RIPS proposal: mapping domain names to categories

Problem description

Oversee.net is an important player in the online advertisement industry. We maintain a large list of internet domains, and part of our monetization strategy is to display keywords on those websites. When a user clicks on a given keyword, we redirect the user to one of our advertising partners.

Our advertising partners allocate their budget based on a category tree. For example, advertiser A may only want to receive users from domains that belong to category 'team sports' or 'soccer'. This means that Oversee.net is classifying domains, i.e. assigning them to one or more categories. The more accurate our classification is, the better we are able to satisfy both our advertisers (by sending them traffic they want) and the user (by redirecting them to a website that will likely help them achieve whatever it is they are trying to do).

The problem then is to optimize the mapping between domains and categories. A domain may map to multiple categories. Each mapping has a score indicating the strength of relationship between domain and category.

Example

Let's say our (fictitious) domain name is *bigbearhuts.com*. Ideally, this domain should have a strong association with categories **hotels**, **travel**, and **winter sports**. Categories are hierarchically ordered, so the association with parent categories such as **sports** should also be significant.

Available data

The following data will be available to the team:

- hierarchically ordered category tree
- about 100K categorized domains for training and testing
- keywords and their click-through rate (CTR)
- edit-distance similarity between keywords and domains

The last two items on this list will allow the team to calculate the strength of relation between keywords to domains. For example, a keyword “snowboard rentals” with a high CTR for our sample domain *bigbearhuts.com* would indicate that “snowboard rentals” (and possibly its components, “snowboard” and “rentals”) are meaningfully related to the domain. Likewise, if a keyword looks similar to a domain name (say, “big bear houses” and *bigbearhuts.com*) we can conclude that the keyword is likely related to the domain.

Potentially helpful pointers and insights

With the above data the team should be able to map keywords to domains. Since the overall task is to map domains to categories, the team may take advantage of keywords and their relation to domains as well.

Most likely, a solution will involve an external knowledge base such as Wikipedia or maybe WordNet. For example, Wikipedia contains categories which in turn are associated with articles. Articles, in turn,

can be represented as a collection of keywords. There is an opportunity to perform the following mapping:

Oversee.net categories → Wikipedia categories → Articles → Keywords ← Domains

The problem can thus be broken up into smaller steps:

- provide a mapping between Oversee.net categories and Wikipedia categories
- represent categories as a collection of keywords
- represent domains as a collection of keywords
- calculate strength of association between domains and categories based on some keyword based similarity measure

Expected outcome

At the end of the project, it would be nice to receive any of the following:

- a meaningful mapping between categories and domains
- a measure of confidence that the mappings are accurate
- any code (team members can pick their programming language of choice)