BIG DATA RECOMMENDATION PROBLEMS IN E-COMMERCE SOLUTIONS FOR SMALL BUSINESS

Michał Bernardelli[1]

Abstract

The dynamic development of e-commerce has increased the demand for efficient algorithms and systems based on statistical analysis. The simplest of them use the web traffic statistics, other use sales parameters. Because of the amazing simplicity, transparency and enhanced features, much popularity was gained by the Google Analytics tool. None of the methods, however, without the appropriate algorithms that automate operations, is suitable for use in real time. Intelligent recommendation systems, such as the mechanism of Collaborative Filtering, significantly contribute to an increase in sales but are generally characterized by poor scalability. Of course with proper computer infrastructure and specialist knowledge, it is possible to gather big volumes of data and analyze them. All sophisticated solutions, however, are rather reserved for large companies, whose activity is based on the Internet.

In this article, Big Data recommendation problems are described. Advantages and disadvantages of several used in practice algorithms are considered in particular emphasis on the suitability for the small e-commerce business. The main point of the article is the proposition of the simple in implementation recommendation algorithm and thereby achievable for small business. What is more, the online test was performed and its results presented as a good performance proof. The actual data were used thanks to the courtesy of Run4Fun.pl. In the test, the aspects of a large amount of data but also their volatility and diversity was taken into consideration.

Key words: Big Data, e-commerce, recommendation algorithm.

JEL Classification: C63, C88, O30.

[1] Warsaw School of Economics, Collegium of Economic Analysis / Szkoła Główna Handlowa w Warszawie, Kolegium Analiz Ekonomicznych, e-mail: michal.bernardelli@sgh.waw.pl

DOI: 10.19253/reme.2017.03.005

1. Introduction

Data mining in the era of highly developed Internet market is unquestionable. Techniques used for data analysis have their roots and theoretical basis mainly in statistics, but they are also strongly associated with the field of computer science. A particularly important aspect of the automation of the analysis is the size of the explored data. With the growing computer computational power, also grows the demand for more accurate analysis. They are usually based on non-aggregated data and due to the big volumes of input data needed to be processed, a proper software, computer infrastructure and/or specialist knowledge is necessary. In many cases, it is even referred as a whole new science, called Big Data (Mayer-Schönberger, Cukier, 2014).

One of the most common Big Data applications in modern personalized marketing is recommendation algorithms. They seem to be especially effective in the Internet industry. In the e-commerce, recommender systems give one of the ways to improve efficiency and increase the so-called conversion rate. There are various of different recommendation algorithms that may be and are used in practice of making product recommendations during a live customer interaction (Sarwarm et al., 2000). Available articles concerning this area of scientific interest, like (Adomavicius, Tuzhilin, 2005) or (Schafer et al., 2001), describe in general effectiveness of recommender systems on the actual customer, which results in designing powerful and scalable algorithms. Those algorithms, however, are out of reach for small business, because of the lack of specialist knowledge, skilled workers, and too expensive infrastructure. The aim of this article is to present a proposition of the simple in implementation recommendation algorithm, which is achievable for small business. Theoretical description of an algorithm is complement by the performance tests, to proof the usefulness of described approach.

This article is constructed as follows. After the introduction, the section gathering basic information about Big Data and links with e-commerce is placed. Next section is devoted to recommendation algorithms, covering the idea behind them, examples of usage and a brief comparison of few existing approaches. An alternative solution, designed for small business, is presented in the fourth section. It also includes results from performance tests. This article ends with conclusions and suggestions for further research.

2. Big data in e-commerce

2.1. Big Data – characteristics

Leading Internet companies like Google, Facebook, Yahoo or Amazon based their businesses on search engines or online shopping. In fact, however, the key aspect of their activities is a collection and processing large volumes of data, needed for example to calculate the search index, to analyze customers' tastes or to suggest

new articles that may be of customers' interest. Very soon, the same solutions were transferred to other, much smaller companies. The reason was rather not to introduce a new service, but thanks to the acquired information to achieve a competitive advantage. At the moment, practically every bank, online shop or supermarket collects data about customers' behavior and analyzes them.

The term Big Data is often identified with large volumes of data. There is no formal definition, but according to any existing so far, the meaning of this term is broader than just the data volume. In this section, the most popular definitions of Big Data are presented. A review of these definitions is taken from (Tabakow et al., 2014).

One of the first definitions of Big Data was that proposed by M. Cox and D. Ellsworth (Cox, Ellsworth, 1997). According to it, the authors treat Big Data as a big data to analyze, the number of which should be maximized in order to extract the information. Another proposal, called "3V" model, was given by an analyst working for the META Group (Laney, 2001). He defined data growth challenges as being three-dimensional:

— increasing volume (amount of data),
— variety (range of data types, sources),
— velocity (speed of data in/out).

In 2012 the same company (which has in the meantime changed its name to Gartner) introduced two additional dimensions relating to Big Data:

— variability,
— complexity.

In 2013 IBM has defined Big Data as a range of data generated from different sources, with a high speed, and in large amounts. IBM has characterized Big Data using four attributes: volume, velocity, variety and veracity. Also, SAS company, describing Big Data, to the volume of data adds additional attributes: variability and complexity. All of those definitions have in common one thing – according to them, Big Data is related not only to the volume of data.

Possibilities of getting information based on the disaggregated data turn out to be much greater than with the use of even complex models, but the limited data sets. Employees of Google Inc. in the article (Halevy, 2009) have noticed, that the analysis of large data sets often with just a few simple models is proving to be extremely effective in comparison to relying on even very sophisticated models, but applied to the data carrying much-limited information.

2.2. Internet – source for Big Data

To realize the magnitude available and sent information on the Internet, it's enough to just quote a few figures from the report of Cisco company (Cisco, 2013):

— in 2012 there were 2.3 billion Internet users, which represents approximately 32% of the population all over the world (7.2 billion people),
— until 2017 the number of Internet users will reach 3.6 billion, which with the projected world population of 7.6 billion will be accounted for over 48% of the population,
— in 2012 amount sent by the average household was estimated on 31.6 gigabytes of data per month,
— the amount of data transmitted by the average household in 2017 will be approximately 74.5 gigabytes per month,
— in 2017 annual Internet traffic will reach 1.4 zettabytes.

Therefore, the increase in the amount of information on the Internet is faster than is predicted by the empirical Moore's Law, according to which technological progress is doubling every two years.

Today's e-commerce is based on database systems with the user-friendly and functional interface, to which access is provided via the Internet. Each click of the user is recorded by the system and stored in a table in a database. There are also saved information such as a timestamp of entering the website, page address, the time of the each recorded user action (like click or scroll), the session number for the identification of the user, the IP address, etc. The increase in the amount collected in this way information, even for small websites is extremely fast and oscillates in the range of tens and hundreds of thousands of records a day. Those data are intended to be used to improve sales, that is to increase the value of the conversion rate (CR). It is the ratio of the number of transactions sales to the total number of visits and is usually expressed as a percentage. The conversion rate is the numerical expression of information on the percentage of customers visiting the store, who finally purchased, and in sales and marketing is considered as a standard key performance indicator (KPI). The statistical analysis of the collected data is intended to show the directions of development of the company that will result in the increased sales. This problem is definitely an example of Big Data, with basic challenges facing the e-commerce:
— large amount of data – even after reducing the data to the period associated with the life cycle of products (in the case of the sports assortment it's a period of 3–24 months), the number of records in the database to be analyzed is at least a few or even several million (in a rather small e-commerce shop);
— a wide variety of data – the behavior and preferences of users are subject to change over time and may depend on the number of potential factors such as time of year, gender, age, education, place of residence;
— data changes in real time – online shops are available 24 hours a day, every day of the week; the exact values of the parameters associated with the user's visits to the websites can change within seconds, which requires re-calculating the indicators used in descriptive statistics or the use of more sophisticated data mining algorithms.

Modern e-commerce is restricted not only to simply online shopping websites for retail sales direct to consumers. There are also numerous other Big Data applications, like online financial exchanges or providing online marketplaces. Especially popular is the real-time bidding (RTB), which means instantaneous online auction systems with advertising inventory that is bought and sold on a per-impression basis (Bernardelli, 2015). Those auctions are conducted using automatic algorithms based on user's demographic information, browsing history, geolocation, and the website information. However, usually, this type of activity is reserved for big companies, which are not of interest of this article.

3. Recommendation algorithms

Recommendation algorithms are an effective form of personalized marketing. In the market reality, small e-commerce businesses are looking for cheap and easy-to-implement ways to improve efficiency and increase the conversion rate. Over the years a number of recommendations algorithms were developed, some more, some less sophisticated, see (Adomavicius, Tuzhilin, 2005), (Sarwarm et al., 2000), (Schafer et al., 2001). The idea behind all of them is to create personalized lists of items for the user. Recommendations algorithms must take into account the following aspects of the e-commerce specifics:

— Big Data – problems associated with the large volumes of data and data changing in real-time,
— the results of the analysis must be determined in fractions of seconds and have high accuracy despite the large amounts of data,
— very limited information on new clients and at the same time blur information related to regular customers.

In general, on websites of modern online stores, various lists of products (items) for the potential customer are placed. Most of them may be described by the following wording:

(a) customers who bought this item also bought …,
(b) proposed, similar to the given products are …,
(c) customers who viewed this item also viewed … .

Implementing a method, which returns the most frequently purchased products from the list (a) is relatively easy and almost every commercially available software for online shopping[2] has built in this functionality. Products from the list (b) may be the result of permanently (by hand) set connections. Despite the obvious disadvantages of such a solution, it is still widely used on the Polish e-commerce market. These connections can be associated for example with the colors of the pro-

[2] Like PrestaShop, Gekosale, Sky-shop and many other freeware solutions.

ducts (eg. other products in green color), their purpose (eg. a similar model, series), completing the set (eg. matching sweatshirt for viewing model pants) and time of delivery / manufacture (ie. other products added this month). The lack of automation, however, disqualifies this method in the situation of more than a few hundred products or fast changing assortment. The problem of finding the items from the list (b) is reduced to the solution of the problem of finding the most similar products. Often used in data mining product similarity measures are based on product's characteristics (color, size, gender, year of manufacture, purpose, used technologies, etc.). At least two of them should be mentioned (Leskovec et al., 2012):

— Jaccard index (Jaccard similarity coefficient) defined for the two sets of properties A and B as

$$J(A, B) = |A \cap B| / |A \cup B|.$$

— cosine similarity defined for the two vectors A and B, which represents the properties of the items, as

$$\cos(A, B) = (A \cdot B) / (\|A\|_2 \|B\|_2),$$

where means Euclidean scalar product.

The list (c) of items is somehow a special case of the (b) problem, since to the assessment of the similarity of products can be also used history of users' visited sites. However, there is an important difference in both approaches, namely, in the case of (b) we can limit analysis only to the products themselves as well as sales history, whereas in the case of (c) we have to deal with profiling users' preferences, and to determine the list of products on the basis of similarity between users, not the products, as in the case (b).

Recommendations algorithms can be divided into two groups. In one group we can place methods exploring characteristics and properties of products (so-called user-to-item methods), whereas the second group consists of the methods based on the customers and their interests (so called item-to-item methods). The second group requires a detailed description of each product and based on that finds the proper match between products using a proper measure like cosine similarity or Jaccard index. The other group requires profiling customers and finding users with similar interests. There are many possibilities of profiling users, for example on the basis of the purchased products, inserted reviews or visited websites. Let N be the number of users in the store, and M the number of available items. With that notation, the standard profiling requires remembering connections between users and items and may be presented in the form of the M-by-N matrix. Searching similarities among products only is less memory consuming[3] because the analogical matrix has sizes M-by-M. This matrix is symmetric, so in the worst case scenario, the number of non-zero elements is equal to $\frac{1}{2}M(M-1)$. In general user-to-item approach is consi-

[3] Assuming that $M < N$, that is number of users is greater than number of items in the store.

dered as potentially more accurate, but also much more difficult than item-to-item approach.

Big companies can afford to use effective, but having high hardware and programming requirements, algorithms. There are three basic classes of used in practical approaches to solving the problem of recommendations: collaborative filtering mechanism, clustering models, and methods based on sophisticated search algorithms. A detailed description of these approaches can be found in the book of Leskovec, Rajaraman, and Ulmann (Leskovec et al., 2012).

One of the most effective recommendation algorithms presently known is used by amazon.com, and it can be classified as an item-to-item collaborative filtering algorithm. This method was designed for tens of millions of users and products. However, it requires many calculations to be done offline with an access to many fast dedicated servers. For this reason, among others, this method is rarely used by smaller companies with a lower volume of turnover and profits. Advanced recommendations algorithms are not a good solution for them because of the complexity level and the high hardware requirements. A reasonable compromise in this situation is the use of publicly available data analysis tools such as Google Analytics, as well as algorithms that create a real-time list of recommended products for a specific client. Recommendation algorithms based on the similarity of the products are a good solution for small and medium-sized businesses. They have lower hardware requirements and run faster than algorithms based on user customization. The disadvantage, however, is generally a less accurate list of products.

4. Alternative solution for small business

To avoid significant investments in infrastructure and employees from the IT sector, with the use of data collected on pages visited by the Internet users, it is possible to create a list of recommended products, using relatively cheap, meaning the time of computations, solution, which will correspond to the needs of e-commerce in Poland. Let us define a matrix of associations between the products, where the strength of an association is measured by the number of visited by the client websites dedicated to specific products. More specifically, let the element p_{ij} of a square matrix P of size M-by-M, where M is the number of available products in the Internet store, be defined as the number of customers who visited at the same time, the website of the product with i index and the website of the product with the j index. Each customer is identified by a unique number of the session. The matrix P is symmetric, so it is enough to know only the values for the elements for $i < j$. Potentially, many elements will have zero value, and the matrix itself will be probably sparse. The reason for this is an existence of rather a big variety of products in each store, which can be divided into categories. For example, in the store with clothes, usually

products for women and men are watched separately. In this case, the elements of the matrix *P* related to many of such pairs of products will be zero.

Since the store's software is based on a database, so the matrix is worth presenting in the form of a database table. A sample *item2item* table structure may be as follows

```
CREATE TABLE IF NOT EXISTS `item2item` (
    `pid1` SMALLINT(6) DEFAULT 0,
    `pid2` SMALLINT(6) DEFAULT 0,
    `value` INT(255) DEFAULT 0
);
```

where *pid1*, *pid2* are the product identifiers, and *value* is the previously described element p_{ij} of the matrix *P*. Because of the matrix symmetry, it must be assumed that the condition *pid1* < *pid2* is fulfilled. On the matrix *P*, represented as an *item2item* table, two basic operations will be performed. The first one is reading *k* most similar products to the given product denoted as *REFERENCE_PROD*. The measure of similarity are the sizes of matrix elements. An exemplary SQL query[4], which returns the searched products, is as follows:

```
(SELECT `pid2` AS pid, `value`
FROM `item2item`
WHERE `pid1` = REFERENCE_PROD)
UNION
(SELECT `pid1` AS pid, `value`
FROM `item2item`
WHERE `pid2` = REFERENCE_PROD)
ORDER BY 2 DESC
LIMIT k
```

The second operation on the table is its update, in the case of the new customer visits the product website in the online store. This update increases by one the value of the records corresponding to the websites visited by the new user. In case the record in the database does not exist yet, a new record with *value = 1* is inserted. Again, as with the select query, updates can be done in a fast an effective way, concerning the structure of the *item2item* table.

An empirical analysis of the proposed recommendation algorithm was made on the basis of the real, historical online data from an Internet store. The estimated number of products used in this study is 3 500, while the number of unique visitors per day is around 30 000. The new session id is given for each new user. Therefore, it is impossible for the small online store, to use recommendation algorithms based on user profiles directly (user-to-item). To get an idea of the magnitude of numbers

[4] MySQL syntax is used.

involving in that kind of an approach, let us estimate a sizes M, N of the matrix P needed to cover a yearly demand of the store:
number of items: $M \approx 3\,500$
number of users: $N = daily\_num\_of\_unique\_users * 365\_days * num\_of\_items \approx$
≈ 38.3 *billion*

It should be emphasized, that database table of this size is rather not possible to maintain for the small business companies unless they have a high-class IT employee. Even on a dedicated server and all available, freeware relational databases, executing a query on the table this size would last nothing less than tens of seconds. Nowadays, use such a table directly to determine the list of recommended products is therefore not possible or at least not practical.

At the same time (one year), the *item2item* table would reach the size of just more than 6 million records (matrix 3 500 by 3 500 elements, but because of the symmetry of the matrix, only about a half of elements are nonzero). Comparing to the user-to-item solution (nearly 135 trillion elements) it is 22 million times more! What's important is that, in the case of item-to-item approach, the estimated number of nonzero elements should be considered as an upper limit. Not all products will be linked to each other because many of such pairs are not lying in the area of interest of one person. A large part of them will never be visited at the same time by a single user. For example, many women will be viewing only products dedicated to them. Thus, the relationship between the products for men and women will be expressed in the matrix by zero entries. Online tests showed that the percentage of nonzero elements in the matrix does not exceed 10. Of course, it depends on the assortment provided by the store. Nevertheless, in many cases, the matrix may be considered as sparse. In the performed tests, the number of nonzero elements was only about half a million, which is significantly less than the pessimistic estimates. For such a database table, queries related with element updates or determining the k most similar products, executed a just fraction of a second. Moreover, table with user data is then not necessary, which means saving space and memory. Relying on users in choosing similar products seems to be the great approximation of more sophisticated and accurate solutions. This could be concluded by the number of redirections from the starting website and websites of the products proposed by the item-to-item algorithm. To be more precise, a number of visited websites after deploying the algorithm increased by the factor 4.3 comparing to the random algorithm and by the factor 2.7 comparing to the approach using basic item characteristics (color, purpose, series). Conversion rate seems to be also greater, but to get statistically significant conclusions, the test should be carried much longer and should take into the consideration the seasonal effects in sales[5].

[5] Also A/B testing is advisable.

5. Conclusions

Recommendation algorithms are proved to be an effective form of personalized marketing. Advanced solutions are not a good solution for small business, because of the high complexity and requirements. One of such an effective, but not commonly available solution, is a user-to-item approach. A proposed in the article algorithm is an example of a concurrent, item-to-item approach. It's cheaper, faster, has low hardware and knowledge requirements, and therefore may be used in many online stores, almost in any existing software. Theoretical analysis and online testing showed many advantages of this approach, which proven to be an excellent alternative for commercial algorithms. However, that kind of algorithm, in general, gives less accurate recommendations. Therefore, it has to be taken into consideration – a proper balance between the accuracy of the results and the amount of work in deploy and maintain the chosen solution.

Of course, the presented algorithm may be easily generalized. Two exemplary ways of improving the performance of an algorithm are adding the non-symmetry of the matrix P and the random items instead of the most popular/similar one. The first idea is the modeling of the situation, where user watching product A is switching to another, similar and recommended by the algorithm, product B, is not necessary the same probable as watching by the same user product B first and then switching to the website with product A. Surely, there are many situations, where the paths are not equivalently used by users. Therefore, it may be better not to assume the symmetry in proposed solution. It will obviously increase the number of non-zero elements in the matrix P but may improve the accuracy of the recommendations. The second proposition is to add some random items on the list of the recommendation products. It is easy to imagine the situation, in which users restrict their attention to the group of products, whereas the other group is completely omitted. Of course, it could be the effect of the low level of attractiveness of those products. However, it could be also the direct consequence of the recommendation algorithm – users are choosing only products from the proposed list of item, leading to the so-called starvation of other products. Adding for example randomly one or two products to the recommendation list could increase a conversion rate and most of all the profit of the online store.

The effectiveness of the recommendation solutions will probably depend on the assortment in the store and many other aspects, like target group of users, competitive market or simply the graphic layout of the store. Nevertheless, with the further development of e-commerce, problems associated with large amounts of data will appear more frequently. At the same time a demand for efficient algorithms, having a solid foundation in the theory of statistics, will be growing. A large part of the problems will be definitely connected with the recommendation, allowing users to find interesting products in the fast increasing set of available items.

References

Journal articles:

1. Adomavicius G., Tuzhilin A. (2005). Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, *IEEE Trans. on Data and Knowledge Engineering*, 17:6, pp. 734–749.
2. Bernardelli M. (2015). Cheater detection in Real Time Bidding system – panel approach. „Roczniki" Kolegium Analiz Ekonomicznych SGH, No. 39, Oficyna Wydawnicza SGH, Warszawa 2015, pp. 11–23.
3. Cisco Visual Networking Index (2013). Global Mobile Data Traffic Forecast Update, 2012–2017, *www.cisco.com*.
4. Cox M., Ellsworth D. (1997). Managing Big Data for Scientific Visualization, *ACM Siggraph*, Vol. 97, pp. 146–162.
5. Halevy A., Norvig P., Pereira F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, doi: 10.1109/MIS.2009.36.
6. Laney D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety, *META Group*.
7. Sarwarm B.M., Karypis G., Konstan J., Riedl J. (2000). Analysis of Recommendation Algorithms for E-Commerce, *ACM Conf. Electronic Commerce*, ACM Press, pp. 158–167.
8. Schafer J. B., Konstan J. A., Reidl J. (2001). E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery*, Kluwer Academic, pp. 115–153.
9. Tabakow M., Korczak J., Franczyk B. (2014). Big Data – defnicje, wyzwania i technologie informatyczne. *Business Informatics*, 1(31):138–153, doi: 10.15611/ie.2014.1.12.

Books:

10. Mayer-Schönberger V., Cukier K. (2014). *Big Data: A Revolution That Will Transform How We Live*, *Work*, *and Think*. Eamon Dolan/Mariner Books.
11. Leskovec J., Rajaraman A.,Ulmann J.D. (2012). *Mining of Massive Datasets*. Cambridge: Cambridge University Press.

ZAGADNIENIA REKOMENDACJI WYKORZYSTUJĄCE BIG DATA DEDYKOWANE DLA MAŁYCH PRZEDSIĘBIORSTW E-COMMERCE

Streszczenie

Dynamiczny rozwój rynku e-commerce spowodował wzrost zapotrzebowania na skuteczne algorytmy i systemy wykorzystujące analizę statystyczną. Najprostsze z nich używają statystyk ruchu internetowego, inne statystyk sprzedaży. Ze względu na niezwykłą prostotę, przejrzystość i funkcjonalność, dużą popularność zyskało narzędzie Google Analytics. Żadna z metod, jednakże, nie nadaje się do wykorzystania w czasie rzeczywistym, bez odpowiednich metod automatyzujących jej działanie. Inteligentne systemy rekomendacji, takie jak mechanizm Collaborative Filtering, znacząco przyczyniają się do wzrostu sprzedaży, ale charakteryzują się na ogół słabą skalowalnością. Oczywiście, mając do dyspozycji rozbudowaną infrastrukturę komputerową i specjalistyczną wiedzę, można gromadzić duże ilości danych i analizować je. Wszystkie zaawansowane rozwiązania są jednak raczej osiągalne dla dużych firm, których działalność koncentruje się w Internecie.

W artykule opisano zagadnienia rekomendacji związane z Big Data. Podkreślono zalety i wady kilku stosowanych w praktyce algorytmów, ze szczególnym uwzględnieniem ich przydatności dla małych firm działających na rynku e-commerce. Celem artykułu jest propozycja prostego w implementacji algorytmu rekomendacji, który byłby dostępny dla małych firm. Co więcej, przeprowadzone zostały testy on-line, których wyniki przedstawiono jako potwierdzenie skuteczności działania algorytmu. Rzeczywiste dane sprzedażowe zostały udostępnione przez firmę Run4Fun.pl. W teście wzięto pod uwagę kwestie dużych wolumenów danych, lecz również ich zmienność i różnorodność.

Słowa kluczowe: Big Data, e-commerce, algorytm rekomendacji.

Klasyfikacja JEL: C63, C88, O30.