Recognizing Unexpected Recurrence Behaviors with Fuzzy Methods in Sequence Databases
Haoyuan Li, Anne Laurent, Pascal Poncelet

To cite this version:

HAL Id: lirmm-00324590
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00324590
Submitted on 24 Mar 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Recognizing Unexpected Recurrence Behaviors with Fuzzy Measures in Sequence Databases

Dong (Haoyuan) Li
LGI2P, École des Mines d’Alès
Parc scientifique G. Besse
30035 Nîmes, France
Haoyuan.Li@ema.fr

Anne Laurent
LIRMM, Univ. Montpellier 2
161 rue Ada
34392 Montpellier, France
laurent@lirmm.fr

Pascal Poncelet
LGI2P, École des Mines d’Alès
Parc scientifique G. Besse
30035 Nîmes, France
Pascal.Poncelet@ema.fr

ABSTRACT
The recognition of unexpected behaviors in databases is an important problem in many real-world applications. In the previous studies, the unexpectedness is mainly stated within the context of the most-studied patterns, association rules, or sequential patterns. In this paper, we first propose the notion of fuzzy recurrence rule, a new kind of rule-based behavior in sequence databases, and then we introduce the problem of recognizing unexpected sequences contradicting the beliefs on fuzzy recurrence rules, with fuzzy measures. We also develop, UFR, an algorithm for discovering unexpected recurrence behaviors in a sequence database. Our approach is evaluated with Web access log data.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—Data mining; I.5.1 [Pattern Recognition]: Models—Fuzzy set

General Terms
Algorithms, Management, Theory

Keywords
Unexpected recurrence behavior, fuzzy recurrence rule, fuzzy measure, sequence data mining

1. INTRODUCTION
During the past years, as the most important topics of data mining, association rules [1] (corresponding to frequent patterns) and sequential patterns [2] have received much attention, such as the work addressed in [11, 5, 12] and [26, 18, 28, 3, 27, 22].

The association rule mining finds the frequent behaviors of the correlation between sets of attributes (also called patterns), as the rules in form “if X then Y” where X and Y are two patterns. An association rule can reflect the information typically like “60% of customers who purchase Coca Cola also purchase potato chips (if Coca Cola then potato chips)”. Different from mining association rules, the purpose of the sequential pattern mining is to find frequent behaviors of sequential data, thus a sequential pattern is a sequence such that “A then B then C then . . . ”, where A, B, C, . . . are patterns. A sequential pattern can help interpreting the information typically like “60% of people purchase beers, then purchase Sci-Fi movies, and then purchase rock music”.

On the other hand, the discoveries (normally belief-driven [23]) of unexpected behaviors contradicting the knowledge on known (normally frequent or predefined) behaviors in databases, becomes more and more interesting for many real-word applications. Since frequent patterns and sequential patterns are the most-studied behaviors in databases, in the existing studies of discovering unexpected behaviors, the unexpectedness is mainly stated within the context of patterns [4, 19], association rules [21], or sequential patterns [24, 16].

We proposed a semantics based framework of unexpected sequence mining in our previous work [16]. For instance, according to the behavior “people purchase Sci-Fi movies, and then purchase rock music”, the behavior “people purchase Sci-Fi movies, and then purchase classical music” can be considered as unexpected since the classical music can be considered as semantically opposite to the rock music. That work has been extended with fuzzy methods in [17].

In this paper, we are interested in the unexpectedness stated by fuzzy recurrence rule, a new kind of rule-based behavior in sequence databases. The fuzzy recurrence rules are in the form “if the sequence $s_n$ repeatedly occurs, then the sequence $s^p_3$ repeatedly occurs”. For instance, a fuzzy recurrence rule can be “60% of customers who often purchase Sci-Fi books then Sci-Fi movies later, also purchase PC games often”. This type of rules reflect the associated correlations between repeatedly occurred elements in sequential data. The unexpectedness on recurrence behaviors is determined by the domain-expert-defined semantic oppositions. For instance, if we consider that the classical music is semantically opposite to PC games, then the fact “1% customers who often purchase Sci-Fi books then Sci-Fi movies later, often purchase classical music” stands for an unexpected recurrence behavior in a customer transaction database.

Such unexpected recurrence behaviors can be interesting for many application domains, including marketing analysis, finance fraud detection, DNA segment analysis, Web content personalization, network intrusion detection, weather
2. RELATED WORK

In order to find more relevant behaviors, the fuzzy set theory has been well adopted for treating the quantitative attributes [25] in databases, such as the work concentrated in [15, 7, 14, 13, 6, 8].

For instance, in the discovery of fuzzy association rules defined in [15], with the usage of fuzzy sets, the form of discovered rules becomes “if X is A then Y is B”, where A and B are fuzzy sets that describe the properties of patterns X and Y, such as “60% of customers who purchase a lot of Coca Cola also purchase a lot of potato chips (if Coca Cola is lot then potato chips is lot)”. In the same manner that the fuzzy association rule is defined, the notion of fuzzy sequential patterns proposed in [7] considers the sequential patterns on quantitative attributes like “60% of people purchase a lot of beers, then purchase many Sci-Fi movies later, then purchase few PC games”, where the sequence is “beer is lot, then Sci-Fi movie is many, and then PC game is few”.

Although most of the existing approaches for mining fuzzy association rules or fuzzy sequential patterns concentrate on developing efficient algorithms, the fuzzy sets are also considered on imprecise data (like [10, 9], but that topic is not covered in this paper). In our approach, we consider the binary-valued attributes in databases as other crisp data mining approaches, however we use fuzzy sets for describing the recurrence behaviors of data, instead of the quantitative attributes.

Unexpected behaviors are generally considered within the framework of subjective interestingness measure. In [23], the notion of unexpectedness is addressed with hard belief and soft belief. A hard belief is a belief that can never be changed by new evidences in data, and any contradiction of such a belief implies data error. A soft belief corresponds to the constraints on data measured by a degree, which can be modified with new evidences contradicting this belief. The interestingness of such new evidences is measured by the change of the degree.

With the unexpectedness measure, a belief-driven approach for for finding unexpected patterns and association rules is proposed in [19, 20, 21]. In that approach, a belief is given from association rule, and the unexpectedness is stated by semantic opposition of patterns. Given a belief $X \rightarrow Y$, a rule $A \rightarrow B$ is unexpected if: (1) the patterns $B$ and $Y$ semantically contradict each other; (2) the support and confidence of the rule $A \cup X \rightarrow B$ hold in the data; (3) the support and confidence of the rule $A \cup X \rightarrow Y$ do not hold in the data. The discovery process is performed within the framework of the a priori algorithm.

[24] proposed an approach for mining unexpectedness with sequence rules transformed from frequent sequences. The sequence rule is built by dividing a sequence into two adjacent parts, which are determined by the support, confidence and improvement from association rule mining. 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 this work considers the unexpected sequences and rules, it is however very different from our problem in the measure and the notion of unexpectedness contained in data.

In our recent work [17], we proposed a belief-driven approach for recognizing fuzzy unexpected sequences corresponding to sequential implication rules. A sequential implication rule is a rule of the form “if the sequence $s_a$ occurs then the sequence $s_j$ occurs latter” so that the beliefs are created with respect to (1) the distance between $s_a$ and $s_j$; (2) the semantics of the implication between $s_a$ and $s_j$, i.e., $s_j$ cannot be replaced by another sequence $s_b$. The fuzzy sets are considered on the distance between the two sequences.

3. FUZZY RECURRENCE RULES AND UNEXPECTED BEHAVIORS

In this section, we first formalize the fuzzy recurrence rules within the common framework of sequence mining model, we then present the belief base on such fuzzy recurrence rules, with which the unexpected recurrence behaviors are therefore proposed and discovered.

3.1 Data Model

We consider the sequential data that consist in binary-valued attributes. Given a limited numbered set of distinct binary-valued attributes $R = \{i_1, i_2, \ldots, i_m\}$, each attribute is an item. An itemset is an unordered collection of items, denoted as $I = \{i_1, i_2, \ldots, i_m\}$, such that $I \subseteq R$. A sequence is an ordered list of itemsets, denoted as $s = I_1 I_2 \cdots I_k$. A sequence database is usually a large set of sequences, denoted as $D$.

Given two sequences $s = I_1 I_2 \cdots I_m$ and $s' = I'_1 I'_2 \cdots I'_n$, if there exist integers $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ such that $I_1 \subseteq I'_1, I_2 \subseteq I'_2, \ldots, I_m \subseteq I'_m$, then the sequence $s$ is a subsequence of the sequence $s'$, denoted as $s \subseteq s'$. If $s \subseteq s'$, we say that $s$ is contained in $s'$, or $s'$ supports $s$. For example, the sequence $s_1 = (a)(b)$ is contained in the sequence $s_2 = (a)(b)(c)$, but not contained in the sequence $s_3 = (ab)(c)$.

Given a sequence database $D$, the support or frequency of a sequence $s$, denoted as $\sigma(s, D)$, is the fraction of the total number of sequences in $D$ that support $s$. Given a user specified threshold of support called minimum support, denoted as min_supp, a sequence $s$ is frequent if $\sigma(s, D) \geq$ min_supp. In addition, we denote the concatenation of $n$ sequences as $s_1 s_2 \cdots s_n$. For example, let $s_1 = (a)(b)$ and $s_2 = (c)(d)$, then we have $s_1 s_2 = (a)(b)(a)(b)$ or $s_1 s_2 = (a)(b)(c)(d)$.

3.2 Fuzzy Recurrence Rules

A recurrence rule is a rule on sequences with form $(s_a, \psi) \Rightarrow (s_b, \theta)$, where $s_a$ and $s_b$ are two sequences, and $\psi, \theta$ are two integers for describing the recurrence behaviors.

The recurrence rule indicates that given a sequence $s$, if $s_a$ is orderly occurred no less than $\psi$ times within $s$, then $s_b$ should occurs in $s$ no less than $\theta$ times, that is, $\frac{(s_a \cdots s_a \subset s) \land (n \geq \psi)}{s_a \cdot s_a \subset s) \land (k \geq \theta)}$. 

prediction, and so on.

The rest of the paper is organized as follows. In Section 2, we introduce the related work. Section 3 presents our proposals of fuzzy recurrence rules and unexpected recurrence behaviors. In Section 4, we develop an effective algorithm UFR for discovering unexpected recurrence behaviors in a sequence database. Finally, we conclude in Section 5.
We call the form \( \langle s, \psi \rangle \) a recurrent sequence, and we have
\[
\langle s, \psi \rangle \subseteq s' \iff (s_0 \ldots s_n \subseteq s') \land (n \geq \psi),
\]
for that the sequence \( s' \) supports the recurrent sequence \( \langle s, \psi \rangle \). We call the recurrent sequence \( \langle s, \psi \rangle \) a \( \psi \)-recurrence sequence. We use the wildcard \( ^* \) for denoting the general-purpose meaning of the support between sequences, that is,
\[
\langle s, ^* \rangle \subseteq s' \equiv s \subseteq s'.
\]
Given a sequence \( s \) and a recurrence rule \( r = \langle s_a, \psi \rangle \Rightarrow \langle s_b, \theta \rangle \), if \( \langle s_a, \psi \rangle \subseteq s \) and \( \langle s_b, \theta \rangle \subseteq s \), then we say that \( s \) supports \( r \), denoted as \( s \models r \). For instance, the recurrence rule \( r = ((a)(b), 3) \Rightarrow ((c)(d), ^* \) depicts that a given sequence \( s \), if \( (a)(b) \) is contained repeatedly in \( s \) no less than 3 times, then \( (c)(d) \) should be contained in \( s \); in other words, if \( (a)(b)(a)(b) \subseteq s \), then \( (c)(d) \subseteq s \). Notice that the occurrences of \( s_a \) must be ordered, that is, for example, given a rule \( r_1 = ((a)(b), 2) \Rightarrow ((c), ^*) \), the sem
\[
s_1 = (a)(a)(c)(b)(b) \text{ does not support } r_1 \text{, but the sequence } s_2 = (a)(b)(c)(a)(b) \text{ supports } r_1; \text{ however, the sequence } s_1 \text{ supports the rules } r_2 = ((a), 2) \Rightarrow ((c), ^*) \text{ and } r_3 = ((b), 2) \Rightarrow ((c), ^*).
\]
Considering the integer \( \psi \), a human-friendly interpretation is more flexible in most applications. For instance, in market basket analysis, to point out that “the customers who often purchase Sci-Fi books often purchase action movies” is more relevant than the conclusion “the customers who purchase classical music often purchase action movies”.

We therefore extend the recurrence rule with fuzzy sets, so-called the fuzzy recurrence rule, in the form \( \langle s_a, \omega_\alpha \rangle \Rightarrow \langle s_b, \omega_\beta \rangle \), where \( \omega_\alpha \) and \( \omega_\beta \) are two fuzzy sets for describing \( s_a \) and \( s_b \). The sequences \( \langle s_a, \omega_\alpha \rangle \) and \( \langle s_b, \omega_\beta \rangle \) are fuzzy recurrent sequences. Given a sequence \( s' \) and a fuzzy recurrent rule \( \langle s, \omega \rangle \), that \( s' \) supports \( \langle s, \omega \rangle \) is defined as
\[
\langle s, \omega \rangle \subseteq s' \iff (s_0 \ldots s_n \subseteq s) \land (\mu_\omega(n) \geq \zeta),
\]
where the fuzzy degree measured by the membership function \( \mu_\omega(n) \) must be superior or equal to a threshold \( \zeta \).

Let us consider the following example.

**Example 1.** Given a set of distinct events \( A, B, C, D, \ldots \), an ordered of events can be represented as the data model of sequence. Assuming that a given event sequence \( s \), if \( s \) supports the recurrent sequence \( \langle (A)(B), 4 \rangle \), then \( s \) supports the subsequence \( \langle C)(D) \rangle \); if \( s \) supports the recurrent sequence \( \langle (A)(B), 9 \rangle \), then \( s \) supports \( (C) \). These behaviors can be described by recurrence rules, such as the rule \( r_1 = \langle (A)(B), 4 \rangle \Rightarrow \langle (C)(D), ^* \rangle \) and the rule \( r_2 = \langle (A)(B), 9 \rangle \Rightarrow \langle (C), ^* \rangle \). Given a sequence \( s_1 \) such that \( \langle (A)(B), 4 \rangle \subseteq s_1 \) and \( \langle (C), D) \subseteq s_1 \), a sequence \( s_2 \) such that \( \langle (A)(B), 8 \rangle \subseteq s_2 \) and \( \langle C \rangle \subseteq s_2 \), we have \( s_1 \not\models r_1 \) and \( s_2 \not\models r_2 \). However, since the recurrent sequences contained in these sequences and rules are close, the sequences \( s_1 \) and \( s_2 \) can be still potentially interesting. On the other hand, considering the fuzzy recurrence rules \( r_1' = \langle (A)(B), \text{rarely} \rangle \Rightarrow \langle (C)(D), ^* \rangle \) and \( r_2' = \langle (A)(B), \text{often} \rangle \Rightarrow \langle (C), ^* \rangle \), corresponding to the rules \( r_1 \) and \( r_2 \) with respect to the fuzzy partitions shown in Figure 1, let the threshold \( \zeta = 0.5 \), then we have \( s_1 \models r_1' \) and \( s_2 \models r_2' \). We can further define more partitions, such as “always” or “rarely”.

In this paper, the fuzzy recurrence rules are considered as having been predefined by domain experts, the discovery of fuzzy recurrence rules will be covered in our future research work.

**3.3 Unexpected Recurrence Behaviors**

We are considering to discover the sequences contained in a database those are semantically opposite to a given set of fuzzy recurrence rules. In order to find such sequences, we propose an approach that constructs a belief base from given fuzzy recurrence rules with semantic constraints, so that each sequence not respecting the belief base is unexpected, from which the unexpected recurrence behaviors can be further discovered.

A belief on recurrence behaviors is a set of constraints that consists of a fuzzy recurrence rule \( (s_a, \omega_\alpha) \Rightarrow (s_b, \omega_\beta) \) and a semantic constraint \( (s_a, \omega_\alpha) \not\equiv \text{sem} (s_b, \omega_\beta) \), where \( \omega_\alpha \) is a fuzzy set for the sequence \( s_a \). The fuzzy recurrence rule implies an association relationship between \( (s_a, \omega_\alpha) \) and \( (s_b, \omega_\beta) \), i.e., if the recurrence of \( s_a \) is \( \omega_\alpha \), then the recurrence of \( s_b \) is \( \omega_\beta \). The semantic constraint implies that the recurrent sequences \( (s_a, \omega_\alpha) \) and \( (s_b, \omega_\beta) \) are semantically opposite to each other. Notice that \( s_a \) and \( s_b \) are not necessary to be different: \( (\text{game}, \text{rarely}) \) and \( (\text{game}, \text{always}) \) are semantically opposite to each other.

We use the triple \( [s_a, \omega_\alpha]; (s_b, \omega_\beta); (s_a, \omega_\gamma) \) for denoting a belief, that constrains: given a sequence \( s \), if \( s \) supports \( (s_a, \omega_\alpha) \), then \( s \) supports \( (s_b, \omega_\beta) \), however \( s \) should not support \( (s_a, \omega_\gamma) \), since \( (s_a, \omega_\gamma) \) semantically be opposite to \( (s_b, \omega_\beta) \), that is,
\[
(\langle s_a, \omega_\alpha \rangle \subseteq s) \land (\langle s_b, \omega_\beta \rangle \subseteq s) \land (\langle s_a, \omega_\gamma \rangle \not\subseteq s).
\]

**Example 2.** Let us consider the instance in Section 1. We know that the customers who purchase Sci-Fi books (noted as book) then Sci-Fi movies (noted as movie) latter like to play PC games (noted as game). Since we consider that PC games and classical music (noted as music) are semantically opposite to each other, the semantic constraint can be \( (\text{game}, \text{often}) \not\equiv \text{sem} (\text{music}, \text{often}) \). Thus the belief can be written as
\[
[(\text{book}(\text{movie}, \text{often})]; (\text{game}, \text{often}); (\text{music}, \text{often})]
\]
if we assume such customers will not often purchase classical music. The fuzzy sets for the purchase of classical music can also be that shown in Figure 1. The above belief describes that the customers who often purchase Sci-Fi books and then Sci-Fi movies, purchase PC games often, however do not purchase classical music often.
If a sequence \( s \) satisfies these constraints, we say that the sequence \( s \) supports the belief \([s_\alpha, \omega_\alpha]; (s_\beta, \omega_\beta); (s_\gamma, \omega_\gamma)\], denoted as \( s \models b \). A sequence \( s \) is unexpected if \( s \) violates a belief \( b \), denoted as \( s \not\models b \). Two cases of violation are considered in our approach according to two factors of the unexpectedness contained in the sequence \( s \).

**Definition 1.** (\( \beta \)-unexected sequence) Given a belief \( b = [(s_\alpha, \omega_\alpha); (s_\beta, \omega_\beta); (s_\gamma, \omega_\gamma)] \), a sequence \( s \) is \( \beta \)-unexpected, denoted as \( s \not\models b \), if \( s \) supports \( (s_\alpha, \omega_\alpha) \) and \( s_\beta \) does not support \( (s_\beta, \omega_\beta) \), that is,

\[
s \not\models b \iff (s_\alpha, \omega_\alpha) \subseteq s \land (s_\beta \subseteq s) \land (s_\beta, \omega_\beta) \subseteq s,
\]

and such a behavior is called \( \beta \)-unexpectedness.

The primary factor of the \( \beta \)-unexpectedness in a sequence \( s \) is that the recurrent sequence \( (s_\beta, \omega_\beta) \) does not occur as expected however at least the sequence \( s_\beta \) occurs in \( s \). For instance, considering the belief in Example 2, noted as \( b \), let \( s \) be a customer transaction sequence, if we have \( \langle\text{book}(\text{movie}), \text{often}\rangle \subset s \) and \( \langle\text{game}, \text{often}\rangle \subset s \), then the sequence \( s \) is expected in the meaning of fuzzy recurrence rule \( \langle\text{book}(\text{movie}), \text{often}\rangle \Rightarrow \langle\text{game}, \text{often}\rangle \) (we discuss the semantic constraint latter); however, if we have \( \langle\text{game}\rangle \subset s \) but not \( \langle\text{game}, \text{often}\rangle \subset s \), for example, we have \( \langle\text{game}, \text{rarely}\rangle \subset s \) instead of \( \langle\text{game}, \text{often}\rangle \subset s \), since \( \langle\text{game}, \text{rarely}\rangle \subset s \) implies \( \langle\text{game}\rangle \subset s \), then \( s \) is a \( \beta \)-unexpected sequence, i.e., \( s \not\models b \).

**Definition 2.** (\( \gamma \)-unexpected sequence) Given a belief \( b = [(s_\alpha, \omega_\alpha); (s_\beta, \omega_\beta); (s_\gamma, \omega_\gamma)] \), a sequence \( s \) is \( \gamma \)-unexpected, denoted as \( s \not\models b \), if \( s \) supports \( (s_\alpha, \omega_\alpha) \) and \( (s_\gamma, \omega_\gamma) \), that is,

\[
s \not\models b \iff (s_\alpha, \omega_\alpha) \subseteq s \land (s_\gamma, \omega_\gamma) \not\subseteq s,
\]

and such a behavior is called \( \gamma \)-unexpectedness.

Respectively, the primary factor of the \( \gamma \)-unexpectedness in a sequence \( s \) is that the semantic constraint \( (s_\beta, \omega_\beta) \not\in \text{sem} \langle s_\gamma, \omega_\gamma \rangle \) is broken, because the recurrent sequence \( \langle s_\gamma, \omega_\gamma \rangle \) occurs in \( s \). Considering again the belief \( b \) in Example 2 and let \( s \) be a customer transaction sequence, if we have \( \langle\text{book}(\text{movie}), \text{often}\rangle \subset s \) and \( \langle\text{music}, \text{often}\rangle \not\subset s \), then the sequence \( s \) is not unexpected in the meaning of semantic constraint \( \langle\text{music}, \text{often}\rangle \not\in \text{sem} \langle\text{music}, \text{often}\rangle \); however, if we have \( \langle\text{music}, \text{often}\rangle \subset s \), then \( s \) is a \( \gamma \)-unexpected sequence, i.e., \( s \not\models b \). Of course, it is not necessary to forbid \( \langle\text{music}\rangle \subset s \); for example, according to this belief, the occurrence of \( \langle\text{music}, \text{rarely}\rangle \) does not imply the \( \gamma \)-unexpectedness.

In our approach, we consider only the belief bases of coherent beliefs, means that the beliefs contained in the belief base do not contradict each others. Given a belief base \( B \), for any two beliefs \( b, b' \in B \), let \( b = [(s_\alpha, \omega_\alpha); (s_\beta, \omega_\beta); (s_\gamma, \omega_\gamma)] \) and \( b' = [(s'_\alpha, \omega'_\alpha); (s'_\beta, \omega'_\beta); (s'_\gamma, \omega'_\gamma)] \), the following constraint must be satisfied:

\[
(s_\alpha \subseteq s'_\alpha) \land (\omega_\alpha = \omega'_\alpha) \Rightarrow (s_\beta \not\subseteq s'_\beta) \lor (\omega_\beta \neq \omega'_\beta).
\]

For example, let us consider two beliefs \( b_1 \) and \( b_2 \). Let \( b_1 = [(\langle a \rangle(b), \text{often}); \langle(c)(d), \text{often}\rangle]; \langle(e)(f), \text{often}\rangle] \) and let \( b_2 = [(\langle a \rangle), \text{often}; \langle(e)(f), \text{often}\rangle]; \langle(c)(d), \text{often}\rangle] \), then \( b_1 \) and \( b_2 \) are in conflict: we have \( \langle\text{a}(\text{b}), \text{often}\rangle \Rightarrow \langle\text{c}(\text{d}), \text{often}\rangle \) and \( \langle\text{a}(\text{b}), \text{often}\rangle \neq \langle\text{e}(\text{f}), \text{often}\rangle \) for \( b_1 \), however for \( b_2 \) we have \( \langle\text{a}(\text{b}), \text{often}\rangle \neq \langle\text{c}(\text{d}), \text{often}\rangle \). In the rest of this paper, we assume that all beliefs in a belief base are coherent.

Given a sequence database \( D \) and a belief base \( B \), the problem of recognizing unexpected recurrence behaviors is to, therefore, find all sequences \( s \in D \) that contain \( \beta \)- and/or \( \gamma \)-unexpectedness corresponding to each belief \( b \in B \).

### 4. UFR: MINING UNEXPECTED FUZZY RECURRENCE BEHAVIORS

UFR stands for mining Unexpected Fuzzy Recurrence behaviors. In this section, we detail the algorithm involved in our approach, which is evaluated in performance study.

#### 4.1 Data Structure

The belief base \( B \) is constructed as a prefix tree structure with 3 blocks \( \alpha, \beta \) and \( \gamma \), where in each block two types of edge are used to represent itemsets and sequences, and between blocks a pair of links are used for representing beliefs.

![Prefix tree structure](image)

**Figure 2:** Prefix tree structure of the belief base.

A belief is in the form \([s_\alpha, \omega_\alpha]; (s_\beta, \omega_\beta); (s_\gamma, \omega_\gamma)] \). In a block, all \( s_\alpha, \omega_\alpha \) parts of each belief are organized as a prefix tree. For example, in Figure 2, block \( \alpha \) contains 3 recurrent sequences \( \langle(a)(b), \omega_1 \rangle, \langle(a)(ab), \omega_2 \rangle \) and \( \langle(ab), \ast \rangle \). All \( s_\beta, \omega_\beta \) parts are contained in \( \beta \) block and all \( s_\gamma, \omega_\gamma \) parts contained in \( \gamma \) block as prefix trees. For example, blocks \( (\beta) \) and \( (\gamma) \) contain recurrent sequences \( \langle(c)(d), \omega_1 \rangle \) and \( \langle(e)(f), \omega_2 \rangle \). The link pair between the 3 blocks shown in Figure 2 represents the belief \([\langle(a)(ab), \omega_2 \rangle; \langle(c)(d), \omega_4 \rangle; \langle(e)(f), \omega_5 \rangle] \).

The verification of each sequence is performed by depth-first traversing the prefix tree in each block.

#### 4.2 Algorithm

The algorithm UFR accepts a belief base \( B \), a sequence database \( D \) and a minimum fuzzy degree threshold \( \zeta \) as input data, and outputs all unexpected sequences in \( D \) with respect to \( B \) and \( \zeta \).

Algorithm 1 shows the verification routine for each sequence \( s \in D \). The algorithm first traverses the prefix tree in block \( \alpha \) to verify the \( (s_\alpha, \omega_\alpha) \) part of each belief \( b \in B \), with depth-first strategy. If the traverse arrives an \( \omega_\alpha \) node, the recurrence of sequence \( s_\alpha \) is examined within the fuzzy set \( \omega_\alpha \) and the minimum fuzzy degree \( \zeta \). If \( \zeta \) is satisfied, the link pair is followed to verify the \( (s_\beta, \omega_\beta) \) part of the same belief \( b \) contained in the prefix tree in block \( \beta \), till to the node \( \omega_\beta \). If the recurrence of \( s_\beta \) does not satisfy \( \omega_\beta \) with respect to \( \zeta \), then the algorithm outputs the sequence \( s \) as...
Figure 2, the UFR-seqveri routine can be illustrated with the tree will be traversed till to the node \( \omega \). If the recurrence of \( s_\beta \) satisfies \( \omega_\beta \) with respect to \( \zeta \), then the algorithm outputs \( s \) as a \( \gamma \)-unexpected sequence. The algorithm continues to traverse the prefix tree in block \( \alpha \) and repeats the above procedure till to the end of the traverse. All unexpected sequences are therefore discovered.

Corresponding to the fuzzy set in Figure 1 and the belief \( [(a)(ab), often) ; (c)(d), rarely) ; (ef)(g), rarely] \) shown in Figure 2, the UFR-seqveri routine can be illustrated with the sequence \( s \) shown in Figure 3, assuming \( \zeta = 0.6 \).

\[
\begin{align*}
(1) & \quad (2) & \quad (3) & \quad (4) & \quad (5) \\
(b)(ab)c(ab)d(ab)c(ab)d & \qquad (1) & \quad (2) & \quad (3) & \quad (4) \\
& \quad (6) & \quad (5) \\
(1) & \quad (2) & \quad (3) & \quad (4) & \quad (5)
\end{align*}
\]

**Figure 3: Finding \( \beta \)-unexpected sequence.**

We have \( \langle(a)(ab), often) \subseteq s \), which is first verified by calling \( seqinc_fuzzy(\langle(a)(ab), often) , s , 0.6) \), and the recurrence of \( (a)(ab) \), which is marked as (1) to (6) above the sequence shown in Figure 3, satisfies the minimum fuzzy degree 0.6. Thus \( \langle c)(d), rarely \rangle \subseteq s \) will be verified by calling \( seqinc_fuzzy(\langle c)(d), rarely \rangle , s , 0.6) \), where the recurrence of \( (c)(d) \) is marked as (1) to (5) below the sequence shown in Figure 3. According to Figure 1, we have the fuzzy degree of \( \langle c)(d) \rangle, 5 \) is 0.5 in the partition “rarely”, so that we have \( \langle c)(d), rarely \rangle \not\subseteq s \), and the sequence \( s \) shown in Figure 3 is \( \beta \)-unexpected to the given belief.

The recognition of \( \gamma \)-unexpectedness is similar to the above illustration.

### 4.3 Performance

Our approach is evaluated with Web access logs. Two types of Web access log are used in our experiments: one is a large access log file of an online forum site (labeled as BBS), and another is a large access log file of a mixed homepage hosting server (labeled as WWW).

**Table 1: Web access logs in the experiments**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BBS</td>
<td>135,562</td>
<td>126,383</td>
<td>15,5591</td>
</tr>
<tr>
<td>WWW</td>
<td>53,325</td>
<td>85,810</td>
<td>8,3507</td>
</tr>
</tbody>
</table>

The composition of the two data sets are listed in Table 1. We first apply a sequential pattern mining algorithm to discover frequent sequences for studying the general behaviors of the data sets. The frequent 4-recurrence sequences and 8-recurrence sequences are shown in Figure 4 and Figure 5.

**Figure 4: Frequent recurrent sequences in BBS.**

**Figure 5: Frequent recurrent sequences in WWW.**

The recurrent sequences in the data sets show that the recurrence behaviors depend on the semantic characteristics of data, for instance, in our experimental data sets, the recurrence behaviors in online forum site are more stronger than those in mixed content Web site.

We generate 15 beliefs for each data set after examining the sequential patterns and frequent 4-recurrence and 8-recurrence sequences discovered in last step, corresponding to 3 groups of 5 beliefs: with “rarely”, “often” and “frequently” (according to Figure 1) appearing in the \( \langle s_\alpha , \omega_\beta \rangle \) part of a belief. Table 2 lists several sample beliefs in our ex-
periments, where the symbols like “(f=4)” represent a forum ID in URL query string.

Figure 6 and Figure 7 show our experimental results.

Table 2: Sample beliefs

<table>
<thead>
<tr>
<th>(s_α, ω_α)</th>
<th>(s_ω, ω_ω)</th>
<th>(s_γ, ω_γ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f=4), rarely</td>
<td>(f=9), rarely</td>
<td>(f=9), often</td>
</tr>
<tr>
<td>(f=0)(f=5), often</td>
<td>(f=8), often</td>
<td>(f=4), often</td>
</tr>
<tr>
<td>(f=5), frequently</td>
<td>(f=4), rarely</td>
<td>(f=9), often</td>
</tr>
<tr>
<td>/li/, rarely</td>
<td>/li/pub/, often</td>
<td>/li/pub/, rarely</td>
</tr>
<tr>
<td>/li/pub/, often</td>
<td>/li/, rarely</td>
<td>/li/doc/, often</td>
</tr>
<tr>
<td>/li/, frequently</td>
<td>/li/doc/, rarely</td>
<td>/li/doc/, often</td>
</tr>
</tbody>
</table>

Figure 6: Unexpected sequences in BBS.

Figure 7: Unexpected sequences in WWW.

With the decrease of the minimum fuzzy degree threshold, the number of unexpected sequences increases. In Figure 6, we find in the “frequently” partition, the number of unexpected sequences is much less than those in the other two partitions, because in the data set the number of long recurrent sequences, such as 8-recurrence sequences, is less. We can also find that the unexpected behaviors focus on the recurrences between “rarely” and “often”. In Figure 7, there is a sharp increase of the number of unexpected sequences in the “often” partition when the minimum fuzzy degree decreases from 0.6 to 0.4, because in the “often” partition, the fuzzy degree 0.5 corresponds to 4-recurrence sequences, so that a lot of unexpected sequences in the “rarely” partition are counted as “often”.

5. CONCLUSION

In this paper, we introduce the problem of discovering unexpected recurrence behaviors in sequence databases. We propose a novel notion, the fuzzy recurrence rules, for depicting the recurrence behaviors of the data, where fuzzy set theory is applied to describe the recurrence of sequences. We present a belief-driven approach for modeling two types of unexpectedness in recurrence behaviors, where the belief consists in a fuzzy recurrence rule and a semantic constraint on the rule. We also develop an effective algorithm UFR, which discovers all unexpected sequences in a sequence database with respect to domain expert specified belief base and minimum fuzzy degree threshold. The experimental results on Web access logs show the usefulness of our propositions.

Our future research includes the discovery of fuzzy recurrence rules in sequential data, we believe that our proposal of this novel rule model on sequences can be interesting for many real-word application domains.

6. REFERENCES


