Aalto University
School of Science
and Technology

Department of Mathematics and Systems Analysis
Mat-2.4177 Seminar on Case Studies in Operations Research

Using Internet search trends for predicting product sales

Midterm report
19.3.2010

Client:
Nokia Markets, Strategy and Business Development
Toni Jarimo

Project Group:
Joonas Ollila (project manager)
Andreas Hübner
Pyry Åvist
Rasmus Hotakainen
Susanna Siitonen

Table of contents

Project status

The project is proceeding according to our original plan. We have investigated the usability of different search terms and found the most effective ones. A model framework for the analysis has been coded in Matlab and can be used for quick analysis of the data. Some surprises have turned up, for instance we noticed that the trends data provider Google adds some noise to the data. This caused some confusion and even greater confusion when we noticed that the noise gets bigger if we use an automatic downloader. Those problems have however been solved and now we are investigating whether we can find some general patterns between different models and countries. Another thing we work on currently is finding different measurements of goodness for the models.

Findings and results

What we have found so far is that there exists a notable (0,7-0,98) correlation between the search trends of a product name and the actual sales of the product. In many cases there is a notable correlation between the sales of two consecutive months, thus the basic model we have used tries to explain the sales at a moment using the trends at that moment and the sales one month before. The most difficult part to predict is the beginning of the sales, as you can see from the figure below.

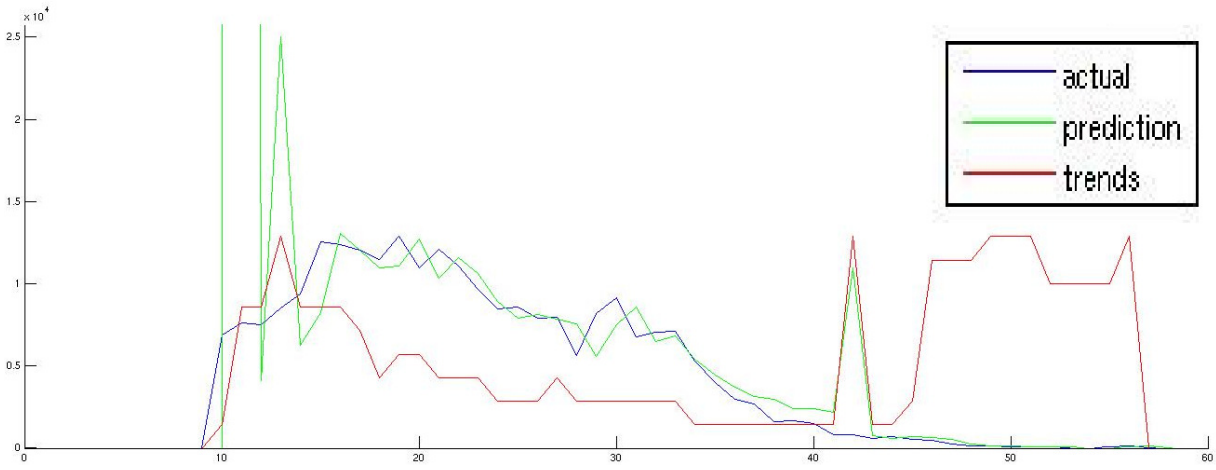

Figure 1. Predicted and actual sales.

Some attention should be paid upon the construction of search terms. In most cases searching with a combination of the brand name (e.g. Nokia) and the product name (e.g. N-Gage) gives the best results. However some product names change from one country to another. Also some products have so long names that the consumers prefer to use the nickname of the product. Wikipedia can be successfully used for acquiring the nicknames. All in all, it is clever to do a quick search on the internet with the intended search query and see what comes up. If there is nothing unusual, go with the combination brand name and product name.

Mathematical methods

The primary mathematical model that has been used in our analysis is a dynamic regression model. A dynamic regression model is a basic multiple regression model with the special characteristic of having the lag(s) of the predicted variable as predictor(s). In other words, if we try to explain the sales of a month y_t, the model could look like this:

$$y_t = \mathrm{ax_t} + \mathrm{by_{t-1}}$$

where $\mathrm{x_t}$ is the sales of that month. This basic model can (in some cases) be improved by adding trends lags and/or sales lags. We are still investigating whether there exists an (even partly) usable model for all the cases.

The other model type that we have used is called an external learning machine, i.e. one kind of a neural network. This model type is more complex than the dynamic regression model and shall be more widely explained in the final report if we find it usable.

New challenges

The dynamics of our sales data change in time

One challenge for us has been that the sales data dynamics are not the same throughout the product life cycle. This originates to the nature of our data, in the beginning the sales vary a lot, this “aggressive” phase is followed by a mature phase with little changes in sales. Our way to get around this is to use only a limited number of data points for building the model. The model framework takes only n

consecutive data points to predict the next one, we are still working on finding an optimal n. By keeping n small enough we can hinder the phases from influencing each other too much.

The changing of dynamics is a reason why we cannot validate our model using the usual 80/20-procedure, i.e. first creating the model with 80 % of the data and then validating it by 20 %. This would lead us to creating the model with data that has different dynamics than the data used for validation. The validation will be done by examining the errors between the predicted sales and the realization – this way the change in dynamics does not affect our validation procedure noteworthy.

Trends data changing from time to time

A new challenge we cannot do so much about is the fact that our trends data changes from time to time. We noticed this accidentally when we were downloading some product trends. It also seems that the trends data gets worse when using an automatic downloader. This led us to download all of the trends manually using Google Insight instead of Google Trends (Insight has more extensive information, even though Google claims the trends information to be identical).

Following steps

The next steps of our projects will be implementing some changes to the model framework, including the new validation method. Simultaneously with this we investigate which sweep length n (i.e. how many data points should be used for building the model) is the most optimal one. When the changes are done we start creating the models for all the data sets and tabulate the results. The final report writing has already started and will continue during the rest of the project.

Updated timetable

The updated timetable of our project can be found below in table 1. The nature of the remaining tasks is such that they can be done simultaneously.

Table 1. Updated timetable

	Week				
Task	12	13	14	15	16
Implementing changes to the model framework					
Creating models for the data sets					
Validating the models					
Writing the final report					

Updated risks

Some of the risks that we included in our project plan can be wiped out at this point. We know that the model gives at least to some extent useful results. We also know now that our data source is inaccurate, so this is not technically a risk anymore. The biggest risks at this point are that the framework has some bugs that affect the results, the project is delayed for a reason or another or that we jump to false conclusions.