Using Semantic Technologies to Mine Customer Insights in Telecom Industry

Rajaraman Kanagasabai, Anitha Veeramani, Le Duy Ngan, Ghim-Eng Yap
Data Analytics Department, Institute for Infocomm Research, A*STAR, Singapore
Email: {kanagasa,vanitha,dnle,geyap}@i2r.a-star.edu.sg

James Decraene, Amy Shi-Nash
R&D Labs, Living Analytics, Singapore Telecommunication Ltd, Singapore
Email: {jdecraene,amyshinash}@singtel.com

Background *:* Many telecommunication companies (telcos) today have actively started to transform the way they do business, going beyond communication infrastructure providers and repositioning themselves as data-driven service providers to create new revenue streams. New business opportunities, notably in terms of market research, can be realised using the telco data especially when it is complemented with external open data sources. Indeed, significant efforts have gone into mining customer insights from the massive mobile data while preserving the end customers' privacy. In this paper, we present a novel industrial application where semantic technologies are successfully used to mine commercial interactions of anonymous mobile customers, to get new aggregated insights from their call records.

Our Method*:* The case study is inspired by an observed semantic gap between the *contextual business categories that can be derived from the customers' call records* and the *industry-standard classification scheme that is often a requirement for consistent ad targeting*. In particular, the *Internet Advertising Bureau (IAB) Contextual Taxonomy*, with 23 Tier-1 classes and 371 Tier-2 classes, is an international standard that is adopted e.g. by the Google Display Network. Mapping from thousands of contextual business categories to such a far more concise taxonomy is clearly a daunting task for the market researchers.

Traditional approaches using machine learning techniques like the Support Vector Machines (SVM) can help automate this task but it is not straightforward to apply them over thousands of categories spread across a wide variety of domain areas. Our experience mapping a total of 2532 contextual business categories to the IAB Taxonomy shows that applying text feature extraction and matching resulted in merely 263 categories being matched (approximately 10% in recall). The challenge is that this is in fact a large-scale multi-class multi-label classification problem that needs sufficient training examples to generalize well. Obtaining such training labels specific to each application is expensive and it is not feasible to repeat this for all applications. In this paper, we leverage domain knowledge in semantic machine learning methodologies and avoid the need to invest in expensive training data.

Using public knowledge bases WordNet, DMOZ, and Yahoo! Answers as our domain ontology model references, we investigated the three different semantic IAB classification methods described below:

I. We employ WordNet features to build an extended text vector for classification, and use it as a baseline for comparison.
II. DMOZ Open Directory is among the largest human-curated directories online. We ingested RDF dumps of DMOZ into AllegroGraph triplestore as our DMOZ ontology model. Low-level

text features were extracted from each contextual category to find its semantically-matching categories in the DMOZ ontology and the DMOZ categories are used to find best IAB classes.

III. Yahoo! Answers is a community-driven Q&A site hosted by Yahoo! Inc. The site categorizes questions and answers in a shallow categorical hierarchy that is similar to IAB, though the category names are very different. We capitalize on this by first searching Yahoo! Answers with the contextual category and using the returned Yahoo! Categories to match IAB classes.

We built a corpus with just 525 contextual business categories and created IAB class assignments by using two human experts to classify independently and a third expert to cross-check the assignment. The corpus was used in our research to fine-tune parameters in all the three classification methods. Following that, we validated the methods on a full set of 2532 contextual categories and manually evaluated their classifications. Method III performed best, followed by Method II and then Method I.

Deployment: A customer insights dashboard product was developed to provide user behavioural insights based on users' geo-location traces and call details records ("who called who"). All records were anonymised via a one-way AES encryption-hashing process and neither personal data nor calls content was used.

A key offering of the product is its market segmentation service where in-depth user profiling was conducted to infer various people traits of interest such as demographics, occupation, housing type, and travel pattern. To illustrate, the example screenshot in Figure 1 shows the distribution of work locations for people who are living within a specific location inside Singapore (marked as a blue cell).
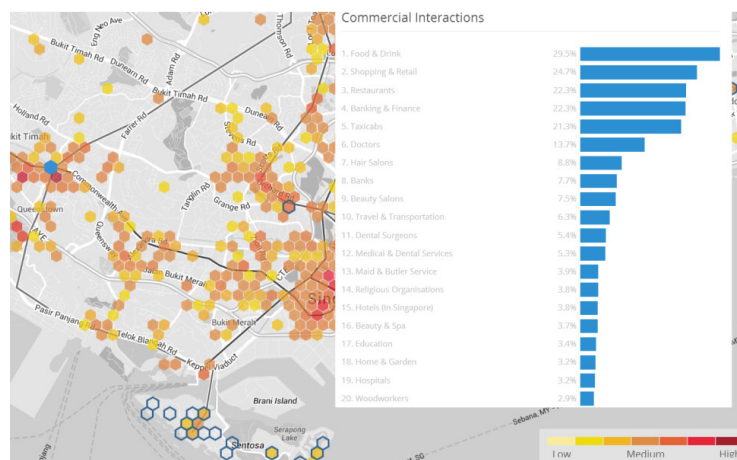

Figure 1 Example screenshot of the customer insights product showing work locations distribution and commercial interactions profile of people living in Buona Vista (blue cell) – the commercial interactions profile were successfully generated using the novel Yahoo! Answers-based semantic IAB classification method presented in this paper

To enrich the customer profiling, we used call detail records to identify user calls to local businesses by identifying the business numbers being called and extracting the associated contextual business category from open data that was available online. We applied Method III (using Yahoo! Answers) to map the arbitrarily-defined 2532 contextual business categories into the IAB Contextual Taxonomy. As shown in Figure 1, the proposed method enables the end product to effectively provide customer insights in terms of market segments based on such commercial interactions of mobile device users.