Discovering Fuzzy Unexpected Sequences with Concept Hierarchies

Haoyuan Li, Anne Laurent, Pascal Poncelet

International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems

DISCOVERING FUZZY UNEXPECTED SEQUENCES WITH CONCEPT HIERARCHIES

Dong (Haoyuan) Li
LGI2P, École des Mines d'Alès, Parc scientifique Georges Besse
30035 Nîmes cedex 1, France
Haoyuan.Li@ema.fr

Anne Laurent
LIRMM, Université Montpellier 2, 161 rue Ada
34392 Montpellier cedex 5, France
laurent@lirmm.fr

Pascal Poncelet
LIRMM, Université Montpellier 2, 161 rue Ada
34392 Montpellier cedex 5, France
poncelet@lirmm.fr

Received 9 October 2008
Revised 15 March 2009

Sequential pattern mining is the method that has received much attention in sequence data mining research and applications, however, a drawback is that it does not profit from prior knowledge of domains. In our previous work, we proposed a belief-driven method with fuzzy set theory for discovering the unexpected sequences that contradict existing knowledge of data, including occurrence constraints and semantic contradictions. In this paper, we present a new approach that discovers unexpected sequences with determining semantic contradictions by using concept hierarchies associated with the data. We evaluate the effectiveness of our approach with experiments on Web usage analysis.

Keywords: Data mining, unexpected sequences, concept hierarchies, soft hierarchy-based similarity.

1. Introduction

With the development of data management and analysis techniques, more and more real-world applications store and process the data in sequence format, including Web usage analysis, telecommunication network monitoring, finance and marketing investigations, science experiments, bioinformatics, and so on. Sequence data mining[15,10] has therefore received much attention, where mining sequential patterns[2] is the method most of the research has been concentrated on, which finds frequent correlations between the elements contained in sequence data.

Up to now, many efficient algorithms have been developed for mining sequential patterns around reducing execution time and memory usage[3,26,28,30,39,40,42]. However, as of most statistical frequency based data mining methods, a drawback is that the sequential pattern mining process does not profit from prior knowledge of data, and often extracts an extremely large number of sequences many of which are obvious or irrelevant with respect to the domain knowledge. To address those issues, the measurement of "interestingness" for data mining has been systematically studied during the past years, which can be found in the survey of McGarry[27]. In the approaches to interestingness measures, a well-focused one is that the discovered patterns or sequences are interesting because they are unexpected to existing knowledge[4,9,18,22,23,31,32,33,29,37,38]. In order to find the unexpectedness in sequence data, we proposed a belief-driven approach in our previous paper[22], where the unexpected sequence discovery depends on the belief consisting of a sequence rule, an occurrence constraint, and a semantic contradiction.

For instance, if the prior knowledge of customer purchase behaviors indicates that in general the customers purchase a `pop music CD` within the next 5 purchases after a purchase of an `action movie DVD`, then a sequence rule can be defined[a] as "`action movie` → `pop music`", with an occurrence constraint that "the intervals between `action movie` and `pop music` should be no more than 5"; if we further consider that the `classical music` *semantically contradicts* the `pop music`, then a semantic contradiction relation that "a purchase of `pop music CD` contradicts the purchase of `classical music CD`" can be applied. We can therefore state the unexpectedness by that: "after purchasing an `action movie DVD`, a customer purchases a `pop music CD` out of the next 5 purchases, or purchases a `classical music CD` within the next 5 purchases." This work is extended with our recent approach[23], where the fuzzy set theory[41] is applied to the occurrence constraint for describing more relevant unexpected sequences.

However, a limit of our previous approaches is that although the beliefs can be specified by domain experts, the enumeration of the complete sets of sequence rules and semantic contradictions based on items[22,23,24] is obviously a hard work. For example, as shown in the above instance, if there exist 10 individual products in each category of `pop music CD`, `classical music CD`, and `action movie DVD`, we have to build 10^3 distinct beliefs to cover all possible combinations of items. Moreover, although the `classical music` can be naturally viewed as contradicting the `pop music`, nevertheless there exist many other genres of music, like the `blues`, `country`, `jazz`, or `rock` that can not be simply considered as contradicting or not to contradict the `pop music`. Another example about this difficulty can be addressed in our work on Web usage mining[24]. In that problem, domain expertise is also required for specifying the semantic contradictions, where the determination of the contradictions between different categories of Web contents is quite subjective. For example, the

[a] According to our proposition of building beliefs, a sequence rule required by a belief can be either extracted from frequent sequences, or defined by domain experts.

contradiction between `politics news` and `technology news` strongly depends on the experiences of users. The detection of the semantic relatedness between concepts in a taxonomy is many discussed in data mining literatures[14,19,35,36].

In the proposed approach, we improve our previous work by using *fuzzy concept hierarchies* in belief construction with the advantages that the semantic contradictions are no longer obligated to build beliefs and the generalized sequence rules can be handled. Hence, in this paper, we use the *semantic relatedness* between sequences, which is a fuzzy degree determined by the *semantic distance* (the *path-length* in a concept hierarchy) and the *semantic similarity*[b] between concepts.

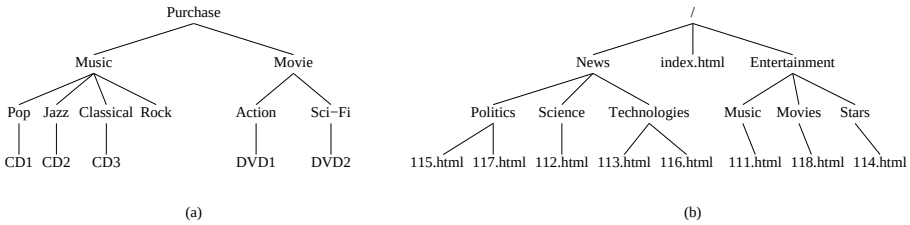

Fig. 1. Semantic hierarchies.

According to the hierarchy shown in Figure 1-a, in the previous example about customer purchases, the rule "`action movie` → `pop music`" requires a purchase of a `pop music CD` after a purchase of an `action movie DVD`. Assume that a customer has purchased `DVD1` (under `Action` of `Movie`), then a purchase of `CD1` (under `Pop` of `Music`) is expected. According to the semantic relatedness between the concepts "`Pop`", "`Jazz`" and "`Classical`", a purchase of `CD2` (under `Jazz` of `Music`) or `CD3` (under `Classical` of `Music`) can be unexpected with respect to the occurrence constraints. Figure 1-b shows an example on Web usage analysis, where the hierarchy corresponds to the Web site structure. The semantic similarity between concepts is not specified, thus the fuzzy degree of relatedness can be computed from the path-length between concepts. For instance, if users are expected to access `politics news` after the access of index page `/index.html`, then the access of other pages like `/Entertainment/Music/111.html`, `/News/Science/112.html` or `/News/Technologies/113.html` have different fuzzy degree of relatedness. In this case, the access of `112.html` or `113.html` can be considered as, for example, "*strong unexpected*" and the access of `111.html` can be considered as "*weak unexpected*" according to their path-length to any `politics news`.

The rest of this paper is organized as follows. The related work is introduced in Section 2. In Section 3, we propose the formalizations of fuzzy unexpected sequences

[b]For instance, defined by domain experts or calculated by lexical database analyzing tools, like the **Java WordNet Similarity Library** (http://grid.deis.unical.it/similarity/) or the **WordNet::Similarity** (http://www.d.umn.edu/~tpederse/similarity.html).

with the soft beliefs built from sequence rules and concept hierarchies. In Section 4, we present our proposed approach *FUSE* (Fuzzy Unexpected Sequence Extraction) for mining fuzzy unexpected sequences with concept hierarchies. In Section 5, we show and discuss the experimental results, and finally we conclude in Section 6.

2. Related Work

Fuzzy set theory has been many applied to discover more relevant association rules and sequential patterns. Indeed, fuzzy sets on quantitative attributes are more usually employed to change the domain of the attributes, employing granules defined by fuzzy sets instead of precise values.

For example, an association rule[1] $X \Rightarrow Y$ depicts the correlation "if X then Y" between patterns X an Y. With fuzzy sets, there is a very extended way of considering fuzzy association rules as "if X is A then Y is B" in considering of various information of attributes, such as the type "if `beer` is `lot` then `potato chips` is `lot`" or "if `age` is `old` then `salary` is `high`"[5,8,11,16,20,21]. In the same manner, the notion of fuzzy sequential patterns[6,17,7,12,13] considers the sequential patterns[2] on quantitative attributes like "60% of young people purchase a lot of beers, then purchase many action movies later, then purchase few PC games", where the sequence represents "`people` is `young`, then `beer` is `lot`, then `action movie` is `many`, and then `PC game` is `few`". Different than many approaches that consider that fuzzy association rules as rules or fuzzy sequential patterns as sequences obtained from fuzzy transactions, i.e., fuzzy subsets of items containing like "`age` is `old`" and "`salary` is `high`", we proposed the notion of fuzzy recurrence rules[25] depicting the relation like "if `DVD` is `often` then `CD` is `often`" in sequence format.

In this paper, we consider the binary-valued attributes in databases as other crisp data mining approaches, however we use fuzzy sets for describing the occurrence and semantics of the unexpectedness. In comparison with association rule and sequential pattern mining, we present a subjective measure for sequence mining.

McGarry[27] 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. Silberschatz and Tuzhilin[29] studied the subjective measures, 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. Silberschatz and Tuzhilin[29] further introduced two types of beliefs, hard belief and soft belief, for addressing unexpectedness. According to their 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 Web site at night, the degree of the belief "users access the Web site 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.

With the unexpectedness measure, Padmanabhan and Tuzhilin[31,32,33] propose a belief-driven approach for finding unexpected association rules. In that approach, a belief is given from association rule, and the unexpectedness is stated by the semantic contradiction between patterns. Given a belief $X \Rightarrow Y$, an association 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.

Spiliopoulou[37] 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. 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. Preliminary Definitions

3.1. *Data Model*

Based on the context of sequential pattern data mining[2], we consider the following definitions of the data model.

Given a set of binary-valued attributes, an *item* is an attribute. An *itemset* is an unordered collection of items sorted by lexical order, denoted as $(i_1 i_2 \dots i_m)$. A *sequence* is an ordered list of itemsets, denoted as $\langle I_1 I_2 \dots I_k \rangle$. A *sequence database* is generally a large set of sequences. Given two sequences $s = \langle I_1 I_2 \dots I_m \rangle$ and $s' = \langle {I'}_1 {I'}_2 \dots {I'}_n \rangle$, if there exist integers $1 \leq i_1 < i_2 < \dots < i_m \leq n$ such that $I_1 \subseteq {I'}_{i_1}, I_2 \subseteq {I'}_{i_2}, \dots, I_m \subseteq {I'}_{i_m}$, then s is a *subsequence* of s', and s' is a *supersequence* of s, denoted as $s \sqsubseteq s'$; further, if $i_m - i_1 = m - 1$, then we say that s is a *consecutive subsequence* of s', denoted as $s \sqsubseteq_c s'$. Denote by $I \in s$ an itemset I contained in a sequence s. For two sequences s and s', if $s \sqsubseteq s'$, then we say that s *is included in* s', or s' *supports* s. In particular, we denote the first itemset in a sequence s as $s^\top$ and the last itemset as $s_\perp$. Thus, given two sequences s and s', we note $s \sqsubseteq^\top s'$ if $s^\top \sqsubseteq {s'}^\top$, $s \sqsubseteq_\perp s'$ if $s_\perp \sqsubseteq {s'}_\perp$, and $s \sqsubseteq_\perp^\top s'$ if $s^\top \sqsubseteq {s'}^\top$ and $s_\perp \sqsubseteq {s'}_\perp$.

The *support* of a sequence s in a sequence database $\mathcal{D}$, denoted as $\sigma(s, \mathcal{D})$, is the fraction of the total number of sequences in $\mathcal{D}$ that support s. Given a minimum

frequency threshold *minimum support*, denoted as σ_{min}, a sequence s is *frequent* if $\sigma(s, \mathcal{D}) \geq \sigma_{min}$. In a sequence database $\mathcal{D}$, if a sequence s is not a subsequence of any other sequence $s' \in \mathcal{D}$, then the sequence s is *maximal*.

The *length* of a sequence s is the number of itemsets contained in this sequence, denoted as $|s|$. The *size* of a sequence s is the total number of items contained in this sequence, denoted as $\|s\|$. An *empty sequence* is denoted as $\emptyset$, we have $s = \emptyset \iff |s| = 0$. The *concatenation* of sequences is denoted as the form $s_1 \cdot s_2$, we have $|s_1 \cdot s_2| = |s_1| + |s_2|$ and $\|s_1 \cdot s_2\| = \|s_1\| + \|s_2\|$. For example, let sequences $s_1 = \langle (a)(b)(c) \rangle$ and $s_2 = \langle (ab)(a) \rangle$, we have $|s_1| = 3$, $\|s_1\| = 3$, $|s_2| = 2$, $\|s_2\| = 3$, $s_1 \cdot s_2 = \langle (a)(b)(c)(ab)(a) \rangle$, $|s_1 \cdot s_2| = 5$, and $\|s_1 \cdot s_2\| = 6$.

3.2. *Belief System on Sequence Data*

In order to find unexpectedness in sequence data, we proposed a belief system on sequences in our previous work[22,23].

A belief on sequences consists of a sequence rule, an occurrence constraint and a semantic contradiction. The rule is defined in the form $s_\alpha \rightarrow s_\beta$, which depicts that given a sequence s, the presence of $s_\alpha \sqsubseteq s$ implies $s_\alpha \cdot s_\beta \sqsubseteq s$. If the implication is satisfied, then we say that the sequence s *supports* the rule $s_\alpha \rightarrow s_\beta$, denoted as $s \models (s_\alpha \rightarrow s_\beta)$.

Let $\tau = [min..max]$ $(min, max \in \mathbb{N}$ and $min \leq max)$ be a constraint on the length of a sequence, that is, given a sequence s, if $min \leq |s| \leq max$, then s satisfies the constraint, denoted as $|s| \models \tau$. Given a sequence $s = s_1 \cdot s' \cdot s_2$, where the length of s' satisfies a constraint τ, that is, $|s'| \models \tau$, then the constraint τ also constrains the occurrences of s_1 and s_2 in the sequence s. Hence, we call the constraint τ the *occurrence constraint* on sequences. With the occurrence constraint, we extend the rule $s_\alpha \rightarrow s_\beta$ to $s_\alpha \rightarrow^\tau s_\beta$, which represents the following relation:

$$(s_\alpha \rightarrow^\tau s_\beta) \Rightarrow (s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s) \wedge (|s'| \models \tau).$$

We call such a rule with occurrence constraint an *occurrence rule*. Notice that given an occurrence constraint $\tau = [min..max]$, when max is not specified (i.e., any integer k such that $k \geq min$), we note $\tau = [min..*]$. In particular cases, for $min = max = 0$, we note $\tau = 0$ and the rule $s_\alpha \rightarrow^0 s_\beta$; for $min = 0$ and $max = *$, we note $\tau = *$ and the rule $s_\alpha \rightarrow^* s_\beta$.

Given two sequences s_1 and s_2, the *semantic contradiction* is denoted as $s_1 \not\approx_{sem} s_2$, which depicts that the sequences s_1 and s_2 semantically contradict each other. A semantic contradiction can be applied onto an occurrence rule for stating that the occurrence of the sequence s_β should not be replace by the occurrence of a sequence s_γ that semantically contradicts s_β. Since the rule $s_\alpha \rightarrow s_\beta$ can be interpreted as the implication $(s_\alpha \sqsubseteq s) \Rightarrow (s_\alpha \cdot s_\beta \sqsubseteq s)$, according to $s_\beta \not\approx_{sem} s_\gamma$ we have the implication $(s_\alpha \sqsubseteq s) \not\Rightarrow (s_\alpha \cdot s_\gamma \not\sqsubseteq s)$. Therefore, considering the occurrence rule with a semantic contradiction, we have the following relation:

$$\{s_\alpha \rightarrow^\tau s_\beta\} \wedge \{s_\beta \not\approx_{sem} s_\gamma\} \Rightarrow (s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s) \wedge (s_\alpha \cdot s' \cdot s_\gamma \not\sqsubseteq_c s) \wedge (|s'| \models \tau),$$

which drives to the definition of belief on sequences.

Definition 1. A *belief* on sequences consists of an occurrence rule $s_\alpha \rightarrow^\tau s_\beta$ and a semantic contradiction $s_\beta \not\approx_{sem} s_\gamma$, where $\tau = [min..max]$ $(min, max \in \mathbb{N}$ and $min \leq max)$ and s_β semantically contradicts s_γ, denoted as $\{s_\alpha \rightarrow^\tau s_\beta\} \wedge \{s_\beta \not\approx_{sem} s_\gamma\}$. If a sequence s satisfies a belief b, denoted as $s \models b$, then we have that $s_\alpha \sqsubseteq s$ implies $s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s$ and $s_\alpha \cdot s' \cdot s_\gamma \not\sqsubseteq_c s$, where $|s'| \models \tau$. □

The task of mining unexpected sequences is to find all the sequences that violate a given set of beliefs in databases.

4. Fuzzy Occurrence of Unexpected Sequences

An *unexpected sequence* is a sequence that violates a belief. In our previous work, the unexpectedness is stated by the violation of the occurrence rule or the semantic contradiction contained in a belief. According to the structure of an occurrence rule, three forms of *unexpected sequences* can be defined as follows.

Definition 2. Given a belief $b = \{s_\alpha \rightarrow^\tau s_\beta\} \wedge \{s_\beta \not\approx_{sem} s_\gamma\}$ and a sequence s where $s_\alpha \sqsubseteq s$:

(1) if $\tau = *$ and there does not exist $s_\beta \sqsubseteq s$ such that $s_\alpha \cdot s_\beta \sqsubseteq s$, then s is an *α-unexpected sequence*, denoted as $s \not\models_\alpha b$;
(2) if $\tau \neq *$ and there exists $s_\beta \sqsubseteq s$ such that $s_\alpha \cdot s_\beta \sqsubseteq s$, and there does not exist $s' \sqsubseteq s$ such that $|s'| \models \tau$ and $s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s$, then s is a *β-unexpected sequence*, denoted as $s \not\models_\beta b$;
(3) if there exists $s_\gamma \sqsubseteq s$ such that $s_\alpha \cdot s_\gamma \sqsubseteq s$ and there exists $s' \sqsubseteq s$ such that $|s'| \models \tau$ and $s_\alpha \cdot s' \cdot s_\gamma \sqsubseteq_c s$, then s is a *γ-unexpected sequence*, denoted as $s \not\models_\gamma b$. □

An unexpected sequence is named by the primary factor that causes unexpectedness: according to an belief $\{s_\alpha \rightarrow^\tau s_\beta\} \wedge \{s_\beta \not\approx_{sem} s_\gamma\}$, an α-unexpected sequence is unexpected because the occurrence of s_β is missing when $\tau = *$, where s_α is the only factor; a β-unexpected sequence is unexpected because the occurrence of s_β violates the constraint τ, thus, s_β is the primary factor of unexpectedness; a γ-unexpected sequence is unexpected because the occurrence of s_γ violates the semantic contradiction $s_\beta \not\approx_{sem} s_\gamma$, in this case, s_γ is the primary factor of unexpectedness. The three forms of unexpectedness is respectively called *α-unexpectedness*, *β-unexpectedness*, and *γ-unexpectedness*.

Example 1. Let us consider a belief on event sequences, where the numbers $\mathtt{11}, \mathtt{12}, \ldots, \mathtt{21}, \mathtt{22}, \ldots, \mathtt{31}, \mathtt{32}, \ldots$ stand for unique event IDs:

$$b = \{\langle(\mathtt{11})\rangle \rightarrow^{[0..2]} \langle(\mathtt{21})\rangle\} \wedge \{\langle(\mathtt{21})\rangle \not\approx_{sem} \langle(\mathtt{31})\rangle\}.$$

This belief b requires the occurrence of event `11` followed by an occurrence of event `21`, but not of event 31, within no more than two intervals. Thus, the event sequence $s = \langle(\texttt{12})(\texttt{22})(\texttt{12})(\texttt{11})(\texttt{12})(\texttt{11})(\texttt{12})(\texttt{21})(\texttt{31})(\texttt{12})\rangle$ is β-unexpected to belief b. The structure of unexpected sequence s are shown in Figure 2. □

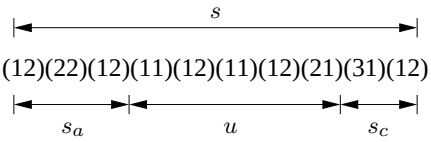

Fig. 2. Structure of an unexpected sequence.

Let u denote the unexpected part of an unexpected sequence s, then the β-unexpectedness and γ-unexpectedness can be represented as the forms $u = s_\alpha \cdot s_d \cdot s_\beta$ and $u = s_\alpha \cdot s_d \cdot s_\gamma$. The satisfiability of the constraint τ between s_α, s_β or s_γ within s can therefore be determined by examining the length of the sequence s_d. In order to handle the fuzzy unexpectedness on the occurrence constraint τ of a belief $b = \{s_\alpha \rightarrow^\tau s_\beta\} \wedge \{s_\beta \not\simeq_{sem} s_\gamma\}$ in a sequence s, we partition the satisfiability of τ between s_α, s_β or s_γ within s into various fuzzy sets, and the best membership degree of the β-unexpectedness and γ-unexpectedness can be determined by a fuzzy membership function $\mu_\tau(|s_d|, \tau, \mathcal{F})$, where $\mathcal{F}$ is a set of fuzzy partitions. We call μ_τ the *fuzzy occurrence degree*. For instance, Figure 3 shows an example about the fuzzy partitions for the β-unexpectedness when $\tau = [0..5]$. Clearly, according to Definition 2, there does not exist any fuzziness in the α-unexpectedness.

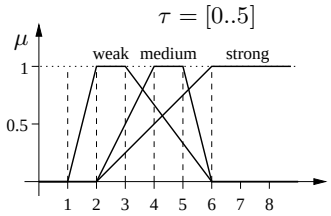

Fig. 3. Fuzzy partitions for the β-unexpectedness with $\tau = [0..5]$.

Example 2. We consider a belief on Web site log files, where `home`, `login`, and `logout` stand for the URL resources visited in a user session:

$$b = \{\langle(\texttt{home})\rangle \rightarrow^{[0..5]} \langle(\texttt{login})\rangle\} \wedge \{\langle(\texttt{login})\rangle \not\simeq_{sem} \langle(\texttt{logout})\rangle\}.$$

We consider three fuzzy sets for the each unexpectedness, they are "weak unexpected" (F_w), "medium unexpected" (F_m) and "strong unexpected" (F_s). In a sequence $s = \langle(\texttt{home})(\texttt{ad1})(\texttt{ad2})(\texttt{ad3})(\texttt{ad4})(\texttt{login})\rangle$, we have $|(\texttt{ad1})(\texttt{ad2})(\texttt{ad3})(\texttt{ad4})| = 4$.

Let $\mathcal{F} = \{F_w, F_m, F_s\}$ According to the fuzzy membership functions shown in Figure 3, we have that $\mu_\tau(4, \tau, \mathcal{F}) = 0.67{:}F_w$, $\mu_\tau(4, \tau, \mathcal{F}) = 1{:}F_m$ and $\mu_\tau(4, \tau, \mathcal{F}) = 0.5{:}F_s$, so that the best description of the sequence s is "medium unexpected". □

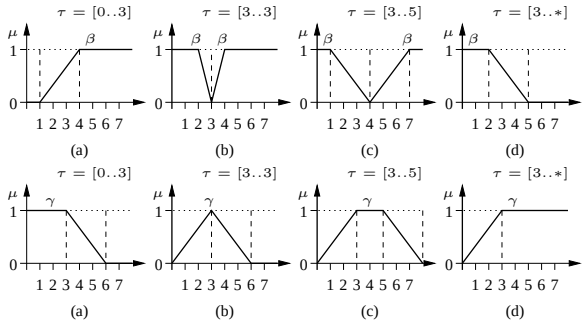

Fig. 4. Fuzzy measure of the "strong unexpected" for the β-unexpectedness and γ-unexpectedness.

For more details of the fuzziness on the occurrence constraint τ, Figure 4 represents "strong unexpected" for β-unexpectedness and γ-unexpectedness with (a) $\tau = [0..3]$, (b) $\tau = [3..3]$, (c) $\tau = [3..5]$ and (d) $\tau = [3..*]$.

5. Soft Belief with Concept Hierarchies

In this section, we extend the notion of belief on sequences to soft belief with concept hierarchies.

A *concept* is a cognitive unit of knowledge, and a group of semantically related concepts can be represented as a hierarchy, defined as follows.

Definition 3. A concept hierarchy $\mathcal{H} = (\mathcal{C}, \preceq)$ of concepts is a finite set $\mathcal{C}$ of concepts and a partial order $\preceq$ on $\mathcal{C}$.

In this definition, the partial order $\preceq$ is a *specialization/generalization relation* on the concepts in the set $\mathcal{C}$. For two concepts $c_\varphi, c_\theta \in \mathcal{C}$, if $c_\varphi \preceq c_\theta$, then we say that the concept c_φ is more general than the concept c_θ, and we also say that the concept c_θ is more specific than the concept c_φ. We write $c_\varphi \prec c_\theta$ if $c_\varphi \preceq c_\theta$ and not $c_\theta \preceq c_\varphi$.

Given a concept hierarchy $\mathcal{H} = (\mathcal{C}, \preceq)$, denote by $c \in \mathcal{H}$ the concept $c \in \mathcal{C}$. A *concept pattern* is an unordered collection $C = (c_1 c_2 \dots c_m)$ of distinct concepts sorted by lexical order, where c_i is a concept and for any $c_i \neq c_j$, $c_i \not\preceq c_j$. A *concept sequence* is an ordered list $S = \langle C_1 C_2 \dots C_k \rangle$ of concept patterns, where C_i is a concept pattern. Denote $C \in S$ a concept pattern contained in a concept sequence S. The specialization relation $\preceq$ can be applied to concept patterns and concept sequences. Given two concept patterns C and C', if for each concept $c \in C$

there exists a distinct concept $c' \in C'$ such that $c \preceq c'$, then we say that the concept pattern C is more general than the concept pattern C' (and C' is more specific than C), denoted as $C \preceq C'$. Given two k-length concept sequences $S = \langle C_1 C_2 \dots C_k \rangle$ and $S' = \langle C_1{}' C_2{}' \dots C_k{}' \rangle$, if for each concept pattern C_i and $C_i{}'$ $(1 \leq i \leq k)$, we have that $C_i \preceq C_i{}'$, than we say that the concept sequence S is more general than the concept sequence S' (and S' is more specific than S), denoted as $S \preceq S'$.

Given a sequence database $\mathcal{D}$ and a concept hierarchy $\mathcal{H}$, each item $i \in \mathcal{D}$ belongs to a concept $c \in \mathcal{H}$, denoted as $i \models c$; if $i \models c_\theta$ and $c_\varphi \preceq c_\theta$, then $i \models c_\varphi$. Let I be an itemset and C be a concept pattern, if for each $i \in I$ there exist a distinct concept $c \in C$ such that $i \models c$, then we say that the itemset I *supports* the concept pattern C, denoted as $I \models C$. Let $S = \langle C_1 C_2 \dots C_m \rangle$ be a concept sequence on $\mathcal{H}$ and $s = \langle I_1 I_2 \dots I_n \rangle$ be a sequence in $\mathcal{D}$, if there exist integers $1 \leq i_1 < i_2 < \dots < i_m \leq n$ such that $I_{i_1} \models C_1, I_{i_2} \models C_2, \dots, I_{i_m} \models C_m$, then we say that the sequence s *supports* the concept sequence S, denoted as $s \models S$.

An occurrence rule of concept sequence is in the form $S_\alpha \rightarrow^\tau S_\beta$, where S_α, S_β are two concept sequences and $\tau = [min..max]$ is a constraint such that $min, max \in \mathbb{N}$ and $min \leq max$. Given a sequence s, if there exists a sequence s' such that $|s'| \models \tau$ and there exist sequences ${s_\alpha}', {s_\beta}' \sqsubseteq s$ such that ${s_\alpha}' \models S_\alpha$, $|{s_\alpha}'| = |S_\alpha|$, ${s_\beta}' \models S_\beta$, $|{s_\beta}'| = |S_\beta|$, and ${s_\alpha}' \cdot s' \cdot {s_\beta}' \sqsubseteq s$, then we say that the sequence s *supports* the rule $S_\alpha \rightarrow^\tau S_\beta$, denoted as $s \models (S_\alpha \rightarrow^\tau S_\beta)$.

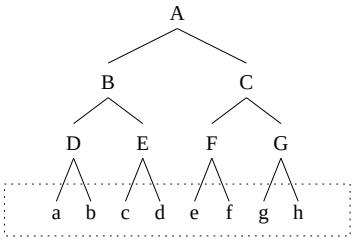

Fig. 5. A concept hierarchy with items.

Example 3. Figure 5 shows a concept hierarchy of concepts and associated items. We have $A \prec B$, $A \prec C$, $B \prec D$, $B \prec E$, $C \prec F$, $C \prec G$, $\{a, b\} \models D$, $\{c, d\} \models E$, $\{e, f\} \models F$, and $\{g, h\} \models G$. With this hierarchy, given a concept occurrence rule $\langle (D) \rangle \rightarrow^* \langle (E)(EF) \rangle$ and a sequence $s = \langle (a)(b)(c)(de) \rangle$, we have $s \models (\langle (D) \rangle \rightarrow^* \langle (E)(EF) \rangle)$ since we have $a \models D$ (or $b \models D$), $c \models E$, and $(de) \models (EF)$.

We now discuss the semantic contradiction relation on concepts and concept sequences. The relation $\preceq$ is monotone to the semantic contradiction relation: for all $c_\varphi, c_\phi \in \mathcal{C}$, if $c_\varphi \not\preceq_{sem} c_\phi$ and $c_\varphi \preceq c_\theta$, then $c_\theta \not\preceq_{sem} c_\phi$; if $c_\varphi \not\preceq_{sem} c_\phi$ and $i_\varphi \models c_\varphi$, $i_\phi \models c_\phi$, then $i_\theta \not\preceq_{sem} i_\phi$. In the same manner, given concept sequences S_φ, S_ϕ, and S_θ, the semantic contradiction relation on concept sequences determines

that, if $S_\varphi \not\approx_{sem} S_\phi$ and $S_\varphi \preceq S_\theta$, then $S_\theta \not\approx_{sem} S_\phi$; if $S_\varphi \not\approx_{sem} S_\phi$ and $s_\varphi \models S_\varphi$, $s_\phi \models S_\phi$, then $s_\varphi \not\approx_{sem} s_\phi$.

We also call a concept sequence a *generalized sequence* and call a occurrence rule of concept sequence a *generalized occurrence rule*. With a concept hierarchy, a soft belief on sequences can therefore be defined as follows.

Definition 4. A *soft belief* on sequences consists of a generalized occurrence rule $S_\alpha \rightarrow^\tau S_\beta$ associated with a concept hierarchy $\mathcal{H}$, where $\tau = [min..max]$ $(min, max \in \mathbb{N}$ and $min \leq max)$, denoted as $\{S_\alpha \rightarrow^\tau S_\beta\} \wedge \{\mathcal{H}\}$. □

We discuss the satisfaction and violation of a soft belief in the next section within the notions of fuzzy unexpected sequences.

6. Fuzzy Unexpected Sequences with Soft Beliefs

In this section, we present the fuzzy unexpected sequences with respect to the soft beliefs and concept hierarchies. With soft beliefs, we consider the fuzziness in unexpected sequences on both of the occurrence and semantics, with respect to given concept hierarchies.

Let us consider again the instance on Web usage analysis addressed in Section 1, where a generalized occurrence rule can be defined as $\langle(\texttt{/})\rangle \rightarrow^{[0..5]} \langle(\texttt{Politics})\rangle$. For example, to build a belief with "`technology news` semantically contradicts `politics news`", the semantic contradiction $\langle(\texttt{Politics})\rangle \not\approx_{sem} \langle(\texttt{Technology})\rangle$ is necessary. However, depending on user class and the hierarchy shown in Figure 6, not only the `technology news` contradicts `politics news`.

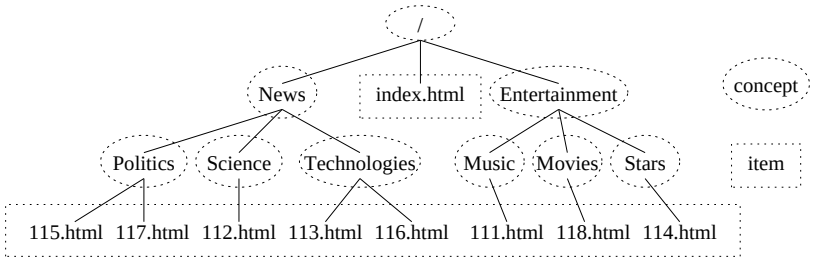

Fig. 6. A concept hierarchy based on Web structure.

The semantic contradiction of two concepts in a hierarchy is determined by the distance and semantic similarity between the concepts. Given a concept hierarchy $\mathcal{H}$ and two concepts $c_i, c_j \in \mathcal{H}$, the *semantic distance* between the concepts c_i and c_j in the hierarchy $\mathcal{H}$ is denoted as $\delta(c_i, c_j, \mathcal{H})$; the *semantic similarity* is defined as a score $\lambda(c_i, c_j)$, where $0 \leq \lambda(c_i, c_j) \leq 1$. For two concepts, we have that the more distance the less importance for relatedness, and the less similarity the more contradiction. Therefore, we propose a simple formula for handling the semantic

contradiction degree between concepts, denoted as $\omega_{sem}(c_i, c_j, \mathcal{H})$, as following:

$$\omega_{sem}(c_i, c_j, \mathcal{H}) = \frac{2 - \lambda(c_i, c_j)}{\delta(c_i, c_j, \mathcal{H})}, \quad (1)$$

where the semantic distance between the concepts c_i and c_j is defined as the *path-length* (i.e., the number of edges) between the nodes c_i and c_j in the hierarchy $\mathcal{H}$, and if $c_i = c_j$, we define $\delta(c_i, c_j, \mathcal{H}) = 1$. In this formula, we have that $0 \leq \lambda(c_i, c_j) \leq 1$ if the semantic similarity between c_i and c_j is defined; otherwise, if the semantic similarity is not defined, we define $\lambda(c_i, c_j) = 1$, so that $\omega_{sem}(c_i, c_j, \mathcal{H})$ is the reciprocal value of the length-path between c_i and c_j in the hierarchy $\mathcal{H}$. In the case that $c_i = c_j$, we define $\lambda(c_i, c_j) = 2$, so that $\omega_{sem}(c_i, c_j, \mathcal{H}) = 0$.

Notice that we consider the semantic contradiction degree $\omega_{sem}(c_i, c_j, \mathcal{H})$ as a value $0 \leq \omega_{sem} < 1$, that excludes the case that $\delta(c_i, c_j, \mathcal{H}) = 1$ when $\lambda(c_i, c_j)$ is undefined.

Table 1. Semantic distance and similarity matrix (path-length : similarity).

	`Politics`	`Science`	`Technology`	`Music`	`Movie`	`Stars`
`Politics`	1:2	2:0.6857	2:0.7183	4:0.4270	4:0.3388	4:0.2996
`Science`	2:0.6857	1:2	2:0.6929	4:1	4:1	4:1
`Technology`	2:0.7183	2:0.9	1:2	4:1	4:1	4:1
`Music`	4:0.4270	4:1	4:1	1:2	2:0.5159	2:0.4274
`Movie`	4:0.3388	4:1	4:1	2:0.5159	1:2	2:0.3392
`Stars`	4:0.2996	4:1	4:1	2:0.4274	2:0.3392	1:2

Table 2. Semantic contradiction degrees between concepts.

$c_i : c_j$	$\delta(c_i, c_j, \mathcal{H})$	$\lambda(c_i, c_j)$	$\omega_{sem}(c_i, c_j, \mathcal{H})$
`Politics : Politics`	1	2	0
`Politics : Science`	2	0.6857	0.65715
`Politics : Technology`	2	0.7183	0.64085
`Politics : Music`	4	0.4270	0.39325
`Politics : Movies`	4	0.3388	0.4153
`Politics : Stars`	4	0.2996	0.4251
`Politics : /`	2	1	0.5
`Politics : News`	1	1	1^*

Example 4. With the hierarchy shown in Figure 6, we have the relations listed in 1, where the semantic similarity between concepts is determined by the JWSL library[34] (assume that the similarities between concepts `Science`, `Technology` and

`Music`, `Movie`, `Stars` are not defined). For instance, the path-length between concepts `Politics` and `Technology` is 2; between `Politics` and `Music` is 4. With the JWSL library we have that the similarity between the concepts `Politics` and `Technology` is 0.7183; between `Politics` and `Music` is 0.4270. Thus, according to Equation (1), the semantic contradiction degrees between `Politics` and other concepts are listed in Table 2, where ω_{sem} between `Politics` and `News` is excluded. □

Given a sequence s, a generalized sequence S, and a concept hierarchy $\mathcal{H}$, where for each concept c contained in S, we have that $c \in \mathcal{H}$. We determine the semantic contradiction degree between s and S on $\mathcal{H}$ in the following manner.

We first consider the *compatible form* constraint on a generalized sequence of concepts and a sequence of items, defined as follows.

Definition 5. Given a generalized sequence S and a sequence s, let $S = \langle C_1 C_2 \dots C_m \rangle$ and $s = \langle I_1 I_2 \dots I_n \rangle$. The *compatible form* is a constraint that there exist integers $1 \leq i_1 < i_2 < \dots < i_m \leq n$ such that $|C_1| \leq |I_{i_1}|, |C_2| \leq |I_{i_2}|, \dots, |C_m| \leq |I_{i_m}|$, denoted as $S \lhd s$, and denote by $S \trianglelefteq s$ the case $|C_1| = |I_{i_1}|, |C_2| = |I_{i_2}|, \dots, |C_m| = |I_{i_m}|$.

In order to determine the semantic contradiction between S and s, we require that $S \lhd s$. Now we consider the semantic contradiction between a concept pattern C and an itemset I (where $|C| \leq |I|$) on a hierarchy $\mathcal{H}$, denoted as $\omega_{pat}(C, I, \mathcal{H})$ and defined as follows. Let $\Omega(c_i, i_j, \mathcal{H}) = max\{\omega_{sem}(c_i, c_j, \mathcal{H}) \mid c_j \in \mathcal{H}, i_j \models c_j\}$ be the maximal semantic contradiction degree between a concept $c_i \in \mathcal{H}$ and an item $i_j \in I$, then the number of the combinations of $\Omega(c_i, i_j, \mathcal{H})$ on the elements in $c_i \in C$ and $i_j \in I$ is the number of permutations of $|C|$ items in I, that is,

$$P(|I|, |C|) = \frac{|I|!}{(|I| - |C|)!}. \tag{2}$$

Let $\mathcal{I}$ be the set of such permutations, we denote the semantic contradiction degree between a concept pattern C and an itemset I as:

$$\omega_{pat}(C, I, \mathcal{H}) = \frac{max\{\sum_{c_i \in C} \Omega(c_i, i_j, \mathcal{H}) \mid i_j \in I', I' \in \mathcal{I}\}}{|C|}. \tag{3}$$

Therefore, given a generalized sequence S and a sequence s, for all subsequences $s' \sqsubseteq s$ such that $S \trianglelefteq s'$, the semantic contradiction degree between S and s, denoted as $\omega_{seq}(S, s, \mathcal{H})$, is defined as the average of the sum of $\omega_{pat}(C_i, I_i, \mathcal{H})$ that is maximal, where C_i and I_i are itemsets contained in S and s', that is,

$$\omega_{seq}(S, s, \mathcal{H}) = \frac{max\{\sum_{1 \leq i \leq \|S\|} \omega_{pat}(C_i \in S, I_i \in s', \mathcal{H}) \mid s' \sqsubseteq s, S \trianglelefteq s'\}}{\|S\|}. \tag{4}$$

Respectively, we define the *semantic relatedness degree* between concepts, denote by $\eta_{sem}(c_i, c_j, \mathcal{H})$, as following:

$$\psi_{sem}(c_i, c_j, \mathcal{H}) = \frac{\lambda(c_i, c_j)}{\delta(c_i, c_j, \mathcal{H})}, \tag{5}$$

and let $\Psi(c_i, i_j, \mathcal{H}) = max\{\psi_{sem}(c_i, c_j, \mathcal{H}) \mid c_j \in \mathcal{H}, i_j \models c_j\}$, in the same manner with the permutation set $\mathcal{I}$ of a given itemset I with respect to a concept pattern C, we define the semantic relatedness degree between C and I as

$$\psi_{pat}(C, I, \mathcal{H}) = \frac{max\{\sum_{c_i \in C} \Psi(c_i, i_j, \mathcal{H}) \mid i_j \in I', I' \in \mathcal{I}\}}{|C|}. \quad (6)$$

Given a generalized sequence S and a sequence s, for all subsequences $s' \sqsubseteq s$ such that $S \trianglelefteq s'$, the semantic relatedness degree between S and s, denoted as $\psi_{seq}(S, s, \mathcal{H})$, is defined as the average of the sum of $\psi_{pat}(C_i, I_i, \mathcal{H})$ that is maximal, where C_i and I_i are itemsets contained in S and s', that is,

$$\psi_{seq}(S, s, \mathcal{H}) = \frac{max\{\sum_{1 \le i \le \|S\|} \psi_{pat}(C_i \in S, I_i \in s', \mathcal{H}) \mid s' \sqsubseteq s, S \trianglelefteq s'\}}{\|S\|}. \quad (7)$$

With the notions of semantic relatedness and contradiction degrees, we formally define the fuzzy unexpectedness of sequences with respect to soft beliefs on a concept hierarchy as follows.

Definition 6. Given a concept hierarchy $\mathcal{H}$, a soft belief $B = \{S_\alpha \rightarrow^\tau S_\beta\} \wedge \{\mathcal{H}\}$ where $\tau = [min..max]$ ($min, max \in \mathbb{N}$ and $min \le max$), a sequence s where there exists $s_\alpha \sqsubseteq s$ such that $s_\alpha \models S_\alpha$, a user defined minimum semantic contradiction degree ω_{min}, and a user defined minimal semantic relatedness degree ψ_{min}:

(1) $\tau = *$: if there does not exist $s_\beta \sqsubseteq s$ such that $s_\alpha \cdot s_\beta \sqsubseteq_c s$ and $\psi_{seq}(S_\beta, s_\beta, \mathcal{H}) \ge \psi_{min}$, then s is a *fuzzy α-unexpected sequence*, denoted as $s \not\models^{\sim}_{\alpha} B$;
(2) $\tau \neq *$: if there exist $s', s_\beta, s_\gamma \sqsubseteq s$ such that $|s'| \not\models \tau$, $s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s$, and $\psi_{seq}(S_\beta, s_\beta, \mathcal{H}) \ge \psi_{min}$, then s is a *fuzzy β-unexpected sequence*, denoted as $s \not\models^{\sim}_{\beta} B$;
(3) if there exist $s', s_\gamma \sqsubseteq s$ such that $|s'| \models \tau$, $s_\alpha \cdot s' \cdot s_\gamma \sqsubseteq_c s$, and $\omega_{seq}(S_\gamma, s_\gamma, \mathcal{H}) \ge \omega_{min}$, then s is a *fuzzy γ-unexpected sequence*, denoted as $s \not\models^{\sim}_{\gamma} B$. □

The fuzzy unexpectedness on semantic relatedness and contradiction can be partitioned into different fuzzy partitions, like "*weak relatedness/contradiction*", "*medium relatedness/contradiction*", or "*strong relatedness/contradiction*" with respect to β-unexpected or γ-unexpected sequences, by fuzzy membership functions $\mu_{sem}(\psi_{seq}, \mathcal{F})$ or $\mu_{sem}(\omega_{seq}, \mathcal{F})$, where $\mathcal{F}$ is a set of fuzzy partitions.

Example 5. For instance, given a soft belief $\{\langle(/)\rangle \rightarrow^* \langle(\texttt{Politics})(\texttt{Movies})\rangle\} \wedge \{\mathcal{H}\}$ corresponding to the hierarchy shown in Figure 6, the sequence $\langle(\texttt{index})(\texttt{117})(\texttt{118})\rangle$ (we ignore file extensions) is an expected sequence. Given a minimum semantic contradiction degree 0.3, according to Table 1, the sequence $\langle(\texttt{index})(\texttt{112})(\texttt{113})\rangle$ is fuzzy γ-unexpected with "medium contradiction" since we have that

$$\omega_{seq}(\langle(\texttt{Politics})(\texttt{Movies})\rangle, \langle(\texttt{112})(\texttt{113})\rangle, \mathcal{H}) = 0.456;$$

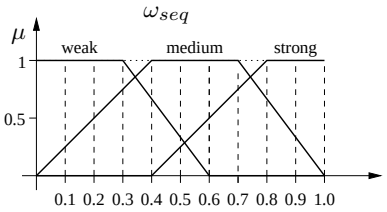

Fig. 7. Fuzzy partitions for the semantic contradiction degree ω_{seq}.

the sequence $\langle(\texttt{index})(\texttt{114})(\texttt{113})\rangle$ is fuzzy γ-unexpected with "weak contradiction" since we have that

$$\omega_{seq}(\langle(\texttt{Politics})(\texttt{Movies})\rangle, \langle(\texttt{114})(\texttt{113})\rangle, \mathcal{H}) = 0.338. \ \square$$

7. *FUSE*: Fuzzy Unexpected Sequence Extraction

In this section, we present the algorithm *FUSE* (Fuzzy Unexpected Sequence Extraction) for extracting fuzzy unexpected sequences in a sequence database. We first introduce the main routine of *FUSE*, then detail the algorithm *HyMatch* (Hierarchy Matching) for computing the semantic relatedness/contradcition degree between a generalized sequence of concepts and a sequence of items.

Notice that in the algorithms we consider a sequence as an object with the properties ψ_{seq}, ω_{seq}, μ_{sem}, μ_{τ}, etc., which correspond to the notions presented in previous sections.

7.1. *Algorithm FUSE*

The main routine of the algorithm *FUSE* is listed in Algorithm 1, which extracts fuzzy unexpected sequences in a sequence database $\mathcal{D}$, with respect to a soft belief set $\mathcal{B}$, a concept hierarchy $\mathcal{H}$, a minimum semantic relatedness degree ψ_{min} and a minimum semantic contradiction degree ω_{min}. The extracted sequences are associated with the best fuzzy degree specified by a fuzzy partition set $\mathcal{F}$ with respect to a minimum occurrence degree $\mu_{\tau_{min}}$.

For each sequence $s \in \mathcal{D}$ and each belief $B \in \mathcal{B}$ as $\{S_\alpha, S_\beta, S_\gamma, \tau\}$, the algorithm first verifies the satisfaction of $s \models S_\alpha$: if not, s cannot be unexpected with respect to B. Then, the algorithm tries to match S_β in the rest of s (from the end of s_α till to the end of s, denote by $s - s_\alpha$ as at line 5), with respect to the occurrence constraint τ. The routine *HyMatch* (see the next section for details) for matches the correspondence of S_β in the rest of s with $s_\beta.\psi_{seq} \geq \psi_{min}$. If the ψ_{seq} property of the returned sequence equals to -1, then matching failed and s cannot be α-unexpected neither β-unexpected; otherwise, if $\tau = *$, then outputs s as α-unexpected, else outputs s as β-unexpected. Finally, the algorithm calls *HyMatch* for matching the correspondence of S_γ in the rest of s with $s_\gamma.\omega_{seq} \geq \omega_{max}$: if the ω_{seq} property of the returned sequence does not equal to -1, then outputs s as γ-unexpected.

Algorithm 1: The algorithm *FUSE*.

Input : $\mathcal{D}$, $\mathcal{B}$, $\mathcal{H}$, $\mathcal{F}$, ψ_{min}, ω_{min}, $\mu_{\tau_{min}}$
Output : all fuzzy unexpected sequences

```
1  foreach s ∈ D do
2      foreach B ∈ B as {S_α, S_β, S_γ, τ} do
3          if ∃s_α ⊑ s such that s_α ⊨ S_α then
4              s_β := HyMatch(S_β, s − s_α, H, τ, ψ_min, 0, F, μ_{τ_min});
5              if s_β.ψ_seq ≠ −1 then
6                  s.μ_sem := s_β.μ_sem;
7                  s.μ_τ := s_β.μ_τ;
8              if τ = ∗ then
9                  output s to {s | s ⊮~_α B};
10             else
11                 output s to {s | s ⊮~_β B};
12             s_γ := HyMatch(S_γ, s − s_α, H, τ, 0, ω_min, F, μ_{τ_min});
13             if s_γ.ω_seq ≠ −1 then
14                 s.μ_sem := s_γ.μ_sem;
15                 s.μ_τ := s_γ.μ_τ;
16                 output s to {s | s ⊮~_γ B};
17         else
18             continue;
```

The fuzzy unexpectedness of the occurrence constraint τ and the semantic relatedness/contradiction is handled by the membership functions μ_τ and μ_{sem} within the routine *HyMatch*.

7.2. *Algorithm HyMatch*

Given a generalized sequence S, a sequence s, a concept hierarchy $\mathcal{H}$, the Algorithm *HyMatch* finds the first highest-scored subsequence $s' \sqsubseteq s$ such that $s' \models S$ with respect to an occurrence constraint τ, a minimum semantic relatedness degree ψ_{min} or a minimum semantic contradiction degree ω_{min}. A set $\mathcal{F}$ of fuzzy partitions is also token into account for handling the fuzzy degrees of μ_τ and μ_{sem}, with respect to a minimum occurrence degree $\mu_{\tau_{min}}$.

The algorithm first verifies the compatible form constraint on S and s, if not $S \lhd s$, then returns an empty sequence (line 3); if $S \lhd s$, the function $seqsat(S, s, \lhd)$ returns the set $\mathcal{S}$ of all maximal subsequences (i.e., without splitting itemsets) of $s'' \sqsubseteq s$ such that $S \lhd s''$ and $|s''| = |S|$. All sequences $s'' \in \mathcal{S}$ that cannot satisfy the constraint τ are removed (line 6). Not difficult to see, the sequence $s'' \in \mathcal{S}$ having the maximal semantic relatedness degree $max\{\psi_{seq}(S, s'', \mathcal{H})\}$ or contradiction degree $max\{\omega_{seq}(S, s'', \mathcal{H})\}$ is also the sequence $s'' \sqsubseteq s$ having the same maximal degree such that $S \trianglelefteq s''$. The algorithm uses the equations proposed in the previous sections

Algorithm 2: The algorithm *HyMatch*.

Input : S, s, $\mathcal{H}$, ψ_{min}, ω_{min}, $\mathcal{F}$, $\mu_{\tau_{min}}$
Output : first highest-scored subsequence $s' \sqsubseteq s$ such that $s' \models S$

```
1  s' := empty_sequence; s'.ψ_seq := −1; s'.ω_seq := −1;
2  if not S ◁ s then
3      return s';
4  M := ∅;
5  S := seqsat(S, s, ◁);
6  S := S \ {s'' | μ_τ(|s''| − |S|, τ, F) < μ_τmin, s'' ∈ S};
7  if ψ_min > 0 and ω_min = 0 then
8      foreach s'' ∈ S do
9          s''.ψ_seq := max{ψ_seq(S, s'', H)};      /* use Equation (5), (6), (7) */
10         s''.ω_seq := −1;
11         if τ = ∗ then
12             if s''.ψ_seq ≱ ψ_min then
13                 M := M ∪ s'';
14         else
15             s''.μ_τ := μ_τ(s''.dist, τ, F);
16             if s''.ψ_seq ≥ ψ_min and s''.μ_τ ≥ μ_τmin then
17                 M := M ∪ s'';
18 else if ψ_min = 0 and ω_min > 0 then
19     foreach s'' ∈ S do
20         s''.ω_seq := max{ω_seq(S, s'', H)};      /* use Equation (1), (3), (4) */
21         s''.ψ_seq := −1;
22         s''.μ_τ := μ_τ(s''.dist, τ, F);
23         if s''.ω_seq ≥ ω_min and s''.μ_τ ≥ μ_τmin then
24             M := M ∪ s'';
25 if M ≠ ∅ then
26     hs := max{abs(s''.μ_τ ∗ s''.ψ_min ∗ s''.ω_max) | s'' ∈ M};      /* highest-score */
27     foreach s'' ∈ M do
28         if abs(s''.μ_τ ∗ s''.ψ_min ∗ s''.ω_max) = hs then
29             return s' := s'';
30 return s';
```

by examining the values of ψ_{min} and ω_{min}: if $\psi_{min} > 0$ and $\omega_{min} = 0$, then compute the semantic relatedness degree of each sequence $s'' \in \mathcal{S}$ for further determining α-unexpected or β-unexpected sequence; if $\psi_{min} = 0$ and $\omega_{min} > 0$, then compute the semantic contradiction degree of each sequence $s'' \in \mathcal{S}$ for further determining γ-unexpected sequence. If the ψ_{seq} or ω_{seq} value of a sequence $s'' \in \mathcal{S}$ satisfies the required condition, and the fuzzy occurrence degree $s''.\mu_\tau \geq \mu_{\tau_{min}}$, then s'' is added to the candidate sequence set $\mathcal{M}$, where $s''.dist$ (line 14 and 21) is the offset of s'' in s, which must correspond to specified occurrence constraint τ.

As shown in Equation (2), totally $P(|I|, |C|)$ queries are needed for computing $\omega_{pat}(C, I, \mathcal{H})$ or $\psi_{pat}(C, I, \mathcal{H})$ of a concept patter C and an itemset I on a hierarchy $\mathcal{H}$. If $|C| = |I|$, then totally $|I|!$ queries must be performed. Therefore, in the worst case, when $|S| = |s| = 1$ and $\|S\| = \|s\|$, totally $\|s\|!$ queries are required. The proof is immediate since we have that $(m+n)! \geq m! + n!$. In the best case, when $\|S\| = \|s\| = |S| = |s|$, that is, s consists of the itemsets of 1 item, $\|s\|$ queries are required. Therefore, for a sequence s such that $\|s\| = |s|$ and a generalized sequence S such that $S \lhd s$, the number of queries is the number of the combinations of $|S|$ itemsets in s, that is, ${}_{|s|}C_{|S|} = \begin{pmatrix} |s| \\ |S| \end{pmatrix} = \frac{|s|!}{|S|!(|s|-|S|)!}$. For instance, if $|s| = 10$ and $|S| = 5$, then totally $\begin{pmatrix} 10 \\ 5 \end{pmatrix} = 252$ queries are required.

8. Experiments

To evaluate our approach, we have performed a serial of experiments to extract fuzzy unexpected sequences from a large log file of an online forum Web server. The sequence database obtained from the Web access log file contains 67,228 sequences corresponding to 27,552 distinct items with average sequence length of about 14 itemsets consisting of 1 item. All experiments have been performed on a Sun Fire V880 system with 8 1.2GHz UltraSPARC III processors and 32GB main memory running Solaris 10 operating system.

First, we examine the fuzzy occurrence of unexpected sequences with 4 groups of 20 beliefs, which correspond to 4 categories of occurrence constraints. All the 20 beliefs are defined by domain experts: `CAT1` stands for 5 beliefs with $\tau = [0..*]$; `CAT2` stands for 5 beliefs with $\tau = [0..X]$ where $X \geq 0$ is an integer; `CAT3` stands for 5 beliefs with $\tau = [Y..*]$ where $Y > 0$ is an integer; and `CAT4` stands for 5 beliefs with $\tau = [X..Y]$ where $Y \geq X > 0$ are two integers. Figure 8 shows the numbers of unexpected sequences with minimum fuzzy occurrence degrees 0.7 and 0.2.

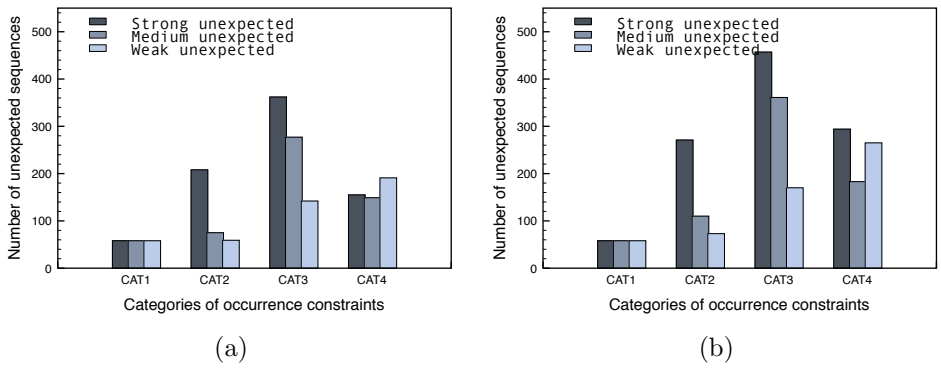

Fig. 8. (a) Minimum fuzzy occurrence degree 0.7. (b) Minimum fuzzy occurrence degree 0.2.

Then, we perform the tests on extracting fuzzy unexpected sequences with 20 soft beliefs, which are manually created from sequential patterns discovered in the data set with examining the concepts of items. The hierarchy used in experiments is built from the Web site structure and URI parameters, which contains 35 concepts with maximal path-length of 8, where the similarities between concepts are defined with expertise domain knowledge. An item-index file is used for mapping each item i to an concept c such that $i \models c$, and a concept-index file is used for indexing the path-length and semantic similarity between any two concepts contained in the hierarchy instead of traversing the hierarchy.

Only one category of occurrence constraint τ is considered with soft beliefs: $\tau = [X..Y]$ where $Y \geq X \geq 0$ are two integers. The soft beliefs are classified to 4 groups with respect to the length of S_β (1, 2, 4, 8), each group contains 5 soft beliefs. The length of S_α is no longer than 2. Since the fuzziness on semantic relatedness/contradiction is determined only by the degree, we did not specify the fuzzy partitions. In order to focus on the performance in considering hierarchies, the range $\tau \pm 2$ is used instead of computing the fuzzy occurrence degree.

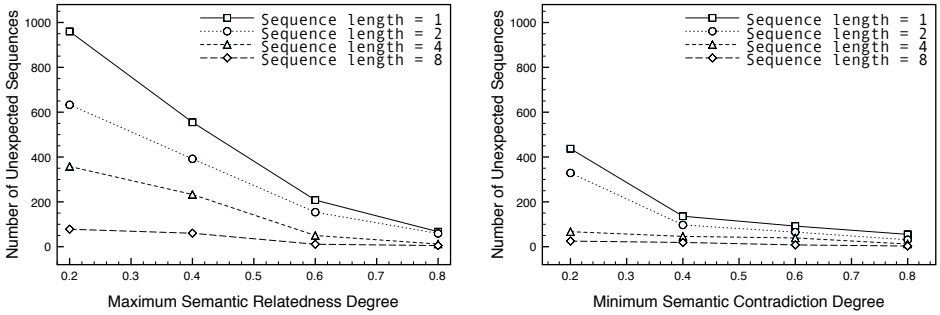

Fig. 9. (a) Fuzzy β-unexpected sequences. (b) Fuzzy γ-unexpected sequences.

Figure 9 shows the numbers of unexpected sequences extracted by using soft beliefs with concept hierarchy. The experimental results on soft beliefs show that the effectiveness of the proposed approach highly depends on the size of the sequence S_β in beliefs. For instance, when $|S_\beta| = 1$, the number of β-unexpected sequences extremely increases with decreasing the minimum semantic relatedness degree ψ_{min}. In fact, according to the combinations of items in a sequence, if $|S_\beta|$ is a small value, then there are higher probability to satisfy the semantic relatedness required for matching S_β. Thus, when $|S_\beta|$ is a small value, the probability to satisfy the semantic relatedness is much lower and much less unexpected sequences are extracted.

The execution time of each test is listed in Table 3, which shows that the time for extracting unexpected sequences significantly increases with the increase of $|S_\beta|$. However, the increase of execution time is slower than ${}_{14}C_1 \rightarrow {}_{14}C_2 \rightarrow {}_{14}C_4 \rightarrow {}_{14}C_8$

Table 3. Execution time of each test by using soft beliefs.

$\|S_\alpha\|$	$\psi_{min}, \omega_{min} : 0.2$	$\psi_{min}, \omega_{min} : 0.4$	$\psi_{min}, \omega_{min} : 0.6$	$\psi_{min}, \omega_{min} : 0.8$
1	22.1 s	22.1 s	20.4 s	19.2 s
2	93.1 s	90.2 s	93.8 s	90.7 s
4	577.8 s	563.3 s	581.8 s	569.7 s
8	2024.2 s	1998.8 s	1994.3 s	1955.2 s

because with the increase of $|S_\beta|$, the satisfaction of τ in the rest of an input sequence (i.e., $s - s_\alpha$ where $s_\alpha \models S_\alpha$) becomes lower, and the step at line 6 in Algorithm 2 avoids matching all combinations of subsequences.

Notice that when we consider the semantics, we can determine the semantic contradiction between two single items, for example, between "login" and "logout". However, for operational conjunction of items with temporal order, the semantic contradiction is hard to be defined, which is still an open problem in semantics data mining.

9. Conclusion

In this paper, we present a novel approach to the discovery of fuzzy unexpected sequences by using soft beliefs with concept hierarchies. In comparison with our previous approaches, the semantics in determining unexpected sequences can be determined by concept hierarchies, instead of specified by domain experts. We also extend the notion of unexpected sequences by unifying the occurrence constraint and semantic contradictions with soft beliefs. We develop the framework *FUSE*, which has been verified with real Web server log file analyzing and the usefulness of hierarchies is shown in the condition of short sequences.

Our future research includes the discovery of fuzzy unexpected sequences or rules in more general cases. For instance, if "`age` is `old` → `salary` is `high`" corresponds to prior knowledge, then "`age` is `young` → `salary` is `high`" or "`age` is `old` → `salary` is `low`" can be considered as unexpected. We concentrate also on studying semantics data mining with integrating fuzzy and natural language processing techniques.

References

1. R. Agrawal, T. Imielinski, and A.N. Swami, "Mining association rules between sets of items in large databases", *Proc. ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207–216.
2. R. Agrawal and R. Srikant, "Mining sequential patterns", *Proc. 11th International Conference on Data Engineering*, 1995, pp. 3–14.
3. J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential PAttern mining using a bitmap representation", *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 429–435.
4. G. Berger and A. Tuzhilin, "Discovering unexpected patterns in temporal data using temporal logic", *Temporal Databases*, Springer, 1998.

5. K.C.C. Chan and W.H. Au, "Mining Fuzzy Association Rules", *Proc. nternational Conference on Information and Knowledge Management*, 1997, pp. 209–215.
6. R.-S. Chen, G.-H. Tzeng, C. C. Chen, and Y.-C. Hu, "Discovery of fuzzy sequential patterns for fuzzy partitions in quantitative attributes", *Proc. ACS/IEEE International Conference on Computer Systems and Applications*, 2001, pp. 144–150.
7. Y.-L. Chen and T. C. K. Huang, "A new approach for discovering fuzzy quantitative sequential patterns in sequence databases", *Fuzzy Sets and Systems* **157** (2006) 1641–1661.
8. M. Delgado, N. Marín, D. Sánchez, and M. Vila, "Fuzzy association rules: general model and applications", *IEEE Transactions on Fuzzy Systems* **11** (2003) pp. 214–225.
9. G. Dong and J. Li, "Interestingness of discovered association rules in terms of neighborhood-based unexpectedness", *Proc. 2nd Pacific-Asia Conference*, 1998, pp. 72–86.
10. G. Dong and J. Pei, *Sequence Data Mining* (Springer, USA, 2007).
11. D. Dubois, E. Hüllermeier, and H. Prade, "A systematic approach to the assessment of fuzzy association rules", *Data Mining and Knowledge Discovery* **13** (2006) pp. 167–192.
12. C. Fiot, A. Laurent, and M. Teisseire, "From Crispness to Fuzziness: Three Algorithms for Soft Sequential Pattern Mining", *IEEE Transactions on Fuzzy Systems* **15** (2007) pp. 1263–1277.
13. C. Fiot, F. Masseglia, A. Laurent, and M. Teisseire, "Gradual Trends in Fuzzy Sequential Patterns", *Proc. Information Processing and Management of Uncertainty*, 2008, pp. 456–463.
14. J. Han and Y. Fu, "Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases", *Proc. Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop*, 1994, pp. 157–168.
15. J. Han and M. Kamber, *Data Mining: Concepts and Techniques* (Morgan Kaufmann Publishers, 2nd edition, 2006).
16. T.-P. Hong, K.-Y. Lin, and S.-L. Wang, "Fuzzy data mining for interesting generalized association rules", *Fuzzy Sets and Systems* **138** (2003) 255–269.
17. Y.-C. Hu, R.-S. Chen, G.-H. Tzeng, and J.-H. Shieh, "A fuzzy data mining algorithm for finding sequential patterns", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **11** (2003) 173–194.
18. S. Jaroszewicz and T. Scheffer, "Fast discovery of unexpected patterns in data, relative to a bayesian network", *Proc. 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005, pp. 118–127.
19. J.J. Jiang and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", *Proc. 17th International Conference on Computational Linguistics*, 1997.
20. C. M. Kuok, A. W.-C. Fu, and M. H. Wong, "Mining fuzzy association rules in databases", *SIGMOD Record* **27** (1998) 41–46.
21. J. Lee and H. Lee-kwang, "An extension of association rules using fuzzy sets", *Proc. International Fuzzy Systems Association World Congress*, 1997, pp. 399–402.
22. D. H. Li, A. Laurent, and P. Poncelet, "Mining unexpected sequential patterns and rules", *Technical Report RR-07027 (2007), Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier*, 2007.
23. D. H. Li, A. Laurent, and P. Poncelet, "Discovering fuzzy unexpected sequences with beliefs", *Proc. 12th International Conference on Processing and Management of Uncertainty in Knowledge-Based Systems*, 2008, pp. 1709–1716.
24. D. H. Li, A. Laurent, and P. Poncelet, "Mining unexpected Web usage behaviors", *Proc. 8th Industrial Conference on Data Mining*, 2008, pp. 283–297.

25. D.H. Li, A. Laurent, and P. Poncelet, "Recognizing unexpected recurrence behaviors with fuzzy methods in sequence databases", *Proc. 5th International Conference on Soft Computing as Transdisciplinary Science and Technology*, 2008, pp. 37–43.
26. F. Masseglia, F. Cathala, and P. Poncelet, "The PSP approach for mining sequential patterns", *Proc. Principles of Data Mining and Knowledge Discovery, 2nd European Symposium*, 1998, pp. 176–184.
27. K. McGarry, "A survey of interestingness measures for knowledge discovery", *Knowledge Engineering Review* **20** (2005) 39–61.
28. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth", *Proc. 17th International Conference on Data Engineering*, 2001, pp. 215–224.
29. A. Silberschatz and A. Tuzhilin, "On subjective measures of interestingness in knowledge discovery", *Proc. 1st International Conference on Knowledge Discovery and Data Mining*, 1995, pp. 275–281.
30. R. Srikant and R. Agrawal, "Mining sequential patterns: generalizations and performance improvements", *Proc. 5th International Conference on Extending Database Technology*, 1996, pp. 3–17.
31. B. Padmanabhan and A. Tuzhilin, "A belief-driven method for discovering unexpected patterns", *Proc. 4th International Conference on Knowledge Discovery and Data Mining*, 1998, pp. 94–100.
32. B. Padmanabhan and A. Tuzhilin, "Small is beautiful: discovering the minimal set of unexpected patterns", *Proc. 6th International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 54–63.
33. B. Padmanabhan and A. Tuzhilin, "On characterization and discovery of minimal unexpected patterns in rule discovery", *IEEE Transactions on Knowledge and Data Engineering* **18** (2006) 202–216.
34. G. Pirrò and N. Seco, "Design, Implementation and Evaluation of a New Similarity Metric Combining Feature and Intrinsic Information Content", *Proc. International Conference on Ontologies, DataBases, and Applications of Semantics*, 2008, pp. 1271–1288.
35. P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *Proc. 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.
36. M.A. Rodríguez and M.J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", *IEEE Transactions on Knowledge and Data Engineering* **15** (2003) 442–456.
37. M. Spiliopoulou, "Managing interesting rules in sequence mining", *Proc. Principles of Data Mining and Knowledge Discovery, 3rd European Conference*, 1999, pp. 554–560.
38. K. Wang, Y. Jiang, and L. V. S. Lakshmanan, "Mining unexpected rules by pushing user dynamics", *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 246–255.
39. J. Wang and J. Han, "BIDE: Efficient mining of frequent closed sequences", *Proc. 20th International Conference on Data Engineering*, 2004, pp. 79–90.
40. X. Yan, J. Han, and R. Afshar, "CloSpan: Mining closed sequential patterns in large databases", *Proc. 3rd SIAM International Conference on Data Mining*, 2003, pp. 166–177.
41. L. A. Zadeh, "Fuzzy sets", *Information and Control* **8** (1965) 338–353.
42. M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences", *Machine Learning* **42** (2001) 31–60.