Similarity of Advertisements with AI techniques

Student Name: Varvara Tzika

Host organization: Marktplaats (www.marktplaats.nl)

Contact person:
Ton Wessling - twessling@marktplaats.nl

Introduction:

Marktplaats.nl is an e-commerce platform on which most content is in the form of advertisements created by common users selling a broad range of items. An advertisement consists of content describing elements such as title, description and attributes. Users can also search for a product in case they want to buy an item from the available advertisements. In Marktplaats anyone (signed in or not) can search for an interesting product. The main procedure is to search for a desired product and matching results will appear.

However, there may be more relevant results that have been omitted because of too specific terms in the search query. For instance, if the user searches for "Mercedes-Benz A-Class 160 Blue" the results may only be advertisements that match the search query in their titles or in their description. The desired result could be Mercedes-Benz A-Class ordered by the relevance. Also, the results could be better if they are related to the advertisement that the user has already seen. A more relevant recommendation can be cars which are in the same class but a different brand.

Many companies had to come up with related solutions to problems like these, as Google News who groups news by story rather than presenting a raw listing of all articles. Also, they record every click or search that every user made and just below the "Top Stories" section users can see a section labeled "*Recommended for your email address*" along with three stories that are recommended to them based on your past click history. [1]
Likewise, Amazon.com based on users past shopping history and site activity recommends books and other items likely to be of interest.

In the music industry, [11] use data from Last.fm and create user profiles based on the genres of the most listened artists. They create communities based on the categorization. So for example one community is hip hop and the members all listen to hip hop. Also, Apple's Genius feature in iTunes classifies songs into potential playlists for users [10].

The previous cases take advantage of a good structure of their big amount of data logs due to the categorization that they create. Also, they can predict users' interests and give better recommendations. They increased their revenues, they keep their users satisfied and they eliminate the time that users spend to find the product that they want.

My opinion is that nowadays it is a demand of the market to provide to their users functionalities like "Recommended to you" or "Relevant advertisements" in case they want to keep their users more time on the site or increase sales. Furthermore, data is increased day by day so a better management is essential. Having large amounts of data creates value for a company and Marktplaats doesn't take full advantage of it.

The big amount of data that Marktplaats has contains this information that we need to relate user's behavior with better recommendations. However, to show to user relevant advertisements based on the previous search requires finding the relevant advertisements first. One of the options is to classify advertisements in categories however, trawling through hundreds or thousands of categories and subcategories of data is no longer an efficient method for finding information. [14]

Recommending similar advertisements to users involves retrieving the most important information of the advertisement which already had seen. Marktplaats advertisements had this information either on their description or in their attributes.
Extracting this knowledge from enormous amounts of data can be achieved with methods from the information retrieval area which is defined as "the area of study concerned with searching for documents,

for information within documents, and for metadata about documents."[5].

Research Question: Is there any way to make the relevancy of advertisements for a big Corporation more accurate and with high performance with the use of information retrieval methods?

Problem analysis:

Our approach, described briefly, is to create an algorithm which will run in every advertisement a user made click and it will search for similar advertisements to recommend them to the user.

As it is mentioned above, advertisement consists of content describing elements such as title, description and attributes. Many attributes of an advertisement are irrelevant with our project. The most important knowledge will be retrieved from the title, the description of the advertisement and the attributes. Examples of bicycles advertisements' attributes are the color, the height, if it is used or not etc.

The existing architecture of Marktplaats helps as to create a scalable solution and it doesn't need big changes. They use a LAMP package architecture which enables us to have a free, open source and easily adaptable distributed solution.

Currently the advertisements and users information are in one database and the log files are in flat texts. My proposed solution will include one module that will communicate with the first database and another module will communicate with the log files. The first module will be responsible for the information retrieval and the second will be responsible to take the previous search query that the user made. We will use this to make a method for relevancy. Also we will create many methods for relevancy based on the first module results. After the creation of the different methods, we will have a list of methods that will make the same thing with different results and after the evaluation of these results we will have this list ordered by the accuracy and performance of every method. With this way we will have many alternative methods for the relevancy of advertisements and a possible failure of one can be replaced by the next method.

To find an existing solution is really difficult because of the big competition between corporations. Such implementations are kept as trade secrets. However we will not start from scratch. To start let's define the problem. Such a problem belongs to the Information Retrieval domain. With the use of information retrieval we can extract the most important terms of every advertisement and then with the use of search engines we will have as result the similar advertisements. Search engines (retrieval model) are responsible to search in our advertisements based on the input and they will produce us the similar advertisements.

A good choice of the retrieval models will be proved once adequate categories are constructed based on the input that we gave. Choosing the best existing model to use for our project will be one of the small challenges of this project because there are so many adequate models implemented.

However we can use different models to perform the categorization-grouping, the chosen model will produce different results bases on the input.The choice of the best input is the big challenge of our project. The different results that the model will give us will affect the performance and the accuracy of the system. However, we can conduct an experiment to find the best method for extracting the most appropriate input. The different methods we can use will be evaluated in the end to find the most accurate similarity of the advertisements and the best solution based on its performance.

Also, another challenge is that due to the big amount of data and because we have to think that in the future it will be doubled, the proposed system has to take scalability in to account. Finally it is desirable that the system will run in real time.

Research Method

Difficulties

The enormous amount of data requires finding a scalable solution and after implementing the proposed model, the performance of the host organization page must not be affected.

- Stopwords like can.

- Polysemy

Literature

- The best retrieval models for the specific case of Marktplaats
- Information Retrieval way to extract and weight the most important terms of an advertisement

Method

Our approach is that we will create a "query modeling" to retrieve the important terms from the advertisement that the user clicked on it. However there is not available any query modeling which is the best in our case. Thus we will create many methods and then we will evaluate which is the best way based on its accuracy and its performance. Then, an existing retrieval model will calculate the most similar advertisements.

Possible ways to create methods for the query modeling are:

1. Every term of the search query that a user made to find the specific advertisement.

2. The most important term of the advertisement

3. Many terms of the advertisement with a different weight based on its importance.

4. The same with 3 added the attribute fields of the advertisement with different way.

5. Combination of the previous.

The two possible tools that I can use are Lucene Apache and Indri. However, the benefits of using Lucene and the fact that Marktplaats already works with apache makes us choose Lucene.

Lucene is a high performance, scalable Information Retrieval (IR) library. Lucene lets you add searching capabilities to your applications. It is a mature, free, open-source project implemented in Java; it's a project in the Apache Software Foundation, licensed under the liberal Apache Software License. As such, Lucene is currently, and has been for quite a few years, the most popular free IR library. [12]

The existing retrieval models that I can use are:

- Okapi BM25F
- Indexing
- Tf.idf

Hypothesis and theory

Our hypothesis is that one of the methods for the query modeling will create relevant advertisements that will not affect the performance and the relevancy will be accurate.

Validation of my hypothesis

To validate who of the methods is the best and how accurate are its results we will create a confusion matrix.

Confusion Matrix is a specific table layout that allows visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). [13]

We can compare the results of every method and the desired results from a chosen set of advertisements and create a confusion matrix. Then, we can have the percentage of methods accuracy.

To validate the performance of every method we will use software tool with benchmarking methods. With this way we will know who of the methods has the better performance.

Expected results of the project:

- The actual proposed model/architecture

- Different methods to make the relevancy of the advertisements.
- The results of my validation.

Required expertise for this project:

- Java programming
- Information Retrieval knowledge
- My SQL

Time line

1. 1/4-10/4: Learn Lucene and choose one of the retrieval models.
2. 11/4-20/5: Creation of different methods for query modeling.
3. 21/5-10/5: Evaluation of my results.
4. 1/6-31/6: Finish the thesis documentation

Bibliography:

1. Author: Abhinandan Das, Mayur Datar, Ashutosh Garg
   Title: Google News Personalization: Scalable Online Collaborative Filtering
   Event: Proceedings of the 16th international conference on World Wide Web: May 8-12, 2007.
   Pages: 271-280.
2. Author: Khan, D
   Title: CAKE – Classifying, Associating & Knowledge Discovery an Approach for Distributed Data Mining (DDM) Using Parallel Data Mining Agents (PADMAs)
   Event: Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on
   a. Pages: 596-601
3. Author: Ankur Narang, Raj Gupta, Anupam Joshi and Vikas K. Garg
   Title: Highly Scalable Parallel Collaborative Filtering Algorithm
   Event: High Performance Computing (HiPC), 2010 International Conference on
   Pages: 1-10

4. Author: Apeh, E.T.; Gabrys, B.; Schierz, A.;

Title: Customer Profile Classification Using Transactional Data
Event: Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on
Pages: 37-43

5. Author: Singh, B.; Singh, H.K.;
Title: Web Data Mining Research: A Survey
Event: Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on
Pages: 1-10

6. Author: Tseng, A.; Petrounias, I.; Chountas, P.;
Title: A Complete Framework for Web Mining
Event: Systems, Man and Cybernetics, 2003. IEEE International Conference on
Pages: 868-873

7. Author: Cooley, R.; Mobasher, B.; Srivastava, J.;
Title: Grouping Web Page References into Transactions for Mining World
Wide Web Browsing Patterns
Event: Knowledge and Data Engineering Exchange Workshop, 1997. Proceedings
Pages: 2-9

8. Author: Shen Hui-zhang; Zhao Ji-di; Yang Zhong-zhi;
Title: A Web Mining Model for Real-time Webpage Personalization
Event: Management Science and Engineering, 2006. ICMSE '06. 2006 International Conference on
Pages: 8-12

9. Author: Cooley, R.; Mobasher, B.; Srivastava, J.;
Title: Web Mining: Information and Pattern Discovery on the World Wide Web
Event: Tools with Artificial Intelligence, 1997. Proceeding. Ninth IEEE International Conference on
Pages: 558-567

10. Author: Sean Owen; Robin Anil;Ted Dunning;Ellen Friedman
Title: Mahood in action
Pages: 28-255

11. Author: Nico Schlitter; Tanja Falkowski.;
Title: Mining the Dynamics of Music Preferences from a Social Networking Site
Event: Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in
Pages: 243 - 248

12. http://en.wikipedia.org/wiki/Lucene

13. http://en.wikipedia.org/wiki/Confusion_matrix

14. Authors :Eric Hatcher, Otis Cospodnetic and Michael McCandless
Title: Lucene in action