20181005 Analytics & Insights team meeting notes (Address Matching)

Notes of meeting held at 14:00-15:00 on Friday 20181005 with:

Jason Ward

Steve Peters MHCLG

Anna Carlsson-Hyslop

Sarah Belghiti

This meeting was held as part of the Landmark address matching topic. The Analytics & Insights (A&I) team have their own address matching algorithm and the aim of the meeting is to understand what for they use it and in high level how it works.

The A&I team do analytics at different address levels. For their analyses they need to join multiple datasets coming from various departments/companies. The key column is the address but it might be differently from one source to another, this is why they need an algorithm to match the addresses to one another.

We use Ordnance Survey's AddressBase Premium as the main source for address matching. See technical specification published here. We are using this under the Department's Public Sector Mapping Agreement Licence terms and conditions. Key points are:

- contains some intelligence on the purpose for which the property is used
- contains pre-built properties
- is an enriched version of the PAF. Contains the 'official GeoPlace issued UPRN, plus links to OS TOIDs.

We store AddressBase Premium in our internal SQL Server 2017.

Address matching is achieved via database scripts, using SQL Server's embedded NLP functions. This is currently in a prototype stage, with database artefacts and an outline strategy for a more sophisticated system in place but not fully implemented.

There's nothing very fancy about the prototype search strategy, it's basically little more than a "bag of words" search. It does not perform well if the search term is very specific, so flats and apartments are not too well dealt with.

The more sophisticated version will extend this by splitting the search term into unique tokens which can be lexically parsed to an extent (eg a string with 2 numbers is probably a flat/house number structure), and using Levenshtein distance to identify spelling mistakes. The next iteration will also use the SQL STRING_SPLIT() function to split the input string into tokens using the space as a delimiter, then convert to a table object with 1 row per token and do some lexical parsing to identify the tokens as "street/st/road/rd" etc, numbers, postcodes etc etc. Then we can make an estimate of "informativeness" for each token and build the search string from the most useful tokens only.

What we have seems to work pretty well, with a recent exercise giving 80% of addresses finding a single match in the database, which we assume is highly probably a correct match. Some manual tweaking is usually required. We can't get 100% accuracy, specially because of new addresses created by assessors.

The algorithm runs quite quickly: a test address returned 25 000 results in 2 seconds

The A&I team tried to use the ONS address matching API (beta) but it doesn't support bulk data (only one row at a time). The A&I team have 28 million rows to match.

Questions for Landmark around the assessor process:

- Where is the model applied in the lodgement process?
- Data quality: how do we handle attribute changes? Is there any link between new and old attributes?

Further information:

- Blog post about the address identifiers: https://www.owenboswarva.com/blog/post-addr1.htm