Information Retrieval

Lab Assignment 3

Paris Mavromoustakos (10407502)
Patrick De Kok (5640318)

December 23, 2012

1 Introduction

In the third lab assignment, the goal is to present a novel IR-based experiment, without actually implementing it. We will propose a system and suggest how it could be applied and evaluated. Since there will be no results or output, our focus will be on the theoretical background and research, analysis and evaluation methods for the proposed IR system.

2 Hypothesis & Research Questions

We are going to propose an IR-based system that will be applied on public transportation for marketing purposes. Our goal is to use passenger information obtained from the Internet to improve the performance (more specifically, the popularity) of public transportation advertisements.

While immense research progress has been made in the field of Internet IR-based advertising, the main challenge of our system will be to choose the appropriate advertisements that will have most impact on the passengers watching it on the bus, tram or any kind of public transportation means' screens. Passengers will have to be "categorized" according to their consumer profiles, their interests and their temporal needs. But how can we turn all these profiles into a "bus" model and pick the advertisements that satisfy the majority of the passengers? And moreover, how will we be able to evaluate the performance of this new advertising system?

Those are all interesting and quite challenging research questions. The fact is that the target audience[3] of our system is fairly random, since the public transportation users are millions of people. However, this gives us the motivation to focus on each single passenger's occasional (temporal) needs, deriving from his/her Internet consumer profile. We believe that recency of information retrieved (for example google search queries) can play a key role, as long as it is helpful enough to categorize a passenger/consumer. We also believe that certain terms from the information retrieved can be weighted as more significant regardless of the likes and interests of the passenger. For example, if we retrieve recent google search queries, we would definately consider the phrase "Cheap houses in Amsterdam" of more importance than just "Houses in Amsterdam". The word "cheap" and other similar terms, can be weighted so as to represent the actual importance or need of a consumer on a certain field of products. It is an indicator that the passenger is hoping to buy something, and is not just looking for some abstract information over the topic (for example a photo search). On the other hand, we could assign small weights to phrases that contain words such as "hate" because most of the time, they show lack of interest on a certain topic.

Our hypothesis is the following: We believe an IR-based system can help increase public transportation's advertisements' popularity, by personalizing advertisements according to the passengers' consumer profiles.

3 Related Work

In terms of Internet-based advertising, we believe that Google AdSense has been the most successful and popular application until now. AdSense is an advertising application, which can target consumers (Internet users) with specific ads that match their interests, or demographics[1]. Google AdSense chooses suitable advertisements for any website that supports Google advertising by retrieving information based on overall site content[1], while it enables customers to choose which field of advertising will be publicated on their browsers. The advertisements shown are also derived from user interaction[1] (the number of clicks on specified categories).The auction system that manages those advertisements has proven to grant Google 97% of its total revenue[2].

Significant progress on the creation of consumer profiles, where a person's likes and interests are stored, has been made with the introduction of Data Management Platforms (DMPs). DMPs are being used in the field of advertising, and more specifically, online advertising, since they collect data from online traffic (user searches, clicks, page visits, etc.) to create databases where each user can be represented by his interests. DMPs are of high importance to Internet-based advertisers who are now able to define their desired target audience, granting them higher precision rates in their advertising campaigns.

Regarding a less advertising and more IR-related part of this experiment, paper *Evaluating Vector-Space and Probabilistic Models for Query to Ad Matching*[4] discusses how queries of web search engine users can be matched with relevant advertisements on the target web pages. Various models are compared and results show that a certain translation model performs optimally when it comes to this task. The proposed translation model categorizes synonyms of product terms (for example shoes/sneakers) in the same class and assigns synonyms non-zero scores when retrieved in documents[4].

Paper *Optimizing Relevance and Revenue in Ad Search: A Query Substitution Approach*[5] proposes a query optimization algorithm that works in two phases; online and offline phase. In the offline phase, the system learns and uses substitute queries related to the queries the user has entered. In that way, weakly composed queries could be substituted by another that optimizes the desired metrics. In the online phase, the system uses *exact match*[5] to retirieve adverstisements that refer to the (possibly) substituted query. This method seems quite helpful in our experiment, since we cannot rely on the well-constructiveness of the queries/statements passengers have submitted on their Google accounts or social network profiles, when trying to create their consumer profiles.

Lastly, an also important field to examine is user intention as derived from queries. How do we know a user is intending to buy a product and how can we take advantage of this fact? Paper *Characterizing Query Intent From Sponsored Search Clickthrough Data*[6] discusses that topic, using online advertisements' clickthrough logs to detect user intent. The interesting part we could use in our experiment is understanding *online commercial intention*[6] given a user query, which represents the willingness of a user to buy or generally use a product.

4 Experimental Setup

Before we start analyzing how our system will be set up, we consider important to describe the existing hardware and software setup in public transportation.

In most European Union countries, most public transportation services encourage passengers to use a check-in card instead of a ticket, like the Oyster card in London, or the OV-chipkaart in the Netherlands. Passenger cards like those can be personalized, containing personal data of the respective owners (Name, date of birth, even e-mail address), so that travelers can gain access to monthly discount packages or other offers.

Moreover, public transportation means like buses, trams, trains and subway trains, already have screens installed, on which, apart from the route, advertisements are displayed. Those advertisement spots can be bought by advertisers from the public transportation authorities for a certain amount of money and be displayed while passengers are en route. Those advertisements are fixed regardless of the passenger's consumer profiles, and that is what we are proposing to change.

Our system will consist of three related parts: the "Profiler", the DMP, and the advertising part. More specifically:

- **Profiler** The profiler will be the first mechanism triggered in the system and is closely related to the DMP. A passenger walks in the bus and checks in with his personal passenger card. The card contains his email, which will be used by the Profiler to search over the Internet for any social media pages, Google search queries or eBay and Amazon accounts, and use the IR model proposed in [4] combined with the user intent model in [6] to create a consumer profile for that passenger.

  In more detail, each single Google search, Amazon or eBay search or Facebook status, comment or Tweet will be retrieved by the Profiler as a document and certain queries will be applied on the sum of all documents to extract each passenger's interests and intentions and store them in the DMP. Passengers who already have an entry in the DMP will have their profiles updated. The queries will already be composed depending on the advertisers and their fields of focus and will be of short length, for example "Air tickets", "Housing", etc.

  The crucial part of the Profiler is how the querying results will create a passenger's consumer profile. For example, does a Google search like "rent bikes in Madrid" have the same value as "bike routes in madrid"? The answer is definately no. We believe that applying a model like the one proposed in [4] for every passenger can give us an idea of his likes and interests, and categorize him according to those. But the significant part is unsderstanding passenger *online commercial intention* to define his needs on products or services. We will focus on that part and consider the recency of retrieved relevant documents as the most important feature. Recent relevant documents retrieved from the profiler will be assigned with a higher score, and will affect the passengers' consumers profiles as follows: each passenger profile will contain a list of his interests and needs, and high ranked topics on that list will be topics of temporal interest. For example, a passenger that likes "Tennis" on Facebook will have "Tennis" in his interest list, but a passenger that recently searched for "cheap tennis rackets" will have "Tennis" in his top ranked interests.

- **DMP** The DMP (Data Management Platform) will be the system where all the passenger profiles will be stored. Passenger profiles will consist of the information stored in the check-in card, and a list of interests and needs, as mentioned above. The DMP will be updated every time a passenger enter a means of public transport with new data retrieved by the profiler. The DMP will be placed on a remote server and all means of transport will be connected to it to exchange information.

- **Advertising** The advertising part includes the mechanism that chooses which advertisements will be displayed in a means of public transport's screens.

  After the Profiling stage and DMP updates on each passenger, a bus is now modeled as an audience of heterogeneous groups of consumers[7]. The Advertising system will have to classify passengers into groups according to their interests and temporal needs, and then decide which group(s) should be targeted by the advertisements displayed. These groups can be created by retrieving the passengers' top ranked interests in the DMP's list, and they should be generalized to a certain point. The labels for these groups will be already defined by the advertisers, but we believe that too specialized topics can lead to an unnecessary rise in the number of groups and can make the classification too complex. Each passenger can be part of more than one groups, but the maximum number of groups should be defined aswell.

  In this part of the system, the DMP can serve computational purposes. For each passenger in a public means, the advertising system could access the DMP to compute the average time spent in that particular route for each passenge, average this time over all present groups of passengers and increase the precision of the advertisements shown. More specifically, if we know that 10 passengers are looking to buy product A and 5 passengers are looking to buy product B, but the second group is estimated to exit the means sooner than the first group, it would be wise to focus on the second group first.

  Given the fact that the time available is very little, a priority queue could be kept in each public means, updated by the DMP at each passenger station. The first group in that queue will be served an advertisement sooner than the rest, while those priorities could be recomputed each time new data (passengers) step in and out of the vehicle.

The general description of the system above is a framework which discusses the basic setup. The models proposed are described in the respective papers as optimal methods, but we could use the evaluation and feedback given by advertisers or passengers to improve the system or some of its parts. Since we are proposing something that has never been implemented yet, we cannot safely state that this would be the best performing model possible.

We are trying to focus on the IR-related part of this experiment, but the truth is that marketing methods and models need to be inserted into the system to make it feasible.

5 Evaluation

The evaluation of this system could be done by three sides: the consumer side, the advertiser side and the public transport services' side. Advertising can bring both an increase to a company's income but also a satisfaction to a customer's needs.

On the consumer side, the result we would like to see is an increase in advertising precision. That means, the advertisements will have made impact if the public transportation passengers use products or services they saw on their trips. But how can we be sure that consumers select products because of our system?

Of course, one way is to report on the products' sell rates, but what if public transportation is not the only means where this product or company is advertised? We suggest that advertisements shown in public transportation could include a special offer code for passengers, with which they could get a discount for a certain product. In this way, we would be able to monitor the sell rates regarding the specific advertising campaign, while an increase in sales through this would confirm that the passengers/consumers are satisfied with our system.

On the advertiser side, this system could provide them with more than just an increase in sales. The DMP is a source of very important data such as, the average age of a consumer group or the area in which a product or service is more wanted. For example, statistics deriving from the DMP could indicate the demand of a specific service in an area, scaling from a part of a city covered by a bus route, to parts of a country covered by train routes. Having that information, companies can power up their advertising campaigns while satisfying a population's needs.

Public transportation services also increase their income if the advertisers choose them to promote their products. We believe that an increase in demand of advertising time in the public transportation screens would be the most trustful evidence that our system satisfies all ends. Passengers learn about products of their interest, buy more, companies make more income, thus interest in that means of advertising is increased. The system that manages the advertisements' prices could be designed similar to an auction system like Google AdWords[8] where an advertisement's price for a company could be defined by the frequency of it being displayed to the public, while advertisers can define in which means, areas(routes), or target audiences they wish to focus their advertising by bidding on advertising time.

6 Conclusion & Remarks

This assignment helped us get creative and propose of an IR-based system is not in use yet. We were able to combine IR with other fields of science such as marketing and economics, and derive to the conclusion that it can be used in numerous applications over many other fields.

Something that may need to be discussed more thoroughly is whether applying this system proposed is a violation of personal data. Creating a DMP over passengers and distributing their profiles to advertisers should be decided by both sides, both public transportation services and their clients. A solution to this could be adding conditions to the purchase of a check-in card for use in public transport. The passenger can always skip using the system by buying a regular ticket.

Lastly, it is a fact that not all passengers of a public transportation vehicle use personal check-in cards. So the profiles deriving from a DMP could represent only a small percentage of the actual load. In this occasion, we understand that there is (still) no means of recognizing every single (anonymous) passenger and only the results of an application of such a system could determine if

there is any improvement in advertising precision.

References

[1] *http://support.google.com/adsense/bin/answer.py?hl=en&answer=9713*

[2] *http://www.wordstream.com/blog/ws/2011/11/16/how-adwords-works*

[3] *http://www.imediaconnection.com/content/28493.asp*

[4] Hema Raghavan, R. Iyer, *Evaluating Vector-Space and Probabilistic Models for Query to Ad Matching*, SIGIR, 2008

[5] Filip Radlinski, Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Lance Riedel, *Optimizing relevance and revenue in ad search: A query substitution approach*, SIGIR, 2008.

[6] A. Ashkan, C. Clarke, E. Agichtein, Q. Guo, *Characterizing Query Intent From Sponsored Search Clickthrough Data*, SIGIR, 2008.

[7] E. F. Fern, *Focus Groups: A Review of Some Contradictory Evidence, Implications, and Suggestions For Future Research*, Advances in Consumer Research Volume 10, 1983.

[8] *http://en.wikipedia.org/wiki/AdWords*