Automatic Discovery of Emerging Trends using Cluster Name Synthesis on User Consumption Data

[Extended Abstract]

T. Chattopadhyay, Santa Maiti, Arindam Pal, Avik Ghose, Arpan Pal
Innovation Lab, Kolkata
Tata Consultancy Services Limited,
India
(t.chattopadhyay,santa.maiti,arindam.pal1,
avik.ghose, arpan.pal) @tcs.com

Shanky Viswanathan, Narendran Sivakumar
Telecommunication Unit
Tata Consultancy Services Limited,
India

ABSTRACT

A business problem for the telecommunication companies is to provide an appropriate promotional coupon to suitable customers. This problem leads to the challenge of identifying behavioral patterns of customers and deliver the right customer engagement at the right time. So there is a need for a system that can enable the telecommunication companies to go for the best marketing strategy by leveraging customer intelligence to drive offer acceptance based on personas. Technically it is possible for the telecommunication companies to recommend suitable advertisements if they can classify the web sites browsed by their customers into classes like sports, e-commerce, social networking, streaming media etc. Another problem is to classify a new website when it doesn't belong to any of the existing clusters.

In this paper, the authors are going to propose a method to automatically classify the websites and synthesize the cluster names in case it doesn't belong to any of the predefined clusters. We have experimented on a small set of data set and the classification results are quite convincing. Moreover, the phrases used to describe a website if it doesn't belong to existing classes are compliant to the phrases obtained from manual annotation. This proposed system uses the Wikipedia data to construct the document for the websites browsed by the customers.

Keywords

Cluster Name Synthesis, Wikipedia, Website Classification, Telecommunication Use Cases.

1. INTRODUCTION

The basic problem addressed in the current paper is motivated by the requirement of the Telecommunication (to be referred as telecom now onwards) companies who want to push personalized and more relevant recommendations/offers to customers based on their usage behavior. As per the business model, telecom companies have some collaboration with some companies who have some promotional coupons to be offered to the prospective customers via these telecom companies. So it is required to push suitable coupons to its most prospective customer so that the turn around is maximized. So the problems addressed in this paper are (i) classify the websites depending on the coupons and (ii) synthesis possible class names for the websites if it doesn't belong to any pre-existing classes.

In the domain of consumer behavior modeling, usually techniques like collaborative filtering based approaches are used [11, 16], which exploit the similarity between user personas. Depending on context, the user is mapped to one or more dominant personas, based on which offers are pushed to the user. The advantage of such targeted marketing is that the conversion rate is much higher, since the user is pushed offers on content or consumable that he is genuinely interested in. The scheme works well in most of the cases. However, there are a few problems with this approach. The major problem is discovery of emerging trends and adding them automatically to the dictionary of behavior. Based on such dictionary updates, marketing experts can design new offer bouquets, an example of which is provided in Table 1. A very pertinent example can be that of the "Whatsapp messenger" [13]. Any categorization system like SimilarWeb [14], would categorize it as an "Internet messenger", thereby putting it into the same bucket as Skype [15], Yahoo Messenger etc. However, that would be a mistake as a better way to classify it would be "Social Networking". It would require a business analyst to manually add the category and classify the new application before the recommendation engine can use the knowledge effectively.

Our paper proposes a method to solve the problem we elucidated above. We use methods of unsupervised machine learning to cluster similar content into logical groups. Followed by this step, we mine the web and use techniques of cluster name synthesis to arrive at category information for each cluster. If any one entry does not logically belong to existing clusters, the system automatically creates a new cluster and adds the content to the same. We prove our algorithm on a telecom provider's data of website URL logs

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

Table 1: Examples of Special Classes

Class	Sub Classes
Streaming	Youtube Netflix
Social Media	Facebook Twitter
Sports	Cricket Football
Uncategorized	

using which we try to come out with category groups of similar websites and also try to synthesize their logical category names automatically. We compare our findings with SimilarWeb [14], to prove the effectiveness of our technique. Such dynamic categories can be used to design dynamic personas based on emerging trends and help marketing companies design much better user centric offers in lesser turn-around time. This will result in faster adoption, higher conversion rate and increased user satisfaction.

2. PROPOSED METHODOLOGY

The coupons provided to the telecom companies are initially classified manually by the coupon provider companies themselves. Some examples of such classes (C) and subclasses (SC) are shown in Table 1. The goal of our project is to initially identify whether any website browsed by the customer of telecom company belongs to any of these existing classes or subclasses or it is uncategorized. If it is not belonging to any of the existing clusters we need to suggest some suitable keywords that describes it best. The proposed method uses multifactorial approach to classify the websites. The factors we have used are defined below:

- Find the Longest Common Sub Sequence (LCS) of each class and subclass type with the URL. Let the string length (strlen) of SC or C $= k$, and string length of LCS $= p$. Then the factor f_1 can be defined as $f_1 = p/k$. Note that, $0 \leq f_1 \leq 1$.
- Use the information used in Wikipedia for the corresponding website to be used as document for the corresponding website (term) .
- Remove stop words and stem them.
- Construct bag of words (BOW) for all URLs. Let it contain n words.
- Construct the feature vector f_i for document d_i by marking which word in BOW is contained how many times in d_i. Thus f_i is of length n and may be highly sparse.
- Construct a term-document matrix, where each row represents a term and each column represents a document. An element of term-document matrix is the *tf-idf* value of corresponding term in corresponding document. *tf-idf* value is obtained by multiplying term frequency with inverse document frequency. *tf-idf* assigns higher score for important words of a document and penalizes words common to all documents.
- Find the clusters by computing the cosine distance. This is represented as feature f_2.
- Also compute the frequency of (C/SC) in d_i. Let it be defined as feature f_3.
- Thus we get 3 factors f_1, f_2, and f_3.
- Then assign weight for each factors by using traditional machine learning methods.
- Define a threshold (t) using the soft max approach. If two websites have a distance at most t, then we define them to belong to the same cluster.
- If the new website has a distance greater than t from any of the existing websites, then we define the new website to be uncategorized.

2.1 Cluster name synthesis

- Sort the key phrases associated with each web site according to their score.
- Each website is represented by the key phrases with having the top three ranks.
- If multiple entries have the same rank, we consider all those phrases as key phrases and the collection of these phrases are used to represent the website.

3. RESULT AND DISCUSSION

We represent our experimental results in two tables. Table 2 shows the similarity score among different websites. We have set $t = 0.2$ as the threshold. If the similarity score among two websites is at most 0.2, we conclude that they belong to the same cluster. In this table, we find that Twitter and Facebook belongs to same cluster as both of them are social networking sites. Similarly, Netflix and YouTube are clustered together in the same cluster as both provide streaming video on demand and also serves as video recommender system. ESPN Cricinfo, Gmail and Zapak are not matching with any of the other websites. Alexa is best matched with Twitter with similarity score 0.23, because both are San Fransisco based companies, though they don't have similarity in their function. This problem can be removed if we carefully select the key words from Wikipedia.

4. CONCLUSIONS

Our proposed method successfully classify different websites dynamically. It can also cluster similar websites that can help to construct a social network graph. It takes the best part of human annotation done for Wikipedia, but we are not taking any humans in the loop. This would help us to identify customer behaviors and can be extended to other applications.

5. REFERENCES

[1] Eagle, Nathan, and Alex Sandy Pentland. *Eigenbehaviors: Identifying structure in routine.* Behavioral Ecology and Sociobiology 63, no. 7 (2009): 1057-1066.

[2] Isoda, Yoshinori, Shoji Kurakake, and Hirotaka Nakano. *Ubiquitous sensors based human behavior modeling and recognition using a spatio-temporal representation of user states.* In Advanced

Table 2: Confusion matrix

	Alexa	Amazon	ESPNcricinfo	Facebook	Gmail	Google	Netflix	Twitter	YouTube	Zapak
Alexa	1	0.04	0	0.02	0	0.05	0.04	0.23	0.11	0.01
Amazon	0.04	1	0	0.09	0.01	0.27	0.2	0.07	0.08	0.02
ESPNcricinfo	0	0	1	0	0	0	0	0	0	0.05
Facebook	0.02	0.09	0	1	0.08	0.14	0.11	0.29	0.25	0.02
Gmail	0	0.01	0	0.08	1	0.05	0.02	0.05	0.11	0.1
Google	0.05	0.27	0	0.14	0.05	1	0.13	0.15	0.11	0.02
Netflix	0.04	0.2	0	0.11	0.02	0.13	1	0.08	0.39	0.01
Twitter	0.23	0.07	0	0.29	0.05	0.15	0.08	1	0.26	0.01
YouTube	0.11	0.08	0	0.25	0.11	0.11	0.39	0.26	1	0.02
Zapak	0.01	0.02	0.05	0.02	0.1	0.02	0.01	0.01	0.02	1

Table 3: Cluster name synthesis

Alexa	Amazon	ESPNcricinfo	Facebook	Gmail	Google	Netflix	Twitter
market	washington	media	software	cross-platform	web	video	social
search	online	cricket	social	webmail	provider	rental	networking
engine	book	espn	networking	google	mountain	silicon	software
research		outlet			view	valley	
amazon		gopher			advertise	nasdaq	
		protocol			world		
		sport			wide		
					portal		

Information Networking and Applications, 2004. AINA 2004. 18th International Conference on, vol. 1, pp. 512-517. IEEE, 2004.

[3] Anmol Madan, Manuel Cebrian, David Lazer and Alex Pentland. *Social Sensing for Epidemiological Behavior Change.* Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp) 2010.

[4] Michal Kosinskia, David Stillwella, and Thore Graepel. *Private traits and attributes are predictable from digital records of human behavior.* Proceedings of the National Academy of Sciences (PNAS) 2013.

[5] Silei Xu and John C.S. Lui. *Product selection problem: improve market share by learning consumer behavior.* Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD) 2014.

[6] K Parsons, *Human thermal environments: the effects of hot, moderate, and cold environments on human health, comfort, and performance,* - 2014

[7] Jyotsana Shukla, *Extreme Weather Events and Mental Health: Tackling the Psychosocial Challenge,* 2013

[8] GW Evans, *Environmental stress,* - 1984

[9] Mauss, Iris B., Robert W. Levenson, Loren McCarter, Frank H. Wilhelm, and James J. Gross. *The tie that binds? Coherence among emotion experience, behavior, and physiology.* Emotion 5, no. 2 (2005): 175.

[10] Levenson, Robert W. *Blood, sweat, and fears.*Annals of the New York Academy of Sciences 1000, no. 1 (2003): 348-366.

[11] Arora, Gaurav, Ashish Kumar, Gitanjali Sanjay Devre, and Amit Ghumare. *MOVIE RECOMMENDATION SYSTEM BASED ON USERS' SIMILARITY.* International Journal of Computer Science and Mobile Computing 3, no. 4 (2014): 765-770.

[12] Wang, Shaoqing, Benyou Zou, Cuiping Li, Kankan Zhao, Qiang Liu, and Hong Chen. *CROWN: A Context-aware RecOmmender for Web News.* In Data Engineering (ICDE), 2015 IEEE 31st International Conference on, pp. 1420-1423. IEEE, 2015.

[13] Anglano, Cosimo. *Forensic analysis of WhatsApp Messenger on Android smartphones.*Digital Investigation 11, no. 3 (2014): 201-213.

[14] Gaikar, Vishal. *Find the Similar Sites on the Internet with SimilarWeb.* (2011).

[15] Bonfiglio, Dario, Marco Mellia, Michela Meo, and Dario Rossi. *Detailed analysis of skype traffic.* Multimedia, IEEE Transactions on 11, no. 1 (2009): 117-127.

[16] Wang, Shiang-Kwei. *The effects of a synchronous communication tool (yahoo messenger) on online learners' sense of community and their multimedia authoring skills.* Journal of Interactive Online Learning 7, no. 1 (2008): 59-74.