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

Project Plan
19.2.2010

Client:
Nokia Markets, Strategy and Business Development
Toni Jarimo

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

Contents

Introduction

The development of Internet search volumes can be examined with various services on the Internet. Some of them (e.g. Google Insight) provide information about the search history in a given area and allow the user to compare different keywords' search volumes. Thus you can examine rising, falling and regular trends. For instance Google Insight allows you to download this information freely from the Internet as a CSV file. In this project we will examine the usefulness of the history of certain search query volumes for predicting product sales. Our hypothesis is that there is a correlation between the search volumes of queries linked to a product and the actual sales volumes of the product (see Figure 1). Interesting questions include among others how long the delay between search volumes and actual product sales is and how strong is the correlation between these two. The aim will be at developing a useful model for predicting product sales.

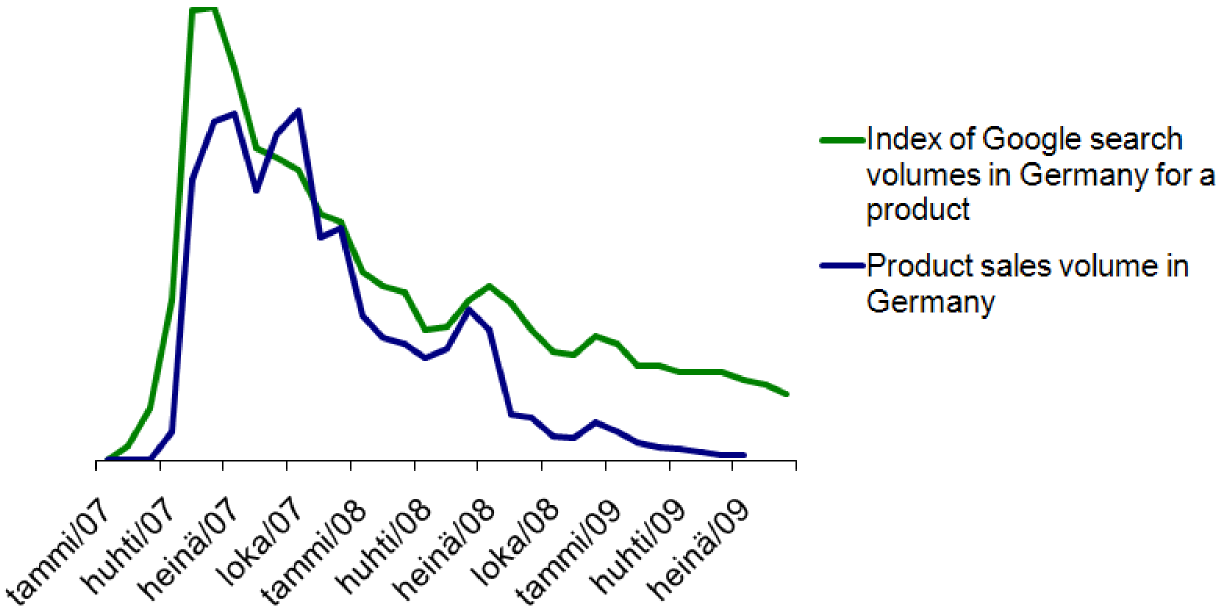


Figure 1. Search trend and sales for a product.

The project is conducted in spring 2010 for the Seminar on Case Studies in Operations Research (Mat-2.4177) at Aalto University. The client of this project is Nokia Markets, Strategy and Business Development and our contact person there is Toni Jarimo.

Objectives and research questions

The main objectives of this work are:

1) To investigate the usability of different data sources in predicting product sales.
   -What sources can be used for predicting product sales?
   -How reliable are the sources?
   -For which geographies are the sources valid?
   -How does the choice of search words affect our results?

2) To build a model for short-term prediction of consumer interests, more precisely for product demand on several markets.
   -Which model is most suitable for this analysis (a (non-)linear regression model, ARMA-model or some kind of dynamic regression model)?
   -What are the parameters?
   -Do the parameters change significantly between areas and/or products?

3) To assess the validity of the model.
   -Is the model usable in real life?
   -What is the time span of our forecast?
   -How accurate are the forecasts (measured with a 95 % confidence interval)?

Approach

As an initial step we search the Internet for research linked to our project. We'll take a deeper look into data mining by reading relevant parts of Alan Porters book Tech Mining. Simultaneously all group members think of possible data sources and search for them. Prospective sources could be Google Trends, Google Search Insight, Twitter Trends and Yahoo Buzz for instance.

We will explore those prospective sources and analyze the revealed information by data mining techniques "in order to find meaningful patterns and rules" (Berry and Linoff 2004: Data mining techniques: for marketing, sales, and customer support). The amount of available data on the Internet grows each day. Users add information about themselves by using public available services like search engines. Those services capture the usage data i.e. search terms and date and time of each request and aggregate the information in anonymous search data collections. Some of these collections are available in public and ready to be explored. The challenge for our research is finding the right sources and applying advanced methods to this data. Thus we will be able to examine large data sets in an effective, semi-automated way.

The next step is to choose our data sources and start planning how we get the data downloaded. Simultaneously we will start figuring out what search queries might be effective. One possibility is to add words like "buy" or "price" in addition to the product name in the search query. A challenge that we face is that the data is proportional, not absolute. This might lead to a model that estimates the market share rather than the actual sales. As a preliminary plan we are going to investigate 20 different products in 22 different countries, thus a program for data downloading is necessary. The countries we are going to investigate are: United Arab Emirates, China, Germany, France, United Kingdom, Italy, Russia, Spain, India, Indonesia, Turkey, Poland, Vietnam, Taiwan, South Africa, Saudi Arabia, Nigeria, Philippines, Egypt, Iran, Argentina and Pakistan.

When the data is downloaded, we'll start refining it and making it usable for the analysis. At this point the usability of different models on the data is analyzed. The product data is monthly whereas the trends data is weekly, thus it will be necessary to create new data points by interpolating. The price of the products we are investigating range from 20 € to 1120 € and they are usually sold during a period of one to three years.

There are a few models that could be used. The linear or nonlinear regression models describe how the sales change with the search volumes. That kind of models could not be used for predicting the future. The ARMA models could be used for describing how the sales or the search volumes develop in time, but not for finding any correlation between these two. The ARMA models could also be used for predicting the future. The best would however be to find some kind of a dynamic regression model that could be used both for describing the correlation and for predicting the future. One good possibility would also be a neural network model, which is basically a blackbox model. The models and their properties are described in table 1.

The model is chosen and then applied on each of the time series. After this the model will be validated and we will see how usable it is in real life. One interesting question is whether the same model can be used for several products of the same company. Our project steps follow the allocation Alan Porter presents in his book Tech Mining (see table 2). The model we are developing is a short term prediction model, i.e. it can maximally estimate the next two months' sales.

Table 1. Models and their properties

Model	Advantage	Disadvantage
Linear or Nonlinear Regression	- Describes how the sales depend on the trends	- No prediction of future
ARMA	- Takes into account the time series aspect - Prediction of future	- Does not describe how the sales depend on the trends
Dynamic Regression	- Takes into account the time series aspect - Describes how the sales depend on the trends - Could possibly predict the future	
Extreme Learning Machine (Neural Network)	- Simple to use - Good generalization performance	-Blackbox

The progress of all activities will be reported to the project manager, client and to necessary extent in the final report. The project manager is responsible for communicating with the client and course staff. Good communication between the group and the client is important in order to produce useful results.

Table 2. Steps in data mining

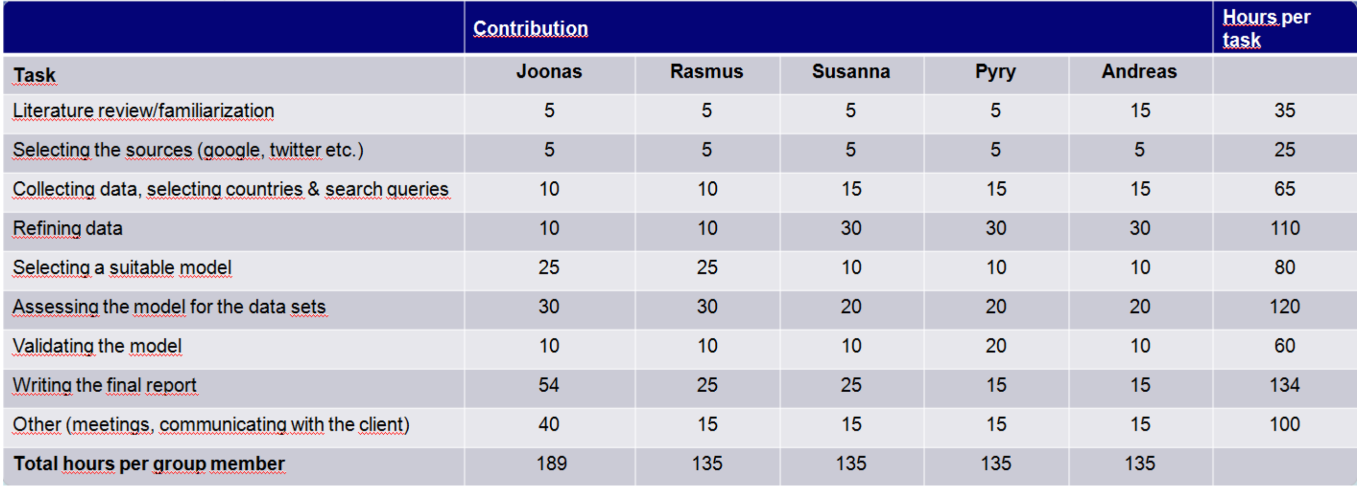
Intelligence	1. Issue Identification 2. Selection of Information Sources 3. Search Refinement and Data Retrieval
Analysis and Design	4. Data Cleaning 5. Basic Analysis 6. Advanced Analysis

Resources

The group consists of five members: Joonas Ollila (project manager), Rasmus Hotakainen, Andreas Hübner, Susanna Siitonen and Pyry Åvist. All of the group members have competence in applied mathematics. To avoid confusion within the tasks, working hours are preliminary assigned to different persons, the distribution is in Table 3. One group meeting per week will take place in the initial state of the project in addition to e-mail communication. Once everyone knows what to do a part of the meetings may be replaced with conference calls (Skype).

In case of technical problems the course staff can be consulted.

Table 3. Working hours

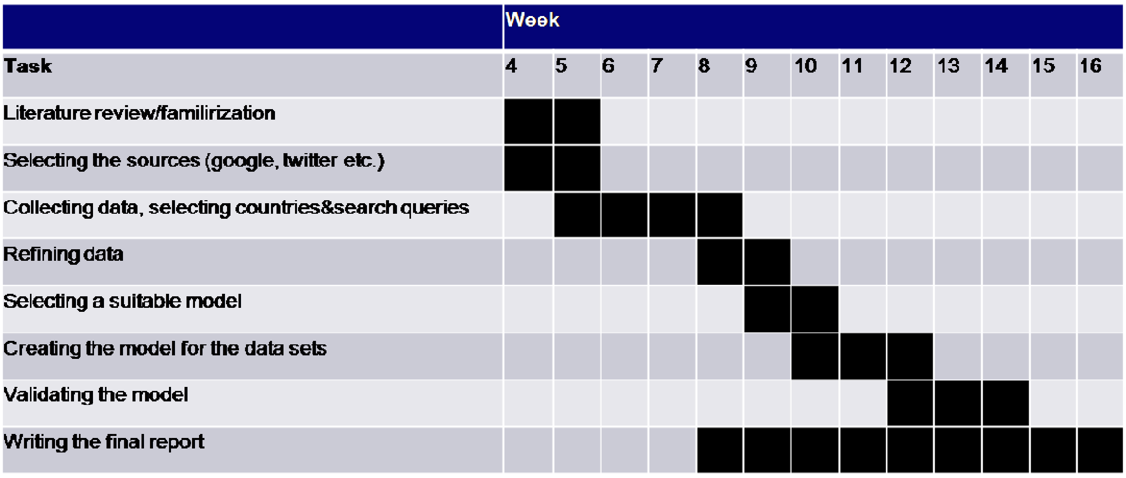


	Contribution					Hours per task
Task	Joonas	Rasmus	Susanna	Pyry	Andreas	
Literature review/familiarization	5	5	5	5	15	35
Selecting the sources (google, twitter etc.)	5	5	5	5	5	25
Collecting data, selecting countries & search queries	10	10	15	15	15	65
Refining data	10	10	30	30	30	110
Selecting a suitable model	25	25	10	10	10	80
Assessing the model for the data sets	30	30	20	20	20	120
Validating the model	10	10	10	20	10	60
Writing the final report	54	25	25	15	15	134
Other (meetings, communicating with the client)	40	15	15	15	15	100
Total hours per group member	189	135	135	135	135	

Schedule

Below in table 4 is a tentative schedule for our project. The nature of the subprojects is such that all have to be (almost) finished before the next one can be started, we cannot for instance build a model without data.

Table 4. Timetable



	Week												
Task	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Literature review/familirization	■	■											
Selecting the sources (google, twitter etc.)	■	■											
Collecting data, selecting countries&search queries		■	■	■	■								
Refining data					■	■							
Selecting a suitable model						■	■						
Creating the model for the data sets							■	■	■				
Validating the model									■	■	■		
Writing the final report					■	■	■	■	■	■	■	■	■

Important dates:
- Project plan presentation (19.2.)
- Midterm report presentation (19.3.)
- Final report presentation (23.4.)

Risks

Below is a table of the most significant risks and how we plan to face and prevent them.

Table 5. Risks

Risk	Prevention
The project is delayed.	The schedule is planned in advance and the responsibilities are clearly assigned to different persons to avoid delays. The progress of the project is followed weekly.
The model does not give useful results.	We preliminary investigate the usability of different models on the data before choosing one. The course staff may be consulted about the model. Our model is constructed such that changes can easily be made to it.
The data sources are not accurate.	This risk cannot be prevented; it is however worth to notice that there is a possibility that the data is not accurate.
The final report is of little use for Nokia.	The project manager will inform our contact person continuously of the progress made and the contact person will inform about the client's needs. This risk is partly linked to the risk that the model does not give any useful results.

Material linked to this project

Tech mining: Exploiting New Technologies for Competitive Advantage (2004, Alan L. Porter, Scott W. Cunningham)

http://google.com/insights/search/#

http://buzzlog.buzz.yahoo.com/overall/