HYPERION GRAY

SiteHound Walk-Through Guide

This is a domain discovery tool, which means that it assists a user with finding websites that are relevant to a particular topic of interest, or domain. SiteHound is designed to conduct domain discovery in a faster, automated, and more scalable way by combining web crawling and machine learning technologies.

1. The first step is to create a workspace. You can name it whatever you like, but we recommend labeling it something that corresponds to the topic or domain you're interested in searching.

2. The next step is to provide the SiteHound system with keywords that are related to the topic, and are likely to be found on the sites you're seeking to discover. We would recommend no more than half a dozen each time (you can run it several times with different keywords). Keep in mind that the more keywords you add, the broader returns you'll receive. Try to pick words that are specific to the domain, and are at least fairly uncommon outside of it. Vague words lead to vague results.

3. After this, you can provide "seed sites," or example sites, for the SiteHound to pick up a 'scent.' This will help the SiteHound understand the types of sites you're hoping to find.

4. Armed with this information, your next step is to turn the SiteHound loose to do an initial search for relevant sites. It will go out to several search engines and query the inputted keywords.
Once it's returned a sufficient number of sites, you can review the results of the initial search and identify which results are relevant and which are irrelevant. This allows the SiteHound to recognize what characteristics matter, without the user needing to explicitly specify these characteristics. This labeling process is crucial, and the quality of the results depends in part on how consistent these training sets get classified.
If the results returned are lacking, you can choose to return to the keywords section and add or subtract keywords, and perform the initial search again. This iterative process lets you quickly tune the system before running it on a larger scale.

5. Once you're satisfied with the positive and negative examples, it's time to build a model, which is where the machine learning component attempts to make sense of the information. You can see the features and quality of this model, and will receive some advice about moving forward or improving the labeling.

6. Now, using this model, smarter crawling is possible, and the SiteHound will undergo a new round of fetching and labeling results, similar to the step 4. You won't need to change the keywords (unless you want to); just label the new results.

7. Again when you're confident with the labeling in the new system's model, you can move on.

8. The next step is to do a large-scale discovery crawl. This takes your search from a few hundred sites to a few thousand sites. You can just let the system run and review the results every so often.

9. Once you have this collection of relevant sites, you can:
   a. pass them along to other systems for further data collection and analysis steps.
   b. view the results as snapshots, and you have the ability to pin sites of interest.
   c. search through the results to find sites that contain a particular keyword.
   d. browse the results by category--for example: a news, blog, or forum post,
      and you can see results based on the language in which they appear.

Hyperion Gray LLC
www.hyperiongray.com | contact@hyperiongray.com