Меры сходства последовательностей

Муратова Анна

Определения

Пусть *I* – конечное множество объектов, а *X* – непустой набор объектов из *I*. **Последовательностью** *S* называется упорядоченное множество $< X_1 \dots X_n >$, где X_i – непустой набор объектов.

Размер последовательности обозначим |S|=n.

Длиной последовательности называется общее число объектов, встречающихся в последовательности $l(S) = \sum_{i=1}^{n} |X_i|$.

Определения

***l*-префиксом** S^l последовательности *S* называется последовательность из первых *l* элементов $< X_1 \ldots X_l >$.

j-й набор объектов X_j последовательности *S* обозначим за *S*[j].

Последовательность $T =< Y_1 \ldots Y_m >$ называется **подпоследовательностью** последовательности $S =< X_1 \ldots X_n >$, если существует $1 \leq i_1 < i_2 < \cdots < i_m \leq n$, что $Y_j \subseteq X_{i_j}$ для всех $j = 1 \ldots m,\ m < n$. Тогда последовательность *S* называется **надпоследовательностью** последовательности *T.*

Определения

Множество всех подпоследовательностей последовательности *S* обозначим за $\varphi(S)$, а $\phi(S) = |\varphi(S)|$. Обозначим за $\varphi(S,T) = \varphi(S) \cap \varphi(T)$ **множество всех общих подпоследовательностей** двух последовательностей *S* и *T* и аналогично $\phi(S,T) = |\varphi(S,T)|$.

Пусть $S = < X_1 \dots X_n >$ – последовательность, а *Y* – набор объектов. **Конкатенацией** набора объектов *Y* с последовательностью *S* $S \circ Y$ называется последовательность $< X_1 \dots X_n Y >$.

Примеры

Пусть $S_1 =< \{a\}\{a,b\}\{e\}\{c,d\}\{b,d\} >$

$S_2 =< \{a\}\{b,c,d\}\{a,d\} >$

$S_3 =< \{a\}\{b,d\}\{c\}\{a,d\} >$

$S_4 =< \{a\}\{a,b,d\}\{a,b,c\}\{b,d\} >$

Примером подпоследовательности последовательности S_1 является $< \{a\}\{b\}\{c,d\} >$.

Длина последовательности $l(S_1) = 8$, а размер последовательности $|S_1| = 5$.

3-префиксом является $S_1^3 =< \{a\}\{a,b\}\{e\} >$, а вторым набором объектов последовательности S_1 является $\{a,b\}$.

Примеры

$$S_1 =< \{a\}\{a,b\}\{e\}\{c,d\}\{b,d\} >$$

$$S_2 =< \{a\}\{b,c,d\}\{a,d\} >$$

$$S_3 =< \{a\}\{b,d\}\{c\}\{a,d\} >$$

$$S_4 =< \{a\}\{a,b,d\}\{a,b,c\}\{b,d\} >$$

Множеством всех подпоследовательностей является

$$\varphi(S_4^2) = \{<>, < \{a\} >, < \{b\} >, < \{d\} >, < \{a,b\} >, < \{a,d\} >, < \{b,d\} >, < \{a,b,d\} >, < \{a\}\{a\} >, < \{a\}\{b\} >, < \{a\}\{d\} >, < \{a\}\{a,b\} >, < \{a\}\{a,d\} >, < \{a\}\{b,d\} >, < \{a\}\{a,b,d\} >$$

Тогда $\phi(S_4^2) = 15.$

Примером конкатенации последовательности S_4^2 с набором последовательности $\{a,b,c\}$ $S_4^2 \circ \{a,b,c\}$ является последовательность $< \{a\}\{a,b,d\}\{a,b,c\} >.$

Меры сходства последовательностей

1. Мера все общие подпоследовательности

Вычисляется как число общих подпоследовательностей, деленное на максимальное число подпоследовательностей *S* и *T*. В обозначениях, приведенных выше, получим

$$sim_{ACS}(S,T) = \frac{\phi(S,T)}{\max\{\phi(S),\phi(T)\}}$$

Мера сходства удовлетворяет следующим условиям:

1) $sim_{ACS}(S,T) \geq 0$

2) $sim_{ACS}(S,T) = 1$, если $S = T$

3) $sim_{ACS}(S,T) = sim_{ACS}(T,S)$

Меры сходства последовательностей

Пример

Пусть

$$S_1 =< \{a\}\{a,b\}\{e\}\{c,d\}\{b,d\} >$$

$$S_2 =< \{a\}\{b,c,d\}\{a,d\} >$$

$$S_3 =< \{a\}\{b,d\}\{c\}\{a,d\} >$$

$$S_4 =< \{a\}\{a,b,d\}\{a,b,c\}\{b,d\} >$$

Множество всех общих подпоследовательностей S_1^4 и S_2^3 равно

$$\varphi(S_1^4, S_2^3) = \{<>, <\{a\}>, <\{b\}>, <\{d\}>, <\{c\}>, <\{c,d\}>, <\{a\}\{a\}>, \\ <\{a\}\{b\}>, <\{a\}\{c\}>, <\{a\}\{d\}>, <\{a\}\{c,d\}>, <\{b\}\{d\}>, \\ <\{a\}\{b\}\{d\}>$$

Тогда сходство последовательностей равно

$$sim_{ACS}(S_1^4, S_2^3) = \frac{\phi(S_1^4, S_2^3)}{\max\{\phi(S_1^4), \phi(S_2^3)\}} = \frac{13}{\max\{56,61\}} = \frac{13}{61} = 0.21$$

Меры сходства последовательностей

2. Мера самая длинная общая подпоследовательность

Здесь рассматривается только самая длинная из подпоследовательностей. Существует два способа ее измерения – использовать размер или длину последовательности.

- Используя размер $$sim_{LCS_{size}}(S,T) = \frac{|LCS(S,T)|}{\max\{|S|,|T|\}}$$
- Используя длину $$sim_{LCS_{length}}(S,T) = \frac{l(LCS(S,T))}{\max\{l(S),l(T)\}}$$

Недостатком данной меры является тот факт, что она не учитывает информацию, содержащуюся во второй по длине, третьей и так далее последовательностях.

Меры сходства последовательностей

Пример

Пусть $S = < \{a, b\}\{a, c\}\{a, f\} >$

$T = < \{a, b, f\}\{c\} >$

- Используя размер $sim_{LCS_{size}}(S, T) = \frac{|LCS(S, T)|}{\max\{|S|, |T|\}} = \frac{2}{3}$
- Используя длину $sim_{LCS_{length}}(S, T) = \frac{l\left(LCS(S, T)\right)}{\max\{l(S), l(T)\}} = \frac{3}{6} = \frac{1}{2}$

Подсчет всех разных отличных друг от друга подпоследовательностей

Рассмотрим метод подсчета числа $\phi(S)$ всех отличающихся подпоследовательностей для данной последовательности *S*. Предположим, что мы хотим расширить данную последовательность $S =< X_1, \dots, X_n >$ набором объектов *Y*. Рассмотрим отношение между $\phi(S)$ и $\phi(S \circ Y)$.

1) В *Y* нет ни одного набора объектов из *S*, тогда число отличающихся подпоследовательностей $S \circ Y$ равно $|\varphi(S)| \cdot 2^{|Y|}$

2) Хоть один объект из *Y* встречается в наборе объектов *S*, тогда число отличающихся последовательностей $S \circ Y$ равно $|\varphi(S)| \cdot 2^{|Y|} - R(S,Y)$, где *R(S, Y)* – корректирующий член, который равен числу повторений подпоследовательностей, которые должны быть убраны из последовательности *S*, конкатенирующей с набором объектов *Y*

Подсчет всех разных отличных друг от друга подпоследовательностей

Набор позиций указывает на позиции наборов объектов, в которых имеются дубликаты из *S*. Пусть дана последовательность $S =< X_1 \ldots X_n >$ и набор объектов *Y.*

L(S, Y) – набор всех максимальных позиций, где набор объектов *Y* имеет максимальное пересечение с разными наборами объектов *S*[*i*], *i=1…n*. Если существует несколько позиций, которые образуют одни и те же дубликаты, то необходимо рассматривать только последнюю из них, самую правую в последовательности.

$$L(S,Y) = \{i \mid S[i] \cap Y \neq \emptyset \wedge \forall j; j > i \wedge S[i] \cap Y \not\subseteq S[j] \cap Y\}$$

Подсчет всех разных отличных друг от друга подпоследовательностей

Пусть *S* последовательность, а *Y* набор объектов, тогда

$\phi(S \circ Y) = \phi(S) \cdot 2^{|Y|} - R(S,Y)$, где

$$R(S,Y) = |\cup_{l \in L} \{\varphi(S^{l-1}) \circ P_{\geq 1}(S[l] \cap Y)\}|$$

и наборы $\varphi(S^{l-1}) \circ P_{\geq 1}(S[l] \cap Y)$ не обязательно не пересекаются.

Теорема 1

Пусть *S* последовательность, а *Y* набор объектов, тогда

$\phi(S \circ Y) = \phi(S) \cdot 2^{|Y|} - R(S,Y)$, где

$$R(S,Y) = \sum_{K \subseteq L(S,Y)} (-1)^{|K|+1} \left(\phi\left(S^{\min(K)-1}\right) \cdot \left(2^{|(\cap_{j \in K} S[j]) \cap Y|} - 1\right)\right)$$

Подсчет всех общих подпоследовательностей

Расширим предыдущие результаты для подсчета всех общих непересекающихся подпоследовательностей между двумя последовательностями *S* и *T*. Пусть мы хотим расширить последовательность *S* набором объектов *Y* и будем рассматривать отношение между $\varphi(S,T)$ и $\varphi(S \circ Y,T)$.

1) Ни один из объектов *Y* не встречается в наборах объектов *S* и *T*, тогда конкатенация набора объектов *Y* с последовательностью *S* не имеет никакого эффекта на набор $\varphi(S,T)$

2) По крайней мере, один из объектов *Y* встречается в одной из последовательностей *S* или *T*, или в обоих. В этом случае, новые общие подпоследовательности могут встретиться в $\varphi(S,T)$.

Повторения могут встречаться, поэтому определим общий корректирующий член для *S* и *T*.

Подсчет всех общих подпоследовательностей

2) По крайней мере, один из объектов *Y* встречается в одной из последовательностей *S* или *T*, или в обоих. В этом случае, новые общие подпоследовательности могут встретиться в $\varphi(S,T)$.

Повторения могут встречаться, поэтому определим общий корректирующий член для *S* и *T*.

$$|\varphi(S \circ Y, T)| = |\varphi(S,T)| + A(S,T,Y) - R(S,T,Y),$$

где *A(S, T, Y)* – число экстра общих подпоследовательностей, которые должны быть добавлены, а *R(S, T, Y)* – корректирующий член.

Тут набор позиций также указывает на позиции, которые образуют дублирующие последовательности.

Подсчет всех общих подпоследовательностей

Пусть $S =< X_1 \dots X_n >, T =< X'_1 \dots X'_m >$ и Y – набор объектов, тогда

$$A(S,T,Y) = \left|\cup_{l \in L(T,Y)} \{\varphi(S, T^{l-1}) \circ P_{\geq 1}(T[l] \cap Y)\}\right|$$

$$R(S,T,Y) = |\cup_{l \in L(S,Y)} \{\cup_{l' \in L(T,Y)} \{\varphi(S^{l-1}, T^{l'-1}) \circ P_{\geq 1}(S[l] \cap T[l'] \cap Y)\}\}|$$

Теорема 2

Пусть $S =< X_1 \dots X_n >, T =< X'_1 \dots X'_m >$ и Y – набор объектов, тогда

$$\phi(S \circ Y, T) = \phi(S,T) + A(S,T,Y) - R(S,T,Y)$$

$$A(S,T,Y) = \sum_{K \subseteq L(T,Y)} (-1)^{|K|+1} \left(\phi\left(S, T^{\min(K)-1}\right) \cdot \left(2^{\left|(\cap_{j \in K} T[j]) \cap Y\right|} - 1\right)\right)$$

$$R(S,T,Y) = \sum_{K \subseteq L(S,Y)} (-1)^{|K|+1} \left(\sum_{K' \subseteq L(T,Y)} (-1)^{|K'|+1} \cdot f(K,K')\right)$$

$$f(K,K') = \phi\left(S^{\min(K)-1}, T^{\min(K')-1}\right) \cdot \left(2^{\left|(\cap_{j \in K} S[j]) \cap \left(\cap_{j' \in K'} T[j']\right) \cap Y\right|} - 1\right)$$

Спасибо за внимание!