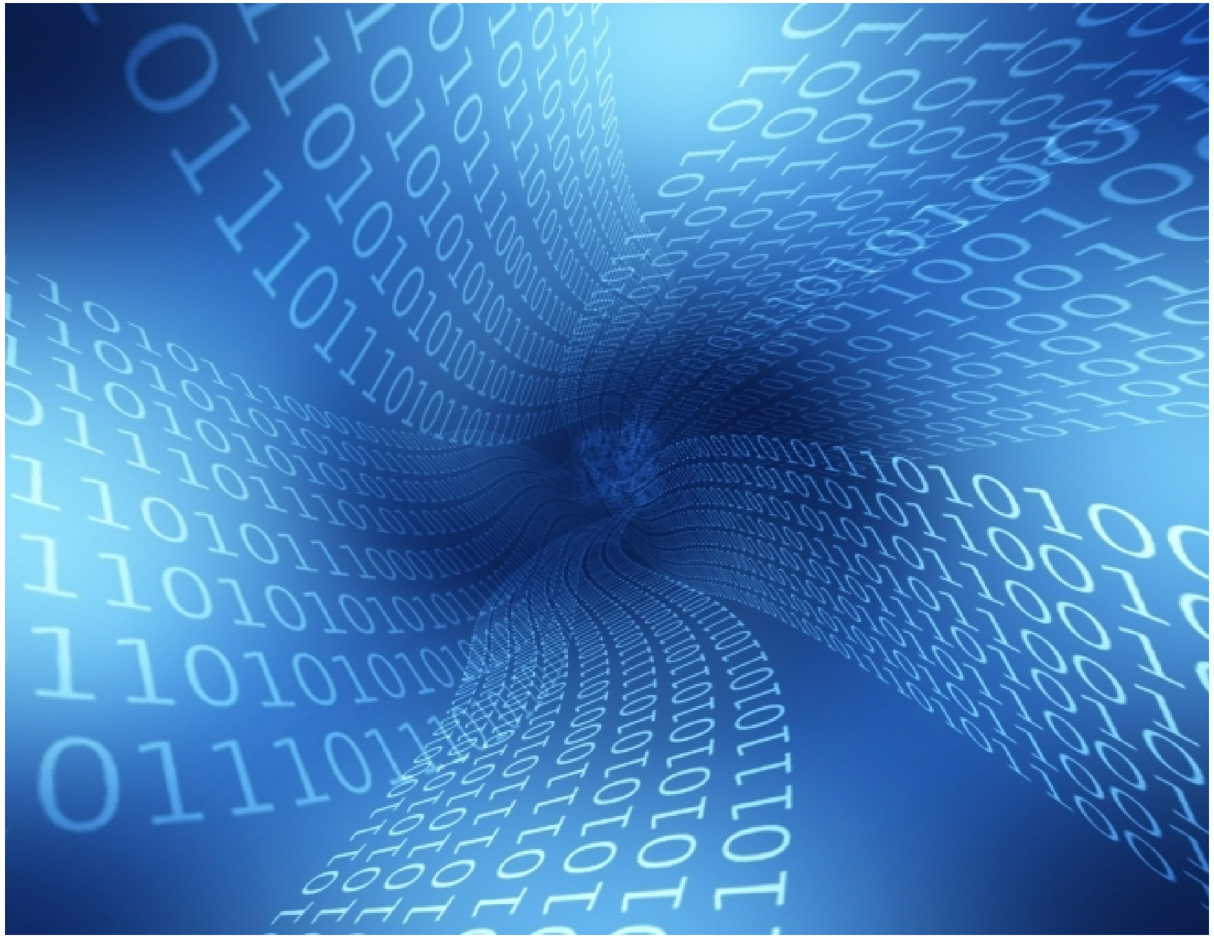


Image Source: https://publicdomainpictures.net/en/view-image.php?image=45242&picture=networking Accessed Nov. 2, 2021.

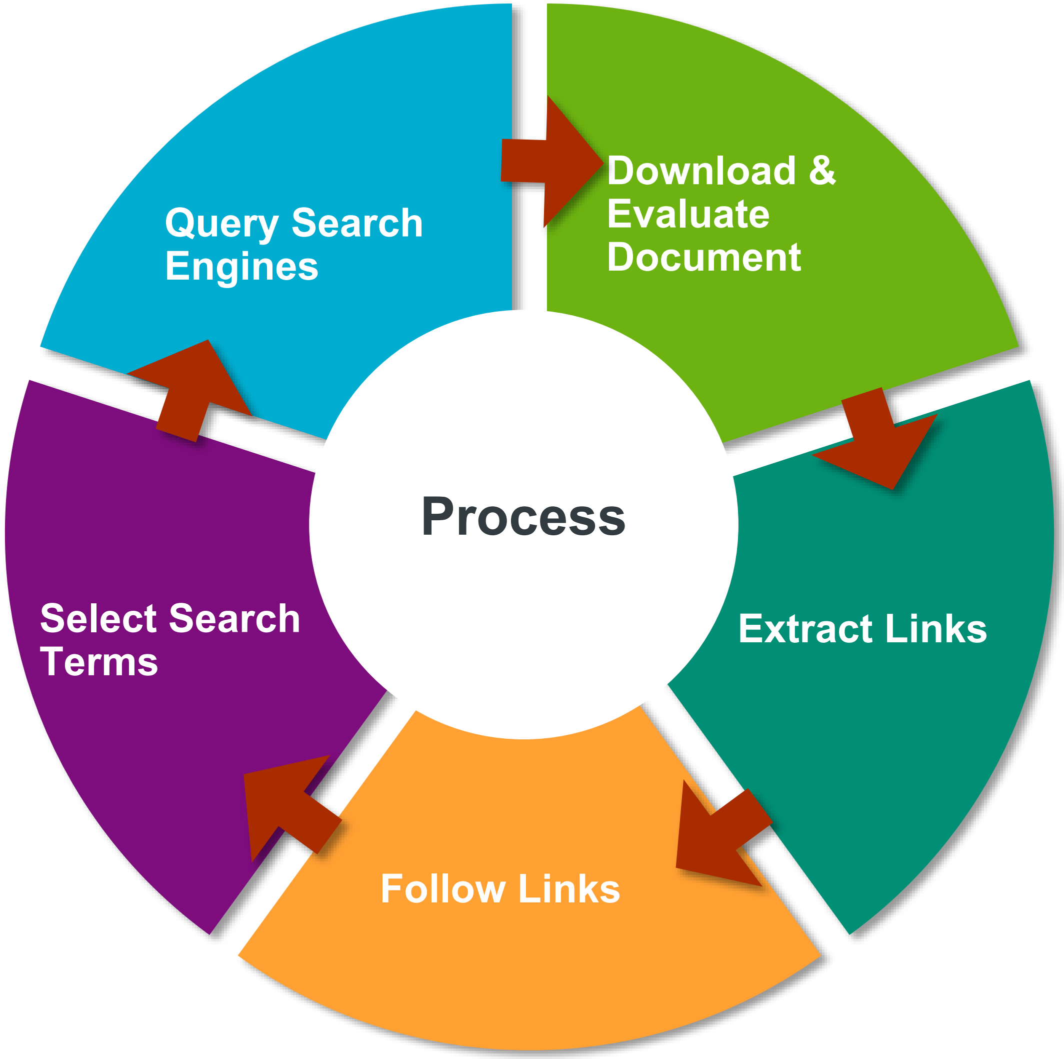
Intelligent Web Crawling

Avondale is an *Intelligent* Web Crawler that uses parameterized techniques and filtering rules to deliver the information pertinent to the task at hand.

- Avondale is a smart agent that analyzes different forms of user input and intelligently produces efficient results, thus significantly reducing the time spent in manual search and analysis.

Intelligent Web Crawling | Process

- User can feed Avondale one or more documents
- Avondale will analyze their topics
- Create its own set of key terms
- Compare that with thousands of websites
- Then model each of the topics and give you the best match for what you are looking for



Capabilities

Text Analysis

- Avondale's text analytics platform helps both the web crawling system determine relevant pages, and lets you input your own pages and have them analyzed.
- The system can derive the main topics, as well as compute key terms, and calculate similarity to other documents.

Document Analysis

- Analyze text documents with natural language processing algorithms
- Avondale analyzes and compares documents using topic modelling based on search keywords and phrases.

Offline Data Analytics

- Avondale can also scan through hard-drives and databases, as well as previous web crawls and apply the same powerful data analytics.
- This allows you to search through other large datastores in the same way that you can search the web.

Image Source: https://sitesmatrix.com/blog/texting-games-games-to-play-over-text/ https://pixabay.com/vectors/search-loupe-document-97587/ https://www.pngall.com/database-png Accessed Nov. 10, 2021.

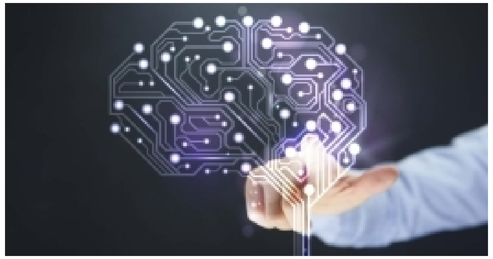
Capabilities

Rank Based Search

- Avondale analyses and ranks information on how valuable it will be to you.
- This is done based not only on the number of keywords it contains, but also on how valuable each keyword is
- If there are any shared phrases or sentences, such as a citation or quote, as well as having a topic that matters to you.

Image-text fusion

- Avondale uses patented algorithms to turn sections of images into "word hashes" that can then be searched for.



Language Support

- Ability to reliably search for documents across languages

Image Source: https://cutecoder.org/software/first-dabbling-machine-learning/ https://commons.wikimedia.org/wiki/File:Pe-tan_with_multilanguage_sign_-_001.jpg Accessed Nov. 10, 2021.

Strengths

Parametrized Metrics

A search can target the entire content of a document, rather than just searching for a few keywords. The search terms can be predefined by the person parameterizing the crawl, or they can be selected from documents of interest based on any number of term ranking algorithms.

Search Engine Integration

The customized query terms are sent by the crawler to an Internet search engine, such as Google® or Bing®. The purpose of this process is to seed the web crawler with URLs that represent good starting places for searching the Internet for relevant content.

High speed evaluation

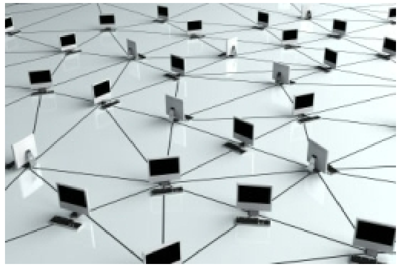
Avondale assesses millions of webpages in a day’s time and can turn days work to hours.

Image Source: https://commons.wikimedia.org/wiki/File:Filter.svg https://commons.wikimedia.org/wiki/File:Hyperlink-internet-search.svg https://www.pngall.com/speed-png Accessed Nov. 10, 2021.

Strengths

Distributed & Scalable

Avondale can perform large-scale, distributed, intelligent web crawls using many document analysis metrics.



Cross Platform

Avondale is a pure Java/Web application program and supports different operating systems.

Image Source: https://www.globo.tech/learning-center/protecting-your-dns-server-against-ddos-attacks/ https://blendervr.limsi.fr/lib/exe/detail.php?id=start&media=os-logos.png Accessed Nov. 10, 2021.

Avondale

Avondale is a versatile capability that has been developed and used across various contexts to assess large amounts of information, which reduces analysts' cognitive burden, increases decision-making confidence, and connects findings to customer need.

Questions?

Exceptional service in the national interest

Thank you.

Image Source: https://www.shutterstock.com/ Accessed Nov. 10, 2021.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.