Using Similarity Scores to Identify Organizations of Interest by Website

September 2024

Authors

Scott Cody, Matthew Ring, Anne Roubal
Westat Insight

Submitted to

U.S. Department of Labor
Chief Evaluation Office
200 Constitution Avenue, NW
Washington, DC 20210

Project Officer

Tara Martin

Submitted by

Westat Insight
1310 North Courthouse Road
Suite 880
Arlington, VA 22201

Project Director

Dr. Allison Hyra

Disclaimer

This report was prepared for the U.S. Department of Labor (DOL), Chief Evaluation Office by Westat Insight and its partners, under contract number 1605DC-18-A-0018. The views expressed are those of the authors and should not be attributed to DOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Acknowledgments

The report was prepared for the U.S. Department of Labor Chief Evaluation Office (CEO). The authors received invaluable support, guidance, and contributions from Tara Martin and Kuang-chi (Kacie) Chang. The authors thank Allison Hyra and Kelsey Gray from Insight Policy Research, who provided constructive feedback throughout the project and April Fales and Carly Mihovich who assisted with manual website review.

Suggested Citation

Cody, S., Roubal, A., & Ring, M. (2024). *Using Similarity Scores to Identify Organizations of Interest by Website.* Westat Insight for U.S. Department of Labor Chief Evaluation Office.

The mission of the U.S. Department of Labor (DOL) includes promoting the welfare of wage earners and job seekers by improving working conditions, advancing opportunities for profitable employment, and ensuring work-related benefits and rights. DOL works closely with employers, benefits providers, unions, trade associations, and other organizations to advance this mission.

Multiple agencies and programs within DOL, such as the Occupational Safety and Health Administration, Employment and Training Administration, and Office of the Solicitor, may have a need to identify different categories of organizations they work with. For example, they may seek to identify employment service providers, benefits providers, local unions, or even specific types of employers. Such identification can support data collection, outreach, compliance, and enforcement activities.

To assist DOL with identifying specific organizations within a category, Insight Policy Research (Insight) worked with the Chief Evaluation Office (CEO) to develop an automated approach for identifying websites of potentially relevant organizations. The automated approach is not designed to replace human judgment; rather, it seeks to prioritize websites for manual review to make human judgment more efficient. The approach can be used when the characteristics of interest (such as the type of organization or the types of occupations the organization employs) are not available in other datasets.

This approach for identifying websites of potentially relevant organizations begins with creating a "training set"—a set of websites from organizations that are known to have the characteristics desired by DOL. Next, the approach identifies potential search terms that can be used in internet searches to find websites like those in the training set. The next step is calculating a similarity score measuring the overlap between text on websites identified via search and text on those in the training set. The similarity score does not guarantee that a website is from an organization with the desired characteristics. However, websites with a higher similarity score are more likely to have the desired characteristics than websites with a lower similarity score—specifically, they are more likely to be captured by the search terms. Finally, a manual review of the websites with the highest similarity score is needed to identify which websites are indeed relevant organizations. This brief describes the automated approach for identifying websites of interest using web scraping[1] and natural language processing (NLP)[2], provides a hypothetical example, and summarizes the lessons learned in applying this process.

[1] Web scraping is a computer software technique that is used to extract relevant data from websites.

[2] Natural language processing is a set of techniques by which a computer program can process text written by humans, such as on webpages or documents.

A. Steps for Identifying and Scoring Potentially Relevant Websites

Key Terms

Characteristics of interest are the characteristics of the types of organizations DOL seeks to identify. For example, DOL may seek to identify local nonprofit organizations that provide job search assistance. In that case, characteristics of interest would include "local," "nonprofit," and "provide job search assistance."

Known websites are the websites from organizations known to have the *characteristics of interest*. These should be identified at the outset of a matching exercise.

Training set is the complete set of *known websites*. This set is used as the benchmark for identifying additional *potentially relevant websites*.

Potentially relevant websites are organizations' websites identified through the automated search. These websites are only *potentially relevant* because the search may also return websites from other types of organizations that are not of interest.

Similarity score measures the degree to which each *potentially relevant website* matches all the websites in the *training set*.

Web crawling is the process by which a computer program accesses websites.

Web scraping is the process by which a computer program extracts data from websites.

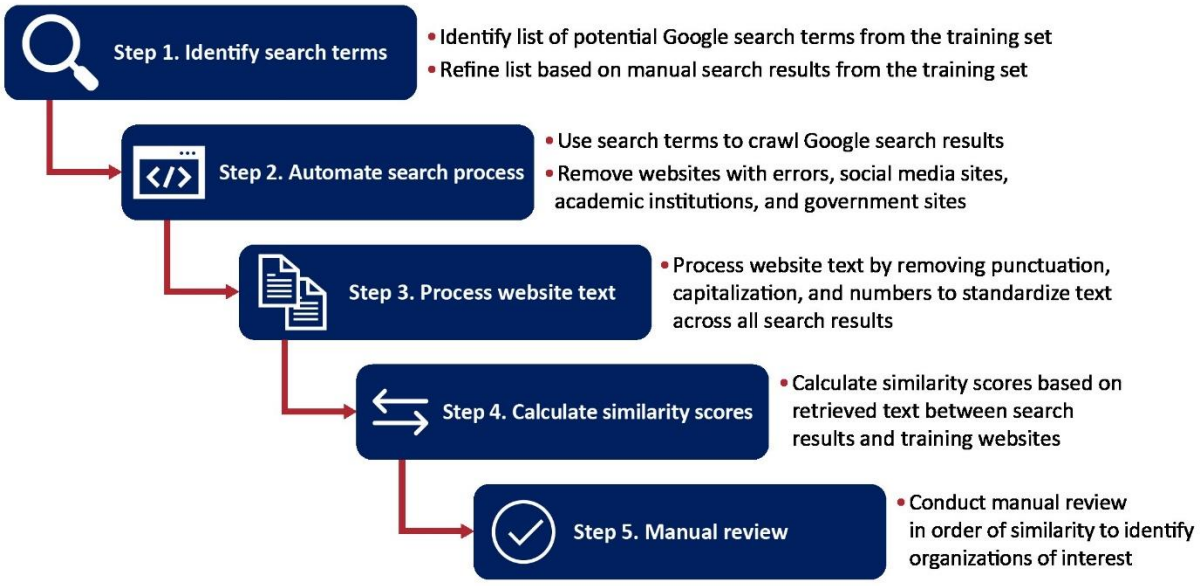
The foundation of our approach is the identification of a training set—a list of websites from organizations known to have the characteristics of interest. In our case, DOL provided a training set of websites. The websites are referred to as the "training set" because they are used to train the algorithm that determines if other websites are similar to those in the training set.

Training sets can vary in size, and there is a tradeoff that occurs when moving from smaller to larger training sets. Having too few websites will result in a model that is trained too narrowly, and the model may miss relevant websites. Adding more websites to a training set creates more confidence that the set captures the variety of organizations of interest. However, as more websites are added, more noise is introduced into the training set, increasing the likelihood that the model will identify irrelevant websites. There is no rule for the optimal number of websites; researchers can "tune" models by increasing the training set if it appears relevant websites are not captured, and decreasing the training set if it appears too many irrelevant websites are captured.

Once the training set is identified, the five-step approach described in figure 1 can begin. The first four steps aid in the development of an algorithm that identifies and prioritizes potentially relevant websites. First, we manually identify relevant search terms. Next, automated web searches identify potentially relevant websites. We then extract text through NLP and use that text to calculate similarity scores. These scores reflect the overlap in text between each potentially relevant website and the training set. Potentially relevant websites are then sorted by similarity score. The final step uses human reviewers to identify the most relevant websites from the list of highest priority websites.

This process is intended to make searching for relevant organizations efficient. Calculating a similarity score and sorting search results by that similarity score enable reviewers to target their efforts to those websites, and subsequently organizations, most likely to be relevant.

Figure 1. Steps for Matching Websites to the Training Set



Step 1. Identify Search Terms and Search Phrases

After a training set is identified, the first step in the process is to manually identify search terms and phrases that can be used to identify potentially relevant websites. Search terms/phrases should—

- Be as unique as possible to the types of organizations of interest.
- Reflect the characteristics of interest.
- Include variations of words to increase the probability of finding hits (e.g., "bicycle" and "bike" may yield different results).

Search terms and phrases can be identified by conducting a manual review of websites in the training set to identify unique words or phrases that seem relevant to the characteristics of interest and are shared by multiple websites in the training set. Analytics such as Google Trends can also help identify search terms unique to certain geographies (e.g., States) or points in time (e.g., a specific month). For example, one could use Google Trends to restrict to the weeks after the election campaign stops in specific states to include specific results about unemployment policies.

We suggest validating search terms before including them in the automated search. Validation steps can include conducting a manual search using the terms and reviewing the results to—

1. Confirm that at least some of the websites in the training set appear within the first 50 search results.
2. Assess whether individual terms or phrases result in the first 50 search results having a high concentration of clearly irrelevant websites.

We suggest only including those search terms or phrases that demonstrate through validation that they can identify a concentration of websites from organizations of interest. This refinement may take a few iterations.

Step 2. Automate Search Process

Web searches may return different results, depending on whether individual search terms or phrases are used or combinations of search terms or phrases are used (e.g. "road" and "bicycle" versus "road bicycle"). We created a set of almost 200 unique combinations of search terms and phrases to be used in our analysis. The optimal number of unique combinations will likely vary by the size and heterogeneity of the training set and the specificity of the potentially relevant websites.

We used a web crawler—a computer program that accesses websites much as a human does—to execute each of these unique search combinations and gather the links that appear in the search results. We developed a tool in the Python open-source programming language to crawl Google and scrape links to search results. The web crawling tool conducted Google searches for each of the search terms or phrases deemed appropriate in step 1. The algorithm retained all URL links from the first five pages of Google results (excluding the links that were Google URLs and any broken links). Depending on the number of searches and the number of relevant websites, this step may result in a list of hundreds or even thousands of websites.

For our analysis, we focused only on the root links of the websites identified. We cleaned our search results by shortening URLs to their root links and removing any duplicates (e.g., if a website's homepage, "about us," and "contact us" pages were identified by the search, we retained only the root URL for the website). For example, if the URL https://www.dol.gov/general/topic/wages/minimumwage was returned, we shortened it to https://www.dol.gov through our automated code.

Additional steps can be taken to clean the search results. For example, if you are confident that educational organizations and government agencies are *not* organizations of interest, you can remove root links pointing to domains with ".edu" and ".gov." You can also manually remove any root links that point to common websites such as Facebook, Wikipedia, or news websites. If there are a series of common root links to remove you could program the code to automatically remove them.

It is important to note that our analysis did not include subpages on organization's websites. While ideally we would match websites based on the total information across all website pages associated with their URL, the resources required for such an effort were beyond the scope of this study. To the extent that organizations differ in what information is contained on the root URL versus subpages, our approach will underestimate the similarity between two websites; however, it is possible to include subpages in the process.

Step 3. Process Website Text

Once we finalized our list of unique root links from potentially relevant websites, we scraped the web page text from each root link. We then standardized these texts using NLP.

In particular, NLP methods were used to remove punctuation, excess spaces after words, stop words, and duplicate words from the text scraped from each website. Stop words are uninformative words, such as "the," "and," or "a," that do not help describe the contents of a website. By removing these words, we increase the efficiency and accuracy of our models by comparing to fewer, more descriptive words.

Step 4. Calculate Similarity Scores

Once the potentially relevant websites were identified and the text from those websites was scraped and processed using NLP, we compared each potentially relevant website to the websites in the training set in Microsoft Excel. For each potentially relevant website, we measured similarity against each of the training set websites' root links. We then calculated the potentially relevant website's median similarity score based on those individual comparisons to the training set websites in Excel.

To measure similarity, we computed Jaccard similarities in Python using only the Pandas package. Jaccard similarities reflect the intersection divided by the union ($\cap/\cup$) for two texts. The number of unique words in common divided by the number of unique words overall provided our metric of similarity. Results ranged from 0 to 1, where 0 implied the texts had no words in common with the training set, and 1 implied the texts were identical. Jaccard similarities only work between two texts (one-to-one), not one text compared with multiple texts (one-to-many). Thus, the median similarity score was reported for the combination of potentially relevant websites to each website in the training set.

After the similarity score for each potential website was calculated, we sorted the database by each website's similarity to the training set. The websites with the highest similarity score are most similar to the training set. Websites above a defined cutoff (e.g., 0.5) could be prioritized for review. The appropriate cutoff should be determined based on distribution of the similarity scores and content expertise. If many similar websites are expected, a higher cutoff would likely be utilized.

Step 5. Manual Review

In the exercise we conducted for DOL, the automated searching for websites resulted in the identification of approximately 1,000 potentially relevant websites. This included websites from many organizations that did, indeed, have the characteristics of interest to DOL. Sorting the list of websites by similarity moved those organizations toward the top of the list.

However, the automated searching also identified organizations that did not have the characteristics of interest, and in some cases, these organizations even had high similarity scores. This is to be expected, especially when it is not possible to identify search terms and phrases used exclusively on websites of relevant organizations.

We conducted a manual review of potentially relevant websites to identify those that came from organizations with the characteristics of interest. We reviewed websites in order of similarity score (highest to lowest). To maximize efficiency, we manually reviewed 50 websites at a time, and we proceeded until we encountered a set of 50 websites in which no website had characteristics of interest. In this case, our review ended after 300 websites.

The manual review included steps that our automated algorithm could not perform. In particular, it allowed us to better assess the context in which search terms and phrases were used. In some cases, the

search terms and phrases we identified were used in a way that was not relevant to DOL. The manual review also allowed us to examine the content on subpages of the website. This additional information could be valuable in understanding whether the website's organization had the characteristics of interest.

B. Example

In this section, we provide a simplified, hypothetical example for implementing the steps for matching websites. In this example, the agency seeks to identify apple-shipping companies. In particular, the characteristics of interest include accepting online sales, shipping apples (alone or with other produce), having a physical storefront, and not being part of a national grocery store or online distributor (e.g., Instacart) chain. For the purposes of this example, we assume that the agency interested in identifying apple shipping companies provided links to the websites of five existing businesses that meet the characteristics of interest and serve as our training websites.

Step 1. Identify Search Terms and Search Phrases

A review of these five training websites results in the identification of key search terms that may yield similar websites:

- Apple delivery
- Gala, Granny Smith, Honeycrisp, McIntosh, and/or Red Delicious
- Fresh and delicious fruit delivery
- Apple shipping
- Visit our orchard
- Click to order

These terms were identified because they seemed uniquely relevant to the five websites and each were used on at least two of the sites. The search terms were validated by confirming that at least half of the training set websites appeared within the first 50 search results. The first page of results (50) also had a high concentration of relevant websites (>50%).

Step 2. Automate Search Process

Using the search terms and phrases, use Boolean logic to create a set of unique combinations to enter into a search engine. Combinations could include the following:

- ["Apple delivery" or "Apple shipping"] and "Visit our orchard"
- "Fresh and delicious fruit delivery" and ["Gala" or "Granny Smith" or "Honeycrisp" or "McIntosh" or "Red Delicious"]
- ["Gala" or "Granny Smith" or "Honeycrisp" or "McIntosh" or "Red Delicious"] and "Click to order"

Test selected combinations to assess how many potentially relevant websites will be generated. Based on that, select a threshold of the number of websites to retain from each search (e.g., the first 30, 50, or 100 websites returned).

Once the threshold is selected, conduct the automated search, retaining the URLs from potentially relevant websites. Clean the resulting set of URLs to extract root links, remove duplicates, and remove irrelevant websites (e.g., Apple.com, a technology company, and Safeway.com, a national grocery chain).

Step 3. Process Website Text

Use web scraping algorithms to extract text from the web pages at the root link of each website retained. Use NLP to clean and format website text.

Step 4. Calculate Similarity Scores

Calculate Jaccard similarity scores for each of the retained websites comparing them to the five training websites. Scores will range from 0 (no similarity) to 1 (exact match). Sort websites in decreasing order of similarity score.

Step 5. Manual Review

Manually review websites starting with the highest similarity score. Use the manual review to assess whether the website does indeed include the characteristics of interest. Depending on how the results are used, the manual review may not need to include all identified websites. In certain circumstances, reviewers can stop reviewing once they continue to encounter only irrelevant websites.

C. Lessons Learned

Our analysis provided DOL with a prioritized list of potentially relevant websites. We also supplied a final set of relevant websites obtained after conducting a manual review of the most similar websites. This resulted in DOL identifying organizations that were previously unknown to the agency in a more efficient and more targeted way than a nonautomated search process.

In developing this approach, we identified three key lessons learned:

1. **Search engine algorithms and terms of use can constrain automated web scraping.** Our identification of potentially relevant websites started with automated Google searches. The goal of these searches was to identify websites of organizations with characteristics of interest. However, features of Google (and other search engines) can affect these results.

   First, Google search results are personalized, including targeted ads and location-specific results. This means that two search terms may produce differing results for different people, hampering reproducibility and generalizability. Using a virtual private network (VPN) or remote desktop to run searches may also affect the results.

   Second, submitting searches uses Google's computational resources. Thus, fast, repeated searches consume too many resources and can lead Google to temporarily block the IP addresses of computers conducting the searches. Because our team had a relatively small

number of searches, we were able to spread our searches out over a period of time to stay within the guidelines of Google's terms of service. However, searches with a large number of search term/phrase combinations may take too long if spread out over time; this could hamper efforts to conduct automated searches.

2. **Subpages may contain details that could better distinguish potentially relevant websites from less relevant websites.** The similarity score developed under this study matched language on the root pages of organizations' websites. We did not scrape data from the subpages on organizations' websites due to resource limitations. This simplified the process and contained costs, but it limited the amount of information we could use when identifying similar organizations. In conducting a manual review of websites, we were able to use the information obtained from subpages. Subpage information increased our confidence when assessing whether a given website was indeed potentially relevant.

3. **Similarity scores alone appear insufficient for differentiating relevant from not relevant websites.** Identifying potentially relevant websites through web scraping and similarity scores is inherently imprecise. This approach focuses on the language websites use. While some language is common to relevant websites, no specific language is (1) used on all relevant websites and (2) used only on relevant websites. This means that using language to identify similar websites will return at least some results of websites that use similar language but are not relevant.

   Given the variation in websites, this finding is not particularly surprising. While the exercise conducted as part of this study is useful in identifying and sorting potentially relevant websites, some level of manual review is necessary to separate relevant websites from nonrelevant websites. Nevertheless, we believe the ability to sort by similarity score—even if they are low—is a useful and time-saving feature. In conducting our manual review, the frequency with which we identified relevant websites declined as we progressed through lower and lower similarity scores, and we did not identify any relevant websites after the site with the 266th-ranked similarity score.