УДК 004.9

Y. Stekh, V. Artsibasov
Lviv Polytechnic National University
CAD Departament

MODELS AND METHODS FOR BUILDING WEB RECOMMENDATION SYSTEMS

Modern Word Wide Web contains a large number of Web sites and pages in each Web site. Web recommendation system (recommendation system for web pages) are typically implemented on web servers and use the data obtained from the collection viewed web templates (implicit data) or user registration data (explicit data). In article considering methods and algorithms of web recommendation system based on the technology of data mining (web mining).

Key words: web page recommendation, collaborative filtering, clustering, pattern extraction, associative rule, evaluation methods.

Сучасна мережа Інтернет містить велику кількість веб-сайтів і сторінок на кожному веб-сайті. Веб-систему рекомендацій (рекомендаційну систему для веб-сторінок), як правило, втілюють на веб-серверах і використовують для даних, отриманих зі збірки проглянутих веб-шаблонів (неявні дані) чи реєстраційних даних користувачів (явні дані). Розглянуто методи і алгоритми рекомендаційних веб-систем, основаних на технології видобування даних (веб-аналіз).

Ключові слова: рекомендаційна веб-сторінка, сукупна фільтрація, кластеризація, добування шаблонів, асоціативне правило, методи оцінки.

Introduction

Recommendation system – a system that work with a certain type of information system of filters that recommend information items that can cause the user interest. A typical recommendation system makes recommendations to users, as input, aggregates and sends them to the recipients in the form of recommendations. This technology allows users to spend a minimum time for the necessary information on the Internet. Recommendation systems compare data collected from users and create a list of items recommended to the user. They significantly improve the retrieval of information for users.

Models of recommendation systems

Web recommendation system takes a set of web pages on the web site and user information as input. The result of the Web recommendation system is a subset of Web pages that best satisfies the information needs of users. Formal statement of the problem of finding web pages on the web recommendation system can be formulated as follows. $U = \{u_1, u_2, \mathbf{K} u_k\}$ – set of users of the site is derived from registration data and server logs. $P = \{p_1, p_2, \mathbf{K}, p_n\}$ – the set of all the Web pages that can be recommended. Let $g(u_i, p_j)$ – utility function to determine the usefulness of the page p_n for the user u_k $g: U \times P \to R$ where R – ordered set of non-negative numbers. The purpose of the Web rscommendation system is that for each user to select a single page, which maximizes the utility function. Data structure Web recommendation system is an important factor which affects the efficiency of the system.

Data structure Web recommendation systems

In Web recommendation systems for web sites used content data, structure data, user information (user data) and usage data.

The content data of Web pages include items and links that are available to users. Text content of Web pages represented as a vector of weighted frequency of occurrence of specific words in the text.

Structure data reflect the organization of the content of the web page. This organization includes information about external and internal structure of the page. Information about the external structure is displayed via hyperlinks that connect one page to another. The internal structure is formed from HTML pages or XML tags. This information is represented as a tree structure with tags, which are taken from the page. These structures are obtained from the use of mapping tools that look through a site from root URL-address to the destination tag and generate a map that illustrates the structure of the site, typically organized in a hierarchical form.

User information can be obtained through its interaction with the web site and depends on the destination Web site. For example, some sites require registration of the user, which usually involves specifying the user name, password, email address and demographic information. Some commercial sites take into account previous user purchases and other explicitly and implicitly represented the interests of users. The most important source of data in recommendation systems have access to the web server logs that are automatically going to web servers and application servers. The server records the date and time of the transaction. It also saves the name and size of the sent file records information about how the file was transferred to, any errors that may have arisen during its transmission, as well as information about the browser that uses the user. Information provided by the web server can be used to construct a data model that consists of several abstractions, including users, pages, click-streams, sessions.

Preprocessing of data

An important objective of the process of creating recommendations are preliminary data processing, which derived from all available sources. The purpose of preprocessing is to build reliable and integrated data sets that are used effectively in a box pattern finding and analysis user steps. Typically, pre-processing for specific data must be performed according to the data included in the recommendation model. For this reason, pre-processing tasks that use in this step are grouped according to type of data used in model recommendation.

Methods for generating recommendations

A key component of web recommendation system is the methods and algorithms to generate recommendations. These methods and algorithms are classified based on the methods of data analysis used to model user. Most methods of finding recommendations are divided into the following classes

collaborative filtering
associative rules
clustering
sequential patterns
semantic Web.

On the basis of the content – recommendations for the products formed, similar to products already ordered by the client or the goods that are ordered like customers.

The formation of recommendations uses the following two approaches:

– Memory-based – recommendations formed the basis for calculating a measure of all the accumulated data. This approach is simpler, has shown high accuracy in practice and has the advantage of incremental account of new data. New transactions are simply added to the database and are included in formation of the forecast, together with existing. This approach is difficult to calculate in terms of time and memory resources. Also, this approach can not provide a descriptive analysis of existing laws to give greater understanding of the available data and explain the prediction.

– Model-based – formed a descriptive model preferences of users, products, and the relationship between them, and then form recommendations based on the model. The advantage of this approach is the availability of the model. This model gives a greater understanding of the generated recommendations and the availability of relationships in the data. The process of generating recommendations is divided into two stages: resource-learning model in deferred mode and a fairly simple calculation based on the recommendations of the existing model in real time. Such models do not support incremental learning. The emergence of new data requires a recalculation of the entire model. This model shows a lower predictive accuracy than a Memory-based

Collaborative filtering algorithm for the similarity between users

The idea of collaborative filtering is reduced to that similar customers make similar purchases, and similar items are bought together customers. The algorithm is reduced to watching a large group of users and the search for her in a smaller group similar to each other, for some metric. For example, the scoring on the same film. Overall ranking list of priorities for this group, and then use the resulting list to all users, members of smaller groups. Therefore, before calculating the similarity of the samples can significantly reduce the number of computations in the formation of the list of preferences for members of smaller groups.

There are two approaches to collaborative filtering:

User-centric filtering – unknown rating services billed on the basis of ratings, which were stamped the same services to users, similar to this one. This approach is implemented in two steps:

Find users who have chosen the same services as the user.

Offer a service with a maximum rating of the services selected by similar users.

Item-centric filtering – unknown rating services billed on the basis of ratings of other similar services already included in the user request. This approach is implemented in two steps:

Construct a matrix of services to determine the degree of similarity between the services.

Using the degree of similarity to offer services similar to the already ordered this user.

Uses a hybrid approach.

Clustering

Clustering – grouping a set of images into groups by similarity of characteristics. In the field of web use are three application clustering: clustering sessions, clustering users and clustering pages. Application of clustering sessions allows clustering user session in which users have similar access model. Most Web applications are very difficult to form clusters of users due to lack of data about users. Because of this are used clustering sessions. Clustering users aims to establish user groups for the similarity of viewing patterns. Clustering of pages for the purposes of prediction is as follows. Given adequately clustered collection. If the user is interested in page p_i, he probably would be interested in the other cluster members which includes a page p_i. Clustering of pages can be grouped into two categories. Such that clustered according to their page content. Such that clustering pages depending on how often they occur together between user sessions. Most used the same function that the similarity in information retrieval (eg, cosine similarity or Euclidean distance). Search for clusters in a set of web pages by using the methods of hierarchical and non-hierarchical clustering.

Associative rules

Associative rules allow to find patterns of related events. For the first time this problem has been proposed to find associative rules to find common patterns of purchases that are made in supermarkets. Because sometimes they are called market basket analysis. The purpose of the analysis is to establish these relationships. If the transaction meets a set of X, then based on this we can conclude that a different set of Y must also meet in the transaction. Establishing such relationships enables us to find very simple and intuitive rules. Search algorithms associative rules are designed to locate all the rules of “if X then Y”, with support and reliability of these rules should be higher than some predefined thresholds, called minimum support respectively (minsupport) and minimal confidence (minconfidence). The task of finding associative rules divided into two subtasks:

Finding all sets of elements that satisfy the threshold minsupport. These sets of elements are called so often encountered.

Generating rules from sets of elements that are found under item with certainty that meets the threshold minconfidence.

In the context of Web mining, this problem reduces to finding correlations between the various links to files available on the server by the client. Each transaction consists of a set of URL-address client access in a single visit to the server. Using the method of finding association rules can be found a correlation between the URL-address. The database contains a significant amount of transactional information.

Therefore, modern methods for finding association rules designed to reduce the search space (scaling) while maintaining the same quality search results.

In the context of Web mining, this problem reduces to finding correlations between the various links to files available on the server by the client. Each transaction consists of a set of URL-address client access in a single visit to the server. Using the method of finding association rules can be found a correlation between the URL-address. The database contains a significant amount of transactional information. Modern database Web search engines are very large, reaching the giga-and terabytes, and tend to further increase. Therefore, to find associative rules require efficient scalable algorithms that can solve the task within a reasonable time. These algorithms include APRIORI algorithm with steps to scale and transform data. Modern methods for finding association rules designed to reduce the search space (scaling) while maintaining the same quality search results.

Conclusion

In this article provides a brief description of models and methods that are used to build Web recommendation systems. In recent years, been made significant progress in the development of web recommendation systems. Introduced new methods and algorithms for making recommendations. Some systems have found practical applications in electronic commerce. The current generation of recommendation systems requires further improvements in the use of scalable algorithms and machine learning methods.

1. Kosala R., Blockeel H. Web mining research: a survey // SIGKDD Exploration. – Vol. 2. – P. 1–15, 1997. 2. Arocena G., Mendelzon A. Weblog: restructuring documents, databases, and webs // Theory and Practice of Object Systems. – Vol. 5. – P. 127–141, 1999. 3. Balabanovi'c M., Shoham Y. Fab: content-based, collaborative recommendation // Communications of the ACM. – Vol. 40. – P. 66–70. 4. Bucher A., Baumgarten M., Anand S., Mulvenn M., Hughes J. Navigation pattern discovery from internet data // Proc. of the WEBKDD'99 Workshop on the Web Usage Analysis and User Profiling, August 15, 1999, San Diego, USA. 5. Agrawal R., Srikant R. Fast algorithm for mining association rules. Proc. of the 20th VLDB Conference, Santjago, Chile, 1994. 6. Bing Liu Web Data Mining: exploring hyperlinks, contents, and usage data. Springer-Verlag, Berlin Heidelberg, 2007. 7. Zang H., Spiliopoulou M., Mobasher B., Giles C., McCallum A., Nasraoui O., Srivastava J., Yen J. Advances in Web Mining and Usage Analysis // Proc. 9th Iternational Workshop on Knowledge Discovery on the Web and 1th Iternational Workshop on Social Networks Analysis, San Jose, August 12–15, 2007. 8. Fisher-Huber] S., Lambrinoudakis C., Pernul G. Trust, Privacy and Security in Digital Business.// Proc. 6th International Conference, Trust Bus 2009, Linz, Austria, September 3–4, 2009. 9. Chakrabarti S. Mining the Web: discovering knowledge from hypertext data. Morgan – Kaufman Publisher, 2002. – 352 p. 10. Johanson P. Design and development of recommender dialogue systems. // Linkopings Studies in Science and Technology, Thesis No. 1079, 2004. – 129 p. 11. Jansen B.J., Spink A., I. Taksa I. Hdbook of research on Web Log Analysis.. IGI Global. – 2009. – 523 p.12. Sharda N. Tourism informatics: visual travel rscommender systems, social communities, and user interface design. IGI Global, 2010, 276 pp. 13. Uchyigit G., Ma M.Y. Personalization technigues and recommender systems. Word Scietific Publishing Co., 2008. – 303 p.