High Average-Utility Itemset Sampling under Length Constraints

Author name

Affiliation

Abstract. This supplementary document presents a first the proofs which were stated in the article titled "High Average-utility Itemset Sampling under Length Constraints". Then, it ends with additional experimental results.

1 Recalls

Definition 1 (Occurrence of a pattern). *Let φ be a pattern defined on a language $\mathcal{L}$ of a database $\mathcal{D}$. If it exists a transaction t_j of $\mathcal{D}$ such that $\varphi \subseteq t_j$, then φ_j is an occurrence of the pattern φ in the transaction t_j. The utility of the pattern φ in the transaction t_j, denoted by $\texttt{uOcc}(\varphi, t_j)$, is equal to 0 if $\varphi \not\subseteq t_j$ or $\varphi = \emptyset$, else $\texttt{uOcc}(\varphi, t_j) = \sum_{e \in \varphi} (q(e, t_j) \times p(e))$.*

There are also utilities that are independent of any database such as length-based utilities. In the following, we consider the length-based utility defined by $\texttt{uLen}_{[m..M]} = 1/|\varphi|$ if $|\varphi| \in [m..M]$ and 0 otherwise, m and M are two positive integers. Thus, a pattern whose length is larger than M or smaller than m will be deemed useless.

Definition 2 (Average-Utility of a pattern under length constraints). *Let $\mathcal{D}$ be a database, $\mathcal{L}$ its language, m and M two integers such that $m \leq M$. The average-utility of the pattern $\varphi \in \mathcal{L}$ in $\mathcal{D}$ under minimum m and maximum M length constraints, denoted by $u^{avg}_{[m..M]}(\varphi, \mathcal{D})$, is the product of the sum of utilities of its occurrences and its length-based utility. Formally, $u^{avg}_{[m..M]}(\varphi, \mathcal{D}) = (\sum_{(j,t) \in \mathcal{D} \wedge \varphi \subseteq t} \texttt{uOcc}(\varphi, t)) \times \texttt{uLen}_{[m..M]}(\varphi)$.*

It is important to note that $u^{avg}_{[m..M]}$ is not a length-based utility.

2 Proofs of the theoretical results

Property 1 (Item weights $\omega^{\bullet}_{\ell}(t[i], t)$). The weights $\omega^{+}_{\ell}(t[i], t)$ and $\omega^{-}_{\ell}(t[i], t)$ of the item $t[i]$, for all $\ell \in [m..M]$, may be formally written as follows:[1]

$$\omega^{+}_{\ell}(t[i], t) = \omega_1(t[i], t) \times \binom{\ell - 1}{|t^i|} + \sum_{\star \in \{+,-\}} \omega^{\star}_{\ell-1}(t[i+1], t)$$

[1] by convention $\binom{k}{n}$ =0 if k>n and 1 if k=0

$$\omega_\ell^-(t[i],t) = \sum_{\star\in\{+,-\}} \omega_\ell^\star(t[i+1],t)$$

with $\omega_1^+(t[i],t) = \mathtt{uOcc}(t[i],t)$ for all $i \in [1..|t|]$ and $\omega_\ell^\star(t[i],t) = 0$ for all $i > |t|$.

Proof (Property 1). Let's start by showing that $\omega_\ell^-(t[i],t) = \sum_{\star\in\{+,-\}} \omega_\ell^\star(t[i+1],t)$. By definition, $\omega_\ell^-(t[i],t)$ is the sum of the utilities of the set of patterns of length ℓ in t^i, $\omega_\ell^-(t[i],t) = \sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell} \mathtt{uOcc}(\varphi,t)$. This set can be split into two parts: the one that contains the patterns starting with the item $t[i+1]$ whose sum of their utilities is equal to $\omega_\ell^+(t[i+1],t)$ by definition, and the one that contains the patterns not starting with $t[i+1]$ and whose sum of their utilities is equal to $\omega_\ell^-(t[i+1],t)$.That implies that $\sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell} \mathtt{uOcc}(\varphi,t) = \omega_\ell^+(t[i+1],t) + \omega_\ell^-(t[i+1],t) = \sum_{\star\in\{+,-\}} \omega_\ell^\star(t[i+1],t)$. (1)
Let's now show that $\omega_\ell^+(t[i],t) = \omega_1(t[i],t) \times \binom{\ell-1}{|t^i|} + \sum_{\star\in\{+,-\}} \omega_{\ell-1}^\star(t[i+1],t)$. We know by definition that $\omega_\ell^+(t[i],t)$ is the sum of the utilities of itemsets of length ℓ in t^i which start with $t[i]$ following the total order relation $>_\mathcal{I}$. Formally, we have: $\omega_\ell^+(t[i],t) = \sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} \mathtt{uOcc}(\{t[i]\} \cup \varphi,t)$. But $\mathtt{uOcc}(\{t[i]\} \cup \varphi,t) = \mathtt{uOcc}(\{t[i]\},t) + \mathtt{uOcc}(\varphi,t)$ by definition. Then, $\omega_\ell^+(t[i],t) = \sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} (\mathtt{uOcc}(\{t[i]\},t) + \mathtt{uOcc}(\varphi,t))$. Which implies: $\omega_\ell^+(t[i],t) = \sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} \mathtt{uOcc}(\{t[i]\},t) + \sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} \mathtt{uOcc}(\varphi,t)$. However, we know on the one hand that $\sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} \mathtt{uOcc}(\{t[i]\},t) = \mathtt{uOcc}(\{t[i]\},t) \times \binom{\ell-1}{|t^i|}$ and that by definition $\mathtt{uOcc}(\{t[i]\},t) = \omega_1^+(t[i],t)$, so $\sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} \mathtt{uOcc}(\{t[i]\},t) = \omega_1^+(t[i],t) \times \binom{\ell-1}{|t^i|}$. On the other hand, $\sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} \mathtt{uOcc}(\varphi,t)$ is the sum of the utilities of the set patterns of length $\ell-1$ in the transaction t^i. From (1), we can also say that $\sum_{\varphi\subseteq t^i \wedge |\varphi|=\ell-1} \mathtt{uOcc}(\varphi,t) = \sum_{\star\in\{+,-\}} \omega_{\ell-1}^\star(t[i+1],t)$. Then we have: $\omega_\ell^+(t[i],t) = \omega_1(t[i],t) \times \binom{\ell-1}{|t^i|} + \sum_{\star\in\{+,-\}} \omega_{\ell-1}^\star(t[i+1],t)$. Hence the result. □

Property 2 (Transaction weight). The weight of a transaction t under minimum m and maximum M length constraints, denoted by $\omega_{[m..M]}^{avgU}(t)$, is the sum of the average-utilities of the occurrences it contains. Formally,

$$\omega_{[m..M]}^{avgU}(t) = \sum_{\ell=m}^{M} \left(\frac{1}{\ell} \sum_{i=1}^{|t|} \omega_\ell^+(t[i],t) \right) = \sum_{\ell=m}^{M} \frac{1}{\ell} \left(\omega_\ell^+(t[1],t) + \omega_\ell^-(t[1],t) \right).$$

Proof (Property 2). By definition, the weight of the transaction t is the sum of the average-utilities of the pattern occurrences it contains. According to Property 1, the weight of the transaction t under the minimum m and maximum M length constraints is nothing more than the sum of the sum of the average-utilities of pattern occurrences that start with the item $t[1]$ and respect the imposed length constraints, $\sum_{\ell=m}^{M} (\frac{1}{\ell} \times \omega_\ell^+(t[1],t))$, and that of the patterns that do not start with the item $t[1]$ but respect the length constraints, $\sum_{\ell=m}^{M}(\frac{1}{\ell}\times\omega_\ell^-(t[1],t))$. However, we know that $\sum_{\ell=m}^{M}(\frac{1}{\ell}\times\omega_\ell^+(t[1],t)) + \sum_{\ell=m}^{M}(\frac{1}{\ell}\times\omega_\ell^-(t[1],t)) = \sum_{m}^{M}\frac{1}{\ell}\times\left(\omega_\ell^+(t[1],t)+\omega_\ell^-(t[1],t)\right)$. Hence the result. $\square$

Lemma 1. *Let ℓ be the length of the itemset to output, $\mathbb{P}_\ell^t(t[i]|\varphi,\ell')$ the probability to draw item $t[i]$ in the transaction t after drawing $\ell-\ell'$ items and storing them in φ, with $e >_\mathcal{I} t[i]$ for all $e \in \varphi$. The probability to draw the $t[i]$ knowing φ and ℓ' can be formulated as follows:*

$$\mathbb{P}_\ell^t(t[i]|\varphi,\ell') = \frac{\sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1} \texttt{uOcc}(\varphi\cup\{t[i]\}\cup\varphi',t)}{\sum_{\varphi'\subseteq t^{i-1}\wedge|\varphi'|=\ell'} \texttt{uOcc}(\varphi\cup\varphi',t)}.$$

Proof (Lemma 1). By definition, the probability to draw the item $t[i]$ of the transaction t after having drawing on it $\ell-\ell'$ items and store them in φ is nothing but the probability of drawing a pattern that begins with $\varphi\cup\{t[i]\}$, according to the order relation $>_\mathcal{I}$, among the set of patterns that start with φ. On the one hand, we know that the set of patterns of length ℓ that start with $\varphi\cup t[i]$ is defined by $\{\varphi''\subseteq t : (\varphi''=\varphi\cup\{t[i]\}\cup\varphi')(\varphi'\subseteq t^i)(|\varphi'|=\ell'-1)\}$. The sum of the utilities of the patterns of this set is equal to $\sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1}\texttt{uOcc}(\varphi\cup\{t[i]\}\cup\varphi',t)$. On the other hand, we know that the set of patterns of length ℓ that start with φ is defined by $\{\varphi''\subseteq t : (\varphi''=\varphi\cup\varphi')(\varphi'\subseteq t^{i-1})(|\varphi'|=\ell')\}$. The sum of the utilities of the patterns of this set is equal to $\sum_{\varphi'\subseteq t^{i-1}\wedge|\varphi'|=\ell'}\texttt{uOcc}(\varphi\cup\varphi',t)$. So $\mathbb{P}_\ell^t(t[i]|\varphi,\ell') = \frac{\sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1}\texttt{uOcc}(\varphi\cup\{t[i]\}\cup\varphi',t)}{\sum_{\varphi'\subseteq t^{i-1}\wedge|\varphi'|=\ell'}\texttt{uOcc}(\varphi\cup\varphi',t)}$. Hence the result. $\square$

Property 3. The probability to draw the item $t[i]$ in the transaction t knowing the itemset φ and the length ℓ', with $|\varphi|=\ell-\ell'$, denoted by $\mathbb{P}_\ell^t(t[i]|\varphi,\ell')$, is given by the following formula:

$$\mathbb{P}_\ell^t(t[i]|\varphi,\ell') = \frac{\left(\sum_{k<i\wedge t[k]\in\varphi}\omega_1(t[k],t)\right)\times\binom{\ell'-1}{|t^i|}+\omega_{\ell'}^+(t[i],t)}{\left(\sum_{k<i\wedge t[k]\in\varphi}\omega_1(t[k],t)\right)\times\binom{\ell'}{|t^{i-1}|}+\left(\sum_{\star\in\{+,-\}}\omega_{\ell'}^\star(t[i],t)\right)}.$$

The probability that the item $t[i]$ is not drawn knowing φ and ℓ' is $1-\mathbb{P}_\ell^t(t[i]|\varphi,\ell')$.

The proofs of these two formulas follow from the fact that the probability of drawing $t[i]$ depends on the utilities of the items already drawn and those of the items which follow it to form a pattern of length ℓ.

Proof (Property 3). From the lemma 1, we have:
$\mathbb{P}_\ell^t(t[i]|\varphi,\ell') = \frac{\sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1}\texttt{uOcc}(\varphi\cup\{t[i]\}\cup\varphi',t)}{\sum_{\varphi'\subseteq t^{i-1}\wedge|\varphi'|=\ell'}\texttt{uOcc}(\varphi\cup\varphi',t)}$.
First, by definition we have $\texttt{uOcc}(\varphi\cup\{t[i]\}\cup\varphi',t) = \texttt{uOcc}(\varphi,t)+\texttt{uOcc}(\{t[i]\}\cup\varphi',t)$. Let $z_i = \sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1}\texttt{uOcc}(\varphi\cup\{t[i]\}\cup\varphi',t)$. That implies that $z_i = \sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1}(\texttt{uOcc}(\varphi,t)+\texttt{uOcc}(\{t[i]\}\cup\varphi',t))$. We then have:
$z_i = \sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1}\texttt{uOcc}(\varphi,t) + \sum_{\varphi'\subseteq t^i\wedge|\varphi'|=\ell'-1}\texttt{uOcc}(\{t[i]\}\cup\varphi',t)$. But

$\sum_{\varphi' \subseteq t^i \wedge |\varphi'|=\ell'-1} \mathtt{uOcc}(\varphi,t) = \mathtt{uOcc}(\varphi,t) \times \binom{\ell'-1}{|t^i|}$ and $\sum_{\varphi' \subseteq t^i \wedge |\varphi'|=\ell'-1} \mathtt{uOcc}(\{t[i]\} \cup \varphi', t) = \omega^{+}_{\ell'}(t[i],t)$ by definition. Then $z_i = \mathtt{uOcc}(\varphi,t) \times \binom{\ell'-1}{|t^i|} + \omega^{+}_{\ell'}(t[i],t)$. We also know that $\mathtt{uOcc}(\varphi,t) = \sum_{k<i \wedge t[k]\in\varphi} \omega_1(t[k],t)$. So we have : $z_i = \left(\sum_{k<i \wedge t[k]\in\varphi} \omega_1(t[k],t)\right) \times \binom{\ell'-1}{|t^i|} + \omega^{+}_{\ell'}(t[i],t)$.

Second, we have $\mathtt{uOcc}(\varphi \cup \varphi', t) = \mathtt{uOcc}(\varphi,t) + \mathtt{uOcc}(\varphi',t)$. By setting $Z_i = \sum_{\varphi' \subseteq t^{i-1} \wedge |\varphi'|=\ell'} \mathtt{uOcc}(\varphi \cup \varphi', t)$, we get then $Z_i = \sum_{\varphi' \subseteq t^{i-1} \wedge |\varphi'|=\ell'} \mathtt{uOcc}(\varphi,t) + \sum_{\varphi' \subseteq t^{i-1} \wedge |\varphi'|=\ell'} \mathtt{uOcc}(\varphi',t)$. But $\sum_{\varphi' \subseteq t^{i-1} \wedge |\varphi'|=\ell'} \mathtt{uOcc}(\varphi,t) = \mathtt{uOcc}(\varphi,t) \times \binom{\ell'}{|t^{i-1}|} = \left(\sum_{k<i \wedge t[k]\in\varphi} \omega_1(t[k],t)\right) \times \binom{\ell'}{|t^{i-1}|}$ et $\sum_{\varphi' \subseteq t^{i-1} \wedge |\varphi'|=\ell'} \mathtt{uOcc}(\varphi',t) = \sum_{\star\in\{+,-\}} \omega^{\star}_{\ell'}(t[i],t)$, so $Z_i = \left(\sum_{k<i \wedge t[k]\in\varphi} \omega_1(t[k],t)\right) \times \binom{\ell'}{|t^{i-1}|} + \sum_{\star\in\{+,-\}} \omega^{\star}_{\ell'}(t[i],t)$.

Finally, $\mathbb{P}^t_\ell(t[i]|\varphi,\ell') = \frac{z_i}{Z_i} = \frac{\left(\sum_{k<i \wedge t[k]\in\varphi} \omega_1(t[k],t)\right) \times \binom{\ell'-1}{|t^i|} + \omega^{+}_{\ell'}(t[i],t)}{\left(\sum_{k<i \wedge t[k]\in\varphi} \omega_1(t[k],t)\right) \times \binom{\ell'}{|t^{i-1}|} + \sum_{\star\in\{+,-\}} \omega^{\star}_{\ell'}(t[i],t)}$. □

Property 4 (Correctness). Let $\mathcal{D}$ be a transactional database having utilities on items with a total order relation $>_{\mathcal{I}}$, and m and M two integers such that $m \leq M$. HAISampler randomly draws a pattern φ from the language $\mathcal{L}(\mathcal{D})$ with a probability equal to $u^{avg}_{[m..M]}(\varphi,\mathcal{D})/Z$ where $Z = \sum_{\varphi' \in \mathcal{L}(\mathcal{D})} u^{avg}_{[m..M]}(\varphi',\mathcal{D})$ is the constant of normalization.

Proof (Property 4). Let m be the minimum and M the maximum length constraints, the probability of drawing the pattern φ of length ℓ in the database $\mathcal{D}$ denoted by $\mathbb{P}_{[m..M]}(\varphi,\mathcal{D})$, and Z a normalization constant defined by $Z = \sum_{\varphi' \in \mathcal{L}(\mathcal{D})} u^{avg}_{[m..M]}(\varphi',\mathcal{D})$.We know that : $\mathbb{P}_{[m..M]}(\varphi,\mathcal{D}) = \sum_{(j,t)\in\mathcal{D}} \left(\mathbb{P}_{[m..M]}(t_j,\mathcal{D}) \times \mathbb{P}_{[m..M]}(\varphi,t_j)\right)$. But $\mathbb{P}_{[m..M]}(t_j,\mathcal{D}) = \frac{\omega^{avgU}_{[m..M]}(t_j)}{Z}$, then

$$\mathbb{P}_{[m..M]}(\varphi,\mathcal{D}) = \sum_{(j,t)\in\mathcal{D}} \left(\frac{\omega^{avgU}_{[m..M]}(t_j)}{Z} \times \mathbb{P}_{[m..M]}(\varphi,t_j)\right). \quad (1)$$

We also know that: $\mathbb{P}_{[m..M]}(\varphi,t_j) = \mathbb{P}_{[m..M]}(\ell|t_j) \times \mathbb{P}^{t_j}_{[m..M]}(\varphi|\ell)$. (2)

But we have: $\mathbb{P}_{[m..M]}(\ell|t_j) = \frac{\omega^{avgU}_{[\ell..\ell]}(t_j)}{\omega^{avgU}_{[m..M]}(t_j)}$ and $\mathbb{P}^{t_j}_{[m..M]}(\varphi|\ell) = \frac{\mathtt{uOcc}(\varphi,t_j)}{\omega^{avgU}_{[\ell..\ell]}(t_j)\times\ell}$ then by substituting the two terms in (2) we obtain $\mathbb{P}_{[m..M]}(\varphi,t_j) = \frac{\omega^{avgU}_{[\ell..\ell]}(t_j)}{\omega^{avgU}_{[m..M]}(t_j)} \times \frac{\mathtt{uOcc}(\varphi,t_j)}{\omega^{avgU}_{[\ell..\ell]}(t_j)\times\ell} = \frac{\mathtt{uOcc}(\varphi,t_j)}{\omega^{avgU}_{[m..M]}(t_j)\times\ell}$.

Now, if we replace $\mathbb{P}_{[m..M]}(\varphi,t_j)$ in (1) by its last expression, we get: $\mathbb{P}_{[m..M]}(\varphi,\mathcal{D}) = \sum_{(j,t)\in\mathcal{D}} \left(\frac{\omega^{avgU}_{[m..M]}(t_j)}{Z} \times \frac{\mathtt{uOcc}(\varphi,t_j)}{\omega^{avgU}_{[m..M]}(t_j)\times\ell}\right) = \frac{1}{Z} \times \frac{\sum_{(j,t)\in\mathcal{D}} \mathtt{uOcc}(\varphi,t_j)}{\ell}$.

But by definition, we have $\frac{\sum_{(j,t)\in\mathcal{D}} \mathtt{uOcc}(\varphi,t_j)}{\ell} = u^{avg}_{[m..M]}(\varphi,\mathcal{D})$, so $\mathbb{P}_{[m..M]}(\varphi,\mathcal{D}) = \frac{u^{avg}_{[m..M]}(\varphi,\mathcal{D})}{Z}$. Hence the result. □

3 Additional experiments

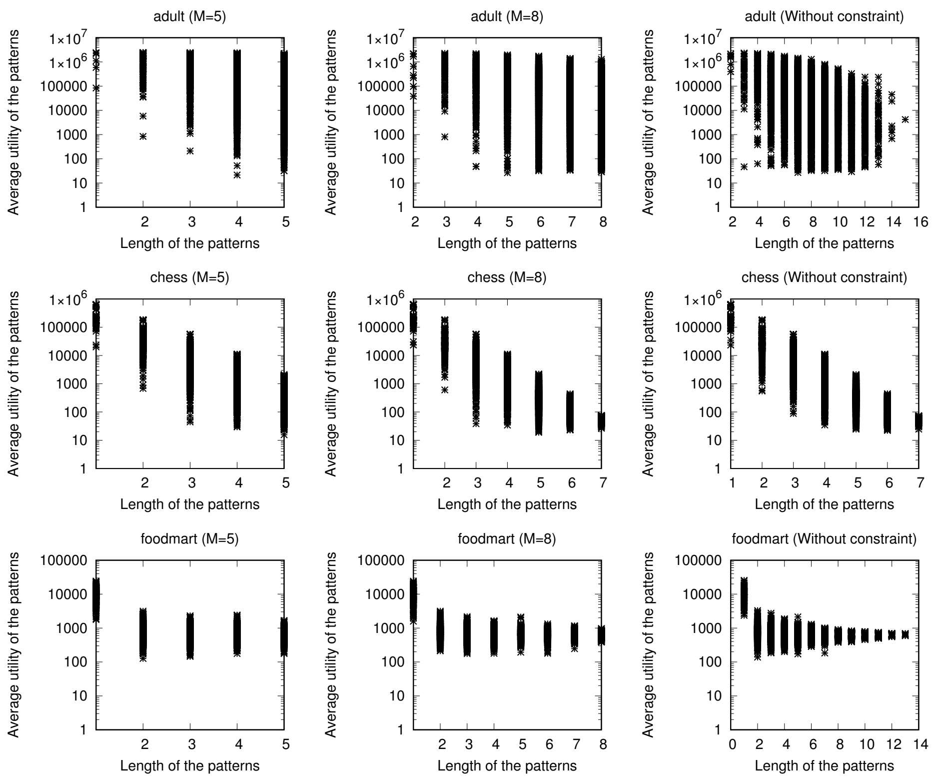

Fig. 1. Dispersion of average utilities of 10,000 sampled patterns