Interest Dashboard Architecture and Classification

Introduction

UP stands for User Personalization. UP is an attempt to in-browser classification of user behavior on the web. Historically, UP was developed as a privacy protecting mechanism of sharing User Profile with interested sites. For example, the search engine may re-rank its search results for "tiger" given that a user is interested in golf, or a news site may change the appearance of its homepage to cater for user interests. Finally, the user himself may want to control what's shared with the site. His interests provide a convenient tool to control the sharing: a user may turn off his health interest, and turn on his sport interest.

We created an Interest Dashboard addon that analyzes user browsing history and surfaces user interests in the dashboard. The purpose of the dashboard is to eventually enable user control over what's shared with the web, and to provide a personalization platform for content recommendation based on user interests. Varios firefox services as well as web sites can query UP API to access user shared interests and personalize user experience accordingly.

Architecture of Interest Dashboard

To support UP functionality a two tier classification system was developed. First, pages visited by a user and stored in browser database are classified into interest taxonomy. Then, based to visitation patterns the user interests are computed. To insure reasonably accurate page-level classification, certain amount of data needs to be downloaded by a browser. Such data is called a classification ruleset (or simply ruleset) and contains patterns indicative of a particular interest category. The interest rules could be of many different forms, as in example below:

```
pattern                    category
-------------------------------------------
nfl.com                → american foortball
nytimes.com:/politics → politics
ebola in title          → health
```

```
golf in URL           → golf
```

A browser processes user history page by page and matches each page URL and title to the ruleset. Interests identified at page level are passed up to user interest categorizer that analyses overall visits made to “topical” pages based on the date, recency and visits count. User interest classification decides if a user has a particular interest, how recent and intense that interest is, and then ranks interests based on recency and intensity. The ranked list of interests, along with pages classified into each, are passed to the Dashboard UI for user interaction.

ruleset + history → Page Interests → User Interests → Ranked Interest → UI

Technical challenges

There are three key requirements that made UP processing challenging:

- User categorization has to be accurate
- Ruleset can not exceed 1M in size
- Telling Interest from Intent

Our in-group-study showed that users are somewhat alerted by interests that they can not associate with, and are significantly irritated if the guessed interest is wrong. The study showed that users will not tolerate more than 2 our 10 errors, and even then users are uncomfortable. Ideally, highly ranked interests should be 100% accurate and perhaps a few low ranked interests may contain occasional errors.

To maintain such level of precision, the ruleset has to contain enough rule patterns to categorizes good portion of pages on the web, for otherwise users will see very small, shallow and potentially noisy interest list. However, the ruleset exists in a browser and, therefore, has severe size limitation, for too large a ruleset may cause memory exhaustion on lower-end machines.

Finally, there is a subtle notion of a long-term interest that spans months of consistent, but low-intensity browsing and a short-lived, but highly intense behavior, caused by a user intent rather than interest. It is not trivial to tell one from another, but is important for potential application of UP.

Tier One Classification - page level interest rules

Interest matching rules are generated automatically from large corpus of pre-classified web pages. Rule generation was developed using so-called Moreover corpus that contains 8 millions of tagged web pages collected from news and blog sources. Moreover collects news from 50K+ websites and blogs. We profiled content volume from 50,000 most popular (according on alexa) English web sites, and identified 2,500 domains among them that generated sizable (more than 50 stories a month) news content. Pages from these sites were extracted into a Mongo database and later analyzed for generation of URL and title patterns that are highly indicative of a particular interest category. Such patterns are called “topical” rules, meaning that a rule pattern matches a given category with prescribed precision level - usually above 90%.

10,000 “topical” rules with highest recall were selected for inclusion in a browser ruleset. The size of the ruleset is below 0.5M uncompressed, which is at acceptable level. Testing on folded data showed that overall precision and recall stays at 90% levels for previously unseen corpus documents, which proves our rule generation strategy successful. However, when auto-generated ruleset was tested on real user histories, the amount of errors on non-news pages (like e-commerce, search queries, social networks) was unacceptable. This was caused by inapplicability of news-based patterns for general web pages. The remedy was to scope news-rules to news-sites only. The rule processing engine was extended to allow for rule scoping, whereby rules are only considered if a page domain (or host) falls into a whitelisted set of domains applicable to genre of the training corpus.

About 1200 content sites were identified and auto-generated rules were scoped to these sites. As a result, ID had dramatically increased recall without loss of precision.

The full ideation, detailed algorithmic description, and test results are here: Generation of browser page classification Ruleset from Corpus statistics

Tier Two Classification - Identifying User Interests and Intents

The assumption behind user interest classification is that user interest in a topic is surfaced by **consistent** visitation to pages from that topic. The degree of interest is assumed to be correlated with an intensity of visitation pattern to topical pages.

We tested various interest-ranking algorithms. A simple aggregation of number of days when a visit(s) to a topical page occured turned to be the best indicator of interest presence and rank. However, a number of highly visited interests were only observed for a limited number of days. Such interests violate our **consistency** assumption: a user’s interest in a given topic (like sports) can’t last just 2 days in a row through the whole year. But ignoring such topics is not correct either, for a user had a strong reason to generate large volume of visits in such short period of time.

Highly voluminous, but short-lived interests appear to be more of intent indicators. For example, buying a flight ticket, booking a vacation, searching for best camera are all examples of “intent” based behavior, where many visits are made to particular sites to achieve the task at hand. The algorithm of intent identification clusters short-lived, high intensity behaviors to distinguish them from potentially low-intensity, but consistent visitation pattern representing an interest.

Detailed description of interest/intent classification is here: User Interest vs. Intent Identification

Code Base

- Interests Dashboard
- Corpus collection and Rule generation