Deductive Systems Inc.

50 Pine Street, Ste. 7N

NY, NY, 10005

Deductive Systems Inc

Technical Documentation for SmartSpider Proto-type

Documentation

SmartSpider and DataMiner Proto-type

SmartSpider:

1. Can extract important features(or raw data) of personal pages on linkedin, twitter, meetup
   1. Linkedin extraction is in /smartspider/transport/linkedin.py,/smartspider/jobs/process\_linkedin.py and associated MongoDB schemas linkedin\_input, linkedin – Two Phase acquisition
      1. Phase 1 – seed a collection of linkedin profiles in linkedin\_input based on paginated queries to bing search engine for linkedin profiles
      2. Phase 2 – process individual linkedin\_input links and pull down all data from linkedin, creating a bonifide first/last name entry in linkedin and a profile entry in raw\_profiles
   2. Twitter extraction is in /smartspider/transport/twitter.py, /smartspider/jobs/process\_twitter.py, and associated MongoDB schemas twitter and twitter\_in – Two Phase acquisition
      1. Phase 1- generate a twitter\_in collection by acquiring list of first/last names from linkedin collection and searching for those names individually via the twitter search people feature
      2. Phase 2 – process each twitter profile and pull down all data into a raw\_profiles entry
   3. Meetup extraction is in /smartspider/transport/meetup.py, /smartspider/jobs/process\_meetup.py
      1. Process each entry in linkedin collection, use bing to query each entry for meetup member links, and pull down the individual meetup member profile and store in raw\_profiles
2. Can construct a meta profile that links pages on all three sites to one person
   1. Each sub-profile is stored in raw\_profiles according to an algorithm including the first/last name, the region, and stemmed attributes associated with the profile.
   2. A meta profile is created, and index based on any number of algorithms. The list at moment is:
      1. Algo0 – cluster meta profile based on matching first/last name, and region
   3. Meta profile also contains a text index of relevant searchable terms
3. Can store meta profiles in a database
   1. All meta profiles and a tag cloud of searchable terms is stored by algorithm

Data Miner:

1. Can search across features to extract a set of sub profiles organized by some common vectorsinto a meta profile. Meta profile contains on a limited bases an extract of the work and personal interest graph.
   1. A Tornado Web service in /smartspider/web/main.py is put together to support two web pages
      1. Home page which lists the tag cloud of common terms and a search bar
      2. Search result which contains a table of meta profile names and associated sub-profiles
      3. Sub-profiles are available for drill down
      4. Meta profiles are available on a drill down basis

Issues Encountered, and Discussion Points

Issues

Linkedin will blacklist blocks of IP addresses as it detects spidering. It has already blacklisted RackSpace, and has pursued Amazon based spidering. A possibly solution is to use rotating VPN addresses(FoxyProxy for instance), or alternatively to engage via the API and a formal relationship.

There are two orthogonal and complex issues requiring natural language processing involved in the solution, that is clustering profiles, and search.

Clustering profiles will require focus on human written profiles and backgrounds to extract meaning and similarity between individual profiles. The key to solving this issue will be to focus on the meaning of the profile.

Searching will require focus on both human and machine structured portions of the profiles (ie skills section), and focus on individual key words that are relevant within the context of the profile. The key to solving this issue is to eliminate all but the most relevant key words in the profile.

The current proto-type is scale limited as it is specifically scoping its acquisition by using the bing search engine and twitter people search to harvest links. This does not scale to a much larger universe, and so as a result we will need to change our link acquisition strategy to pull all links from meetup, linkedin, twitter.

Discussion Points

Not everyone who has a linkedin profile also has twitter, or meetup profiles. Approximately 10% of total linkedin profiles I’ve processed (16K) also have potentially pairing with twitter and meetup profiles.

Pairing meetup and twitter profiles with linkedin profiles via full name and region provides a good first approximation match of candidates.

In the case they do not, it can be because:

1. Region is not provided
2. Multiple twitter/meetup profiles are available for the same name/region match
3. Profiles exist for the same name/region but are for actually different people

For these cases, we need to assess the capabilities of similarity/clustering based methods on background summaries, or if they don’t exist, the entire profile itself. A first cut of a basic clustering mechanism available as part of the algo0 custering delivered in this proto-type. In general the results do not appear to be that promising as the information in background summaries in twitter tend to be significantly more sparse than linkedin, possibly we will need to consider extracting tweets as part of this process.