WebUser: mining unexpected web usage

Haoyuan Li, Anne Laurent, Pascal Poncelet

WebUser: mining unexpected web usage

Dong (Haoyuan) Li*

LGI2P, École des Mines d'Alès,
Parc scientifique Georges Besse, 30035 Nîmes, France
E-mail: Haoyuan.Li@ema.fr
*Corresponding author

Anne Laurent and Pascal Poncelet

LIRMM, Université Montpellier 2,
161 rue Ada, 34392 Montpellier, France
E-mail: laurent@lirmm.fr
E-mail: poncelet@lirmm.fr

Abstract: Web usage mining has been much concentrated on the discovery of relevant user behaviours from Web access record data. In this paper, we present WebUser, an approach to discover unexpected usage in Web access log. We present a belief-driven method for extracting unexpected Web usage sequences, where the belief system consists of a temporal relation and semantics constrained sequence rules acquired with respect to prior knowledge. Our experiments show the effectiveness and usefulness of the proposed approach. Further, discovered rules of unexpected Web usage can be used for Web content personalisation and recommendation, site structure optimisation, and critical event prediction.

Keywords: data mining; web usage mining; log analysis; unexpected usage; sequence rules; concept hierarchies.

Reference to this paper should be made as follows: Li, D., Laurent, A. and Poncelet, P. (xxxx) 'WebUser: mining unexpected web usage', *Int. J. Business Intelligence and Data Mining*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Dong (Haoyuan) Li is a PhD candidate in Computer Science at the University of Montpellier 2, France. He received his MSc Degree in Computer Science at the University of Montpellier 2, France and his BE Degree in Mechanical Engineering at the Zhejiang University, China. His research interests include knowledge based data mining and its applications.

Anne Laurent received her PhD Degree in Computer Science at the University of Paris 6, France. She is an Assistant Professor at the University of Montpellier 2, France. As a member of the TATOO team in the Laboratory LIRMM, she works on data mining, sequential pattern mining, tree mining, both for trends and exceptions detections and is particularly interested in the study of the use of fuzzy logic to provide more valuable results, while remaining scalable.

Pascal Poncelet received his PhD Degree in Computer Science at the Nice Sophia Antipolis University, France. He is a Professor at the University of Montpellier 2, France and the head of the TATOO team in the Laboratory LIRMM. He was a Professor and the head of the data mining research group in the computer science department at the École des Mines d'Alès, France. His research interest can be summarised as advanced data analysis techniques for emerging applications.

1 Introduction

With the explosive growth of the World Wide Web, data mining techniques are more and more concentrated for the discovery of relevant user behaviours from web access log data. A great deal of research work has been performed on porting data mining techniques to web usage analysis, in order to improve the personalisation, recommendation, and even effectiveness of web sites (Büchner and Mulvenna, 1998; Eirinaki and Vazirgiannis, 2003; Ezeife and Lu, 2005; Huang et al., 2006; Masseglia et al., 2008; Missaoui et al., 2007; Mobasher et al., 2002; Mobasher, 2007; Spiliopoulou et al., 1999; Srivastava et al., 2000; Yen and Lee, 2006), by exploring the question: *what resources are frequently visited by whom during which periods?*

Among existing approaches, the methods for association rules (Agrawal et al., 1993) and sequential patterns (Agrawal and Srikant, 1995) mining have been well adapted for answering the above question. The sequential patterns extracted from web access logs are typically the relationships like "on a customer support forum site, 40% of users visit the *TopicList* page, then the *Search* page, then the *Login* page, and then the *PostTopic* page", or like "in an online store, 10% of customers visit the page of notebook cases after having added a notebook computer to the shopping cart". This kind of relationships reflects the most general and reasonable user behaviours during their navigations in the web, however, not limited to such frequent behaviours, the studies of unexpected usage become influential since they might reflect important unknown user behaviours. In fact, when we regularly perform statistical frequency based data mining approaches on web access logs, the redundancy of newly discovered behaviours increases with the growth of already discovered behaviours, and the decision makers will be more and more interested in exploring the behaviours that have been never discovered, in order to, for example, further develop website structures and improve user experiences.

To illustrate our motivation, let us consider an online news website. The titles of all latest news are listed on the static home page *index* and ordered by categories; the latest previous news can be visited from the static category index pages like *politics*, or *technology*, etc.; the dynamic page *read* provides the detail of a news with the method like *read?20080114002*. Assume that the following user behaviours exist in the access log of this website:

1 45% of users visit *index*, then various *read*, then *politics*, then various *read*, then *technology*, then various *read*, and then other categories and various *read*, etc.

2 10% of users visit *index*, then *politics*, then various *read*

3 1% of users visit *index*, then *technology*, then various *read*

4 0.05% of users only visit once *read*.

With sequential pattern mining algorithms, we can find the frequent sequences representing the behaviour (1) with a suitable minimum support threshold, but it is quite hard to find the sequences representing the behaviours (2), (3) and (4) because:

1 Most existing approaches for sequential pattern mining do not consider the missing elements, neither the semantic contradictions between elements (e.g., between *politics* and *technology*) in a sequence. The complex constraint based approaches like SPIRIT (Garofalakis et al., 1999) may find the sequences representing the behaviours (2) and (3) by specifying *politics* and *technology* as constraints, however the premise is that we must know the patterns like *politics* or *technology* before the extraction, and an important drawback is that we cannot find all sequences representing the behaviour (3) by saying 'the categories contradicting *politics*', since the constraint 'not *politics*' implies all categories different to *politics*.

2 According to the model of sequential patterns, the sequences representing the behaviours (2), (3) and (4) are contained in the sequences representing the behaviour (1). In fact, the existence of the behaviours (2), (3) and (4) can be detected by existing approaches, like mining closed sequential patterns (Yan et al., 2003) that distinguishes the support value of each frequent sequence instead of finding maximal frequent sequences, with comparing the support value of each frequent sequences, however it is also difficult to indicate which sequences they are for further studies.

The rest of this paper is organised as follows. In Section 2, we introduce the related work. In Section 3, we first list several preliminary concepts of sequence data mining, then we propose a formal definition of the user session sequences contained in web access logs, with which we further propose two categories of sequence rules of web usage, so-called the web usage rules. In Section 4, we first propose a belief-driven method for extracting unexpected web usage, then we extend the extraction process with semantic hierarchies of concepts. Section 5 evaluates our proposed approach with experimental case studies. Finally, we conclude in Section 6 with a short discussion and a list of perspectives.

2 Related work

There exist a great deal of web access log analysing tools, e.g., Webaliser (Barrett, 1997–2009), offer statistics based web access analysis. However, these tools are principally based on simple requests, for example, number of page views, number of hist, etc., for offering the information contained in log files. For resolving this problem, many approaches have been focused on using data mining techniques for extracting additional knowledge on the web usage, where the essential is to transfer the log files for applying the data mining algorithms (Kosala and Blockeel, 2000; Srivastava et al., 2000), and pattern (sequential pattern) mining is broadly used for discovering user navigation behaviours.

For instance, Spiliopoulou and Pohle (2001) developed WUM that measures the success of a site's components and obtain concrete indications of how the site should be improved by mining navigation patterns via generalised sequence of user sessions; Ezeife and Lu (2005) proposed a sequential pattern mining based approach to analyse web access log with the notion of web access pattern tree; Huang et al. (2006) developed a Navigational Pattern mining (NP-miner) algorithm for discovering frequent sequential patterns on the proposed Navigational Pattern Tree: according to historical patterns, the NP-miner scans relevant sub-trees of the Navigational Pattern Tree repeatedly for generating candidate recommendations; Dalamagas et al. (2007) provided a set of mining tasks for user navigation patterns and a set of personalisation tasks that customise the organisation of the topic directory according to these patterns for certain user groups; Masseglia et al. (2008) proposed a data mining process in order to automatically discover the densest periods where user behaviours were completely hidden on the log files and cannot be extracted by traditional approaches since they are frequent on particular periods rather than frequent on the whole log; Yen and Lee (2006) proposed an incremental algorithm for mining web access patterns, etc. All the above approaches are based on the principle of pattern or sequential pattern mining, where the statistical frequency is the primary measure for extracting and generating user navigation patterns and rules that represent user behaviours in web usage analysis.

Not difficult to see, most existing approaches to web usage mining are frequency based. Li et al. (2008) presented a belief driven approach to find unexpected usage from user session sequences in web access logs, which contradict prior knowledge on user navigation behaviours. In this paper, we extend this approach with concept hierarchies and propose the framework WebUser for mining unexpected web usage.

McGarry (2005) systematically investigated the interestingness measures for data mining, which are classified into two categories: the objective measures based on the statistical frequency or properties of discovered patterns, and the subjective measures based on the domain knowledge or the class of users.

Subjective measures were studied by Silberschatz and Tuzhilin (1995), in particular the unexpectedness and actionability. The term unexpectedness stands for the newly discovered patterns or sequences that are surprising to users. For example, if most of the customers who purchase action movies purchase pop music, then the customers who purchase action movies but purchase classical music are unexpected. The term actionability stands for reacting to the discovered patterns or sequences to user's advantage. For example, for the customers who purchase action movies without purchasing any kind of music, it is actionable to improve the promotion of pop music, even though it is unexpected. Therefore, in many cases, the unexpectedness and actionability exist at the same time, however, clearly, some actionable patterns or sequences can be expected and some unexpected patterns or sequences can also be non-actionable (Silberschatz and Tuzhilin, 1995). Two types of beliefs are further introduced, hard belief and soft belief, for addressing unexpectedness. According to the authors' proposition, the hard belief is a belief that cannot be changed by new evidences in data, and any contradiction of such a belief implies data error.

For example, in the web access log analysis, the error '404 Not Found' can be considered as a contradiction of a head belief: "the resources visited by users must be available"; however, the soft belief corresponds to the constraints on data that are measured by a degree, which can be modified with new evidences in data that contradict such a belief and interestingness of new evidences is measured by the change of the degree. For example, when more and more users visit the website at night, the degree of the belief 'users access the website at day time' will be changed. The computation of the degree can be handled by various methods, such as the Bayesian approach and the conditional probability.

Many unexpectedness based approaches have therefore been proposed. Liu and Hsu (1996) studied the unexpected structures of discovered rules based on pattern similarity based on the attribute name and value, which has been extended by Liu et al. (2001) to find unexpected information in the context of Web content mining. Suzuki and Shimura (1996), Suzuki (1997), Suzuki and Zytkow (2005) systematically studied exception rules in the context of association rule mining, where an association rule can be classified into two categories: a *common sense rule*, which is a description of a regularity for numerous objects, and an *exception rule*, which represents, for a relatively small number of objects, a different, regularity from a common sense rule. Dong and Li (1998) roposed neighbourhood-based interestingness in association rules, which is based on the distance between rules and the neighbourhoods of rules. Padmanabhan and Tuzhilin (1998, 2000, 2006) proposed a semantics-based belief-driven approach to discover unexpected patterns in the context of association rules, where the unexpectedness is determined from domain-experts-defined logically contradiction of patterns. Wang et al. (2003) studied unexpected association rules with respect to the value of attributes. Jaroszewicz and Scheffer (2005) proposed a Bayesian network based approach to discover unexpected patterns, that is, to find the patterns with the strongest discrepancies between the network and the database.

In the context of web mining, Spiliopoulou (1999) presented a belief-driven approach to find unexpected sequence rules based on the notion of generalised sequences. The sequence rule is built by dividing a sequence into two adjacent parts, which are determined by the support, confidence and improvement. A belief on sequences is constrained by the frequency of the two parts of a rule, so that if a sequence respects a sequence rule but the frequency constraints are broken, then this sequence is unexpected. Although that work considers the unexpected sequences and rules, it is however very different to our problem in the measure and the notion of unexpectedness contained in data.

3 Sequence rules on web usage

3.1 Preliminary concepts

We discuss the extraction of unexpected web usage within the context of sequence data mining, which is first stated with the problem of sequence data mining (Agrawal and Srikant, 1995; Dong and Pei, 2007).

Given a set $R = \{i_1, i_2, \ldots, i_n\}$ of a limited number n of distinct binary-valued attributes, an attribute is an *item*. An *itemset* is an unordered collection $I = (i_1, i_2, \ldots, i_m)$ of items. A *sequence* is an ordered list $s = \langle I_1 I_2 \ldots I_k \rangle$ of itemsets. A *sequence database* is generally a large set D of sequences. Given two sequences $s = \langle I_1 I_2 \ldots I_m \rangle$ and $s' = \langle I'_1 I'_2 \ldots I'_n \rangle$, if there exist integers

$1 \leq i_1 < i_2 < \cdots < i_m \leq n$ such that $I_1 \subseteq I'_{i_1}, I_2 \subseteq I'_{i_2}, \ldots, I_m \subseteq I'_{i_m}$, then we say that the sequence s is a *subsequence* of the sequence s', denoted as $s \sqsubseteq s'$. If $s \sqsubseteq s'$, then we say that *s is included in* the sequence s', or s' *supports* s. If a sequence s is not included in any other sequences, then s is *maximal*. The *length* of a sequence s is the number of itemsets contained in the sequence, denoted as $|s|$. An *empty sequence* is denoted as $\emptyset$, where $|\emptyset| = 0$. The *concatenation* of sequences is denoted as the form $s_1 \cdot s_2$, so that we have $|s_1 \cdot s_2| = |s_1| + |s_2|$.

Example 1: The sequence $s_1 = \langle (a)(b) \rangle$ is included in the sequence $s_2 = \langle (a)(d)(b,c) \rangle$ since $(a) \subseteq (a)$ and $(b) \subseteq (b,c)$. However, s_1 is not included in the sequence $s_3 = \langle (a,b)(d) \rangle$.

Given a sequence database D, the *support* of a sequence s in D, denoted as $\sigma(s, D)$, is the number of the sequences $s' \in D$ such that $s \sqsubseteq s'$. Given a minimal frequency threshold, denoted as σ_{min}, a sequence s is *frequent* if $\sigma(s, D) \geq \sigma_{min}$. A *sequential pattern* is a maximal frequent sequence.

3.2 Session sequences

We consider the web access log in the NCSA Common Logfile Format (NCSA HTTPd Development Team, 1995), which is supported by most mainstream web servers. The Common Logfile Format (CLF) is defined as follows:

```
remotehost"rfc931""authuser"[date]"request""status"bytes.
```

A web access log file is generally an ASCII text file, each line contains a CLF log entry that represents a request from a remote client machine to the web server.

According to the concepts of item, itemset, and sequence, we propose the notion of *session sequence* for representing the user session contained in web access log entries. Notice that we only consider the *remotehost*, *date*, and *request* fields in our approach for the general-purpose of protecting user privacy.

Definition 1: Let $\mathcal{L}$ be an ordered list of web access log entries and $\ell \in \mathcal{L}$ be a log entry consisting of the properties $\{ip, time, url, query\}$. A session sequence is a sequence

$$s = \langle (ip, S_0)(\ell_1.url, S_1) \cdots (\ell_n.url, S_n) \rangle,$$

such that:

1 for any two integers $1 \leq i, j \leq n$ and $i \neq j$, we have $\ell_i.ip = \ell_j.ip$ (denoted as ip)

2 for any two integers $1 \leq i < j \leq n$, we have $\ell_i.time < \ell_j.time$

3 for any two integers $1 \leq i < j \leq n$, we have $\ell_j.time - \ell_i.time \leq \mu_{max}$, where μ_{max} is the maximum idle time of a session.

S_0 is the global parameter set of the session sequence s. S_i $(1 \leq i \leq n)$ is the local parameter set of the log entry ℓ_i.

Given a session sequence s of n $(n > 0)$ log entries, the sequence can be represented as $s = \langle I_0 R_1 R_2 \ldots R_n \rangle$, where $I_0 = (ip, S_0)$ stands for the identification a session and $R_1 = (\ell_1.url, S_1), R_2 = (\ell_2.url, S_2), \ldots, R_n = (\ell_n.url, S_n)$ stand for the requests contained in session. *Notice that in $R_i = (\ell_i.url, S_i)$, the index i corresponds to the position of the log entry in the user session.* The global parameter set S_0 of the session sequence s can be empty or contain additional information that can be associated with this user session, such as *geographical region*, *time period*, *season* and even *weather*. The local parameter set S_i $(1 \leq i \leq n)$ can also be empty or contain additional information of the log entry ℓ_i, which is mainly considered as the HTTP query of the request.

Example 2: Let us consider the session sequence shown as follows:

$$\langle (\mathit{10.0.0.8}, \mathit{23h}, \mathit{fr})(\mathit{index.php})(\mathit{open.php}, p = \mathit{203}, g = \mathit{5}) \rangle .$$

This sequence represents a user session consisting of two access log entries. The remotehost field of this session is 10.0.0.8, the date field is translated to 23 h, and we known the remote host is located in France. The page index.php without HTTP query was first accessed, i.e., the request field is 'index.php'; the page open.php with HTTP query $p = 203$ and $g = 5$ was accessed later, which corresponds to the request field '*open.php*?$p = 203$&$g = 5$'.

With the formalisation of session sequences, we can apply association rule or sequential pattern mining algorithms for discovering the most general user behaviours of websites.

3.3 Web usage rules

In our proposed approach, the usage behaviours in web access data are represented as the *web usage rules*, which are sequence rules of the elements contained in session sequences. To be precise, we propose two categories of web usage rules, including *occurrence rules* and *class rules*, that correspond to different rule structures.

An occurrence rule of web usage reflects to the correlation between the occurrences of the requested resources in session sequences, which is defined as follows.

Definition 2: The occurrence rule of web usage is a rule in the form $s_\alpha \rightarrow^\tau s_\beta$, where $s_\alpha = \langle R_{\alpha_1} R_{\alpha_2} \ldots R_{\alpha_m} \rangle$, $s_\beta = \langle R_{\beta_1} R_{\beta_2} \ldots R_{\beta_m} \rangle$ are two subsequences contained in the user sessions in web access log such that $1 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_m < \beta_1 < \beta_2 < \cdots < \beta_m \leq n$, and $\tau = [min..max]$ $(min, max \in \mathbb{N}$ and $min \leq max)$ is a constraint on the intervals between s_α and s_β.

For any session sequence $s = \langle I_0 R_1 R_2 \ldots R_n \rangle$ that confirms the rule, there exist $s_\alpha = \langle R_{\alpha_1} R_{\alpha_2} \ldots R_{\alpha_m} \rangle$ and $s_\beta = \langle R_{\beta_1} R_{\beta_2} \ldots R_{\beta_m} \rangle$ such that $s_\alpha \cdot s_\beta \sqsubseteq s$ and $min < (\beta_1 - \alpha_m) < max$.

In an occurrence rule, if the resources $R_{\alpha_1}, R_{\alpha_2}, \ldots, R_{\alpha_m}$ are requested, then the resources $R_{\beta_1}, R_{\beta_2}, \ldots, R_{\beta_m}$ are also requested later in the same user session within the interval range $[min..max]$. If the interval range cannot be specified, we use a wild-card '$*$' for denoting the constraint τ, and we call such a rule a *simple* *occurrence rule*, denoted as $s_\alpha \rightarrow^* s_\beta$. According to the Definition 2, we have $\beta_1 - \alpha_m \geq 0$ for a simple occurrence rule.

The extraction of sequence rules with interval constraint is still a work-in-progress problem, while there exists various related research (Mannila et al., 1997; Das et al., 1998; Höppner and Klawonn, 2001; Harms and Deogun, 2004; Xing et al., 2008). Therefore, in this paper, we propose a simple frequent sequence rule mining algorithm, shown in Algorithm 1, for extracting simple occurrence rules only. Domain experts are necessary for building occurrence rules with the interval constraint τ from the website workflow (e.g., Cardoso and Lenic, 2006) or structure, and also from available simple occurrence rules.

In Algorithm 1, the concepts of closed sequential patterns are issued by Yan et al. (2003). Notice that (1) the global parameter sets of session sequences are eliminated in the mining process; (2) the aim of Algorithm 1 is not to propose a complete approach for mining sequence rules, thus other measures like *confidence* is not addressed.

Algorithm 1 Extracting simple occurrence rules

```
1. Perform closed sequential pattern mining with a minimum support threshold to the session sequence
   database $D$, let the result set be $D'$.
2. Find all maximum sequences $s_{max}$ in $D'$ and group each of them with all its subsequences contained
   in $D'$.
3. For each group:
   (a) Sort the closed sequential patterns by descended order of support values.
   (b) Compute the ratio of support values of each two neighbor closed sequential patterns, and find
       the minimal value that is superior to the given threshold.
   (c) Assuming the minimal value in the precedent step is obtained between closed sequential patterns
       $s_i$ and $s_{i+1}$, output $s_i$ as $s_\alpha$ and the rest from $s_{i+1}$ to the end of $s_{max}$ as $s_\beta$ of a simple occurrence
       rule $s_\alpha \rightarrow^* s_\beta$.
4. If more than one rule is extracted from the same group, keep the rule with minimal $s_\beta$.
```

Example 3: Let integers $1, 2, 3, \ldots$ be the items standing for the requested URLs and HTTP query parameters of web accesses. Assume that the closed sequential patterns $\langle(1)\rangle$, $\langle(1)(2)\rangle$, $\langle(1)(2)(1)\rangle$, and $\langle(1)(2)(1)(3)\rangle$ are frequent with the support values $0.7, 0.6, 0.4$, and 0.3 respectively. According to Algorithm 1, with a given threshold 0.5 of the ratio in the step 3-(b), we have the rule $\langle(1)(2)\rangle \rightarrow^* \langle(1)(3)\rangle$ extracted because the value $0.4/0.6$ is minimal.

In order to discover the correlations between frequently requested resources and user classes, we propose the class rule of web usage.

Definition 3: The class rule of web usage is a rule in the form $I_\alpha \Rightarrow s_\beta$ where I_α is a set of global parameters and $s_\beta = \langle R_{\beta_1} R_{\beta_2}, \ldots, R_{\beta_m} \rangle$ is a subsequence contained in the user sessions in web access log.

For any session sequence $s = \langle I_0 R_1 R_2 \ldots R_n \rangle$ that confirms the rule, we have $I_\alpha \subseteq I_0$, $s_\beta \sqsubseteq s$, and $1 \leq \beta_1 < \beta_2 < \cdots < \beta_m \leq n$.

In a class rule, the user class is indicated by the subset I_α of the global parameters of session sequences. The rule depicts that if a user session belongs to the class I_α, then the resources $R_{\beta_1}, R_{\beta_2}, \ldots, R_{\beta_m}$ are requested in the session. We propose Algorithm 2 for extracting the class rules of web usage. Notice that in our proposal, the specification of the global parameter set of session sequences is performed by domain experts in order to obtain the best relevance between user classes and web usage.

Algorithm 2 Extracting class rules

```
1. Perform general sequential pattern mining with a minimum support threshold to the session sequence database $D$.
2. For each sequential pattern $p$:
    (a) Group all session sequences in $D$ that supports $p$ to $D'$.
    (b) Perform frequent pattern mining with a minimum support threshold to $D'$.
    (c) For each frequent pattern $I$, compute the confidence with the fraction of its support value in $D'$ on its support value in $D$.
    (d) If the confidence value is superior to given threshold, output $I$ as $I_0$ and $p$ as $s_\beta$ of a class rule $I_0 \Rightarrow s_\beta$.
```

The occurrence rules and class rules of web usage provide objective views of the behaviours recorded in web access log. With the occurrence rules, we are able to indicate the correlations between the most frequently visited contents; with the class rules, we can indicate what kind of users (from what location, in which period, etc.) access what resources frequently. This information is meaningful not only for web content personalisation and recommendation, but also for detecting abnormal user access events. In the next section, we further present the rest of our approach for mining unexpected web usage, where the rules proposed in this section are the base.

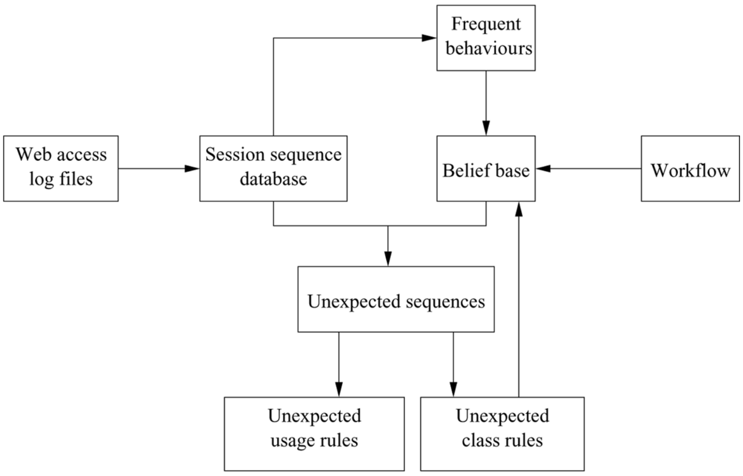
4 Mining unexpected web usage

In this section, we present our approach WebUser (Web Unexpected Sequence Rules), a belief-driven framework for mining unexpected web usage in session sequence databases, as shown in Figure 1.

4.1 Belief and unexpected web usage

To find unexpected usage in web access log, we propose the notion of *belief* on web usage, which consists of a web usage rule (i.e., occurrence rule or class rule) and a semantic constraint. Therefore, a session sequence is *unexpected* if this sequence contradicts a belief.

Figure 1 The WebUser framework



The formal definition of the belief based on the occurrence rule of web usage is proposed as follows.

Definition 4: A belief on web usage occurrence consists of an occurrence rule $s_\alpha \rightarrow^\tau s_\beta$ and a semantic constraint $s_\beta \not\simeq_{sem} s_\gamma$, denoted as $\{s_\alpha \rightarrow^\tau s_\beta\} \wedge \{s_\beta \not\simeq_{sem} s_\gamma\}$. Let $s_\alpha = \langle R_{\alpha_1} R_{\alpha_2} \dots R_{\alpha_m}\rangle$, $s_\beta = \langle R_{\beta_1} R_{\beta_2} \dots R_{\beta_m}\rangle$, and $\tau = [min..max]$ $(min, max \in \mathbb{N}$ and $min \le max)$, for a session sequence $s = \langle I_0 R_1 R_2 \dots R_n\rangle$ that confirms the belief, we have that $s_\alpha \sqsubseteq s$ implies $s_\alpha \cdot s_\beta \sqsubseteq s$ but $s_\alpha \cdot s_\gamma \not\sqsubseteq s$, where $1 \le \alpha_1 < \alpha_2 < \dots < \alpha_m < \beta_1 < \beta_2 < \dots < \beta_m \le n$ and $min < (\beta_1 - \alpha_m) < max$.

Given a belief $b = \{s_\alpha \rightarrow^\tau s_\beta\} \wedge \{s_\beta \not\simeq_{sem} s_\gamma\}$, a session sequence $s = \langle I_0 R_1 R_2 \dots R_n\rangle$ is unexpected if:

1 the interval constraint $\tau = *$ is violated, that is, we have $s_\alpha \sqsubseteq s$ however $s_\alpha \cdot s_\beta \not\sqsubseteq s$, and we call this case as α-*unexpected* since the factor of unexpectedness is s_α

2 the interval constraint $\tau = [min..max]$ is violated, that is, we have $s_\alpha \cdot s_\beta \sqsubseteq s$ however $(\beta_1 - \alpha_m) < min$ or $(\beta_1 - \alpha_m) > max$, and we call this case as β-*unexpected* since the factor of unexpectedness is s_β

3 the semantic constraint $s_\beta \not\simeq_{sem} s_\gamma$ is violated, that is, we have $s_\alpha \cdot s_\gamma \sqsubseteq s$ and $min < (\beta_1 - \alpha_m) < max$, and we call this case as γ-*unexpected* since the factor of unexpectedness is s_γ.

Example 4: Let us consider a website of online news. Assume that most users visit the website home page *index* and then no more than ten visits of other pages before visiting the politics news index page *politics*. This behaviour can be described as a occurrence rule

$$\langle(index)\rangle \rightarrow^{[0..10]} \langle(politics)\rangle\,.$$

If the technology news index page *technology* is not considered to be visited too early before the visit of *politics*, then we can specify the semantic constraint

$$\langle(politics)\rangle \not\simeq_{sem} \langle(technology)\rangle\ .$$

Hence, finally we have the following belief

$$\{\langle(index)\rangle \rightarrow^{[0..10]} \langle(politics)\rangle\} \wedge \{\langle(politics)\rangle \not\simeq_{sem} \langle(technology)\rangle\}$$

on an occurrence rule of web usage. In this instance, *politics* is visited too late or *technology* is visited too early respectively corresponds to β-unexpected or γ-unexpected. Further studies of the unexpected session sequences stated by this belief, for example finding corresponded time periods is able to provide useful reference of content personalisation.

Algorithm 3 briefly outlines the identification process of unexpected web usage based on the occurrence rules. Given a session sequence database D, the algorithm first identifies unexpected session sequences and regroup them as D' for each belief. Then the algorithm finds sequential patterns in order to describe the associations between the frequent requested resources and unexpectedness. The algorithm also extracts frequent patterns I_φ from the global parameter sets of all unexpected sequences for generating unexpected rules on user classes for studying the correlations between user classes and unexpectedness, which can be further measured by computing the confidence value.

Algorithm 3 Extracting unexpected Web usage with a belief on occurrence rule in a session sequence database D

```
1. For each session sequence $s \in D$:
    (a) Match $s_\alpha$ in $s$ and record the index $\alpha_m$.
    (b) If $\tau = *$, match $s_\beta$ from $\alpha_m + 1$. If not matched then identify $s$ as $\alpha$-unexpected and quit the routine.
    (c) Match $s_\beta$ in $[(\alpha_m + 1)..(\alpha_m + min)]$ (if $min = 0$, skip) and $[(\alpha_m + 1 + max)..end]$. If matched then identify $s$ as $\beta$-unexpected.
    (d) Match $s_\gamma$ in $[(\alpha_m + 1 + min)..(\alpha_m + 1 + max)]$. If matched then identify $s$ as $\gamma$-unexpected.
2. For each unexpectedness, let identified unexpected session sequences be the set $D'$.
3. Find and output sequential patterns in $D'$.
4. Find and output frequent patterns $I_\varphi$ in the set of all $I_0$ in each session sequence in $D'$.
```

In our proposed approach, we compute the confidence of such an *unexpected usage rule* '$I_\varphi \Rightarrow unexpectedness$' by the fraction of the number of sequences in D' that contain I_φ on the number of sequences in D that contain I_φ.

Unexpectedness can be also stated by the class rules of web usage, which is defined as follows.

Definition 5: A belief on web usage class consists of a class rule $I_\alpha \Rightarrow s_\beta$ and a semantic constraint $s_\beta \not\simeq_{sem} s_\gamma$, denoted as $\{I_\alpha \Rightarrow s_\beta\} \wedge \{s_\beta \not\simeq_{sem} s_\gamma\}$. Let $s_\beta = \langle R_{\beta_1} R_{\beta_2} \ldots R_{\beta_m} \rangle$, for a session sequence $s = \langle I_0 R_1 R_2 \ldots R_n \rangle$ that confirms the belief, we have that $I_\alpha \subseteq S_0$ implies $s_\beta \sqsubseteq s$ but $s_\gamma \not\sqsubseteq s$.

Given a belief $\{I_\alpha \Rightarrow s_\beta\} \wedge \{s_\beta \not\simeq_{sem} s_\gamma\}$ a session sequence $s = \langle I_0 R_1 R_2 \ldots R_n \rangle$ is unexpected if and only if the semantic constraint $s_\beta \not\simeq_{sem} s_\gamma$ is violated, that is, we have $I_\alpha \subseteq S_0$ and $s_\gamma \not\sqsubseteq s$. Thus, a belief on the class rule of web usage only states γ-unexpected sequences.

Example 5: Considering again last example, we already knew that the politics news politics is semantically different to the technology news technology. If from 08 h to 23 h, 60% of users visit the news listed on index then those listed on politics, then by exploring the users who visit technology instead of politics (a belief $\{(daytime) \Rightarrow \langle (index)(politics) \rangle\} \wedge \{\langle (index)(politics) \rangle \not\simeq_{sem} \langle (index)(technology) \rangle\}$) may find that the frequent period of the unexpected usage is 23 h to 08 h of the second day, for example 80%. This information can therefore be used for indicating the periods during which the usage of site is different. In fact, a class rule like '$(night) \Rightarrow \langle (science) \rangle$' can be further discovered if the visits of science news (*science*) are frequently associated with the visits of technology in the night.

Algorithm 4 shows the extraction of the unexpected web usage based on the class rules.

Algorithm 4 Extracting unexpected Web usage with a belief on class rule in a session sequence database D

```
1. For each session sequence $s \in D$:
    (a) If $I_0 \in s$ does not contain $I_\alpha$, quit the routine.
    (b) Match $s_\gamma$ in $s$. If matched then identify $s$ as $\gamma$-unexpected, otherwise quit the routine.
2. Let identified unexpected session sequences be the set $D'$.
3. Find and output sequential patterns $s_\varphi$ (with eliminating $s_\gamma$) in $D'$.
4. Find and output frequent patterns $I_\varphi$ in the set of all $I_0 \setminus I_\alpha$ in each session sequence in $D'$.
```

Given a session sequence database D, similar to Algorithm 3, the algorithm first identifies unexpected session sequences and regroup them as D' for each belief. Then the algorithm finds sequential patterns s_φ in D' with eliminating the subsequence s_γ in each sequence, and finds frequent patterns I_φ from the global parameter sets of all unexpected sequences with eliminating in elements in the set I_α (i.e., $I_\varphi \cap I_\alpha = \emptyset$). The *unexpected class rule* '$I_\varphi \Rightarrow s_\varphi$' can be generated for depicting the behaviours associated with the unexpected web usage.

Unexpected class rules are *unobservable*, since we cannot find such rules only except I_φ and s_φ is enough frequent and occur together. In fact, in our proposed approach, although the pattern I_φ and the sequence s_φ are not required to be contained in same sequences, such rules permit positioning new strategies for website promotions. Related problems of discovering unobservable events are issued by Ohsawa and McBurney (2003).

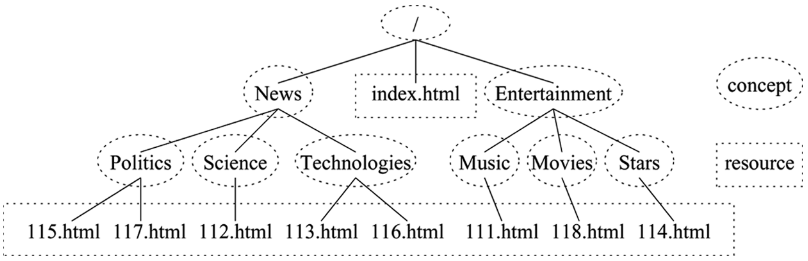
4.2 Unexpected web usage with concept hierarchies

In precedent sections, we propose a belief-driven approach for extracting unexpected web usage. Examples 4 and 5 shows the usefulness of our proposal. However, a drawback is that it is difficult to specify all semantic constraints for each belief by domain experts when the belief base is large, and it is also difficult to specify the web usage rules (in stead of extracting such rules) from workflow or site structure since the number of resources can be large even within a same topic. For instance, in Example 4 we illustrate a rule on index pages of different news topics, but it is difficult to specify the visits of news content because each content of a news topic may have unique URL or query parameter in the request.

On the other hand, the extracted web usage rules often lack human-readbale interpretations in semantics, since the resources may have meaningless filenames or be accessed by HTTP query with only numbers. Moreover, the number of extracted rules might be extremely large for a website with high volume page views, however the visited topics are relatively much fewer.

To resolve these issues, we further propose the discovery of unexpected web usage with semantic hierarchies. In fact, the development of the semantic web and natural language processing techniques makes it possible to extract the topic of web content in semantics, which are usually represented as the hierarchies of concepts. Furthermore, the structure of a website is also organised as a hierarchy, and which can be easily integrated with a semantic hierarchy of concepts. For instance, Figure 2 shows a semantic hierarchy of concepts corresponding to website structure and the semantics of resources.

Figure 2 A semantic hierarchy of concepts of website structure and content semantics



Given a semantic hierarchy H of concepts, let C denote a *concept*. Considering a session sequence $s = \langle I_0 R_1 R_2 \dots R_n \rangle$ with a semantic hierarchy H on the resources present in the website, each request R_i $(1 \leq i \leq n)$ can be generalised to a concept $C_i \in H$, denoted as $C_i \prec R_i$. Given a *concept sequence* $S = \langle C_1 C_2 \dots C_n \rangle$ and a session sequence $s = \langle I_0 R_1 R_2 \dots R_n \rangle$, if for any R_i $(1 \leq i \leq n)$ we have $C_i \prec R_i$, then we say that the session sequence s supports the concept sequence S, denoted as $s \preceq S$.

With semantic hierarchies of concepts, we can extend the web usage rules to *concept web usage rules* as follows.

Definition 6: The concept occurrence rule of web usage with a semantic hierarchy H is a rule in the form $S_\alpha \to^\tau S_\beta$, where $S_\alpha = \langle C_{\alpha_1} C_{\alpha_2} \dots C_{\alpha_m} \rangle$, $S_\beta = \langle C_{\beta_1} C_{\beta_2} \dots C_{\beta_m} \rangle$ are two concept sequences on H such that $1 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_m < \beta_1 < \beta_2 < \cdots < \beta_m \leq n$, and $\tau = [min..max]$ ($min, max \in \mathbb{N}$ and $min \leq max$) is a constraint on the intervals between S_α and S_β.

For any session sequence $s = \langle I_0 R_1 R_2 \dots R_n \rangle$ that confirms the rule, there exist $s_\alpha = \langle R_{\alpha_1} R_{\alpha_2} \dots R_{\alpha_m} \rangle$ and $s_\beta = \langle R_{\beta_1} R_{\beta_2} \dots R_{\beta_m} \rangle$ such that $S_\alpha \preceq s_\alpha$, $S_\beta \preceq s_\beta$, $s_\alpha \cdot s_\beta \sqsubseteq s$, and $min < (\beta_1 - \alpha_m) < max$.

Definition 7: The concept class rule of web usage with a semantic hierarchy H is a rule in the form $I_\alpha \Rightarrow S_\beta$ where I_α is a set of global parameters and $S_\beta = \langle C_{\beta_1} C_{\beta_2} \dots C_{\beta_m} \rangle$ is a concept sequence on H.

For any session sequence $s = \langle I_0 R_1 R_2 \dots R_n \rangle$ that confirms the rule, we have $I_\alpha \subseteq I_0$ and there exists $s_\beta = \langle R_{\beta_1} R_{\beta_2} \dots R_{\beta_m} \rangle$ such that $s_\beta \sqsubseteq s$, $S_\beta \preceq s_\beta$, and $1 \leq \beta_1 < \beta_2 < \cdots < \beta_m \leq n$.

Example 6: Let us consider the semantic hierarchy shown in Figure 2. The rules $\langle (index)(115) \rangle \to^* \langle (114)(113) \rangle$ and $\langle (index)(117) \rangle \to^* \langle (114)(116) \rangle$ are different in sequence of requests, but in semantics they are the same one:

$$\langle (/)(politics) \rangle \to^* \langle (star)(technology) \rangle \, ;$$

the rules $(night) \Rightarrow \langle (113)(115) \rangle$ and $(night) \Rightarrow \langle (116)(117) \rangle$ are also the same one as:

$$(night) \Rightarrow \langle (technology)(politics) \rangle \, .$$

In our proposed method, we build a index of all resources in a website to maintain a resource-to-concept mapping for fast lookups. With this manner, an extra step is appended to Algorithms 1 and 2 to convert extracted rules to concept rules.

With semantic hierarchies of concepts, the specification of semantic constraint for determining the opposition in a belief defined in Definitions 4 and 5 is no longer obligated. Further, in order to make our approach flexible, we consider two criteria for measuring the semantic opposition between concepts: *concept distance* and *semantic relatedness*. Given a semantic hierarchy H, the concept distance between two concepts $C_i, C_j \in H$ is measured by the minimal path-length between them, denoted as $\delta(C_i, C_j, H)$ (we define that $\delta(C_i, C_j, H) = 1$ when $C_i = C_j$); the semantic relatedness between them is a *score* ($0 < score < 1$ if defined, $score = 1$ for unknown and $score = 2$ for self) specified by domain experts or computed from WordNet (Pedersen, 2008), denoted as $\lambda(C_i, C_j)$. For instance, in Figure 2, the distance between the concepts '*politics*' and '*technology*' is 2, between the concepts '*politics*' and '*music*' is 4.

We propose a formula for computing the semantic opposition between two nodes in a concept hierarchy H, denoted as $\omega_{sem}(C_i, C_j, H)$, as following:

$$\omega_{sem}(C_i, C_j, H) = \frac{2 - \lambda(C_i, C_j)}{\delta(C_i, C_j, H)} . \tag{1}$$

For two concepts, we have that the more distance the less importance for relatedness, and the less similarity the more contradiction. Table 1 lists the concept distance and semantic relatedness between several concepts shown in Figure 2. According to equation (1), we have the following Table 2 semantic opposition values between the concept 'politics' and others in the hierarchy shown in Figure 2 and the distance and relatedness listed in Table 1.

Table 1 Concept distance and semantic relatedness matrix ($distance : relatedness$)

	politics	*science*	*technology*	*music*	*movie*	*star*
politics	$1:2$	$2:0.3$	$2:0.2$	$4:0.1$	$4:0.4$	$4:0.5$
science	$2:0.3$	$1:2$	$2:0.9$	$4:1$	$4:1$	$4:1$
technology	$2:0.2$	$2:0.9$	$1:2$	$4:1$	$4:1$	$4:1$
music	$4:0.1$	$4:1$	$4:1$	$1:2$	$2:0.6$	$2:0.7$
movie	$4:0.4$	$4:1$	$4:1$	$2:0.6$	$1:2$	$2:0.8$
star	$4:0.5$	$4:1$	$4:1$	$2:0.7$	$2:0.8$	$1:2$

Table 2 Semantic opposition between concepts

C_i	C_j	δ	λ	ω_{sem}
politics	politics	2	2	0
politics	science	2	0.3	0.85
politics	technology	2	0.2	0.9
politics	music	4	0.1	0.475
politics	movie	4	0.4	0.4
politics	star	4	0.5	0.375
politics	/	2	1	0.5

For concept sequences S and S', we propose the following formula for determining the semantic opposition between them with the average value of the ω_{sem} between all concepts contained in S and S', denoted as $\omega_{seq}(S, S', H)$:

$$\omega_{seq}(S, S', H) = \frac{\sum_{C_i \in S, C_j \in S'} \omega_{sem}(C_i, C_j, H)}{\|S\|}, \tag{2}$$

where $\|S\|$ stands for the total number of concepts contained in S.

The semantic opposition between two sequences of web access requests is computed by equation (2) with concept sequences. Once we are able to determine the semantic opposition with hierarchies, the beliefs proposed in Definitions 4 and 5 on web usage can be replaced by corresponded web usage rules and a semantic hierarchy of concepts.

Given a concept occurrence rule $S_\alpha \rightarrow^\tau S_\beta$ specified in Definition 6, a semantic hierarchy H of concepts, and user defined minimum semantic opposition threshold ω_{min}, a session sequence $s = \langle I_0 R_1 R_2 \ldots R_n \rangle$ is unexpected if:

1 for *α-unexpected* ($\tau = *$), there exists $s_\alpha \sqsubseteq s$ such that $S_\alpha \preceq s_\alpha$ however does not exist $s_\alpha \cdot s_\beta \sqsubseteq s$ such that $S_\beta \preceq s_\beta$

2 for *β-unexpected* ($\tau = [min..max]$), there exists $s_\alpha \cdot s_\beta \sqsubseteq s$ such that $S_\alpha \preceq s_\alpha$ and $S_\beta \preceq s_\beta$, however $(\beta_1 - \alpha_m) < min$ or $(\beta_1 - \alpha_m) > max$

3 for *γ-unexpected*, there exists $s_\alpha \cdot s_\gamma \sqsubseteq s$ such that $S_\alpha \preceq s_\alpha$, $S_\gamma \preceq s_\gamma$, $\omega(S_\beta, S_\gamma, H) \geq \omega_{min}$, and $min < (\beta_1 - \alpha_m) < max$.

Given a concept class rule $I_\alpha \Rightarrow S_\beta$ specified in Definition 7, a concept hierarchy H, and user defined minimum semantic opposition threshold ω_{min}, a session sequence $s = \langle I_0 R_1 R_2 \ldots R_n \rangle$ is *γ-unexpected* if there exists $s_\gamma \sqsubseteq s$ such that $S_\gamma \preceq s_\gamma$ and $\omega(S_\beta, S_\gamma, H) \geq \omega_{min}$.

In order to find unexpected web usage with a concept hierarchy, we extend the extraction processes listed in Algorithms 3 and 4 by two extra routines for (1) matching a concept sequence in a session sequence, which is listed in Algorithm 5; (2) finding semantic opposition of a concept sequence, which is listed in Algorithm 6.

Algorithm 5 Match a concept sequence S in a session sequence s

```
1. Initialize position counters $i$ and $j$.
2. While $i \leq endof(S)$ and $j \leq endof(s)$:
    (a) If $j = endof(s)$, then match failed and return.
    (b) If $i = endof(S)$, do backward match till to the nearest match of $C_1$ and return head and tail positions.
    (c) If $C_i \prec R_j$, then $i := i + 1$ and $j := j + 1$. Continue the loop.
    (d) If $C_i \not\prec R_j$, then $j := j + 1$. Continue the loop.
```

Algorithm 6 Find semantic opposition of a concept sequence S on hierarchy H in a session sequence s with respect to minimum semantic opposition threshold ω_{min}

```
1. Initialize position counters $i$ and $j$.
2. While $i \leq endof(S)$ and $j \leq endof(s)$:
    (a) If $j = endof(s)$, then match failed and return.
    (b) If $i = endof(S)$, do backward match till to the nearest match of $C_1$ and return head and tail positions.
    (c) Let the concept of $R_j$ in $H$ be $C_j$ i.e., $C_j \prec R_j$.
    (d) If $\omega_{sem}(C_i, C_j, H) \geq \omega_{min}$, then $i := i+1$, $j := j+1$, and append $C_j$ to $S'$. Continue the loop.
    (e) If $\omega_{sem}(C_i, C_j, H) < \omega_{min}$, then $S'' := S'$ and append $C_j$ to $S''$.
    (f) If $\omega_{seq}(S, S'', H) \geq \omega_{min}$, then $i := i + 1$, $j := j + 1$, and append $C_j$ to $S'$. Continue the loop.
    (g) Otherwise $j := j + 1$. Continue the loop.
```

Algorithms 5 and 6 are used for replacing the 'match sequence' statements in Algorithms 3 and 4 in order to enable the concept hierarchy. Notice that because of the complexity issue, Algorithm 6 finds the first subsequence of s having $\omega_{seq} \geq \omega_{min}$, instead of the subsequence with the highest semantic opposition value. To determine the semantics of concept combination is still an open problem (for instance, double positive is still positive, but double negative is also positive), however Algorithm 6 can still provide acceptable accuracy when $|S|$ is small.

The concept hierarchies of websites can be specified by web masters, however it is impractical to define hierarchies for large websites by hand. There is a very extended unsupervised way of building concept hierarchies with respect to website structure. For example, the approach proposed by Han and Fu (1994). Further, a web spider can follow the links from the index page to other sections, where the *title* tag of web pages can provide useful information for determining the concepts. Moreover, we also can build site hierarchy from web access log mining: the site structure can be directly extracted from the URLs of accessed resources, or can be generated by computing the dependency of HTTP query fields.

Example 7: Let us consider the following data contained in web access log entries.

```
192.168.1.10 - - [11/Jan/2009:17:40:00 +0100] "/viewtopic.php?f=2&t=198" 200 123

192.168.1.10 - - [11/Jan/2009:17:40:00 +0100] "/viewtopic.php?f=2&t=211" 200 234

192.168.1.11 - - [11/Jan/2009:17:40:27 +0100] "/viewtopic.php?f=3&t=212" 200 345

192.168.1.11 - - [11/Jan/2009:17:40:32 +0100] "/viewtopic.php?f=2&t=213" 200 567

192.168.1.10 - - [11/Jan/2009:17:49:21 +0100] "/viewtopic.php?f=3&t=220" 200 789
```

With discovering the dependency between the fields `f` and `t`, we can build a hierarchy where the node labelled 2 contains sub-nodes 198, 211, and 213; the node labelled 3 contains sub-nodes 212 and 220. It is not difficult to further replace the labels by the titles contained in web pages.

5 Experimental case studies

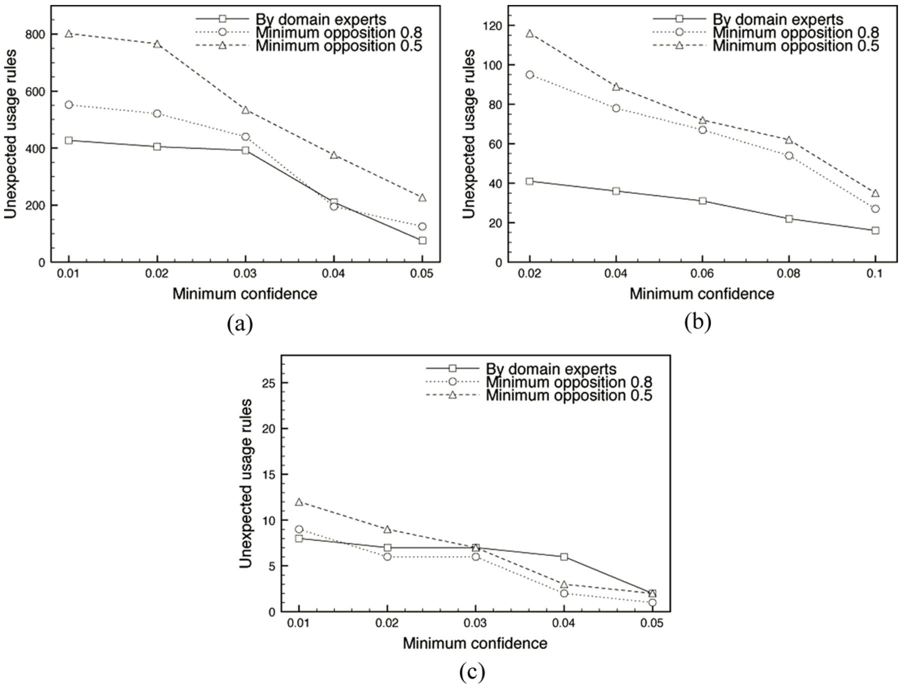
To evaluate of our approach, we performed a serial of experiments on three web access log files, including a very large log file of a BSD UNIX online discussion forum (labelled as BSD, 11 GB), a large log file of a customer support forum of an online game provider (labelled as BBS, 1 GB), and a small log file of a university library Web portal (labelled as LIB, 200 MB). All log files are converted to session sequence databases. The global parameter sets of session sequences are fixed to contain hour periods (from 00 h to 23 h), day periods (from Monday to Sunday). Due to privacy issues, user location information is not included in global parameter sets, other sensible information, such as session ID or login name in HTTP query fields, is also removed.

We generate web usage rules from for each data set, including ten occurrence rules and ten class rules from the most frequent closed sequential patterns, where one semantic constraint is specified for each rule. From each data set, we create one semantic hierarchy of concepts according to topics (BSD, BBS) or site structure (LIB), and then we generate ten concept occurrence rules and ten concept class rules from the most frequent sequential patterns with respect to semantic hierarchies. Therefore, for each data set, totally 20 beliefs with/without hierarchies are used for extracting unexpected web usage. For example, the following belief corresponds to an expected browsing order of the BBS data set, where $t = 2$ corresponds to the access of the discussion topic 'user terms' and $t = 5$ corresponds to 'user manual', such that the site designer wishes that users may read the agreement terms before reading the manual of the forum:

$$\{\langle(/)\rangle \rightarrow^{[0..5]} \langle(t=2)(t=5)\rangle\} \wedge \{\langle(t=2)(t=5)\rangle \not\simeq_{sem} \langle(t=5)(t=2)\rangle\}.$$

We first discover unexpected usage rules from ten beliefs on occurrence rules for each data set, the results are shown in Figure 3. In the experiments, we compared the number of unexpected usage rules extracted from domain experts specified beliefs and from hierarchies where minimum opposition is fixed to 0.8 and 0.5. In the data set BSD, the results are similar between domain experts specified beliefs and hierarchies-enabled beliefs with minimum opposition 0.8. In the data set BBS, the results are similar between hierarchies-enabled beliefs with minimum opposition 0.8 and 0.5. In the data set LIB, the results are similar between domain experts specified beliefs hierarchies-enabled beliefs with minimum opposition both of 0.8 and 0.5.

Figure 3 Unexpected usage rules discovered in: (a) BSD; (b) BBS and (c) LIB (see online version for colours)



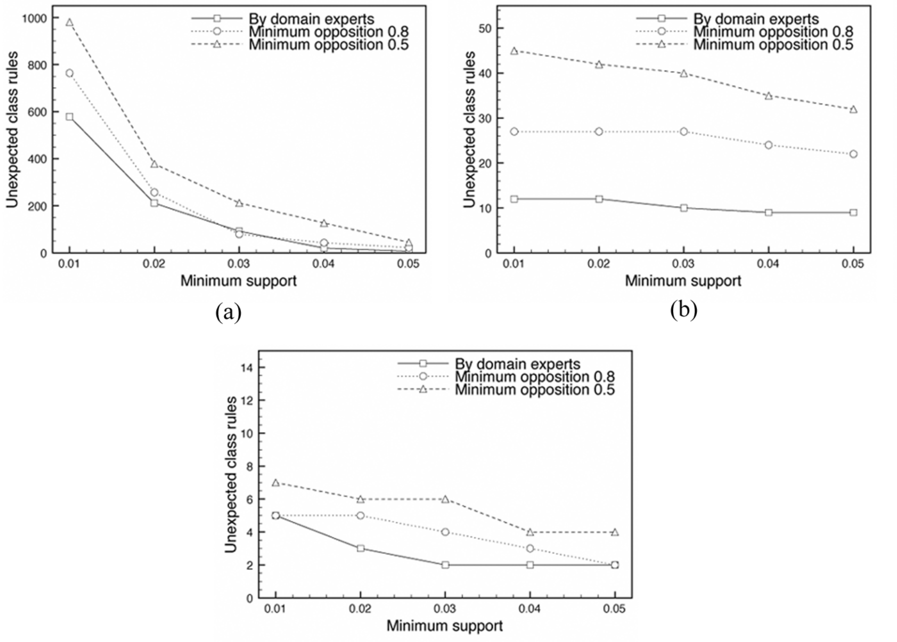
We also discover unexpected class rules from ten beliefs on occurrence rules for each data set, the results are shown in Figure 4. In order to not generate too much unexpected rules $I_\varphi \Rightarrow s_\varphi$, the minimum support for extracting s_φ is fixed to 0.5, which produces less sequential patterns. In the figures, the minimum support is used for extracting I_φ.

To illustrated discovered unexpected rules, for example, in the data set BBS, we wish that users of the forum 3 (discussions on an online game noted as G3) view several threads in this forum, i.e., $\langle(f=3)\rangle \rightarrow^* \langle(f=3)\rangle$, however we know from prior knowledge on playing games that the players of G3 may be not interested in the game discussed in the forum 6 (noted as G6), thus the semantic constraint $\langle(f = 3)\rangle \not\simeq_{sem} \langle(f = 6)\rangle$ can be added. With this belief, we discovered the frequent behaviours $(Sunday) \Rightarrow \gamma$-unexpectedness and γ-unexpectedness $\rightarrow \langle(f = 7)\rangle$, which can be further combined as an unexpected class rule $(Sunday) \Rightarrow \langle(f = 7)\rangle$, where forum 7 discusses a game noted as G7. Moreover, from expertise knowledge given by the game provider, we know that the players of G7 seldom play the game noted as G5, then the following belief can be generated from the discovered behaviours for further discoveries:

$$\{\langle(Sunday)(f = 3)\rangle \rightarrow^{*} \langle(f = 7)\rangle\} \wedge \{\langle(f = 7)\rangle \not\simeq_{sem} \langle(f = 5)\rangle\}.$$

The data sets BSD and LIB show good relevance of hierarchy-enabled beliefs, however the result in data set BBS is quite irrelevant. The main cause of this behaviour is that when we consider the semantics, we can determine the opposite semantics of one item, for example, the opposition between ‘beer’ and ‘Cola’, or between ‘login’ and ‘logout’. However, for operational conjunction of items, which depends on the operations, the semantics is of indeterminable. For instance, for temporal order between items, the determination of semantics is more harder than unordered collection. We might be able to indicate the semantic opposition between two itemsets, however we cannot exactly say that what is the opposition between sequences (it is possible to determining semantics within sequences in very strict cases, for example, natural language processing). The combination problem of semantics is still an open problem in semantics-driven data mining.

Figure 4 Unexpected class rules discovered in (a) BSD; (b) BBS and (c) LIB



(a)

(b)

(c)

6 Conclusion

In this paper, we present the approach WebUser for mining unexpected web usage in web access log with semantic hierarchies. We formalise the notion of session sequence and propose different sequence rules of session sequences for describing web usage. We specify the belief on web usage with such rules and semantic constraints, which can be replaced by semantic hierarchies of concepts. The unexpected web usage can therefore be extracted in the form of rules. The approach WebUser is evaluated with different types of web access log, that shows its effectiveness and usefulness.

Our perspectives include adopting our approaches to various application domains, such as text processing and classification, and so on. We are concentrate on studying the semantics of sequence data, we believe that is valuable to our future research direction. We plan to develop an approach for discovering sequence rules with interval constraints, instead of specifying occurrence rules by domain experts. We are also interested in generating concept hierarchies from web access log data, as described in the end of Section 4. Finally, we are developing a general purposed knowledge driven web usage mining framework in the context of sequence and sequential rule mining.

References

Agrawal, R., Imielinski, T. and Swami, A.N. (1993) 'Mining association rules between sets of items in large databases', *Proc. 1993 ACM SIGMOD International Conference on Management of Data*, Washington, DC, USA, pp.207–216.

Agrawal, R. and Srikant, R. (1995) 'Mining sequential patterns', *Proc. 11th International Conference on Data Engineering*, Taipei, Taiwan, pp.3–14.

Barrett, B.L. (1997–2009) Webalizer, *http://www.webalizer.org/*

Büchner, A.G. and Mulvenna, M.D. (1998) 'Discovering internet marketing intelligence through online analytical web usage mining', *SIGMOD Record*, Vol. 27, No. 4, pp.54–61.

Cardoso, J. and Lenic, M. (2006) 'Web process and workflow path mining using the multimethod approach', *International Journal of Business Intelligence and Data Mining*, Vol. 1, No. 3, pp.304–328.

Dalamagas, T., Bouros, P., Galanis, T., Eirinaki, M. and Sellis, T.K. (2007) 'Mining user navigation patterns for personalizing topic directories', *Proc. 9th ACM International Workshop on Web Information and Data Management*, Lisbon, Portugal, pp.81–88.

Das, G., Lin, K-I., Mannila, H., Renganathan, G. and Smyth, P. (1998) 'Rule discovery from time series', *Proc. 4th International Conference on Knowledge Discovery and Data Mining*, New York City, New York, USA, pp.16–22.

Dong, G. and Li, J. (1998) 'Interestingness of discovered association rules in terms of neighborhood-based unexpectedness', *Proc. 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining*, Melbourne, Australia, pp.72–86.

Dong, G. and Pei, J. (2007) *Sequence Data Mining (Advances in Database Systems)*, Springer, USA.

Eirinaki, M. and Vazirgiannis, M. (2003) 'Web mining for web personalization', *ACM Transactions on Internet Technology*, Vol. 3, No. 1, pp.1–27.

Ezeife, C.I. and Lu, Y. (2005) 'Mining web log sequential patterns with position coded pre-order linked WAP-Tree', *Data Mining and Knowledge Discovery*, Vol. 10, No. 1, pp.5–38.

Garofalakis, M.N., Rastogi, R. and Shim, K. (1999) 'SPIRIT: sequential pattern mining with regular expression constraints', *Proc. 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, UK, pp.223–234.

Han, J. and Fu, Y. (1994) 'Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases', *Proc. 12th National Conference on Artificial Intelligence*, Seattle, WA, USA, pp.157–168.

Han, J. and Kamber, M. (2006) *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann Publishers, San Francisco.

Harms, S.K. and Deogun, J.S. (2004) 'Sequential association rule mining with time lags', *Journal of Intelligent Information Systems*, Vol. 22, No. 1, pp.7–22.

Huang, Y-M., Kuo, Y-H., Chen, J-N. and Jeng, Y-L. (2006) 'NP-miner: a real-time recommendation algorithm by using web usage mining', *Knowledge Based Systems*, Vol. 19, No. 4, pp.272–286.

Höppner, F. and Klawonn, F. (2001) 'Finding informative rules in interval sequences', *Proc. 4th International Conference on Intelligent Data Analysis*, Cascais, Portugal, pp.123–132.

Jaroszewicz, S. and Scheffer, T. (2005) 'Fast discovery of unexpected patterns in data, relative to a bayesian network', *Proc. 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA, pp.118–127.

Kosala, R. and Blockeel, H. (2000) 'Web mining research: a survey', *SIGKDD Explorations*, Vol. 2, No. 1, pp.1–15.

Li, D.H., Laurent, A. and Poncelet, P. (2008) 'Mining unexpected web usage behaviors', *Proc. 8th Industrial Conference on Data Mining*, Leipzig, Germany, pp.283–297.

Liu, B. and Hsu, Y. (1996) 'Post-analysis of learned rules', *Proc. 13th National Conference on Artificial Intelligence and 8th Innovative Applications of Artificial Intelligence Conference*, Portland, Oregon, USA, pp.828–834.

Liu, B., Ma, Y. and Yu, P.S. (2001) 'Discovering unexpected information from your competitors' web sites', *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, pp.144–153.

Mannila, H., Toivonen, H. and Verkamo, A.I. (1997) 'Discovery of frequent episodes in event sequences', *Journal of Data Mining and Knowledge Discovery*, Vol. 1, No. 3, pp.259–289.

Masseglia, F., Poncelet, P., Teisseire, M. and Marascu, A. (2008) 'Web usage mining: extracting unexpected periods from web logs', *Data Mining and Knowledge Discovery*, Vol. 16, No. 1, pp.39–65.

McGarry, K. (2005) 'A survey of interestingness measures for knowledge discovery', *The Knowledge Engineering Review*, Vol. 20, No. 1, pp.39–61.

Missaoui, R., Valtchev, P., Djeraba, C. and Adda, M. (2007) 'Toward recommendation based on ontology-powered web-usage mining', *IEEE Internet Computing*, Vol. 11, No. 4, pp.45–52.

Mobasher, B. (2007) 'Data mining for web personalization', in Brusilovsky, P., Kobsa, A. and Nejdl, W. (Eds.): *The Adaptive Web: Methods and Strategies of Web Personalization*, Springer, pp.90–135.

Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. (2002) 'Using sequential and non-sequential patterns in predictive web usage mining tasks', *Proc. 2002 IEEE International Conference on Data Mining*, Maebashi City, Japan, pp.669–672.

NCSA HTTPd Development Team (1995) *NCSA HTTPd Online Document: TransferLog Directive*, http://hoohoo.ncsa.illinois.edu/docs/setup/httpd/TransferLog.html

Ohsawa, Y. and McBurney, P. (Eds). (2003) *Chance Discovery (Advanced Information Processing)*, Springer, Berlin.

Padmanabhan, B. and Tuzhilin, A.A. (1998) 'A belief-driven method for discovering unexpected patterns', *Proc. 4th International Conference on Knowledge Discovery and Data Mining*, New York City, New York, USA, pp.94–100.

Padmanabhan, B. and Tuzhilin, A.A. (2000) 'Padmanabhan-2000-Small', *Proc. 6th International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA, pp.54–63.

Padmanabhan, B. and Tuzhilin, A. (2006) 'On characterization and discovery of minimal unexpected patterns in rule discovery', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 2, pp.202–216.

Pedersen, T. (2008) WordNet::Similarity. *http://www.d.umn.edu/~tpederse/similarity.html*

Silberschatz, A. and Tuzhilin, A.A. (1995) 'On subjective measures of interestingness in knowledge discovery', *Proc. 1st International Conference on Knowledge Discovery and Data Mining*, Montreal, Canada, pp.275–281.

Spiliopoulou, M. (1999) 'Managing interesting rules in sequence mining', *Proc. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Lyon, France, pp.554–560.

Spiliopoulou, M. and Pohle, C. (2001) 'Data mining for measuring and improving the success of web sites', *Data Mining and Knowledge Discovery*, Vol. 5, Nos. 1–2, pp.85–114.

Spiliopoulou, M., Pohle, C. and Faulstich, L. (1999) 'Improving the effectiveness of a website with web usage mining', *Proc. Web Usage Analysis and User Profiling, International WEBKDD'99 Workshop*, San Diego, California, USA, pp.142–162.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P-N. (2000) 'Web usage mining: discovery and applications of usage patterns from web data', *SIGKDD Explorations*, Vol. 1, No. 2, pp.12–23.

Suzuki, E. and Shimura, M. (1996) 'Exceptional knowledge discovery in databases based on information theory', *Proc. 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, pp.275–278.

Suzuki, E. (1997) 'Autonomous discovery of reliable exception rules', *Proc. 3rd International Conference on Knowledge Discovery and Data Mining*, Newport Beach, California, USA, pp.259–262.

Suzuki, E. and Zytkow, J.M. (2005) 'Unified algorithm for undirected discovery of exception rules', *International Journal of Intelligent Systems*, Vol. 20, No. 7, pp.673–691.

Wang, K., Jiang, Y. and Lakshmanan, L.V.S. (2003) 'Mining unexpected rules by pushing user dynamics', *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington DC, USA, pp.246–255.

Xing, Z., Pei, J., Dong, G. and Yu, P.S. (2008) 'Mining sequence classifiers for early prediction', *Proc. 8th SIAM International Conference on Data Mining*, Atlanta, Georgia, USA, pp.644–655.

Yan, X., Han, J. and Afshar, R. (2003) 'CloSpan: mining closed sequential patterns in large databases', *Proc. 3rd SIAM International Conference on Data Mining*, San Francisco, CA, USA, pp.166–177.

Yen, S-J. and Lee, Y-S. (2006) 'An incremental data mining algorithm for discovering web access patterns', *International Journal of Business Intelligence and Data Mining*, Vol. 1, No. 3, pp.288–303.