Open Data Business Licence Usage to Improve Local Search Engine Content Ranking

Robert F. Lytle
Founder, rel8ed.to
1 St. Paul Street
St. Catharines, ON L2R 7L2
+1 (905) 321-0466
rlytle@rel8ed.to

ABSTRACT

This paper addresses the use of Open Data business licence records in the course of local web search. It assesses the feasibility of increasing the search ranking of authorised service providers and rank reduction or removal of providers with an invalid licence status. A case study was conducted to identify usable results returned with a local search across multiple providers. Result records were then analysed against an applicable Open Data set to determine the usability of the search results for end-users seeking service. A model for adjusting search result ranking is proposed for further analysis. Finally, findings based on the analysis lead to additional proposed research efforts in this space.

Keywords

Open Data; Search Results; Ranking Algorithm

1. INTRODUCTION

Local search engines occupy a growing niche within the overall web search space. Serving a target market of end-users with immediate desire to engage services, this variant of web search has captured significant investment from both global search providers and regional providers with limited geographic coverage. For all providers, the challenge is the same: to present high quality results to a customer who will immediately engage services from companies found through the local search process.

Local search engine providers use a variety of techniques to index and present information to end-users. Raw source data is typically provided from an internal dataset, gathered via web indexing, or purchased as a dataset or feed. Once the provider's algorithm identifies the best matches, additional information may be appended to records for display and ranking purposes. This extra information may come from additional internal datasets, realtime feeds and API's, or the target site's metadata/structured data. In most cases, providers also offer businesses the opportunity to create, claim, or update their own records. For some providers, this is their primary business model.

Data quality and relevancy concerns are significant in local search. Global providers excel at collecting relevant, current web content and data from partners. Regional providers can often provide better coverage of small businesses given local business relationships but do not have the benefit of the larger provider's data networks. A number of poor search results are observed due to business-provided content or missing context data. The prevalence of advertising-based models and pay-for-listing approaches can lead to search rankings that serve the advertising company and agency first, and the end-user second or not at all.

Open Data presents an opportunity to improve results for local search providers by adding key missing element data to display results, and presenting information that may be used for result ranking purposes. In particular, government licencing data presents a key opportunity to enhance display while offering a potential filtering source based on available regulatory information. Information from an authoritative source may serve the end-user with a better overall service experience, but may impact a searched company's advertising strategy, the search provider's current business model, and the ecosystem of agencies providing Search Engine Optimisation services.

Analysis note: Given the potentially sensitive nature of the analysis results, which includes the possible identification of companies operating outside licencing agency regulations, the results of this research are presented in aggregate with suppression of individual corporations named in the search results. The author also chooses to suppress the usability rating of the specific local search engine vendors, as this paper is not focused on comparative engine assessment and the presence of confounding factors may have marginally impacted search results during the analysis.

2. OPEN DATA FITNESS FOR SEARCH

2.1 Availability and Use of Licence Data

Canadian business licencing data is available from both provincial and municipal sources. Licence coverage leans heavily towards the municipal data provider given the prevalence of local governing bodies and licencing agencies. Notable exceptions to this are the professional service classes such as lawyers, real estate or investment agents, etc. which tend towards provincial licencing agencies.

Header data is base information used to identify a business. This data includes name, address, and identity or registration number. These are typically the strongest keys used by the local search provider's algorithms and are required to support the process of matching web search result records to the company's licence record. Header data can also figure prominently in the display of results. Most Open Data sources contain header data at a minimum, with significant availability of name and address and strong availability of identity number. The presence of this data is a positive indicator of the utility of the licence data for inclusion in search results and ranking strategies. A brief sample of example Open Data sets applicable to local search in Canada appears in Table 1.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.2890486

Table 1. Sample Canadian Open Licence Datasets with applicability to Local Search

Dataset	Name	Addr	ID
Nanaimo Business Registry	x	x	x
Toronto Business Licences/Permits	x	x	x
Toronto Dinesafe Restaurants	x	x	x
Victoria Business Licences	x	x	

Key: Name=Business Name; Addr=Address; ID=Identification Number

2.2 Relevancy of Enriched Data

Enriched data must improve the end-user experience to be of value. Open Data sources contain a significant number of variables, and the feature selection of elements for display and processing is the most important early step in assessing a dataset's utility for local search.

Enriched information is broken into two categories: 1) data for display to help distinguish and interact with search results. Available Open Data elements in this class include alternate names, contact information (phone, email), website, and descriptions of products and services. Depending on the primary purpose of the local search engine (ease of use, detailed results, generate advertising revenue, deliver premium products, etc.), the provider may choose to limit the amount of data displayed; 2) data for algorithmic use in search and ranking. Available Open Data elements in this class include enhanced location data, category and licence information, firmographic metrics on corporation size, and financial or payment interaction information.

A special class of elements, total record metrics, provides content metadata about the record for use in algorithms. These elements may be generated by the data provider, created prior to loading of the raw material into the search provider's engine, or during the search transaction itself. Examples include licence count, number of locations and owners, and presence of the search record in multiple datasets. These elements are typically not of interest to the end-user but may factor heavily in the capability of the system to determine the best result and ranking. Total record metrics factor into the proposed ranking solution in Section 3.2.

2.3 Stakeholders in the Local Search Process

A local search engine serves multiple stakeholders: 1) the service-seeking end-user, who is attempting to locate a business to engage the service; 2) the information-seeking end-user, who is attempting to gather information about the business; 3) the business itself, whose marketing department is eager to increase customer contact via search engine ranking; 4) the local search provider, who is attempting to increase revenue through custom search services, advertising revenue, and potentially services revenue from businesses seeking to be listed; 5) a large marketing industry for Search Engine Optimisation (SEO) focused on helping businesses achieve higher search ranking results using acquired knowledge of current search platform behaviour. Each of these constituents could be impacted by a change to algorithms that values verified regulatory data over current methods of ranking. Service-seeking end-users would potentially be the biggest winners. The other four groups would experience some level of impact, with the business experiencing the highest risk. Using this paper's proposed ruleset changes, an estimated 25% of search results would be adjusted significantly up or down, a large potential impact on stakeholders.

3. CASE STUDY

3.1 Assessment of Toronto Tow Truck Search Results

The City of Toronto Municipal Licencing and Standards - Business Licences and Permits dataset (the Registry) contains nearly 200,000 records of various city licences, covering the period from 1995 through current with a daily update schedule [1]. Contained within the dataset are approximately 2750 active and 1900 inactive Tow Truck Operators. This dataset was loaded and cleansed for this analysis.

Five local search providers (google.ca, bing.ca, yp.ca, yelp.ca, goldbook.ca) were selected for this study. Each provider delivers a Local Search capability. The two Global Providers (Google and Bing) offer location-sensitive search views, which operate in a functionally-equivalent manner to the Regional Providers' (Yellow Pages, Yelp, and GoldBook) full search view. All providers were queried with the phrase “TOW TRUCK” and the location “NORTH YORK, ON”. The theory behind this search criteria is twofold: 1) a stranded end-user requires contact information for a reputable personal car tow service with urgency and enters a typical string into a local search engine. For this end-user, sorting through poor results is not practical – tow service is a crisis need and the best contact information is required immediately; 2) the end-user uses the neighbourhood name, not the official name of the city administering the licence. This might be the most typical behaviour of an end-user under duress: to name the locale through memory, locate a landmark or local sign, or engage a passerby. The challenge for each of the local search engines is to return the most relevant, usable results.

A detailed analysis of each search engine's ability to identify the Toronto neighbourhood “North York” (an amalgamated municipality in 1998) was not undertaken. However, observation indicates inconsistent treatment of North York as a location. The Toronto city dataset is tagged as Toronto for all addresses, however the boundaries of the licencing agency fully cover North York as well as Toronto's five other amalgamated municipalities.

Up to 15 valid returned search results from each engine were analysed based on search rank and the ability to locate the company within the Toronto Licence Registry. Any result with an address falling outside the bounds of Toronto was rejected as invalid, as the Toronto Licence Registry would not have contained these companies. Some providers returned < 15 results meeting the locality test. A total of 54 unique search results were returned across all engines, and duplicates are ignored in this analysis by using the top-found ranking for a given result. If Engine One returns a result in position two and Engine Two returns the result in position six, position two is selected for analysis purposes.

Search results pointing to companies located in the Registry were assessed by quartile for current licence status (Active/Inactive). Search results not located in the Registry were further assessed for whether they provide the service of personal car towing. A sample using a fictitious name: “M & M Tow Truck Builder” was ranked second in two of five engines, but was not found on the Business Licences list. Although the company's business name contains the term “Tow Truck”, its focus is customising and maintaining trucks, not providing tow services. Finally, companies were assessed to identify whether the business still exists, via external federal and provincial registry checks, public search engine queries, and in some cases direct contact using the information listed in the search results.

Tow companies with licences were also assessed for record metrics, in this case the total number of active licences assigned to the company. This information is not analysed here, however the metrics were captured for use in a proposed filtering and ranking solution in Section 3.2.

As displayed in Table 2, the final usable rating is achieved when a result points to an active company providing personal car towing services, with either a valid tow licence or no licence indicator. This broad leeway is granted due to the potential for match logic failures between the search result and the Business Licences list.

Table 2. Aggregate Search Results Analysis

Quartile	Active Licence	Inactive Licence	Not In Registry	Offers Car Towing	Business Closed	Usable
1	31%	8%	62%	62%	15%	54%
2	33%	7%	60%	60%	7%	53%
3	21%	36%	43%	71%	7%	43%
4	33%	8%	58%	67%	0%	58%
Average	30%	15%	56%	65%	7%	52%

Mean percentage of all provider search results

The study indicates that only 52% of the search results from all providers link to a usable tow company. Findings of particular note: 1) regardless of licence status, only 65% of the results point to businesses providing the personal towing service that is most likely sought by an end-user engaging a local search service; 2) in the top quartile of results, 15% of the returned companies were apparently no longer in operation or did not have a usable website or contact method for engagement their service.

Taken as a whole, the ability to find a usable tow truck service through local search is possible, but the end-user will likely have a less-satisfactory experience either finding the service (the business closure and the Offers Car Towing risk) or receiving poor service (the unregulated operator risk).

Table 3. Global Search Engine Results Only

Quartile	Active Licence	Inactive Licence	Not In Registry	Offers Car Towing	Business Closed	Usable
1	17%	0%	83%	67%	17%	67%
2	50%	0%	50%	67%	0%	67%
3	17%	33%	50%	83%	0%	50%
4	40%	20%	40%	80%	0%	60%
Average	30%	13%	57%	74%	4%	61%

To test for poor result bias by regional local search engines, the aggregate analysis is split out by global and regional search engines in Tables 3 and 4.

For the global providers, an overall advantage is seen in the area of provision of personal car towing service. A smaller advantage is also seen in the measure of business existence, possibly due to less-frequent web indexing schemes present in some of the regional providers. These factors contribute to a more-favorable overall usability rating for the global providers. However, this relative improvement is not significant enough to limit the result ranking issue only to local providers. At best, the global providers are still presenting search results in which nearly 40% of the companies cannot provide a usable service given this measure.

Table 4. Regional Search Engine Results Only

Quartile	Active Licence	Inactive Licence	Not In Registry	Offers Car Towing	Business Closed	Usable
1	38%	13%	50%	50%	13%	50%
2	38%	13%	50%	63%	13%	50%
3	25%	38%	38%	50%	13%	25%
4	25%	0%	75%	63%	0%	63%
Average	31%	13%	53%	56%	9%	47%

Note that in the first quartile the global providers fare even worse in presenting licenced operators compared to the regional providers. This may be a symptom of a potential anti-competitive algorithm postulated in Luca's work on global search providers and their ranking approaches [2].

3.2 Confounding Factors

Approximately 25% of the Toronto Licence dataset contains Numbered Provincial Corporations without a DBA name. This most likely impacted the original licence match result, returning a higher false positive rate than would normally be acceptable for a larger study. Use of a linking system to identify the real name as appears in search results for all companies would yield higher accuracy in the measures.

The sample size of the survey was limited due to the need to capture truly meaningful search results in a specific target neighbourhood of Toronto. Search result quality for the scenario degraded significantly after the first page of results for each provider. While a larger locality search (using "TORONTO" instead of "NORTH YORK") may have brought in additional companies to increase sample size, the poor performance of the first quartile of results would have remained the same.

To assess business existence, the contact and web information in search results was used to affirm corporate status. While it is possible that poor contact information could result in a false positive for this measure, for the purpose of the measured scenario the outcome is the same: the end-user cannot easily contact the company to engage the service.

Taken as a group, these factors do not significantly impact the overall findings. However, addressing these issues would lead to additional detailed findings for algorithm design assessment.

4. PROPOSED RANKING SOLUTION

Given the large number of unusable results observed for both global and regional providers, a solution that factors in business licencing data would serve the end-user appropriately in this scenario. To craft a modified ruleset, an assessment of known data at time of search is required. Only data that is available within the records of the search provider would be used for this algorithm. To address usability, the information required would be some combination of the following: 1) whether the company provides a personal tow service; 2) whether the company is still in operation; 3) if the business is licenced; 4) how many licences it possesses as a proxy for the capability to provide quick service.

Of these four criteria, the first two are problematic for large-scale adoption. Affirming tow truck service for each company might require implementation of a new structured data schema for use in the company web pages, or addition of new flags within the search provider's system. Affirming the business is still open might require the implementation of additional web indexing capability. Only the licence designator and licence count usability measures are readily available and reliable, via Open Datasets. A company possessing active licences is likely to provide the service, and would be more likely to remain in operation when compared with the overall company set found in unfiltered search results. Therefore, licence count is selected as the key indicator of usability for this algorithm.

A mechanism of filtering and re-ranking search results is proposed with the ranking rules: 1) filter out or significantly down-rank any results that point to companies with an expired licence; 2) re-order the results for businesses with licences based on the number assigned to the company, with the highest number of licences favoured at the top of the list; 3) use the existing provider ranking algorithm to order the rest of the results for which no licence recod is available; and 4) optionally display the results of companies with expired licences.

4.1 Ruleset results

The initial search results performance of the five search providers identified one regional provider with a significant lack of licenced businesses in their database. The proposed algorithm was tested against the search results of the remaining four providers (two global, two regional).

Table 5. Re-ranked search results with the proposed ruleset

New Position	Result
1-3	Companies with valid licence record, ranked by # of licences in descending order
4-12	Companies with no licence record, ranked by current search algorithm
13+	(optional) Companies with expired licence record, ranked by current search algorithm

Table 5 demonstrates that using this reordering ruleset on the analysed data, all four tested providers successfully re-ranked licenced companies to the top of the list, with at least the first 3 results for each provider containing only licenced businesses. This ruleset accomplished its primary purpose: to ensure the end-user can find a usable tow truck provider quickly. The second two quartiles displayed the results of companies judged less-usable due to the non-presence of licence data. This group contains companies that provide the personal towing service as well as those companies that provide related services, ranked by the search provider's current algorithms.

Table 6 demonstrates the significant increase in usability in the key first quartile. All of the entries are deemed usable, with the vast majority of results pointing to businesses with verified licences. The dropoff in registry entries in the second and third quartiles supports the theory that presence of a licence is a moderately successful proxy for companies that actually provide a personal tow service, versus towing-related companies. These entries will still show in the results list, supporting the capability to display related services. A search provider using this approach might further consider visually tagging the results with a badge or other indicator of the licence, further enabling self-selection by end-users who are seeking other tow-related services.

Table 6. Aggregate Search Results with New Rules

Quartile	Active Licence	Inactive Licence	Not In Registry	Offers Car Towing	Business Closed	Usable
1	93%	0%	7%	100%	0%	100%
2	31%	0%	69%	50%	19%	50%
3	0%	0%	100%	40%	10%	40%
4	0%	100%	0%	88%	0%	0%

The inclusion of the companies with an expired licence in the fourth quartile of the results is a straightforward solution that would still allow search ranking of these correctly-indexed records from the perspective of traditional search indexing approaches.

4.2 Ruleset impact

An important assessment based on these results is the impact on the formerly-ranked companies, with implications to those companies' online marketing strategy

Table 7. Search Result Adjustment with New Rules

	Global P1	Global P2	Regional P3	Regional P4	Average Change
Dropped	20%	0%	22%	13%	14%
Static	27%	33%	0%	13%	18%
Rank Up	20%	22%	50%	25%	29%
Rank Down	33%	44%	28%	50%	39%
Quartile Move	13%	33%	28%	25%	25%

Percentage of result items impacted for each provider

The impact to four search provider whose results were tested is shown in Table 7. The provider impact is assessed by 1) the percentage of search results dropped into the last quartile due to inactive licences; 2) the percentage of results that stayed static in the original rank position; 3) the percentage of results that ranked up and down; and 4) the percentage of results that moved up or down into a new quartile.

Quartile movement is extremely important for the SEO industry, as significant focus is placed on moving client company search results into specific quartiles (top five results, second page of results, etc.). In particular, the top 5 search results have been shown to account for upwards of 70% of user interactions as referenced in Relevance's recent work on click-through rates [3]. Therefore, any quartile movement (up or down) would create significant impact on advertising strategy results.

The largest impact of this ruleset is the dropping of an average 14% of records from the search result, due to the identification of an expired licence. For the purpose of this test those results were moved to the end of the results. Note that this heavy penalisation in ranking could cause significant impact to the company's online advertising strategy. These results represent companies in some cases with valid websites that would no longer be visible in the search engine's primary offering, regardless of any other content ranking strategy in place. A premium would therefore be placed

on ensuring the government licencing records are accurate and current (and accessible to the search provider), and the company website properly indicates the correct name and other data required to perform a licence match.

The observed impact as represented by results moving into a new quartile is also significant as a proxy for the amount of overall change to end-user experience introduced by this ruleset. A secondary observation is the significant up/down rank movement and potential impact on the provider advertising model, and the searched businesses and SEO providers who serve them. This re-ranking algorithm therefore has promise for end-user experience (finding a usable service provider) but at the expense of impact to other stakeholders in the search process.

5. CONCLUSION

Local search provides a valuable service to end-users, who are seeking information leading to quick engagement with services. In the case of a search for licenced businesses, the enrichment of search ranking with available Open Data can lead to a better end-user experience. Using a proposed re-ranking scheme, the top results in the search engine would be very likely to point to usable, licenced providers of the correct service sought by the end-user.

However, a change in search ranking strategy could have a moderate impact on some current search provider advertising and SEO firm strategies. A premium would be put on providers retrieving accurate licencing data, and businesses exhibiting diligent compliance with regulations and proper tagging of their websites with all names used in the course of business. The approach described in this paper projects both positive and negative impact on stakeholder groups, and would benefit from further analysis.

Search engine providers should first assess their primary market focus when considering the approach of licenced data usage. If the leading mission of the provider is to provide the most usable results to end-users, an active programme to secure authoritative information and champion the necessary changes to algorithms and customer site tagging would drive change in the search marketplace. However, it is important to note that the profit motive of advertising and services may in fact advocate for a lesser focus on results usability to ensure the viability of the provider's business model.

End-users should bear in mind that local search is fundamentally a free service for the consumer, and the ready availability of these services will continue to rely on entire search engine ecosystem, including the advertising support of companies – even those whose licencing status is unknown or lapsed. A balance in the marketplace should be reached to support the needs of all stakeholders in the local search process to ensure its ongoing capability to provide meaningful service in the market.

6. FURTHER WORK

Several findings present opportunities for additional research as a result of this analysis.

Search engine results can be assessed at a deeper level to determine improvement steps for removal of non-existent businesses, or better handling of locality and neighbourhoods.

Additional work on matching and linking corporation names, particularly in the case of Numbered Corporations, would lead to higher confidence in the search results for companies who appear to be non-licence holders.

Additional research on available Open Datasets would identify the potential for identifying business closure or additional category information to increase the usability of search results.

Development of a structured data tagging scheme for accurate licence matching would afford additional flexibility for companies who are licenced under one name but operate under another.

Research into additional algorithms that factor in licencing data as part of, rather than a replacement for, current logic may yield a better balance between end-user experience and large changes to current engine behaviour.

7. REFERENCES

[1] City of Toronto. 2016. Municipal Licencing and Standards - Business Licences and Permits. Retrieved January 8, 2016 from http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=83a7c060155d0310VgnVCM1000003dd60f89RCRD

[2] Michael Luca, Timothy Wu, Sebastian Couvidat, Daniel Frank, William Seltzer. 2015. *Does Google Content Degrade Google Search? Experimental Evidence*. Working Paper 16-035. Harvard Business School.

[3] Relevance, Inc. 2013. A TALE OF TWO STUDIES: Establishing Google & Bing Click-Through Rates. (2013). Retrieved February 7, 2016 from http://connect.relevance.com/a-tale-of-two-studies-establishing-google-bing-click_through-rates-relevance