

Google Local Search Research Paper On Business Type Extraction via @sejournal, @martinibuster

Google published a research paper describing how it extracts “services offered” information from local business sites to add it to business profiles in Google Maps and Search. The algorithm describes specific relevance factors and confirms that the system has been successfully in use for a year.

What makes this research paper especially notable is that one of the authors is Marc Najork, a distinguished research scientist at Google who is associated with many milestones in information retrieval, natural language processing, and artificial intelligence.

The purpose of this system is to make it easier for users to find local businesses that provide the services they are looking for. The paper was published in 2024 (according to the Internet Archive) and is dated 2023.

The research paper explains:

“...to reduce user effort, we developed and deployed a pipeline to automatically extract the job types from business websites. For example, if a web page owned by a plumbing business states: “we provide toilet installation and faucet repair service”, our pipeline outputs toilet installation and faucet repair as the job types for this business.”

The System Uses BERT

Google used the BERT language model to classify whether phrases extracted from business websites describe actual job types. BERT was fine-tuned on labeled examples and given additional context such as website structure, URL patterns, and business category to improve precision without sacrificing scalability.

Developing A Local Search System

The first step for creating a system for crawling and extracting job type information was to create training data from scratch. They selected billions of home pages that are listed in Google business profiles and extracted job type information from tables and formatted lists on home pages or pages that were one click away from the home pages. This job type data became the seed set of job types.

The extracted job type data was used as search queries, augmented with query expansion (synonyms) to expand the list of job types to include all possible variations of job type keyword phrases.

Second Step: Fixing A Relevance Problem

Google's researchers applied their system on the billions of pages and it didn't work as intended because many pages had job type phrases that were not describing services offered.

The research paper explains:

"We found that many pages mention job type names for other purposes like giving life tips. For example, a web page that teaches readers to deal with bed bugs might contain a sentence like a solution is to call home cleaning services if

you find bed bugs in your home. They usually provide services like bed bug control. Though this page mentions multiple job type names, the page is not provided by a home cleaning business.”

Limiting the crawling and indexing to identifying job type keyword phrases resulted in false positives. The solution was to incorporate sentences that surrounded the keyword phrases so that they could better understand the context of the job type keyword phrases.

The success of using surrounding text is explained:

“As shown in Table 2, JobModelSurround performs significantly better than JobModel, which suggests that the surrounding words could indeed explain the intent of the seed job type mentions. This successfully improves the semantic understanding without processing the entire text of each page, keeping our models efficient.”

SEO Insight

The described local search algorithm is purposely excluding all information on the page and zeroing in on job type keyword phrases and surrounding words and phrases around those keywords. This shows the importance of how the words around important keyword phrases can provide context for the keyword phrases and make it easier for Google’s crawlers to understand what the page is about without having to process the entire web page.

SEO Insight

Another insight is that Google is not indexing the entire web page for the limited purpose of identifying job type keyword phrases. The algorithm is hunting for the keyword phrase and surrounding keyword phrases.

SEO Insight

The concept of analyzing only a part of a page is similar to Google's Centerpiece Annotation where a section of content is identified as the main topic of the page. I'm not saying these are related. I'm just pointing out one feature out of many where a Google algorithm zeroes in on just a section of a page.

The Extraction System Can Be Generalized To Other Contexts

An interesting finding detailed by the research paper is that the system they developed can be used in areas (domains) other than local businesses, such as "expertise finding, legal and medical information extraction."

They write:

"The lessons we shared in developing the largescale extraction pipeline from scratch can generalize to other information extraction or machine learning tasks. They have direct applications to domain-specific extraction tasks, exemplified by expertise finding, legal and medical information extraction.

Three most important lessons are:

- (1) utilizing the data properties such as structured content could alleviate the cold start problem of data annotation;*
- (2) formulating the task as a retrieval problem could help researchers and practitioners deal with a large dataset;*
- (3) the context information could improve the model quality without sacrificing its scalability."*

Job Type Extract Is A Success

The research paper says that their system is a success, it has a high level of precision (accuracy) and that it is scalable. The research paper says that it has already been in use for a year. The research is dated 2023 but according to the Internet Archive (Wayback Machine), it was published sometime in July 2024.

The researchers write:

“Our pipeline is executed periodically to keep the extracted content up-to-date. It is currently deployed in production, and the output job types are surfaced to millions of Google Search and Maps users.”

Takeaways

- **Google’s Algorithm That Extracts Job Types from Webpages**
Google developed an algorithm that extracts “job types” (i.e., services offered) from business websites to display in Google Maps and Search.
- **Pipeline Extracts From Unstructured Content**
Instead of relying on structured HTML elements, the algorithm reads free-text content, making it effective even when services are buried in paragraphs.
- **Contextual Relevance Is Important**
The system evaluates surrounding words to confirm that service-related terms are actually relevant to the business, improving accuracy.
- **Model Generalization Potential**
The approach can be applied to other fields like legal or medical information extraction, showing how it can be applied to other kinds of knowledge.
- **High Accuracy and Scalability**
The system has been deployed for over a year and

delivers scalable, high-precision results across billions of webpages.

Google published a research paper about an algorithm that automatically extracts service descriptions from local business websites by analyzing keyword phrases and their surrounding context, enabling more accurate and up-to-date listings in Google Maps and Search. This technique avoids dependence on HTML structure and can be adapted for use in other industries where extracting information from unstructured text is needed.

Read the research paper abstract and download the PDF version here:

Job Type Extraction for Service Businesses

Featured Image by Shutterstock/Shutterstock AI Generator

<https://www.searchenginejournal.com/googles-local-job-type-algorithm-detailed-in-research-paper/548590/>