Ranking Sequential Patterns with Respect to Significance

Robert Gwadera and Fabio Crestani

Universita della Svizzera Italiana
Lugano, Switzerland

Abstract. We present a reliable universal method for ranking sequential patterns (itemset-sequences) with respect to significance in the problem of frequent sequential pattern mining. We approach the problem by first building a probabilistic reference model for the collection of itemset-sequences and then deriving an analytical formula for the frequency for sequential patterns in the reference model. We rank sequential patterns by computing the divergence between their actual frequencies and their frequencies in the reference model. We demonstrate the applicability of the presented method for discovering dependencies between streams of news stories in terms of significant sequential patterns, which is an important problem in multi-stream text mining and the topic detection and tracking research.

1 Introduction

1.1 Motivation

Frequent sequential pattern mining, introduced in [1], has established itself as one of the most important data mining frameworks with broad applications including analysis of time-related processes, telecommunications, bioinformatics, business, software engineering, Web click stream mining, etc [9]. The problem is defined as follows. Given a collection of itemset-sequences (sequence database of transactions) and a minimum frequency (support) threshold, the task is to find all subsequence patterns, occurring across the itemset-sequences in the collection, whose frequency is greater than the minimum frequency threshold. The main focus of the research on sequential pattern mining has been on devising efficient algorithms for discovering frequent sequential patterns (see [9] for a review). Although state of the art mining algorithms can efficiently derive a complete set of frequent sequential patterns under certain constraints, the main problem is that the set of frequent sequential patterns is still too large for effective usage [9]. The two most effective methods for reducing the large set of frequent sequential patterns have been: *closed sequential pattern* mining [12] and *maximal sequential pattern* mining [5]. However no methods for assessing interestingness of sequential patterns have been proposed while such methods are very important to advance the applicability of frequent sequential pattern mining. By comparison, such methods have been proposed for subsequence patterns in the sliding window model [6] and for itemsets (see [11] for a review).

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 286–299, 2010.

1.2 Overview of the Method

We approach the problem by first building a probabilistic reference model for the collection of itemset-sequences and then deriving an analytical formula for the relative frequency for sequential patterns. Given such a model we discover sequential patterns that are *under-represented* or *over-represented* with respect to the reference model, where a pattern is under-represented if it is too infrequent in the input collection of itemset-sequences and a pattern is over-represented if it is too frequent in the input collection of itemset-sequence. According to this notion a sequential pattern is significant if the probability that it would occur by chance a specific number of times, in the reference model, is very small. Note that the frequency of occurrence alone is not enough to determine significance, i.e., an infrequent sequential pattern can be more significant than a frequent one. Furthermore an occurrence of a subsequence pattern may be *meaningless* [6] if it occurs in an sequence of an appropriately large size. Our algorithm for ranking sequential patterns with respect to significance works as follows: (I) we find frequent sequential patterns using *PrefixSpan* [10] for a given minimum support threshold; (II) we compute their frequencies and variances of the frequencies in the reference model and (III) we rank the frequent sequential patterns with respect to significance by computing the divergence (*Z-score*) between the empirical (actual) frequencies and frequencies in the reference model. Given the reference model a presence of significant divergence between the actual and computed frequency of a sequential pattern indicates that there is a dependency between itemsets/items in that pattern. In order to capture these dependencies our reference model consists of two sub-models: (I) *sequence-wise reference model*: treats itemsets as alphabet symbols and represents an independence model where itemsets occur independently of their order in an itemset-sequence and (II) *itemset-wise reference model*: provides decorrelated frequencies of itemsets for the sequence-wise reference model. By decorrelated frequencies we mean that given an attribute (item) a_1 and attribute a_2 the frequency of itemset (a_1, a_2) is computed using a *maximum entropy model*, where the marginal empirical probabilities are preserved. The reason we use such a model for itemsets is that unlike in the case of frequent itemset mining, we do not consider empty itemsets (empty attribute sets) and therefore the independence model for itemsets [3] is inappropriate as an itemset-wise reference model. In particular, using the independence model for sparse non-empty itemsets (the average number of ones in a row is much smaller than the number of attributes) would artificially overestimate the probability of the empty itemset causing a distortion of proper proportions of probabilities of non-empty itemsets. Note, that the sequence-wise reference model can be easily extended to *Markov models* in the spirit of [7]. The reason we consider the sequence-wise model to be independence model in this paper is because of the following reasons: (I) it is the model of choice if the Markov reference model is not known; (II) it has an intuitive interpretation as a method for discovering dependencies and (III) it leads to exact polynomial formulas for computing the frequencies of sequential patterns.

1.3 Multi-stream of News Stories and the Reference Model

We demonstrate the applicability of the presented method for discovering dependencies between streams of news stories, which is an important problem in multi-stream text mining and the topic detection and tracking research [2]. For this purpose we generated a collection of itemset-sequences from a multi-stream of news stories that was gathered from RSS feeds of major world news agencies [8]. Every itemset-sequence in that collection consists of stream identifiers of stories in a cross-stream cluster of news stories reporting the same news event, where the sequence is ordered according to the timestamps of the stories. Every itemset contains stream identifiers of documents published within the same time granularity. As an example itemset-sequence in that collection consider [(AP, MSNBC), UPI] that corresponds to three articles on the same news event (e.g., an earthquake in Italy), where the first two of them were published by AP and MSNBC within the same time granularity and followed by an article by UPI. Thus, clearly the empty itemset () does not occur in our data set. We stated the following research questions with respect to this collection of itemset-sequences: (I) what is the relationship between frequency, significance and content similarity in the discovered significant sequential patterns? and (II) what are the dependencies between the news sources in terms of sequential patterns of reporting the same news events?

As an example of the application of the reference model consider a case where the input collection of itemset-sequences contains a frequent sequential pattern s =[(AP, MSNBC), UPI], that consist of two itemsets s_1 =(AP, MSNBC) and s_2 =UPI that are correlated by occurring frequently together. Then since the sequence-wise reference model assumes independence between the elements, the frequency of s computed from the sequence-wise reference model will be much smaller then its actual frequency leading to a high significance rank of s. Furthermore, s_1 =(AP, MSNBC) contains two items a_1 =AP and a_2 =MSNBC which are correlated by occurring frequently together in the same itemsets. Then there are two possibilities for computing the frequency of s in the sequence-wise reference model: (I) we use the empirical frequency of s_1 or (II) we use a frequency of s_1 provided by the itemset-wise reference model. Then since the itemset-wise reference model provides decorrelated frequencies of itemsets while preserving marginal frequencies of the items (the publishing rates of AP and MSNBC), the frequency of s_1 computed from the itemset-wise reference model will be smaller that its empirical frequency leading to an even higher significance rank of s.

1.4 Related Work and Contributions

Thus, we present a reliable universal method for ranking sequential patterns with respect to significance that builds on the previous work [6], where a framework for assessing significance of subsequence patterns in the sliding window model was presented. The challenges of analysing itemset-sequences with respect to the previous work on sequences in [6] stems from the following facts: (I) itemset-sequences have variable sizes; (II) itemset-sequences contain itemsets (unordered sets) and (III) we do not consider empty itemsets. We address the first problem by modeling the frequency of an itemset-sequence using a probabilistic *discrete mixture model* and we approach the second and third problem by using an appropriate *maximum entropy* itemset-wise reference model.

To the best of our knowledge this is the first algorithm for ranking sequential patterns with respect to significance while there has been an extensive research on mining frequent sequential patterns (see [9] for a review).

The paper is organized as follows. Section 2 reviews theoretical foundations, Section 3 defines the problem, Section 4 presents the sequence-wise reference model, Section 5 presents the itemset-wise reference model, Section 6 presents the algorithm for ranking sequential patterns with respect to significance, Section 7 presents experimental results and finally Section 8 presents conclusions.

2 Theoretical Foundations (Review)

In this section we review some concepts that are necessary in order to explain our framework.

2.1 Sequential Pattern Mining

In this section we review the problem of sequential pattern mining [1]. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ be a set of items (alphabet). A subset $\mathcal{I} \subseteq \mathcal{A}$, where $\mathcal{I} = \{a_1, a_2, \ldots, a_{|\mathcal{I}|}\}$ is called an *itemset* or *element* and is also denoted by $(a_1, a_2, \ldots, a_{|\mathcal{I}|})$. An *itemset-sequence* $s = [s_1, s_2, \ldots, s_m]$ is an ordered list of itemsets, where $s_i \subseteq \mathcal{A}$. The size of the itemset-sequence is denoted by $|s|$ and the length of itemset-sequence s is defined as $l = \sum_{i=1}^{m} |s_i|$. An itemset-sequence $s = [s_1, s_2, \ldots, s_m]$ is a *subsequence* of itemset-sequence $s' = [s'_1, s'_2, \ldots, s'_{m'}]$, denoted $s \sqsubseteq s'$, if there exist integers $1 \leq i_1 \leq i_2 \ldots \leq i_m$ such that $s_1 \subseteq s'_{i_1}$, $s_2 \subseteq s'_{i_2}, \ldots, s_m \subseteq s'_{i_m}$. We also say that s' is a *supersequence* of s and s is *contained* in s'. Given a *collection of itemset-sequences* $\mathbf{S} = \{s^{(1)}, s^{(2)}, \ldots, s^{(|\mathbf{S}|)}\}$ the *support* (frequency) of an itemset-sequence s, denoted by $sup_{\mathbf{S}}(s)$, is defined as the number of itemset-sequences $s^{(i)} \in \mathbf{S}$ that contain s as a subsequence. The *relative support* (relative frequency) $rsup_{\mathbf{S}}(s) = \frac{sup_{\mathbf{S}}(s)}{|\mathbf{S}|}$ is the fraction of itemset-sequences that contain s as a subsequence. Given a relative support threshold *minRelSup* an itemset-sequence s is called a *frequent sequential pattern* if $rsup_{\mathbf{S}}(s) \geq minRelSup$. The problem of mining sequential patterns is to find all frequent sequential patterns in $\mathbf{S}$ given *minRelSup*. The support has an *anti-monotonic* property meaning that $sup_{\mathbf{S}}(s) \geq sup_{\mathbf{S}}(s')$ if $s \sqsubseteq s'$. A pattern s is called a *closed frequent sequential pattern* if none of its frequent supersequences has the same support. A pattern s is called a *maximal frequent sequential pattern* if none of its frequent supersequences is frequent. Table 1 presents an example collection of itemset-sequences, where itemset-sequence $id = 1$ has size $s = 3$, length $l = 4$ and consists of three elements (itemsets): $(1, 3)$, 1 and 1. Given $minRelSup = 0.5$, $s = [(1, 3), 1]$ is a frequent sequential pattern that is contained in itemset-sequences: $id = 1, 3$, where $rsup_{\mathbf{S}}(s) = 0.5$.

Table 1. A collection of itemset-sequences

id	itemset-sequence
0	$[2, 0]$
1	$[(1, 3), 1, 1]$
2	$[2, 3, (0, 2)]$
3	$[(1, 3), (2, 3), 0, 1]$

2.2 Significance of Subsequence Patterns

In this section we review the framework introduced in [6]. Let $e = [e_1, e_2, \ldots, e_m]$ be a sequence of symbols. Let $\Omega_n(e|w) = \sum_{i=1}^{n} I_i$ be a random variable that represent the actual frequency (support) of size w windows containing at least one occurrence of e as a subsequence in an event sequence of size $n + w - 1$ (n shifts of the window), where I_i is an indicator function equal to 1 if the i-th shift contains e. Clearly $\mathbf{E}[\Omega_n(e|w)] = nP^{\exists}(e|w)$, where $P^{\exists}(e|w)$ is the probability that a window ending at a given position in the event sequence contains at least one occurrence of e as a subsequence. The superscript $\exists$ means "at least one occurrence as a subsequence" and is used to distinguish this probability from a probability of e as a string. Clearly, $I_1, I_2, \ldots, I_n$ is a sequence of dependent random variables because a given subsequence pattern occurring in the input sequence may occur in many consecutive windows depending on its span. Therefore, because of the sliding window overlap $\Omega_n(e|w)$ does not have a *Binomial* distribution meaning $\mathbf{Var}[\Omega_n(e|w)] \neq nP^{\exists}(e|w)(1 - P^{\exists}(e|w))$. Let $\overline{\Omega}_n(e|w) = \frac{\Omega_n(e|w)}{n}$ be a random variable that represents the actual relative frequency of size w windows containing at least one occurrence of e as a subsequence in an event sequence, where $\mathbf{E}[\overline{\Omega}_n(e|w)] = P^{\exists}(e|w)$ and $\mathbf{Var}[\overline{\Omega}_n(e|w)] \leq \frac{1}{n}P^{\exists}(e|w)(1 - P^{\exists}(e|w))$.

Let $\mathcal{W}^{\exists}(e|w)$ be the *set* of all distinct windows of length w containing at least one occurrence of pattern e as a subsequence. Then $P^{\exists}(e|w) = \sum_{x \in \mathcal{W}^{\exists}(e|w)} P(x)$, where $P(x)$ is the probability of string x in a given Markov model. $\mathcal{W}^{\exists}(e|w)$ can be enumerated using an enumeration graph. The enumeration graph for a subsequence pattern $e = [e_1, e_2, \ldots, e_m]$ is shown in Figure 1. In particular for the 0-order Markov reference model $P^{\exists}(e|w)$ can be expressed as follows

$$P^{\exists}(e|w) = P(e) \sum_{i=0}^{w-m} \sum_{\sum_{k=1}^{m} n_k = i} \prod_{k=1}^{m} (1 - P(e_k))^{n_k}, \qquad (1)$$

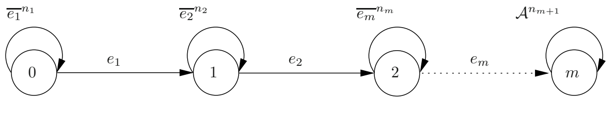


Fig. 1. Enumeration graph for a subsequence pattern $e = [e_1, e_2, \ldots, e_m]$, where $\overline{e} = \mathcal{A} - e$ and $\mathcal{A}$ is the alphabet. The exponents $n_1, \ldots, n_{m+1}$ above the self-loops denote the number of times the corresponding self-loops are selected.

where $P(e) = \prod_{i=1}^{m} P(e_i)$ and $P(e_i)$ is the probability of symbol e_i in the reference model. Then $P^{\exists}(e|w)$ is the probability of getting from state 0 to m in w steps. Paper [6] also presented an efficient $O(w^2)$ dynamic programming algorithm for computing $P^{\exists}(e|w)$ from (1). It was shown that if $\mathbf{Var}[\overline{\Omega}_n(e|w)] > 0$ then $\overline{\Omega}_n(e|w)$ satisfies the *Central limit theorem* (CLT) and this fact was used to set a lower and upper significance thresholds for $\overline{\Omega}_n(e|w)$.

3 Problem Definition

The problem of ranking sequential patterns (itemset-sequences) with respect to significance can be defined as follows.

Given: (I) collection of itemset-sequences $\mathbf{S} = \{s^{(1)}, s^{(2)}, \ldots, s^{(n)}\}$, where $s^{(i)} = [s_1^{(i)}, s_2^{(i)}, \ldots, s_{|s^{(i)}|}^{(i)}]$, $s_t^{(i)} \subseteq \mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ and $M = \max_{1 \le i \le n} |s^{(i)}|$ and (II) minimum relative support threshold *minRelSup* for sequential patterns.

Task: rank the discovered sequential patterns with respect to significance.

Note that in our method the main purpose of the support threshold for sequential patterns is to limit the search space of possible significant patterns.

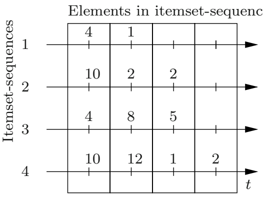
4 Sequence-Wise Reference Model

The sequence-wise reference model treats itemsets as alphabet symbols and represents an independence model where itemsets occur independently of their order in an itemset-sequence. In order to present the sequence-wise reference model we introduce the *element-wise* representation of a collection of itemset-sequences, that is a sequence of itemset-sequences $\mathcal{R} = [r^{(1)}, r^{(2)}, \ldots, r^{(|\mathcal{R}|)}]$ over an itemset alphabet $\Xi = \{\xi_1, \xi_2, \ldots, \xi_{|\Xi|}\}$, where $r^{(i)}$ is the i-th itemset-sequence and $r_t^{(i)} \in \Xi$ is the element (itemset) at time point t. As an example, Figure 2 presents the element-wise sequence of itemset-sequences for the collection from Table 1, where $\Xi = \{0, 1, 2, 3, (1,3), (2,3)\}$ and the itemsets are represented as decimal numbers. Note that for the sequence-wise reference model Ξ is provided by the itemset-wise reference model and includes all non-empty subsets of $\mathcal{A}$.

4.1 Generative Process

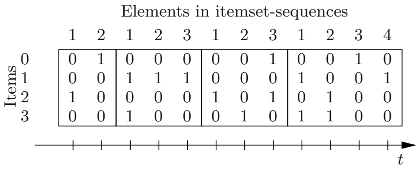
Now consider the sequence-wise reference model as a generative process, that generates itemset-sequences in $\mathcal{R}$ as follows:

1. it first generates the size of the itemset-sequence from a distribution $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_M]$, where α_m is the probability of generating an itemset-sequence of size m and
2. it generates a sequence of itemsets $r^{(i)}$ of size m from distribution $\theta = [\theta_1, \theta_2, \ldots, \theta_{|\Xi|}]$, provided by the itemset-wise reference model, where $\theta_j = P(r_t^{(i)} = \xi_j)$, for $\xi_j \in \Xi$.



Itemset-sequences				
1	4	1		
2	10	2	2	
3	4	8	5	
4	10	12	1	2

Fig. 2. Element-wise sequence of itemset-sequences representing the collection of itemset-sequences from Table 1. The streams correspond to itemset-sequences, where a decimal number at a given time point corresponds to an itemset (e.g., 10 = 2^{1+3} for itemset (1, 3)).



Items	1	2	1	2	3	1	2	3	1	2	3	4
0	0	1	0	0	0	0	0	1	0	0	1	0
1	0	0	1	1	1	0	0	0	1	0	0	1
2	1	0	0	0	0	1	0	1	0	1	0	0
3	0	0	1	0	0	0	1	0	1	1	0	0

Fig. 3. Item-wise multi-attribute sequence representing the collection from Table 1. The streams correspond to items, every time point corresponds to an itemset, where digit 1 in the i-th stream means a presence of the i-th item.

Let $P(r^{(i)}, m)$ be the joint probability of a particular itemset sequence $r^{(i)}$ of size m to be generated by the process. Then given the independence assumption of the presented generative process we factorize $P(r^{(i)}, m)$ as follows

$$P(r^{(i)}, m) = \alpha_m \cdot P(r^{(i)}|m), \tag{2}$$

where $P(r^{(i)}|m) = \prod_{t=1}^{m} P(r_t^{(i)})$ is the probability of a particular itemset-sequence $r^{(i)}$ to be generated given the size m.

We compute the parameters of the sequence-wise reference model from $\mathbf{S}$ as follows: (I) $\alpha_m = \frac{N_n(|s^{(i)}|=m)}{n}$ (ML estimator), where $N_n(|s^{(i)}| = m)$ is the number of occurrences of itemset-sequences of size m in $\mathbf{S}$ and (II) θ is computed form the itemset-wise reference model, that is presented in Section 5 and whose purpose is to provide decorrelated frequencies of itemsets. Note that we could compute θ_j as $\theta_j = \frac{N_n(\xi_j)}{n_\Xi}$ (ML estimator), where n_Ξ is the number of itemsets in $\mathbf{S}$ and $N_n(\xi_j)$ is the number of occurrences of itemset ξ_j in $\mathbf{S}$. However the ML estimator for the itemsets does not provide decorrelated frequencies.

4.2 Relative Frequency

In this section we consider occurrences of a sequential pattern as a *subsequence* (gaps between elements of the sequential pattern are allowed) in the sequence-wise reference model represented by its element-wise representation $\mathcal{R}$. Let $\overline{\Omega}_n(s)$ be a random variable representing the actual relative frequency of a sequential pattern s occurring as a subsequence in $\mathcal{R}$. Recall that the relative frequency of a sequential pattern s is equal to the fraction of itemset-sequences in $\mathcal{R}$ that contain it as a subsequence. This means that even if s occurs many times in a given itemset-sequence $s' \in \mathcal{R}$ we count it only as one occurrence. Let $\Omega_n(s) = \sum_{i=1}^{n} I_i$ be a random variable that represent the actual frequency of s occurring as a subsequence in $\mathcal{R}$ ($sup_{\mathcal{R}}(s)$), where I_i is an indicator function equal to 1 if the i-th itemset-sequence contains s. Then clearly, $I_1, I_2, \ldots, I_n$ is a sequence of independent random variables because occurrences of a pattern as a subsequence in itemset-sequences are independent of each other. Therefore, $\Omega_n(s)$ has the *Binomial distribution* and $\overline{\Omega}_n(s) = \frac{\Omega_n(s)}{n}$ is a *Binomial proportion*, where $\mathbf{E}[\overline{\Omega}_n(s)] = P^{\exists}(s)$, $\mathbf{Var}[\overline{\Omega}_n(s)] = \frac{1}{n}P^{\exists}(s)(1-P^{\exists}(s))$ and $P^{\exists}(s)$ is the probability that s exists as a subsequence in an itemset-sequence in $\mathcal{R}$. Thus, clearly $\Omega_n(s)$ and $\overline{\Omega}_n(s)$ both satisfy CLT. However, since itemset-sequences have variable sizes, for a given s, $P^{\exists}(s)$ depends on the distribution of the sizes of itemset-sequences in $\mathcal{R}$.

Let $P^{\exists}(s,m)$ be the joint probability that an itemset-sequence s of size $|s|$ exists as a subsequence in another itemset-sequence s' of size $m \geq |s|$ in $\mathcal{R}$. Then following (2) we factorize $P^{\exists}(s,m)$ as follows

$$P^{\exists}(s,m) = \alpha_m \cdot P^{\exists}(s|m), \tag{3}$$

where $P^{\exists}(s|m)$ is the probability that s occurs given an itemset-sequence of size m in $\mathcal{R}$. In order to obtain the formula for $P^{\exists}(s)$ we marginalize from (3) as follows:

$$P^{\exists}(s) = \sum_{m=|s|}^{M} \alpha_m \cdot P^{\exists}(s|m). \tag{4}$$

Thus, $P^{\exists}(s)$ is expressed as a *discrete mixture model*, where the *mixing coefficients* $(\alpha_1, \alpha_2, \ldots, \alpha_M)$ model the fact that an occurrence of s as a subsequence in an itemset-sequence s' depends on the size of s' and may possibly occur in any itemset-sequence $s' \in \mathcal{R}$ for which $|s'| \geq |s|$. In other words, $P^{\exists}(s)$ is a weighted combination of contributions from itemset-sequences of all possible relevant sizes in $\mathcal{R}$.

Finally, the formula for $P^{\exists}(s|m)$ for an itemset-sequence $s = [s_1, s_2, \ldots, s_{|s|}]$ given an itemset-sequence of size m in $\mathcal{R}$ can be obtained as follows. Let $\mathcal{X}_i = \bigcup_{\xi_j \in \Xi, s_i \subseteq \xi_j} \xi_j$ be the set of all supersets of itemset s_i in itemset alphabet Ξ. Then clearly, the enumeration graph for $\mathcal{W}^{\exists}(s|m)$ can be obtained from the enumeration graph for a sequence of items $e = [e_1, e_2, \ldots, e_m]$ by substituting $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{|s|}$ for items $e_1, e_2, \ldots, e_m$ in Figure 1. Then the formula for $P^{\exists}(s|m)$ can be obtained from (1) by substituting marginal probabilities of itemsets $P_{\mathcal{M}}(s_i) = P(\mathcal{X}_i)$ for probabilities of items in (1). Thus, $P^{\exists}(s|m)$ in a 0-order Markov sequence-wise reference model, can be obtained from (1) as follows:

$$P^{\exists}(s|m) = P_{\mathcal{M}}(s) \sum_{i=0}^{m-|s|} \sum_{\sum_{k=1}^{|s|} n_k = i} \prod_{k=1}^{|s|} (1 - P_{\mathcal{M}}(s_k))^{n_k}, \tag{5}$$

where $P_{\mathcal{M}}(s) = \prod_{i=1}^{|s|} P_{\mathcal{M}}(s_i)$ and $P_{\mathcal{M}}(s_i)$ is the marginal probability computed from the itemset-wise reference model. Formula (5) can be computed in $O(m^2)$ using the dynamic programming algorithm given in [6]. Thus computing $P^{\exists}(s)$ from (4) takes $O(M^3)$ time.

The presented sequence-wise reference model can be easily extended to more application specific models. As a first extension, we could assume that itemset-sequences are generated using a *Markov model* and use the algorithm for computing $P^{\exists}(s|m)$ from [7] for Markov models. As another extension, we could assume that the distribution of itemsets θ depends on the size of an itemset-sequence (e.g., itemsets having large cardinality are more likely to occur in shorter itemset-sequences).

5 Itemset-Wise Reference Model

The itemset-wise reference model treats itemsets as binary vectors and provides decorrelated frequencies of itemsets. In order to present the itemset-wise reference model we introduce the *item-wise* representation, that is a multi-attribute binary sequence $\mathcal{B} = \{b^{(1)}, b^{(2)}, \ldots, b^{(|\mathcal{A}|)}\}$ of size $|\mathcal{A}|$, where: $b^{(i)}$ is a binary sequence corresponding to attribute (item) $a_i \in \mathcal{A}$ and $b_t^{(j)} \in \{0,1\}$ is the value at time point t. Thus, $\mathcal{B}$ represents $\mathbf{S}$ as a sequence of time ordered itemsets. Figure 3 presents the item-wise multi-attribute sequence for the collection from Table 1.

Note that we do not consider empty itemsets (binary vectors consisting of all zeros in $\mathcal{B}$) because a lack of attributes is meaningless in our framework. Therefore the streams in $\mathcal{B}$ are inherently dependent, i.e., $P(b_t^{(1)}, b_t^{(2)}, \ldots, b_t^{(|\mathcal{A}|)}) \neq \prod_{j=1}^{(|\mathcal{A}|)} P(b_t^{(j)})$ and the independence model is inappropriate in our framework.
Therefore we build a *maximum entropy* model of the form [4]

$$P(b_t^{(1)}, b_t^{(2)}, \ldots, b_t^{|\mathcal{A}|}) = Z \left(\prod_{i=1}^{|\mathcal{A}|} \mu_i^{I^{(i)}} \right) \mu_c^{|\mathcal{A}| - \sum_{i=0}^{|\mathcal{A}|} I^{(i)}}, \tag{6}$$

where Z is the normalizing constant, $I^{(i)}$ is an indicator function equal to one if $b_t^{(i)} = 1$. We build (6) using *Generalized Iterative Scaling* (GIS) algorithm by finding μ_i for $i = 1, \ldots |\mathcal{A}|$, μ_c and Z under constraints that empirical marginal probabilities of items are preserved, i.e, $\sum P(b_t^{(1)}, b_t^{(2)}, \ldots, b_t^{|\mathcal{A}|}) I^{(i)} = P(b_t^{(i)} = 1)$, where μ_c is the *correction feature* that ensures that the number of parameters (features) for every binary vector in (6) is constant. Let $Sum = \sum_{b_t^{(i)} \in \{0,1\}, \sum_i b_t^{(i)} > 0} P(b_t^{(1)}, b_t^{(2)}, \ldots, b_t^{(|\mathcal{A}|)})$, let $p^{(i)} = P(b_t^{(i)} = 1)$ and let $p_{\mathcal{M}}^{(i)} = Z \cdot u_i \sum_{n_j \in \{0,1\}} \left(\prod_{j \neq i} \mu_j^{n_j} \right) \mu_c^{|\mathcal{A}| - 1 - \sum_{j \neq i} n^{(j)}}$ be the marginal probability of attribute i computed from the model given the current estimates of the parameters.

The iterative scaling algorithm proceeds as follows: (I) initialization: $\mu_i = 1$, $\mu_c = 1$, $Z = \frac{1}{Sum}$; and (II) iteration: **repeat for** i = 1 **to** $|\mathcal{A}|$ **do begin** $\mu_i^{n+1} = \mu_i^n \left(\frac{p^{(i)}}{p_{\mathcal{M}}^{(i)}} \right)^{\frac{1}{|\mathcal{A}|}}$, $\mu_c^{n+1} = \mu_c^n \left(\frac{1}{Z \cdot Sum} \right)^{\frac{1}{|\mathcal{A}|}}$, $Z = \frac{1}{Sum}$ **end until for** i = 1 **to** $|\mathcal{A}|$ $\frac{|p_{\mathcal{M}}^{(i)} - p^{(i)}|}{p^{(i)}} < \varepsilon$. Thus, the maximum entropy model satisfies our requirements: (I) it preserves empirical marginal probabilities; (II) it is defined only for all

Table 2. Comparison of marginal probabilities of the itemsets of size two and the probability of the empty itemset for the collection from Table 1 obtained using the independence model and the maximum entropy model

Itemset	Independence model	Maximum Entropy model
(0,2)	8.33e-02	2.35e-01
(1,3)	1.39e-01	3.81e-01
(2,3)	1.11e-01	3.08e-01
()	1.94e-01	0

non-empty subsets of $\mathcal{A}$ and (III) it gives as much independence to the attribute streams as possible given the constraints.

Table 2 presents marginal probabilities of the itemsets of size two from Figure 3 obtained using the independence model and the maximum entropy model. Thus, Table 2 shows the following facts: (I) although the empty itemset does not occur in Figure 3 the independence model assigns a bigger probability ($1.94e{-}01$) to the empty itemset than to the occurring itemsets of size two; and (II) the ME model, as expected, assigns greater probabilities to the occurring itemsets than the independence model.

6 Ranking Algorithm

Given a collection of itemset-sequences $\mathbf{S} = \{s^{(1)}, s^{(2)}, \ldots, s^{(n)}\}$, where $s^{(i)} = [s_1^{(i)}, s_2^{(i)}, \ldots, s_{|s^{(i)}|}^{(i)}]$, the ranking algorithm proceeds as follows:

1. run *PrefixSpan* for a given value of $minRelSup$ to obtain a set of frequent sequential patterns $\mathcal{F}$.
2. compute $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_M]$, where $\alpha_m = \frac{N_n(|s^{(i)}|=m)}{n}$ and $N_n(|s^{(i)}| = m)$ is the number of itemset-sequences of size m in $\mathbf{S}$.
3. for every frequent sequential pattern $s = [s_1, s_2, \ldots, s_{|s|}]$, where $s \in \mathcal{F}$ and $rsup_{\mathbf{S}}(s) \geq minRelSup$ do the following:
   (a) compute the marginal probability vector $[\theta_1, \theta_2, \ldots, \theta_{|s|}]$ ($\theta_i = P_{\mathcal{M}}(s_i)$) for elements of s from the itemset-wise reference model.
   (b) compute $P^{\exists}(s)$ from (4) and compute the significance rank as follows $sigRank(s) = \frac{\sqrt{n}\left(rsup_{\mathbf{S}}(s) - P^{\exists}(s)\right)}{\sqrt{P^{\exists}(s)(1-P^{\exists}(s))}}$.

The reference model will be violated in $\mathbf{S}$ in two cases: (I) the sequence-wise reference model is violated by correlated itemsets and (II) the itemset-wise reference model is violated by correlated items in itemsets.

7 Experiments

In this section we present our experimental results on the multi-stream of news stories of size 224062 stories that have been retrieved, via RSS feeds, from the following thirteen news sources: ABC news (ABC), Aljazeera (ALJ), Associated Press (AP), British Broadcast Co. (BBC), Canadian Broadcast Co. (CBC), Xinhua News Agency (CHV), Central News Agency Taiwan (CNE), CNN, MSNBC, Reuters (REU), United Press International (UPI), RIA Novosti (RIA) and Deutsche Welle (DW). The stories were retrieved over a period of thirteen months from the 12-th of November 2008 to the 3rd of January 2010. We implemented a clustering algorithm that uses a time-window of a given duration (e.g., 24 hours) and is an incremental variant of a non-hierarchical document clustering algorithm using a similarity measure based on nearest neighbors. We ran the algorithm for the following parameters: (I) the time-window size $w = 24$ hours; (II) the document similarity threshold $\tau_d = 0.5$ that is used to identify nearest neighbors for a new arriving document to the window and (III) the time quantization step size $Q_t = 1$ hour. As a result we obtained a collection of itemset-sequences $\mathbf{S}$ of size $|\mathbf{S}| = 32464$, where there are 109964 itemsets, the maximum itemset-sequence size $M = 25$, the average itemset-sequences size is 3.5 and the average itemset size is 1.2.

7.1 From Clusters to Itemset-Sequences

Let $\mathcal{D} = \{d^{(1)}, d^{(2)}, \ldots, d^{(|\mathcal{D}|)}\}$ be a multi-stream of news stories (documents), where $d_t^{(i)}$ is a document in stream i at a *time point* t and has three attributes: (I) the exact publishing timestamp $d_t^{(i)}.timestamp$; (II) stream identifier $d_t^{(i)}.stream$ $= i$; and (III) text content $d_t^{(i)}.content$. The publishing timestamp $d_t^{(i)}.timestamp$ is unique in each stream $d^{(i)}$. Let $\mathcal{C} = [d_1, d_2, \ldots, d_{|\mathcal{C}|}]$ be a cluster of documents (reporting the same event in our case) defined as a sequence of documents ordered with respect to publishing timestamp $d_i.timestamp$. We convert $\mathcal{C}$ to an itemset-sequence $s = [s_1, s_2, \ldots, s_{|s|}]$, where $s_i \subseteq \mathcal{A}$ and $\mathcal{A} = \{0, 1, \ldots, |\mathcal{D}| - 1\}$ is the set of all stream identifiers of the news sources in $\mathcal{D}$. As a result of the conversion each itemset s_i contains stream identifiers of documents with the same timestamp ($d_i.timestamp$) and the itemset-sequence s is ordered with respect to the timestamps of the itemsets. As an example consider itemset-sequence $[(1,3),1,1]$ in Table 1, where $s_1 = (1,3)$ means that two documents: the first from source 1 and the second from source 3 were published (within the time granularity Q_t) before a document from streams 1 and 1 respectively. Furthermore, for every itemset-sequence, we recorded content similarity between the stories corresponding to its elements in terms of the *cosine similarity measure*. In order to asses the nature of content similarity between documents in a given itemset-sequence s we define the average content similarity $AvgSim_{\mathbf{S}}(s)$ and the variance of the content similarity $VarSim_{\mathbf{S}}(s)$ between documents in an itemset-sequence s of length l occurring as a subsequence over the whole collection of itemset-sequences $\mathbf{S}$ are expressed as follows

$$AvgSim_{\mathbf{S}}(s) = \frac{2 \cdot sup_{\mathbf{S}}(s)}{l^2 - l} \sum_{s' \in \mathbf{S}, s \sqsubseteq s'} \sum_{k=1}^{l} \sum_{i=j_1}^{j_k - 1} sim(s'_i, s'_{j_k}) \qquad (7)$$

Table 3. Baseline: Top-20 most frequent sequential patterns of size greater than one

	Pattern	$rsup_{\mathbf{S}}$
1	[AP, MSNBC]	1.78e-01
2	[MSNBC, UPI]	1.20e-01
3	[BBC, UPI]	1.06e-01
4	[AP, UPI]	1.06e-01
5	[REU, UPI]	1.06e-01
6	[AP, ABC]	9.03e-02
7	[BBC, ALJ]	8.78e-02
8	[REU, BBC]	8.71e-02
9	[CNN, UPI]	8.56e-02
10	[REU, CNE]	8.55e-02
11	[REU, MSNBC]	8.41e-02
12	[BBC, CNE]	8.15e-02
13	[ABC, UPI]	8.09e-02
14	[ABC, MSNBC]	8.07e-02
15	[CNE, UPI]	7.87e-02
16	[AP, REU]	7.83e-02
17	[BBC, REU]	7.51e-02
18	[MSNBC, REU]	7.49e-02
19	[MSNBC, ABC]	7.20e-02
20	[CNE, BBC]	7.05e-02

Table 4. Top-20 most significant sequential patterns for *minRelSup* = 0.01

	Pattern	*sigRank*
1	[BBC, ALJ, ALJ, ALJ]	25.5
2	[ALJ, ALJ, ALJ, CNE]	19.1
3	[CNE, ALJ, ALJ, ALJ]	18.6
4	[BBC, CNE, ALJ, ALJ]	18.1
5	[ALJ, ALJ, CNE, ALJ]	17.7
6	[CNE, ALJ, ALJ, CNE]	16.4
7	[BBC, ALJ, ALJ, UPI]	16.3
8	[BBC, CNE, ALJ, ALJ]	16.1
9	[AP, MSNBC]	15.7
10	[ALJ, ALJ, BBC, ALJ]	15.1
11	[ALJ, CNE, ALJ, ALJ]	14.5
12	[CNE, BBC, ALJ, ALJ]	14.2
13	[BBC, ALJ, ALJ, BBC]	14.1
14	[ALJ, BBC, ALJ, ALJ]	13.9
15	[BBC, ALJ, ALJ, CNN]	13.2
16	[ALJ, ALJ, CBS]	13.1
17	[ALJ, ALJ, ALJ, UPI]	12.9
18	[REU, ALJ, ALJ, ALJ]	12.8
19	[ALJ, ALJ, ALJ, BBC]	12.7
20	[BBC, ALJ, CNE, ALJ]	12.4

and

$$VarSim_{\mathbf{S}}(s) = \frac{2 \cdot sup_{\mathbf{S}}(s)}{l^2 - l} \sum_{s' \in \mathbf{S}, s \sqsubseteq s'} \sum_{k=1}^{l} \sum_{i=j_1}^{j_k - 1} (AvgSim_{\mathbf{S}}(s) - sim(s'_i, s'_{j_k}))^2, \quad (8)$$

where $j_1 \leq j_2 \, \ldots \leq j_l$ are the positions where s occurs in s' as a subsequence and $sim(d_i, d_j)$ is the *cosine similarity* or *content similarity* between documents i and j. Thus, (7) computes the average content similarity over all itemset-sequences containing s as a subsequence. We also use $StdDevSim_{\mathbf{S}}(s)$ to denote $\sqrt{VarSim_{\mathbf{S}}(s)}$.

7.2 Baseline: Most Frequent Patterns

As a baseline against which we compare the performance of the ranking algorithm we use the top-20 most frequent sequential patterns of size greater than one, where we also removed patterns containing the same symbol, which corresponds to frequent updates of the same news event. Table 3 presents the top-20 most frequent sequential patterns of size greater than one.

7.3 Significant Patterns

In the first experiment we rank the top-20 most frequent patterns from Table 3 with respect to significance. Figure 4 presents the results. As it turns out the most frequent pattern in Table 3 is also the most significant one but for the following patterns there is not any obvious relationship between the significance rank and the frequency rank. The dependency between AP and MSNBC can be explained by the fact that as we saw in the recorded stories MSNBC is reusing some content from AP.

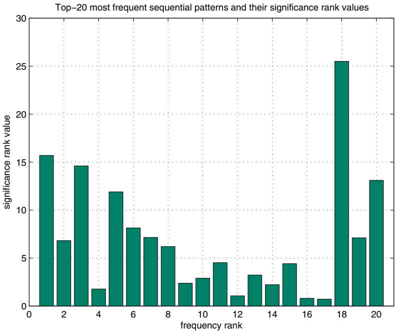


Fig. 4. Frequency rank (x-axis) versus significance rank (y-axis) for the top-20 most frequent sequential patterns from Table 3

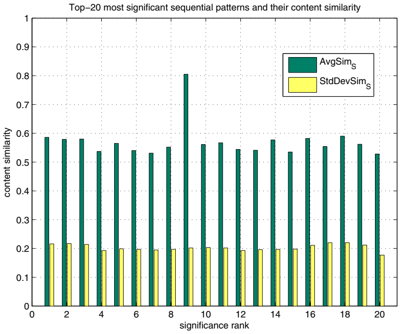


Fig. 5. Significance rank (x-axis) versus content similarity (the average and the standard deviation) (y-axis) for the top-20 most significant sequential patterns from Table 4

In the second experiment we set $minRelSup = 0.01$ and found the most significant over-represented ($sigRank > 0$) sequential patterns. Table 4 presets the top-20 most significant sequential patterns, where among the top patterns we removed patterns containing the same symbol and patterns having significant supersequences. Note however that the top-20 most significant patterns for the whole collection may not be the same since the patterns in Table 4 were obtained using $minRelSup = 0.01$. In general the lower the value of $minRelSup$ the higher the chance that the reference model will discover long significant patterns having low support. By comparing the results from Table 3 and from Table 4 we can make the following observations: (I) the most significant patterns are generally longer than the most frequent ones since the sequence-wise reference model leverages rank of correlated longer patterns and (II) there is a prevalence of patterns involving BBC in the first position and ALJ in the following positions. The dependency between BBC and ALJ may be related to the fact that, as we found out from the BBC web site, BBC signed a news exchange agreement with ALJ and as the pattern suggests this exchange seems to be really "one-way" from BBC to ALJ. Furthermore, ALJ tends to provide many updates of the same news event. Also, although [AP, MSNBC] is the most frequent pattern it has significance rank nine in Table 4 as a result of the reference model leveraging rank of the longer patterns involving BBC and ALJ.

Figure 5 presents a graph of the significance rank (x-axis) versus the average content similarity $AvgSim_{\mathbf{S}}$ and the standard deviation $StdDevSim_{\mathbf{S}}$ (y-axis) for the top-20 most significant sequential patterns from Table 4. Figure 5 shows two facts: (I) the average content similarity is above the document similarity threshold $\tau_d = 0.5$ and (II) the value of the standard deviation is relatively low for all patterns. These results suggest that temporally correlated news streams tend to be also correlated with respect to their content.

8 Conclusions

We presented a reliable general method for ranking frequent sequential patterns (itemset-sequences) with respect to significance. We demonstrated the applicability of the presented method on a multi-stream of news stories that was gathered from RSS feeds of the major world news agencies. In particular we showed that there are strong dependencies between the news sources in terms of temporal sequential patterns of reporting the same news events and content similarity, where the frequency and significance rank are correlated with the content similarity.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: ICDE, pp. 3–14 (1995)
2. Allan, J.: Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publishers, Norwell (2002)
3. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: Generalizing association rules to correlations. In: Proceedings ACM SIGMOD International Conference on Management of Data, May 1997, pp. 265–276 (1997)
4. Darroch, J., Ratcliff, D.: Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics 43(5), 1470–1480 (1972)
5. Guan, E., Chang, X., Wang, Z., Zhou, C.: Mining maximal sequential patterns. In: 2005 International Conference on Neural Networks and Brain, pp. 525–528 (2005)
6. Gwadera, R., Atallah, M., Szpankowski, W.: Reliable detection of episodes in event sequences. In: Third IEEE International Conference on Data Mining, November 2003, pp. 67–74 (2003)
7. Gwadera, R., Atallah, M., Szpankowski, W.: Markov models for discovering significant episodes. In: SIAM International Conference on Data Mining, pp. 404–414 (2005)
8. Gwadera, R., Crestani, F.: Mining and ranking streams of news stories using cross-stream sequential patterns. In: CIKM 2009: Proceedings of the 18th International Conference on Information and Knowledge Management, Hong Kong, October 2009, pp. 1709–1712 (2009)
9. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 15(1) (2007)
10. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q.: Mining sequential patterns by pattern-growth: The prefixspan approach. TKDE 16(11) (November 2004)
11. Tatti, N.: Maximum entropy based significance of itemsets. KAIS 17(1), 57–77 (2007)
12. Yan, X., Han, J., Afshar, R.: Clospan: Mining closed sequential patterns in large datasets. In: SDM, pp. 166–177 (2003)