Персистентные гомологии для марковских цепей и их применение к анализу текста на естественном языке

Кушнарева Л.П., Кузьминых Д.В.

В работе предложен новый метод введения персистентных топологических инвариантов для марковских цепей. Продемонстрирован пример использования этих инвариантов для прикладной задачи классификации текстов.

Ключевые слова: Топологический анализ данных, персистентные гомологии, марковские цепи, текст на естественном языке

Введение

Персистентные гомологии - один из основных инструментов топологического анализа данных - молодой дисциплины, которая исследует возможности выделения внутренней структуры в экспериментальных данных различной природы и введения топологических инвариантов на этой структуре. Это позволяет применять методы из алгебраической топологии для анализа данных и решения связанных с ними прикладных задач. В статье [1] приведен пример того, как можно использовать топологические методы для анализа изображений. В частности, автор сопоставляет каждому небольшому фрагменту изображения вектор в пространстве большой размерности. И оказывается, что векторы, соответствующие фрагментам изображений, полученным из реальных фотографий, лежат в окрестности небольшого подмногообразия пространства векторов, соответствующих произвольным (случайным) фрагментам. Таким образом, появляется возможность описать на языке алгебраической топологии, чем осмысленные изображения отличаются от бессмысленных.

Марковские цепи позволяют моделировать процессы в большом количестве приложений. Однако, по описанию марковских цепей в явном виде может быть трудно оценить "структурное" сходство или различие

таких процессов. Это приводит к вопросу о том, могут ли марковские цепи быть классифицированы аналогично тому, как это сделано для комплексов и поверхностей методами алгебраической топологии. В данной работе мы предлагаем метод введения топологических инвариантов для марковских цепей, в качестве основы используя идеи из статьи [**1**].

Мы преобразуем методологию, используемую в [**1**], так, чтобы она была применима к марковским цепям и введем на них инвариант - аналог персистентных гомологий, который позволяет выделять отличительные признаки различных марковских цепей. В качестве основного примера марковской цепи в данной работе будет использоваться марковская цепь, построенная на основе текста на естественном языке. В частности, будет показано, каким образом введенные на марковских цепях топологические инварианты позволяют отличать цепи, построенные по осмысленным текстам, от цепей, построенных по случайно сгенерированным текстам с той же частотой слов.

1. Предварительные сведения. Персистентность

Для удобства читателя, выпишем базовые определения теории категорий, на которые опирается определение персистентного объекта.

Определение. Категория $\underline{K}$ - это совокупность класса $Ob(\underline{K})$, элементы которого называются объектами категории $\underline{K}$, и класса $Mor(\underline{K})$, элементы которого называются морфизмами категории $\underline{K}$. Объекты и морфизмы категории должны быть связаны между собой следующими условиями:

1) Каждой упорядоченной паре объектов $a, b \in \underline{K}$ сопоставлено некоторое множество $H_{\underline{K}}(a, b)$ морфизмов категории $\underline{K}$.

2) Каждый морфизм категории $\underline{K}$ принадлежит одному и только одному из множеств $H_{\underline{K}}(a, b)$.

3) В классе $Mor(\underline{K})$ введена операция композиции. Композиция морфизмов $\alpha \in H_{\underline{K}}(a, b)$, $\beta \in H_{\underline{K}}(b, c)$ дает морфизм $\alpha\beta \in H_{\underline{K}}(a, c)$. Операция ассоциативна, т.е. $\alpha(\beta\gamma) = (\alpha\beta)\gamma$ для любых трех морфизмов $\alpha \in H_{\underline{K}}(a, b)$, $\beta \in H_{\underline{K}}(b, c)$, $\gamma \in H_{\underline{K}}(c, d)$.

4) В каждом множетве $H_{\underline{K}}(a, a)$ содержится такой морфизм 1_a, что $\alpha 1_a = \alpha$ и $1_a\beta = \beta$, для произвольных морфизмов $\alpha \in H_{\underline{K}}(x, a)$,

$\beta \in H_{\underline{K}}(a, y)$. Такой морфизм 1_a называется тождественным или единичным морфизмом объекта a.

Определение. Одноместным ковариантным функтором из категории $\underline{K}$ в категорию $\underline{L}$ называется отображение $F : \underline{K} \to \underline{L}$, удовлетворяющее условиям:

1) Для всякого $a \in Ob(\underline{K})$, $F(a) \in Ob(\underline{L})$
2) Для всякого $\alpha \in H_{\underline{K}}(a, b)$, $F(\alpha) \in H_{\underline{L}}(F(a), F(b))$
3) Для всякой единицы $1_a \in \underline{K}$, $F(1_a) = 1_{F(a)}$
4) Для $\alpha \in H_{\underline{K}}(a, b)$, $\beta \in H_{\underline{K}}(b, c)$ имеет место равенство $F(\alpha\beta) = F(\alpha)F(\beta)$.

В дальнейшем будем под "функтором"понимать одноместный ковариантный функтор, если не сказано иное.

Определение Пусть $\underline{C}$ - произвольная категория, а P - частично упорядоченное множество. Наделим P категорной структурой следующим образом: возьмем P в качестве множества объектов и будем считать, что из объекта x в объект y из C существует морфизм если и только если $x \leq y$. Тогда функтор $\Phi : \underline{P} \to \underline{C}$ мы будем называть P-персистентным объектом.

Более конкретно, это означает наличие семейства$\{c_x\}_{x \in P}$ объектов в $\underline{C}$ вместе с морфизмами ϕ такими, что $\phi : c_x \to c_y <=> x \leq y$, и $\phi_{yz}\phi_{xy} = \phi_{xz} <=> x \leq y \leq z$.

2. Основная конструкция

Воспользуемся следующими определениями марковской цепи и матрицы переходов из [3] :

Определение. Пусть $(\Omega, \mathcal{A}, P)$ - конечное вероятностное пространство и $M = (\xi_0, ..., \xi_n)$- последовательность случайных величин со значениями в конечном множестве X. Последовательность $(\xi_0, ..., \xi_n)$ называется конечной марковской цепью, если выполнено условие $P(\xi_{k+1} = a_{k+1} | \xi_k = a_k, ..., \xi_0 = a_0) = P(\xi_{k+1} = a_{k+1} | \xi_k = a_k)$ для любых $a_0, ..., a_k$ таких, что $P\{\xi_k = a_k, ..., \xi_0 = a_0\} > 0$.

Определение. Множество X называется пространством состояний цепи, а матрица $||p(x, y)||$, $x, y \in X$, где $p(x, y) = P(\xi_k = y | \xi_{k-1} = x)$ -

матрицей переходных вероятностей. Если матрица переходных вероятностей не меняется со временем, соответствующая ей марковская цепь называется стационарной.

В дальнейшем в данной работе мы будем под "марковской цепью" всегда подразумевать конечную стационарную марковскую цепь M с фиксированным начальным состоянием I, если не сказано иное.

Определение.Путем ω в марковской цепи M назовем произвольную последовательность переходов между её состояниями $I, x_{i_1}, ..., x_{i_n}, ...$, такую, что вероятности переходов $p(x_{i_j}, x_{i_{j+1}}) > 0$ для всех $j \in \mathbb{Z}_+$. Пространство всех путей для данной марковской цепи M с фиксированным начальным состоянием I будем обозначать $\Omega_I(M)$.

Объединение пространств путей со всеми возможными начальными состояниями марковской цепи M будем обозначать $\Omega(M) = \bigcup\limits_{x \in X} \Omega_x(M)$, где X - множество состояний исходной марковской цепи.

Определение. Построим по произвольной марковской цепи M взвешенный ненаправленный граф $\Gamma_i = (V, W_i)$ следующим образом:

- В качестве вершин графа Γ_i возьмем состояния исходной марковской цепи M.

- Определим вспомогательную функцию $\xi(\{x_1, x_2\}) : \Omega_I(M) \to \mathbb{Z}_+ \cup \{\infty\}$, где $\{x_1, x_2\}$ - неупорядоченная пара вершин (состояний) исходной марковской цепи, следующим образом: функция ξ сопоставляет каждому возможному пути с началом в I (лежащему в марковской цепи M), количество раз, которое ребра (x_1, x_2) и (x_2, x_1) (т.е. произвольное ребро с концами в x_1 и x_2 без учета направления) встретились в данном пути.

  Далее, с помощью функции ξ введем весовую функцию на каждом ненаправленном ребре $e = \{x_1, x_2\}$:

  $$\omega_i(e) = \omega_i(\{x_1, x_2\}) = P(\xi(\{x_1, x_2\}) \geq i).$$

  Эта функция сопоставляет каждой паре вершин x_1 и x_2 вероятность пройти по ребру с концами в этих вершинах (в любом из двух направлений) не менее, чем i раз, начиная путь из вершины I.

Определение.Для каждого графа Γ_i, построенного в предыдущем определении и произвольного $p \in [0, 1]$, построим невзвешенный ненаправленный граф Γ_i^p следующим образом:

- В качестве вершин графа Γ_i^p возьмем вершины графа Γ_i (которые, в свою очередь, соответствуют состояниям исходной марковской цепи M).

- Будем считать, что ребро $e = \{x_1, x_2\}$ присутствует в графе Γ_i^p тогда и только тогда, когда вес соответствующего ребра $\{x_1, x_2\}$ в графе Γ_i не меньше, чем p: $w_i(\{x_1, x_2\}) \geq p$.

Полученный таким образом граф Γ_i^p будем называть фильтрацией графа Γ_i по вероятности p.

Лемма 1. *Пусть графы Γ_i^a и Γ_i^b построены по одной и той же марковской цепи с фиксированным начальным состоянием, $i \in \mathbb{N}$, а вероятности a и b удовлетворяют неравенству $0 \leq a < b \leq 1$. Тогда имеет место включение $\Gamma_i^a \supseteq \Gamma_i^b$.*

Доказательство. Графы Γ_i^a и Γ_i^b имеют одно и то же множество вершин по построению. Далее, для каждого произвольно взятого ребра e графа Γ_i^b выполняется неравенство $w_i(e) \geq b > a$. Следовательно, e также является ребром и в Γ_i^a, и мы имеем включение по ребрам. □

Лемма 2. *Пусть графы Γ_i^p и Γ_{i+1}^p также построены по одной и той же марковской цепи с фиксированным начальным состоянием, а $i \in \mathbb{N}$. Тогда для любого $p \in [0, 1]$ имеет место включение $\Gamma_i^p \supseteq \Gamma_{i+1}^p$.*

Доказательство. Множества вершин указанных графов, опять же, совпадают по построению. Далее, если некоторое ребро e принадлежит графу Γ_{i+1}^p, то имеет место неравенство $P(\xi(e) \geq i) \geq P(\xi(e) \geq i + 1) \geq p$. Следовательно, e также присутствует и в графе Γ_i^p, что дает искомое включение.

□

Теорема 1. *Рассмотрим объединение множеств*

$$P_{Th} = \bigcup_{i \in \mathbb{N}} \bigcup_{e \in \Gamma_i} \omega_e = \bigcup_{i \in \mathbb{N}} \bigcup_{e \in \Gamma_i} P(\xi_e \geq i) = \{p_1, \ldots, p_N\},$$

где $0 \leq p_1 \leq ... \leq p_N \leq 1$.

Тогда имеет место таблица включений:

$$\begin{array}{ccccccc}
\Gamma_1^{p_1} & \supseteq & \Gamma_1^{p_2} & \supseteq \ldots \supseteq & \Gamma_1^{p_n} & \supseteq \ldots \\
\cup| & & \cup| & & \cup| & \\
\Gamma_2^{p_1} & \supseteq & \Gamma_2^{p_2} & \supseteq \ldots \supseteq & \Gamma_2^{p_n} & \supseteq \ldots \\
\cup| & & \cup| & & \cup| & \\
\vdots & & \vdots & & \vdots &
\end{array}$$

Доказательство. Вертикальные включения - следствие утверждения 2; горизонтальные включения - следствие утверждения 1. □

Замечание. Из указанных утверждений также следует, что для любых натуральных m и n, таких, что $m \geq n$ и любых вещественных a и b, таких, что $0 \leq a \leq b \leq 1$, имеют место следующие включения:

$$\begin{array}{ccc}
\Gamma_n^a & \supseteq & \Gamma_n^b \\
\cup| & & \cup| \\
\Gamma_m^a & \supseteq & \Gamma_m^b
\end{array}$$

.

Определение. Построим по каждому графу $\Gamma_k^{p_j}$ абстрактный симплициальный комплекс $K_k^{p_j}$ следующим образом:

- Каждой вершине v_i графа $\Gamma_k^{p_j}$ сопоставим (взаимно однозначно) нульмерный симплекс $\{\Delta_i\}$ и включим его в комплекс $K_k^{p_j}$.
- Если вершины $\{v_{i_1}, ..., v_{i_n}\}$ графа $\Gamma_k^{p_j}$ образуют клику, добавим в комплекс $K_k^{p_j}$ $n-1$-мерный симплекс, посмтроенный на соответствующих им вершинах $\{\Delta_{i_1}, ..., \Delta_{i_n}\}$.
- Никаких других симплексов, кроме добавленных по правилам, перечисленным выше, комплекс $K_k^{p_j}$ не содержит.

Замечание. Поскольку каждая клика графа содержит все свои подклики, каждый симплекс комплекса $K_k^{p_j}$ также содержит все свои подсимплексы, а значит, $K_k^{p_j}$ действительно является симплициальным комплексом.

Теорема 2. *Для любых натуральных m и n, таких, что $m \geq n$ и любых $a, b \in [0, 1]$, таких, что $b \geq a$, следующая диаграмма коммутативна:*

$$\begin{array}{ccc} H_l(K_n^a) & \to & H_l(K_n^b) \\ \downarrow & \searrow & \downarrow \\ H_l(K_m^a) & \to & H_l(K_m^b) \end{array} \qquad (*)$$

где H_l - l-я группа симплициальных гомологий с коэффициентами в целых числах.

Доказательство. Доказанные вложения для графов индуцируют вложения клик, а значит, и вложения цепей симплексов соответствующих симплициальных комплексов:

$$\begin{array}{ccc} C_l(K_n^a) & \supseteq & C_l(K_n^b) \\ \cup| & & \cup| \\ C_l(K_m^a) & \supseteq & C_l(K_m^b) \end{array} \qquad (**)$$

где $C_l(K_n^a)$ - группа l-цепей комплекса K_n^a.

Эти вложения, в свою очередь, можно рассматривать как соответствующие гомоморфизмы - проекции. А именно, если $C_l \supseteq C_l'$, то мы имеем отображение $f : C_l \to C_l'$, однозначно определенное следующим образом: f тождественно на элементах базиса C_l' (как подмножестве C_l), а остальные элементы базиса C_l переводит в ноль.

Заметим, что поскольку проекции, очевидно, коммутируют, то при замене вложений в (**) на эти проекции, мы получаем коммутативную диаграмму.

Будем в дальнейшем для простоты обозначать все отображения, построенные таким образом для различных цепей, одним и тем же символом - f. Каждое конкретное построенное по такому принципу отображение, очевидно, восстанавливается однозначно, если мы знаем, какие именно группы цепей представляют собой его область определения и область значения.

Заметим, что для построенного таким образом гомоморфизма - проекции f выполняется соотношение:

$$f\delta(\Delta) = f(\sum_i (-1)^i [v_0, ..., \hat{v_i}, ..., v_n]) =$$
$$\sum_i (-1)^i f([v_0, ..., \hat{v_i}, ..., v_n]) = \delta f(\sigma),$$

где $\Delta = [v_0, ..., v_n]$ - произвольный n-мерный симплекс.

Таким образом, следующие диаграммы коммутативны:

$$
\begin{array}{ccccccccc}
... & \to & C_{k+1}(K_n^a) & \stackrel{\delta}{\to} & C_k(K_n^a) & \stackrel{\delta}{\to} & C_{k-1}(K_n^a) & \stackrel{\delta}{\to} & ... \\
 & & \downarrow f & \searrow & \downarrow f & \searrow & \downarrow f & & \\
... & \to & C_{k+1}(K_m^a) & \stackrel{\delta}{\to} & C_k(K_m^a) & \stackrel{\delta}{\to} & C_{k-1}(K_m^a) & \stackrel{\delta}{\to} & ...
\end{array}
$$

и

$$
\begin{array}{ccccccccc}
... & \to & C_{k+1}(K_n^a) & \stackrel{\delta}{\to} & C_k(K_n^a) & \stackrel{\delta}{\to} & C_{k-1}(K_n^a) & \stackrel{\delta}{\to} & ... \\
 & & \downarrow f & \searrow & \downarrow f & \searrow & \downarrow f & & \\
... & \to & C_{k+1}(K_n^b) & \stackrel{\delta}{\to} & C_k(K_n^b) & \stackrel{\delta}{\to} & C_{k-1}(K_n^b) & \stackrel{\delta}{\to} & ...,
\end{array}
$$

где δ - граничное отображение, $C_k(K_n^a)$ - группа k-мерных цепей комплекса (K_n^a).

По свойству функториальности симплициальных гомологий, указанные цепные отображения индуцируют отображения в гомологиях

$f_* : H_k(K_n^a) \to H_k(K_m^a)$,

$f_* : H_k(K_n^a) \to H_k(K_n^b)$,

и $f_* : H_k(K_n^a) \to H_k(K_m^b) \ \ \forall k \in \mathbb{Z} \cup \{\infty\}$

и коммутативность диаграммы (*).

□

Следствие 1. *Пусть $\underline{C}$ - категория групп гомологий с гомоморфизмами в качестве морфизмов. Далее, пусть $\underline{P}$ - категория вещественных чисел из отрезка $[0, 1]$ с отношением порядка в качестве морфизмов. Тогда существует функтор $\Phi : \underline{C} \to \underline{P}$, являющийся персистентным объектом.*

Доказательство. Непосредственно следует из предыдущей теоремы и определения персистентного объекта (Определение 3).

А именно, для каждой пары вероятностей a, b такой, что $a \leq b$, мы построили соответствующий гомоморфизм $H_l(K_n^a) \to H_l(K_n^b)$ для всех натуральных l и n. Таким образом, мы уже построили искомый функтор, а предыдущая теорема показывает его персистентность.

□

3. Численный эксперимент

3.1. Набор данных

Для эксперимента использовалось 50 первых блогов из открытого набора данных “The Blog Authorship Corpus”. Этот набор данных состоит из текстов интернет-блогов 19,320 авторов с сайта blogger.com. Каждый блог включает в себя не менее 200 вхождений часто используемых английских слов и разбит на отдельные посты. Для более подробной информации о наборе данных отсылаем читателя к [5].

3.2. Предварительная обработка данных

Для всего корпуса (состоящего из выбранных 50 блогов) как цельного объекта, была проделана следующая предварительная обработка:

- Были удалены все дополнительные символы, кроме латинских букв и пробелов;
- Был применен Porter stemming algorithm для нормализации формы слов;
- Был составлен частотный словарь 99 наиболее часто встречающихся в корпусе слов;
- Слова, не попавшие в словарь, были названы “редкими” и заменены всюду в тексте на специальное слово “RARE_WORD”. Это слово было включено в частотный словарь с частотой, равной сумме частот встречаемости отдельных “редких” слов.

  Таким образом, был получен общий для всего корпуса словарь из 100 слов.

После этого для каждого блога в отдельности было проделано следующее:

- Были вычислены частоты встречаемости каждого слова из общего для всего корпуса словаря, включая “редкое слово”, в данном конкретном блоге.

- По полученным частотам был сгенерирован случайный текст с тем же словарем и той же длины, что и данный блог, с той же вероятностью встречаемости каждого слова.

  Отметим, что таким образом, для каждого блога был построен случайный текст, неотличимый от него с помощью простых алгоритмов, не учитывающих порядок слов (таких, как bag of words).

- По каждому тексту - и случайному, и неслучайному - была построена марковская цепь. А именно, каждому слову x частотного словаря было сопоставлено ровно одно состояние марковской цепи, которое мы обозначаем для простоты тем же символом. В качестве вероятности $p(x, y)$ перехода из состояния x в состояние y была назначена вероятность встретить слово y в данном тексте сразу после слова x.

  Но поскольку мы не можем знать точно указанные вероятности, в качестве их приближений были приняты эмпирические вероятности, вычисленные отдельно на основе каждого текста.

- Для каждой полученной таким способом марковской цепи методом Монте-Карло были вычислены матрицы смежностей соответствующего графа Γ_1. Подробности этого вычисления описаны в следующем параграфе.

3.3. Оценка весов графа Γ_1; построение графов Γ_1^p для различных значений p и вычисление их гомологий.

Для вычисления графа Γ_1 для каждой марковской цепи запускалось 100 случайных блужданий из каждого возможного начального состояния. Блуждание останавливалось, если мы пришли в конечное слово (после которого ничего не встречается) или совершили более 1000 шагов (ограничение введено для того, чтобы программа всегда заканчивала работу). Затем подсчитывалось, сколько проходов по каждому ребру было сделано в процессе каждого блуждания. Результаты усреднялись сначала по всем проходам, а затем - по всем начальным состояниям.

Далее, на основе каждого графа Γ_1 были вычислены по 8 графов $\Gamma_1^{p_1}, ...\Gamma_1^{p_8}$ для восьми равно отстоящих друг от друга значений p_i. Затем были вычислены размерности нулевых групп гомологий (то есть, количество компонентов связности) каждого из этих графов. Из этих восьми чисел Бетти был составлен вектор. Каждый такой вектор был дополнен

числом 100 (максимально возможное количество компонент связности) в качестве константы. Далее мы назначили метку 1 векторам, которые соответствуют текстам, написанным человеком, и 0 - векторам, которые соответствуют текстам, сгенерированным случайно.

Таким образом, мы сопоставили каждому тексту 9 целых чисел от 0 до 100 (включая константу) в качестве признаков и один бинарный параметр в качестве метки.

Приведем в качестве примера векторы признаков, полученных на основе текстов пяти первых блогов:

[5, 29, 61, 75, 80, 88, 91, 96, 100],

[1, 43, 65, 77, 86, 93, 94, 97, 100],

[3, 34, 54, 74, 85, 89, 92, 95, 100],

[4, 35, 60, 78, 80, 88, 93, 96, 100],

[4, 41, 67, 81, 86, 91, 93, 95, 100]

И векторы признаков, полученных на основе соответствующих им случайно сгенерированных текстов:

[1, 17, 52, 65, 80, 84, 90, 93, 100],

[2, 23, 63, 72, 78, 87, 92, 93, 100],

[1, 32, 55, 73, 84, 87, 91, 93, 100],

[3, 29, 57, 72, 82, 86, 92, 94, 100],

[3, 37, 64, 76, 81, 88, 91, 94, 100]

Отметим, что каждая отдельная компонента векторов из первой группы в среднем больше соответствующей компоненты векторов из второй группы, что наталкивает на мысль о линейной разделимости соответствующих классов. Поэтому мы использовали их в качестве признаков для обучения линейного классификатора - логистической регрессии.

Мы перемешали все вектора случайным образом и разбили их на два набора - тренировочный и тестовый в отношении 9:1. После этого на тренировочном наборе была обучена логистическая регрессия методом стохастического градиентного спуска с регуляризатором $C = 1$ и максимальным количеством итераций 1000.

3.4. Результат эксперимента

Эксперимент показал, что вектора из тестового множества, соответствующие осмысленным и бессмысленным текстам, линейно разделимы с точностью от 0.96 до 1 с помощью логистической регрессии.

4. Выводы

В данной статье мы ввели новые инварианты - персистетные гомологии на марковских цепях - и доказали их основные свойства. Был также приведен пример их эффективного использования в решении одной из важных прикладных задач, а именно - классификации текстов на естественном языке.

Тем не менее, возможности применения разработанного нами инструмента не ограничиваются только данной конкретной прикладной задачей. Как уже было отмечено, он потенциально может оказаться полезным и во многих других областях, где требуется находить инварианты марковских цепей. Но это уже тема для следующих исследований.

5. Благодарности

Авторы выражают благодарность А.А. Ирматову за поддержку в работе - помощь в уточнении формулировок и определений, а также отдельных деталей доказательств.

Список литературы

[1] G. Carlsson, "Topology And Data", *Bulletin (new series) of the American Mathematical Society*, **46**:2 (April 2009), 255–308.

[2] Н.Бурбаки, *Теория множеств*, пер. с фр. Г.Н.Поварова и Ю.А.Шихановича, ред. В.А.Успенского, Мир, М., 1965.

[3] А.Н. Ширяев, *ВЕРОЯТНОСТЬ*. Т. 1, 4-е изд., переработ. и доп, МЦНМО, М., 2007.

[4] А.Т.Фоменко, Д.Б.Фукс, *Курс гомотопической топологии*, Наука, М., 1989.

[5] J. Schler, M. Koppel, S. Argamon and J. Pennebaker, "Effects of Age and Gender on Blogging", Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs (2006), 2006.

Persistent homology of Markov chains and its application to natural language processing

Kushnareva L., Kuzminykh D.

In this work we introduce new persistent topological invariants for Markov chains. We also demonstrate the way of using these invariants for natural language processing task.

Keywords: Topology data analysis, persistent homology, Markov chains, natural language processing