Similitud topológica de usuarios

Aplicación a los sistemas de recomendación

Jesús Serrano
UIMP
Santander, España
serranopriego@posgrado.uimp.es

Fernando Díez
Universidad Autónoma de Madrid
Madrid, España
fernando.diez@uam.es

Alejandro Bellogín
Universidad Autónoma de Madrid
Madrid, España
alejandro.bellogin@uam.es

ABSTRACT

La topología es la rama de las Matemáticas que analiza los conjuntos desde un punto de vista diferente, estudiando su forma, transformaciones, relaciones, etc. Tradicionalmente, en el estudio y desarrollo de sistemas de recomendación, los algoritmos implementados, las métricas empleadas, los métodos de evaluación, etc. están basados en el uso de los datos desde una perspectiva Euclídea, la cual difiere completamente de la representación topológica de los mismos. En el presente artículo exponemos una línea de trabajo en la que los perfiles de usuarios pasan a contemplarse como espacios topológicos. Para poder comprender a los usuarios desde un punto de vista topológico nos hemos centrado en su caracterización usando sus correspondientes números de Betti. Estos números representan una propiedad característica de los conjuntos. A partir de esta y otras propiedades se definen caracterizaciones como, por ejemplo, los códigos de barras. Cada código de barras representa, desde un punto de vista topológico, un perfil de usuario. Mediante el cálculo de similitudes basadas en códigos de barras comparamos perfiles. Esta nueva perspectiva del tratamiento de datos es parte de lo que se denomina Análisis Topológico de Datos y abre nuevas perspectivas para el tratamiento de perfiles de usuario y la mejora de la personalización de productos y servicios.

KEYWORDS

Sistemas de Recomendación, Análisis Topológico de Datos, Modelado de usuario, Códigos de Barras, Números de Betti.

1 Introducción

El Análisis Topológico de Datos (o TDA, por sus siglas en inglés) es un área de la topología computacional que desarrolla técnicas para el análisis de datos desde la perspectiva de la Topología, rama de las Matemáticas que estudia las propiedades invariantes de los conjuntos cuando se aplican transformaciones continuas como traslaciones o rotaciones [1]. Las propiedades que tienen estos conjuntos que no cambian al aplicar este tipo de transformaciones se llaman invariantes topológicas. Una de las invariantes que más se estudia en topología es la conexión, de forma que se puede clasificar la forma de un espacio por la conexión entre sus elementos.

En el trabajo que presentamos estudiamos las propiedades de los conjuntos de usuarios de un sistema de recomendación desde el punto de vista de la forma que tienen dichos conjuntos. Convencionalmente, los perfiles de los usuarios se representan mediante secuencias de valores, en forma de vector, que describen las preferencias de estos para cada uno de los ítems del sistema. En sistemas basados en contenido o contextuales, los perfiles incorporan, adicionalmente, valores referidos a atributos del usuario o del contexto. Todos estos conjuntos de valores se combinan haciendo uso de diferentes algoritmos para resolver las tareas habituales de recomendación, como el cálculo de similitudes entre usuarios o la predicción de ratings. Pues bien, a diferencia de este uso convencional de los datos, exploramos una perspectiva diferente basada en la forma de los conjuntos y en las características de los mismos. Nos centraremos en el cálculo de los denominados números de Betti, los cuales representan características invariantes como son el número de componentes conexas o el número de agujeros en la estructura. En la Figura 1 se muestra la metodología que sigue el TDA para, a partir de un conjunto de datos S (en forma de nube de puntos en un espacio n-dimensional), aproximar una representación sólida K, con la que se obtienen sus invariantes topológicas.

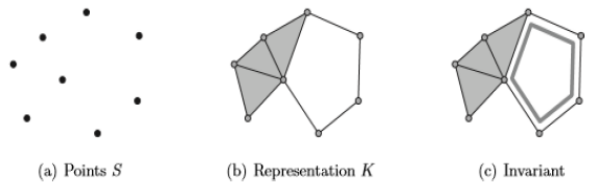

Figura 1. Cálculo de invariantes de un conjunto [2].

En (a) se representa el conjunto de datos en el espacio. En (b) se muestra la estructura asociada al conjunto de puntos. En (c) se calculan sus invariantes topológicas (aquí se muestra un ciclo).

A partir de las propiedades invariantes que caracterizan los conjuntos se pueden realizar diferentes tratamientos de datos empleando los algoritmos, técnicas, métricas, etc. existentes en el contexto del TDA. En particular, en este artículo exploramos el uso de una nueva forma de representación de perfiles de usuario mediante técnicas TDA que dan lugar a estructuras que los caracterizan, denominadas códigos de barras, así como al cálculo de similitudes entre usuarios mediante una métrica específica

diseñada para comparar los códigos de barras. Además hemos comparado las métricas basadas en la similitud con códigos de barras frente a baseline estándar obteniendo resultados que, aun siendo mejorables, resultan prometedores.

2 Antecedentes

La topología computacional se ha aplicado en distintos ámbitos en el mundo de la ciencia, como podemos comprobar en [2]. Por ejemplo, en el contexto de procesamiento de imágenes, se suele considerar una imagen como un rectángulo formado por píxeles, blancos y negros, donde los negros representan la imagen y los blancos el fondo. Una vez pixelada la imagen, se pueden definir diferentes tipos de conexiones entre píxeles. Por ejemplo, dos píxeles están conectados si comparten aristas o si comparten aristas o vértices, etc., produciendo como resultado algo similar a un esqueleto de la imagen. Una vez obtenido el mismo se está en condiciones de estudiar el objeto representado desde un punto de vista topológico.

En cartografía, una práctica común consiste en la deformación de un mapa, en la que están delimitadas ciertas zonas, generando áreas deformadas de cada zona en función de una propiedad (por ejemplo, deformar el mapa de los Estados Unidos en función de los votos existentes en cada estado). A este tipo de mapas se les denomina cartogramas. Esta técnica ofrece una mayor información visual sobre lo que se quiere representar. Estas prácticas se realizan empleando transformaciones topológicas que deforman las regiones del mapa de acuerdo con ciertas propiedades de las transformaciones y de los conjuntos sobre las que se aplican. Un ejemplo común de este tipo de técnicas es la deformación que se aplica sobre una taza para demostrar su equivalencia topológica (homeomorfa) a una rosquilla (Figura 2). El concepto de homeomorfismo se resume en la equivalencia por transformaciones como, por ejemplo, un cambio de escala, un giro o una traslación de puntos.

Figura 2. Taza y rosquilla homeomorfos [2].

En biología, el estudio de la forma de las proteínas puede ayudar en la comprensión de enfermedades y el desarrollo de medicamentos [3]. Para estudiar las proteínas se representa cada uno de sus átomos en el espacio y se analizan sus formas y conexiones. Este estudio puede realizarse con un enfoque geométrico o topológico. El enfoque geométrico se centra en la forma que tiene la molécula viéndola como una figura tridimensional. Sin embargo, el enfoque topológico se centra más en las conexiones existentes entre los átomos.

3 Fundamentos teóricos

Como decíamos antes, la Topología es la rama de las matemáticas que estudia las propiedades de los denominados espacios topológicos, que permanecen inalteradas bajo transformaciones o deformaciones continuas. En un espacio topológico podemos construir estructuras y analizar sus propiedades invariantes, en particular los números de Betti.

Una de las estructuras que frecuentemente se emplean son los n-simplex. Un n-simplex x es un conjunto de puntos geométricamente independientes de un espacio R^n. Sobre los n-simplex (también denominados simplex), podemos identificar caras, como subconjuntos de puntos que forman parte del propio simplex, los cuales son, a su vez, nuevos simplex. Se denomina Complejo Simplicial K en R^n a una colección de simplex en R^n tales que:

1. Cada cara de un simplex de K está en K.
2. La intersección de cualesquiera dos simplex de K es una cara de cada uno de ellos, y por tanto, está en K.

La Figura 3 muestra un ejemplo de complejo simplicial en el que se identifican 0-simplex (cada uno de los puntos del conjunto), 1-simplex (cada arista del conjunto), 2-simplex (cada triángulo sombreado). El cuadrado blanco del centro es un agujero de dimensión 1.

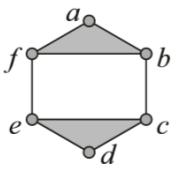

Figura 3. Complejo simplicial [2].

Para construir un complejo simplicial sobre un conjunto de puntos existen diferentes alternativas basadas, generalmente, en el recubrimiento de los puntos mediante secuencias de bolas centradas en cada punto y de radios incrementales. Dependiendo del método escogido generamos diferentes complejos simpliciales. Uno de los más sencillos de generar, incluso en dimensiones altas, es el complejo simplicial de Vietoris-Rips [4]. Consiste, básicamente, en ir uniendo puntos con aristas, a partir de la nube inicial de puntos, a medida que vamos incrementando el radio de las bolas centradas en cada uno de los puntos y estas se van intersecando. En la Figura 4 se ejemplifica el proceso.

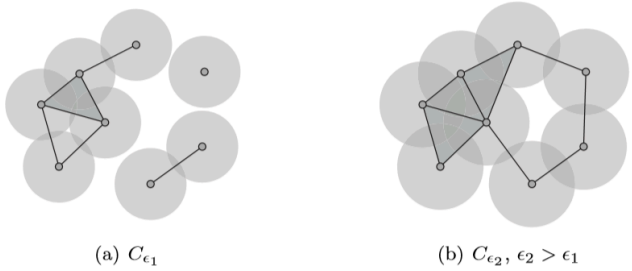

Figura 4. Creación de complejos C_{ε_i}, con radios ε_i [2].

En (a) se generan aristas entre los puntos cuyas correspondientes bolas de radio ε_1 se intersecan. Para dicho radio en la nube de

puntos se generan 3 complejos simpliciales no conexos entre sí: un 0-simplex formado por un punto aislado, un 1-simplex, formado por dos puntos unidos y un 5-simplex formado por 5 puntos unidos por aristas. En (b) el radio de las bolas se incrementa de modo que $\varepsilon_2 > \varepsilon_1$ y se comprueba cómo se generan aristas entre los tres complejos simpliciales anteriores dando lugar a un único complejo simplicial sin partes no conexas (desunidas). A los complejos simpliciales podemos asociarles los valores cuantificables que mencionábamos anteriormente: el número de componentes que son independientes de otras (no conexas), el número de agujeros en las superficies y el número de huecos entre las estructuras. En topología a estas propiedades se les da nombres concretos. Son los números de Betti de dimensión cero, uno y dos, respectivamente. Estos valores representan propiedades invariantes muy importantes que ayudan a distinguir entre complejos simpliciales. Comparando los correspondientes valores de distintos objetos sólidos podemos determinar si tienen, o no, la misma topología (es decir, si son equivalentes desde un punto de vista topológico). La idea que proponemos es representar los perfiles de los usuarios como esos objetos sólidos, de forma que sean comparables entre ellos mediante algún tipo de medida de similitud que expondremos más adelante.

El siguiente paso, una vez definidos los complejos simpliciales, es calcular los correspondientes números de Betti. Y para ello vamos a emplear una representación gráfica denominada *códigos de barras*. Las barras son representaciones gráficas de los intervalos de los radios sobre los que persisten las características topológicas (componentes, agujeros y huecos). El conjunto de estas barras caracteriza la forma en que la topología del objeto cambia a medida que se va representando el complejo simplicial. La forma en que se elaboran estos códigos se puede encontrar explicada con detalle en [5] y, en la Figura 5, se muestra, de manera gráfica, la visualización del código de barras asociado a un conjunto de puntos a medida que construimos su complejo simplicial asociado. El código de barras hemos de interpretarlo del siguiente modo:

1. Cada instancia de cada componente, cada agujero y cada hueco se representa con una línea.
2. La posición y longitud de cada barra representa el ciclo de vida de cada componente, agujero y hueco.
3. El inicio de la barra representa el radio en que la instancia correspondiente comienza su vida.
4. El otro extremo de la barra representa el radio en que la instancia correspondiente cesa su vida.

En el ejemplo de la Figura 5, inicialmente hay catorce puntos y catorce barras con origen en la abscisa 0 (en color verde). Al alcanzar radio 0.75 dos puntos se unen mediante una arista y, en consecuencia, quedan trece barras correspondientes a trece componentes (doce puntos y una arista uniendo dos puntos). Y así sucesivamente. A medida que los radios crecen se van uniendo componentes y van desapareciendo barras, hasta quedar una sola barra, a partir de la abscisa 2.2, de la única componente conexa que queda a partir de dicho radio. También podemos observar dos barras entre las abscisas 2.4 y 3.7 (en color morado). Estas barras representan los agujeros de dimensión 1. No hay huecos por lo que no habría barras asociadas.

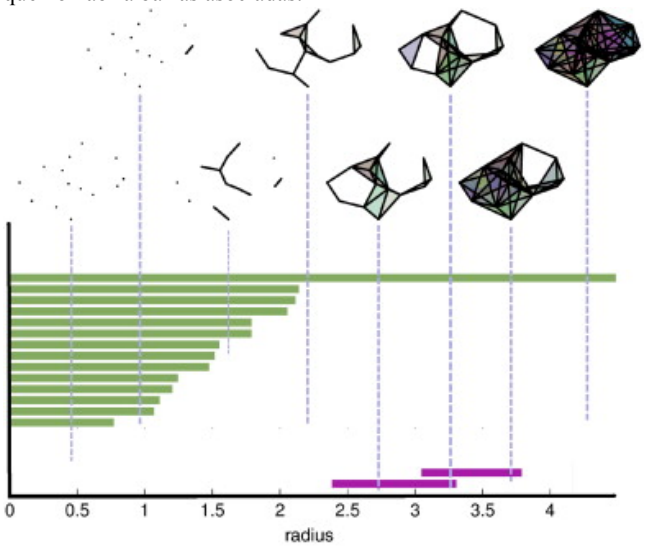

Figura 5. Códigos de barras de un complejo simplicial [3].

3.1 Similitudes basadas en códigos de barras

Una vez hemos calculado los códigos de barras procedemos con el estudio de la similitud entre dos objetos representados por dichos códigos. Basándonos en la similitud de Jaccard, definida para dos conjuntos no vacíos M y N como $S_J(M,N) = \frac{|M \cap N|}{|M \cup N|}$, podemos calcular una medida de similitud basada en esta expresión que promedie las similitudes máximas entre todos los posibles pares de similitudes de Jaccard entre barras de ambos conjuntos. Esta similitud promedio del solapamiento de códigos de barras (*Barcode Overlapping*) se define como

$$S_{BO}(A,B) = \frac{1}{|A|+|B|}\left[\sum_{a\in A}\sup_{b\in B}\frac{a\cap b}{a\cup b} + \sum_{b\in B}\sup_{a\in A}\frac{a\cap b}{a\cup b}\right] \qquad (1)$$

Puede darse la circunstancia de que los conjuntos a comparar carezcan de agujeros y huecos. Estos casos se pueden gestionar ampliando la definición anterior a los casos correspondientes a alguno de los conjuntos A o B, o ambos A y B, vacíos. De este modo se define la siguiente expresión general de la similitud basada en códigos de barras extendida (*Barcode Overlapping Extended*)

$$S_{BOE}(A,B) = \begin{cases} S_{BO}(A,B) & , si\ A \neq \emptyset\ y\ B \neq \emptyset \\ 1 & si\ A = \emptyset\ y\ B = \emptyset \\ 0 & en\ casos\ contrarios \end{cases} \qquad (2)$$

Los valores de similitudes de las diferentes invariantes (componentes, agujeros o huecos) se pueden combinar mediante alguna función monótona creciente (por ejemplo el promedio) que mantenga el orden de clasificación entre objetos, es decir, si un par de objetos es más similar en todos los diferentes códigos de

barras que otro par de objetos, la similitud unificada resultante debería ser mayor para el primer par.

4 Experimentación y conclusiones

En este artículo, además de exponer brevemente el marco general teórico que estamos interesados en explorar, explicamos un ejemplo sencillo que hemos desarrollado con el fin de comprender el problema. Hemos empleado datos del dataset MovieLens 100k, formado por 100000 ratings, 943 usuarios (miembros del conjunto $\mathcal{U}$) y 1682 items. Convencionalmente los perfiles de los usuarios se representan mediante vectores de ratings n-dimensionales. En nuestro caso consideraremos los usuarios como colecciones de puntos (ítem, valor) en el plano. A partir de dichas colecciones de puntos podemos construir sus complejos simpliciales asociados a cada usuario, obteniendo posteriormente sus invariantes asociadas a los códigos de barras generados para cada uno de ellos. En la Figura 6 se muestra el proceso de construcción del complejo simplicial para un usuario con 3 ratings.

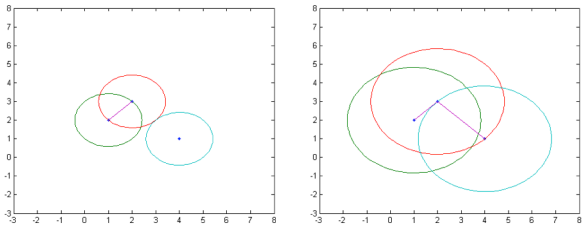

Figura 6. Construcción del complejo simplicial de un usuario

El cálculo de la similitud entre usuarios basada en su códigos de barras se ha realizado en Matlab empleando la librería JavaPlex, dedicada al cálculo de invariantes de complejos simpliciales [6]. Se pueden hacer muchas consideraciones sobre el modo en que se representan los usuarios. En nuestro caso hemos optado por la más intuitiva, aun no siendo probablemente la mejor desde el punto de vista del rendimiento del sistema.

Una vez obtenidos los complejos simpliciales asociados a cada usuario calculamos sus códigos de barras y sus invariantes. La Figura 7 muestra el código de barras de la invariante de dimensión 0 para el usuario anterior. Este usuario no presenta invariantes de dimensiones 1 (agujeros) ni 2 (huecos).

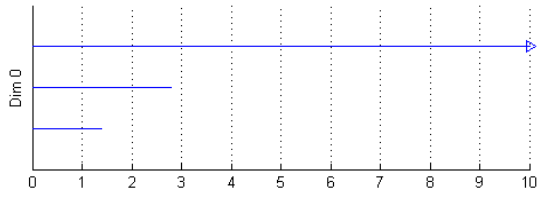

Figura 7. Código de barras de dimensión 0.

Con los códigos de barras de todos los usuarios calculados, estamos en condiciones de obtener, para cada par de usuarios $\forall\, u, v \in \mathcal{U};\ u \neq v$ del conjunto, su similitud $S_{BOE}(u, v)$.

Hemos restringido el cálculo de similitudes a usuarios con menos de 50 ratings, dado que es muy costoso computacionalmente calcular códigos de barras para usuarios con valores mayores. Esto supone un contratiempo para estudiar la eficiencia de la métrica S_{BOE}, pues nos estamos restringiendo a usuarios con perfiles relativamente pequeños. En [7] se analizan alternativas para optimizar el cálculo de complejos simpliciales, de momento dejamos el estudio de esas u otras alternativas como trabajo futuro. En nuestro caso, hemos hecho estimaciones del cálculo de error RMSE [8] con diferentes parametrizaciones del algoritmo obteniendo la comparativa que se muestra en la tabla siguiente, usando un algoritmo básico kNN.

Similitud	S_{BOE} 50 20	S_{BOE} 100 15	Coseno	Pearson	Random
RMSE	1.1616	1.1482	1.0791	1.5728	1.1217

Tabla 1. Métricas de error

Como se puede comprobar, los valores obtenidos para las similitudes basadas en los códigos de barras tienen un comportamiento razonablemente competitivo, incluso frente a métricas muy consolidadas como Coseno o Pearson, basadas en correlaciones entre los objetos comparados [8, 9], lo cual puede interpretarse como que esta métrica puede ser buena cuando las métricas colaborativas funcionan peor, cuando hay pocos datos. Además, es interesante la comparación con el método random, el cual devuelve un valor aleatorio entre un máximo y un mínimo para expresar la similitud entre usuarios, para mostrar que esta similitud no está devolviendo datos al azar.

En adelante queremos estudiar el comportamiento de esta métrica de manera más exhaustiva cuando los usuarios tienen pocos datos. Esto puede ser particularmente interesante para problemas de cold-start. Hemos de complementar este análisis con otros datasets, otros algoritmos basados en teoría de grafos, optimización u obtención de invariantes, así como otras métricas de similitud y de ranking habituales en el área.

AGRADECIMIENTOS

Investigación en el marco del proyecto BIBECA (RTI2018-101248-B-I00) financiado por MINECO/FEDER.

REFERENCIAS

[1] James R Munkres. 1984. *Elements of algebraic topology*. (volume 2). Addison-Wesley Menlo Park.

[2] Afra Zomorodian, 2012. Topological data analysis. *Advances in applied and computational topology*, 70:1–39.

[3] Gabriell Máté, Andreas Hofmann, Nicolas Wenzel, and Dieter W Heermann, 2014. A topological similarity measure for proteins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1838(4):1180–1190.

[4] Afra Zomorodian. 2010. Fast construction of the Vietoris-Rips complex. *In Computers and Graphics, 34:263–271*.

[5] Robert Ghrist, 2008. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*. Vol. 45(1), 61-75.

[6] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams, 2014. JavaPlex: A research software package for persistent (co)homology. *Proceedings of ICMS 2014, Lecture Notes in Computer Science 8592, 129–136.* Software available at http://appliedtopology.github.io/javaplex/

[7] Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., & Harrington, H. A. (2017). A roadmap for the computation of persistent homology. EPJ Data Science, 6(1), 17.

[8] Xia Ning, Christian Desrosiers, George Karypis. A comprehensive survey of neighborhood-based recommendation methods. *Recommender Systems Handbook*, (2nd. ed.) 37–76.

[9] Marko Balabanović and Yoav Shoham. *Fab: content-based, collaborative recommendation*. Communications of the ACM, Vol. 40(3), 66-72.