SEQUENCE MINING ON WEB ACCESS LOGS: A CASE STUDY

Carlos Soares [a] Edgar de Graaf [b] Joost N. Kok [b]

Walter A. Kosters [b]

[a] *LIACC/Faculty of Economics, University of Porto*
Rua de Ceuta 118, 6 andar, 4050-190 Porto, Portugal
csoares@liacc.up.pt
[b] *LIACS, University of Leiden, Leiden, The Netherlands*
{edegraaf,joost,kosters}@liacs.nl

Abstract

We present a case study in which sequence mining algorithms were applied to web access log data. The data are from a portal that is targeted for business users. In this portal, like in many others, content is described using a set of descriptors, such as keywords, category and type. We investigate whether representing content by the type rather than its identifier enables existing sequence mining methods to obtain interesting patterns. Rather than a more traditional approach based on measures such as support and confidence, we analyze results from an application perspective. This enables us to identify opportunities for improving and extending these methods.

1 Introduction

Understanding the behavior of users in websites is essential for successful e-business (e.g., pricing [3] and improving loyalty [10]). The most common approach to this task is based on the analysis of web access logs, which store clickstream information, i.e., information about the requests performed by the users of a website. Given the information in web access logs, which can be regarded as sequences of clicks, a natural approach to analyse the navigation behavior of users is the use of sequence mining methods [8].

The case study in this work is PortalExecutivo.com (PE), a Portuguese web portal targeted to business executives. The goal of PE is to become an essential information tool for its customers by facilitating the access to a large body of relevant content. One interesting feature of PE is that although only registered users can access content items, guest users (i.e., non-registered users) are able to navigate within the structure of the portal and view summaries of those items. Like for any portal, understanding the behaviour of its users is essential for PE to provide a better service to them. Improvements can be made regarding services (e.g., if email alerts are not used, maybe they should not be provided), structure (e.g., relocation of a sub-tree of the content hierarchy which is not commonly accessed), meta-data (i.e., meta-data describing content, such as keywords and categories) and type and quantity of content (e.g., increasing the number of items from a particular source which is very popular). Two questions that are relevant to analyze the behavior of users are: Which are the paths most frequently followed by the users? Are there differences in the navigation patterns of guest and registered users?

The use of data mining techniques to address this kind of questions is usually referred to as web usage mining (e.g., [9]). The use of sequence mining techniques for web usage mining is increasingly popular (e.g., [5]). One method of analysis would be to do process mining as discussed in [13, 14]. The browsing behaviour will result in sequences of clicks and these sequences can be interpreted as Petri nets describing this behaviour. However, probably there are multiple types of users with different Petri nets and it is difficult to separate these users in different workflows.

In this paper we present a case study of the application of sequence mining algorithms to address those questions. However, most papers focus on research issues (e.g., proposing and comparing methods) and evaluate results in terms of technical measures (e.g., confidence and support) which relate only indirectly to application goals. By taking on this application perspective we are able to identify advantages and disadvantages of this approach, focusing our analysis of results on a business point-of-view rather than a more technical perspective. We also investigate the effect of using an alternative representation of items in the sequences. The representation is based on the *meta-data* (i.e., descriptors such as type, category, keywords) which are often used to describe content in portals. Although the paper focuses on a case study, we believe that conclusions are applicable to a wide range of websites.

A summary of the necessary background is given in Section 2. More detail concerning the application is given in Section 3. In Section 4 we present results, which are discussed in Section 5. Some conclusions are given in Section 6.

2 Background

In this section, we provide background concerning the Web and Sequence mining, which puts our work in a suitable context.

2.1 Web Usage Mining

The amount of information available on the web makes the use of semi-automatic tools essential to obtain knowledge that is useful to users. One approach to this problem is *web mining*, which consists of the application of data mining methods for that purpose. A good overview of web mining can be found in [9].

One of the sub-categories into which web mining is commonly divided is *web usage mining*. It is defined as the "application of data mining techniques to discover usage patterns from Web data" [12]. Most of the time this is achieved by analyzing access log data that represents clicks that users make and is generated by the web server. A number of problems arise in the use of this type of data, which include the accurate identification of the user and navigation that does not generate requests to the server (e.g., pages stored in local cache). A number of different web usage mining applications have been reported, such as personalization, user profiling and site modification.

2.2 Sequence Mining

The goal of sequence mining is to identify interesting sequential patterns in a database of sequences [1]. Data is stored as sequences in many applications, such as in biology [4], shopping baskets [1] and web access logs [11]. Here, we will limit our description to basic concepts which are required to understand the rest of the paper. More information can be found in many papers dedicated to this area (e.g., [8, 2, 4]).

We assume we have a database D of sequences of items, where each item is an atomic representation of an object of interest in the application (e.g., a request for a content item in a web portal). We call a given sequence $d = (d_1, d_2, \ldots, d_m)$ a *super-sequence* of a sequence $s = (s_1, s_2, \ldots, s_k)$ if $k \leq m$ and for each s_i $(1 \leq i \leq k)$ there is a d_{j_i} $(1 \leq j_i \leq m)$ with $s_i = d_{j_i}$ and $j_{i-1} < j_i$ $(i > 1)$. We denote this with $s \prec d$. The sequence s is called a *sub-sequence* of d. Commonly a sequence $d \in D$ is said to support a *pattern* s if the pattern is a sub-sequence in the sequence d:

$$supp(s, d) = \begin{cases} 1 & \text{if } s \prec d; \\ 0 & \text{otherwise,} \end{cases}$$

We then can define the *sequence_support* of pattern s in database D as

$$sequence\_support(s, D) = \sum_{d \in D} supp(s, d)$$

We call s a *frequent subsequence* if its sequence support is larger than or equal to a user-defined threshold *minsupp*. This defines *sequential patterns* on sequences of items. Another definition of

sequential patterns was given by Agrawal et al. in [1], in which they define sequential patterns on sequences of item sets.

The most common task addressed in sequence mining research is the identification of frequent patterns. PrefixSPan as described in [8] searches for those patterns with *support* larger than or equal to a given support threshold *minsupp*. The algorithm starts with all frequent sub-sequences of size one. For each sub-sequence a projected database is created. The *projected database* is a database of pointers to the first item occurring after the current pattern, also called the *prefix*. A sequence is only in the projected database if it contains the prefix. The set of frequent sub-sequences is extended with all frequent sub-sequences of size two by only looking in the projected database. Again for each frequent sub-sequence of size two a corresponding projected database is created. This process continues recursively until no extension is frequent anymore.

More recently, other tasks have been addressed, such as classification of sequences. A few methods have been proposed to search for maximally discriminating patterns, such as PrefixTWEAC [4] and correlated association rule mining [15]. The main parameter of PrefixTWEAC, used in this work, is the number of k maximal discriminating patterns (i.e., the k most different patterns). This k can be used to prune, even though the difference measure doesn't have the anti-monotone property. This technique is described in more detail in [7].

2.3 Sequence Mining to Mine Web Access Log Data

Sequence mining algorithms are commonly used to analyze access log data [5, 11]. However, experimental evaluation in most papers is done from a technical perspective (e.g., computational efficiency). For understandable reasons, there are very few examples where the results of a sequence mining algorithm are evaluated from the point-of-view of the application. A notable exception is the work of Spiliopoulou and Pohle [11]. They apply a sequence mining algorithm to the logs of a website containing a database of German schools. The structure of the site is changed based on the patterns found and the impact of those changes is quantified using suitable metrics (contact and conversion efficiency).

The questions addressed by these authors are similar to the ones that are investigated in this work. Besides identifying frequent patterns in navigation sequences, the navigation patterns of customers and non-customers patterns are also compared. A heuristic in two phases is proposed: first frequent patterns are identified in customer sessions. Then, these patterns are applied to non-customer sessions to find interesting differences.

3 Problem Definition

PortalExecutivo.com (PE) is a Portuguese web portal targeted to business executives. The business model of PE is subscription-based, which means that only paying users have full access to content through web login. However, users can freely browse the site's structure and glimpse the kind of content which is available. Although some of the content is provided by PE, the majority originates from a large number of partners, which include several publications (The Economist, Wired, etc.) and companies (Accenture, Portugal Telecom, McKinsey, etc.). Value is added not only by concentrating content in a single access point but also by structuring and interrelating content items in a way that makes it easier for the user to find relevant information. This is achieved by providing several mechanisms (e.g., a search engine) that are based on quite a rich set of meta-data fields that are associated with each item, including keywords, categories, source and authors.

The goal of PE is to facilitate the access of its members to relevant content. To pursue this objective, it is essential to obtain as much information about the profile and the behavior of its users as possible. A number of different questions arise, including the two which we address in this work:

- Which are the paths most frequently followed by the users?
- Are there differences in the navigation patterns of guest and registered users?

Answers to the first question can be used, for instance, to identify problems in the structure that cause inefficient navigation. Additionally, those patterns can be used to identify content which a

users	3,237
sessions	71,547
accesses	1,784,642
content items	18,959
content types	651

Table 1: Basic statistics for the access log data.

Field	1st click	2nd click	3rd click	...
Title	home page	Analysis of National Budget	The Xmas trap	...
Type	navigation	article	news	...
Source	—	PE	clipping service	...

Table 2: Alternative representations for a given session using different fields from the meta-data describing content items.

user is expected to be interested in and, thus, provide recommendations. As for the second question, differences between guest and registered users could be used, for instance, to adapt the site to the behavior of the former, with the goal of convincing them to register.

Commonly, web usage mining is carried out with data from web access logs generated by the web server. However, the pre-processing of access logs required to obtain data which is ready for mining is not a trivial task [12]. Problems for which there is really no definite answer include the identification of sessions (i.e., sequences of clicks that constitute one session of the user on the site) and the identification of the user within and across sessions. However, in this case study we are able to avoid these issues by the use of access data which is stored by the content management system of PE in a table of a relational database. There are 6 fields in this table: content id(entifier), session id, user id, time stamp, title and type of content item. Our analysis has focused on two of these, namely the id, which uniquely identifies content items, and the type, which aggregates items into groups, such as articles, news items, homepage, etc. This table can be regarded as a clean version of the access logs. We have used data from the period between May 2002 and November 2005, for which a number of statistics is given in Table 1.

4 Mining Different Representations of Sequences

In a typical application of sequence mining to web access logs, each sequence is a session represented as a set of content item ids. Given the large number of different content items (18,959), not many sequences are expected to achieve the minimum support. In spite of this, it may be expected that in such a portal, some content items are very popular (e.g., an interview with Jack Welch or an analysis of the national budget) and, thus, patterns relative to these items could exist.

As mentioned earlier, content items in PE are described not only by the id (or title) but also by a set of descriptors (i.e., meta-data) which include author, category, type and keywords. These descriptors can provide an alternative representation of content items, as shown in Table 2. These alternative representations can be exploited for sequence mining. Different patterns will be obtained that provide different perspectives on the navigation patterns of the users. For instance, with the original representation it is possible to obtain patterns such as: "users that access the analysis of the national budget for 2004, followed by an interview to the Minister of Finance". By representing content items using their type or source, patterns such as "users that read news, followed by articles" and "users that read content from the clipping service, followed by one from Wired", respectively, can be obtained. Alternatively, given that meta-data descriptors usually provide a higher-level representation of the content, the domain is usually smaller (e.g., 651 different content types), which means that more patterns are expected to have higher support and, thus, more knowledge may be obtained. On the other hand, this approach is more useful for characterization of the navigation patterns than for recommendation of content. For instance, given the first pattern above, if a user accesses the analysis of the national budget for 2004, the portal may recommend the interview with the Minister of Finance to this user. Using the second pattern for this purpose is more difficult because, if a user reads a news item, then it is necessary to choose which, from the

many articles available, to recommend.

4.1 Results of Frequent Itemset Mining

To address the first question, we have used our implementation of the PrefixSpan algorithm for frequent sequence mining [8]. All patterns with a minimum support of 100 were analyzed.

The results confirm our expectations. We obtained approximately 5,000 patterns and very few were interesting. Most of them were related to navigation pages (i.e., non-leaf nodes in the site structure). Our hypothesis that some content items could be sufficiently popular to appear in sequences was confirmed in very few cases only (e.g., an analysis of the national budget for 2004, following the homepage of the portal).

Next, we applied PrefixSpan to the sequences in which the items were represented as the type of content (2nd line in Table 2). As expected, many more interesting patterns were obtained. For instance, we observed that in 1,261 sessions, the user browses through several articles and then prints one of them. Additionally, in 859 sessions, the user accesses the stock exchange quotes, followed by news. Although further investigation is required to investigate these patterns, they can trigger changes in the way PE provides its service to these users. The first pattern may indicate that these users search for content to read while returning home. If this is the case, then these are clearly users that could benefit from a personal recommendation mechanism in the portal triggered at the most suitable time (i.e., when the user usually returns home). The second pattern suggests that it could be a good idea to have a page where both quotes and recent news items (particularly, financial news) could be provided to these users.

4.2 Results of Sequence Difference Mining

Concerning the second question, with the goal of understanding the difference between browsing behaviour of registered and unregistered users (i.e., guests), we have addressed it with the Prefix-TWEAC algorithm [4]. We have set the time window parameter to 10 and we have analyzed the 100 *maximal discriminating patterns* obtained with a minimum support of 100. These patterns *do* occur for one group of users, but *do not* occur for another group.

As in the earlier set of experiments, we started by representing items by their ids. Again, the results were not outspoken. First of all they are not very discriminating: the maximally discriminating pattern matches 171 sessions of registered users and 298 guest sessions. Additionally, the patterns obtained are not very interesting from an application point-of-view.

Using the alternative representation, where the algorithm is applied to sequences of types of content items, more interesting results were obtained. For instance, as could be expected, registered users have longer threads of articles, news items and other functionalities than guests. A more interesting observation, from a business perspective, is that guest users access the areas of the partners less often than registered users. Given that this is one of the main features of the portal, we would like guest users to be more exposed to it. Therefore, the partners should be made more visible to guest users, for instance, by highlighting the most recent contributions or even by making selected content available for free.

5 Discussion

We start by discussing the representation of the sequences. Next we identify a few problems which could lead to variants of sequence mining algorithms.

5.1 Representation of Items

In general, our results show that for descriptive applications, in which the goal is to understand the behavior of the users, more interesting knowledge can be obtained by representing the items in sequences using more abstract representations, when these are available. In our case, we have used meta-data containing alternative descriptions of the items. However, this approach may not be suitable if the goal is to provide recommendations. That is, given a frequent sequence (`A`,`B`,`C`)

and a user who browsed (A,B), suggest C. In this case, the most common approach, in which items are represented by their ids, may be better.

Additionally, in the current definition of sequential patterns it is possible to have an unlimited amount of items, in the sequence in the database, between two items from the pattern. So if the pattern is (A,B) then the sequence in the database containing the pattern could be (A,C,C,C,C,C,B). This means that the patterns that are obtained hide some information. This may be an advantage, if the loss consists of irrelevant details. For instance, one of the patterns observed was that many users access the stock exchange quotes, followed by news. This is an interesting pattern, independent of how much browsing the user does in between these two types of content. On the other hand, the missing information may be important for the interpretation of the pattern. For instance, in the other pattern mentioned above, concerning users that browse through several articles and then print one item, it may actually mean that the item that is printed is not an article. This may be addressed by looking for patterns consisting of adjacent items or by imposing a constraint on the size of the gap or the size of the pattern [2, 6]. Alternatively, this issue could be addressed by extending the representation of the rules with some statistics describing the distribution of the items in the gaps. For instance, given the pattern (A,B,C) with matching sequences:

(A,C,B,C)
(A,C,D,B,D,C)
(A,B,C)

the pattern could be complemented with histograms describing the size of the gaps in the sequences. In this case, the histogram for the subsequence of the pattern (A,B) would be 1/0 (meaning 1 sequence with gap size 0), 1/1 and 1/2, and for the subsequence (B,C), 2/0 and 1/1. Additionally, information about the items in the gaps would also be useful. In this case, concerning subsequence (A,B), the gaps in the sequences contain C twice immediately after A and both C and D appear once each immediately before B. Although this kind of information could be collected in the sequence mining process, given the large amount of different statistics that may be interesting, it would better be implemented as a tool for post-processing sequential patterns. A related approach has been proposed earlier, in which patterns are represented as aggregate trees [11].

The use of an alternative representation for items based on meta-data introduces another issue. Many redundant patterns of the form (A,B), (A,A,B), (A,B,B), (A,A,B,B), etc. were obtained. Again, results could be simpler to understand by post-processing the patterns. For instance, these patterns could be represented by a single regular pattern (A{1,2},B{1,2}).

5.2 Suggestions for Sequence Mining Research

We have identified four alternative sequence mining tasks which could be useful for the analysis of access log data but are possibly also applicable in other domains.

Mining frequent repetitive subsequences The goal is to find subsequences that repeatedly appear within sequences. An an example the subsequence (A,B,C) is a frequent repetitive pattern in the following data:

sequence 1: ... A,B,C ... {e_1 elements} ... A,B,C ... {e_2 elements} ... A,B,C ...
sequence 2: ... A,B,C ... {e_3 elements} ... A,B,C ...
...

One application in access log mining would be to find repetitive behaviour of users that may, for instance, indicate that different organization of the content or a recommendation system could decrease the number of clicks required to reach necessary information. This method would require two new parameters: minimum number of repetitions and range of valid e_i values.

Mining frequent periodic subsequences The goal is the same as the previous one but imposing constraints on the variation of the values of e_i for each pattern. For all e_i of all sequences matching a pattern, $max(e_i) - min(e_i)$ must be smaller than a given threshold, which is a parameter to the algorithm. Alternatively, the function assessing the quality of patterns, which usually takes only the support into account, could also include a term that reflects the amount of variation of the e_i values.

Mining frequent subsequences of elements represented as sets In this case, the goal is to find subsequences of elements which are described using sets of properties [1]. As an example, ($\mathtt{A}_1$,$\mathtt{C}_1$) is a frequent pattern in the following data, in which each item is represented by a set of fields $\langle \mathtt{XY}\ldots \rangle$:

sequence 1: $\ldots \langle \mathtt{A}_1\mathtt{B}_1 \rangle, \langle \mathtt{A}_2\mathtt{C}_1 \rangle \ldots$
sequence 1: $\ldots \langle \mathtt{A}_1\mathtt{B}_2 \rangle, \langle \mathtt{B}_3\mathtt{C}_1 \rangle \ldots$
...

Its application in access log mining could be to identify navigation patterns combining different representations for content items (e.g., content id, category, keyword, author).

Mining frequent subsequences of structured elements The goal is the same as in the previous task but the patterns are built taking into account constraints on the values of sets. As an example, assume the constraint $C \leftarrow A$, which means that including in a pattern a value concerning the field `C` is possible only if the pattern already restrains the values of field `A` for the same item. Then, in the following data, ($\langle \mathtt{B}_1 \rangle \langle \mathtt{A}_2 \rangle$), ($\langle \mathtt{A}_1 \rangle \langle \mathtt{A}_2 \rangle$) and ($\langle \mathtt{A}_1\mathtt{C}_2 \rangle \langle \mathtt{A}_2 \rangle$) are frequent patterns but not ($\langle \mathtt{C}_2 \rangle \langle \mathtt{A}_2 \rangle$)

sequence 1: $\ldots \langle \mathtt{A}_1\mathtt{B}_1\mathtt{C}_1 \rangle \langle \mathtt{A}_2 \rangle \ldots$
sequence 2: $\ldots \langle \mathtt{A}_1\mathtt{B}_1\mathtt{C}_2 \rangle \langle \mathtt{A}_2 \rangle \ldots$
sequence 3: $\ldots \langle \mathtt{A}_1\mathtt{B}_2\mathtt{C}_2 \rangle \langle \mathtt{A}_2 \rangle \ldots$
...

In access log analysis, this could be used to identify time-references navigation patterns, e.g., if `A` is the category and `C` is the weekday, we would be able to obtain patterns such as "users who consult the stock exchange section on friday" but not patterns such as "users who access something on friday".

6 Conclusions

This paper describes the application of frequent sequence mining methods to web access log data. First, we exploit an alternative representation for the items that constitute the sequences. We compare the common approach, in which items are represented by their ids, with a higher-level representation, which indicates the type of content (e.g., navigation page, article, news item). Although we have focused on an alternative representation which is specific to the application at hand, similar approaches can be followed in other access log analysis problems, because many web sites are currently supported by content management systems which characterize content using this kind of meta-data.

Rather than proposing a new algorithm, we apply existing algorithms. However, we provide an analysis of the results from an application perspective. We observe that, if the goal is to understand user behavior, it is possible to obtain better patterns with the alternative representation. We also discuss the applicability of frequent sequence mining to web access log data in light of that analysis. We identify some advantages and shortcomings. Although this is not a novelty in the community, this application confirms that better post-processing methods are required.

Finally, based on this application, we identify variants of frequent mining tasks which can be useful. We expect that our observations will trigger research on methods to address those tasks. Although they are identified in the context of access log mining, we believe that they may be relevant in different sequence mining applications.

Acknowledgements

The first author thanks the financial support of Fundação Calouste Gulbenkian, POSC/EIA/-58367/2004/Site-o-Matic Project (Fundação Ciência e Tecnologia) co-financed by FEDER and the European Research Training Networks SegraVis (HPRN-CT-2002-00275). The second author is financed by the Netherlands Organization for Scientific Research (NWO) MISTA Project (grant no. 612.066.304).

References

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In *ICDE*, pages 3–14, 1995.

[2] C. Antunes and A. L. Oliveira. Generalization of pattern-growth methods for sequential pattern mining with gap constraints. In *MLDM*, pages 239–251, 2003.

[3] W. Baker, M. Marn, and C. Zawada. Price smarter on the net. *Harvard Business Review*, 79(2):122–127, 2001.

[4] E. de Graaf and W.A. Kosters. Efficient feature detection for sequence classification in a receptor database. In *BNAIC*, pages 81–88, 2005.

[5] C.I. Ezeife and Y. Lu. Mining web log sequential patterns with position coded pre-order linked wap-tree. *Data Min. Knowl. Discov.*, 10(1):5–38, 2005.

[6] M. Leleu, C. Rigotti, J.-F. Boulicaut, and G. Euvrard. Constraint-based mining of sequential patterns over datasets with consecutive repetitions. In *PKDD*, pages 303–314, 2003.

[7] S. Nijssen and J.N. Kok. Multi-class correlated pattern mining. In *KDID*, pages 165–187, 2005.

[8] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *ICDE*, pages 215–224, 2001.

[9] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos. Web usage mining as a tool for personalization: A survey. *User Model. User-Adapt. Interact.*, 13(4):311–372, 2003.

[10] F.F. Reichheld and P. Schefter. E-loyalty. *Harvard Business Review*, 78(4):105–113, 2000.

[11] M. Spiliopoulou and C. Pohle. Data mining for measuring and improving the success of web sites. *Data Min. Knowl. Discov.*, 5(1/2):85–114, 2001.

[12] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[13] W.M.P. van der Aalst, T. Weijters, and L. Maruster. Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.

[14] T. Weijters and W.M.P. van der Aalst. Process mining: Discovering workflow models from eventbased data. In *BNAIC*, pages 283–290, 2001.

[15] A. Zimmermann and L. De Raedt. Corclass: Correlated association rule mining for classification. In *Discovery Science*, pages 60–72, 2004.