

Estructures de grafs amb equivalències d'arestes aplicades a l'anàlisi de dades relacionals

Laura Rodríguez Navas

laura.rodriguez.navas@upc.edu

Universitat Politècnica de Catalunya

Resum

La teoria de les 2-estructures [1] proporciona una infraestructura matemàtica per a la descomposició de grafs. Permet representar els grafs en una única estructura algebraica, una 2-estructura. En aquest treball d'investigació es contribueix en l'estudi de les 2-estructures. Per un costat, es dissenya i s'implementa un conjunt d'algoritmes per a la creació i la visualització de les 2-estructures i per un altre costat, s'estableixen resultats per a enfortir l'anàlisi de dades relacionals mitjançant l'ús d'aquestes estructures.

Introducció

En els últims anys s'ha produït un increment significatiu en la capacitat d'emmagatzemar i compartir dades. Segons IBM (International Business Machines), el 90 % de les dades del món s'han generat en els últims anys. La importància estratègica de les dades massives no consisteix en la quantitat sinó en les aplicacions potencials que ofereixen. Alguns exemples. Les dades en blocs i xarxes socials s'aprofiten per a dissenyar estratègies polítiques. Les dades poden ajudar a gestionar ciutats o recursos naturals, a estudiar el canvi climàtic o a promoure el desenvolupament de països. També en la caracterització de malalties complexes a escalar molecular combinada amb l'historial mèdic i del tractament amb proves diagnòstiques o d'imatge afavoreix oportunitats sense precedents per a personalitzar la medicina. I altres.

Davant de l'allau de dades o tsunami de dades, en aquest document s'hi reflecteixen alguns punts teòrics que en milloren l'anàlisi.

Dades Relacionals

Les dades relacionals amb les que es treballa inicialment es poden incloure en un dels tres tipus de fitxers:

- ARFF (Attribute-Relation File Format)

Arxius de text en format ASCII que descriuen una llista d'instàncies que comparteixen un conjunt d'atributs.

- TXT

Arxius de text estàndard que contenen text sense format.

- DB

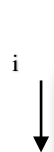
Arxius de bases de dades genèriques.

Els fitxers es transformen en taules SQLite [2]. Les taules d'aquest tipus de les bases de dades relacionals s'utilitzen en aquest treball ja que alhora d'iniciar una línia d'investigació és molt important la senzillesa i la generalització que aporten.

Recorreguts en una taula SQLite

Una vegada s'han transformat les dades d'un fitxer a taula SQLite, es recorren les files de la taula per extreure'n les dades. Amb les dades es crea un graf dirigit complet¹. Cada element no repetit de la taula representarà un node del graf. I cada node estarà enllaçat amb la resta de nodes. Una arista discontinua representa un enllaç.

Després es realitza un segon recorregut que identifica les parelles de nodes que apareixen a la taula.



| | A | B | C | D |
|---|---|---|---|---|
| 1 | a | b | d | e |
| 2 | a | f | d | e |
| 3 | c | c | c | c |

Figura 1. Exemple de taula SQLite.

Sent i , un índex que recorre les columnes de la figura 1 i j , un índex que recorre les files de la figura 1, llavors per a cada parella de valors (i, j) , $i \neq j$, existirà un enllaç entre els elements de la parella. Una arista contínua representa un enllaç. El graf resultant s'anomena graf de Gaifman [3] o graf primer.

Exemple 1. A la figura 2 es pot veure el graf de Gaifman resultant de la figura 1.

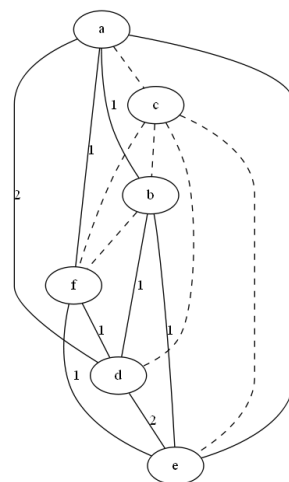


Figura 2. Graf de Gaifman.

¹ un graf és dirigit i complet quan entre totes les parelles de nodes del graf existeix una arista.

Grafs

En aquest treball, a partir dels grafs de Gaifman, es desenvolupen quatre tipus de grafs:

- Graf pla

Graf de Gaifman que conté dues classes d'equivalències¹. La primera classe d'equivalència representa les parelles de nodes que apareixen a la taula (arestes contínues). La segona classe d'equivalència representa les parelles de nodes que no apareixen a la taula (arestes discontinúes). El graf de la figura 2 representa el graf pla de la taula de la figura 1.

- Graf pla amb llindar

Graf pla que també conté dues classes d'equivalències. Una classe d'equivalència representa les arestes amb el nombre d'equivalències² per sota el llindar³ (arestes discontinúes). L'altra classe d'equivalència representa les arestes amb el nombre d'equivalències és superior o igual a el llindar (arestes contínues).

Exemple 2. Considerant la taula de la figura 1, la figura 3 mostra el graf pla resultant amb llindar igual a dos. Les arestes etiquetades amb el valor igual o superior al llindar són les arestes contínues. I les arestes etiquetades amb el valor inferior al llindar són les arestes discontinúes.

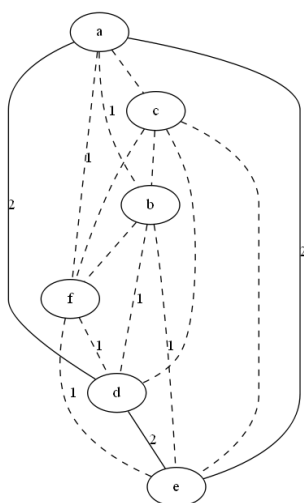


Figura 3. Graf pla amb llindar.

- Graf lineal

Graf pla que conté més de dues classes d'equivalències. Cada classe d'equivalència li correspon un color diferent.

Exemple 3. Considerant la taula de la figura 1, la figura 4 mostra el graf lineal resultant.

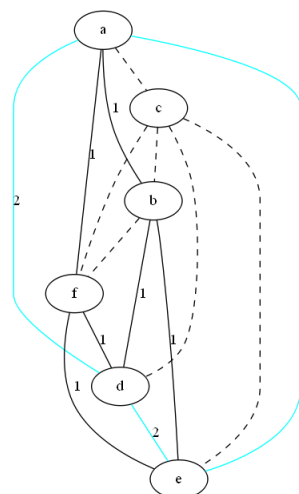


Figura 4. Graf lineal.

- Graf exponencial

Graf lineal que agrupa les arestes en diferents classes d'equivalències. Cada classe d'equivalència comprèn un interval exponencial, que s'inicia en 2^0 i acaba en 2^9 . L'interval exponencial indica l'agrupació de nombres d'equivalències que representarà cada classe d'equivalència. Cada interval li correspon un color diferent.

¹ Les classes d'equivalències representen les arestes acolorides d'un graf. Dues arestes són del mateix color si pertanyen a la mateixa classe d'equivalència.

² Nombre d'aparicions d'una parella de valors en una taula.

³ El llindar és un nombre enter introduït per l'usuari.

Clans

Un graf es divideix en clans primers per a formar una 2-estructura. La idea amb la que es basa aquesta descomposició consisteix en trobar subconjunts del graf, anomenats clans, en els quals els elements continguts en cada clan es relacionen de la mateixa manera amb tots aquells elements fora del clan. Formalment, un subconjunt $X \subseteq D$ d'un graf g és un clan si cada node $y \notin X$ "veu" els nodes de X de la mateixa manera i cada dos nodes $x_1, x_2 \in X$ "veuen" cada node $y \notin X$ de la mateixa manera per a tot $x_1, x_2 \in X$ i $y \notin X$: $(y, x_1) R (y, x_2)$ i $(x_1, y) R (x_2, y)$.

Exemple 4. A la figura 5 cada color diferent representa una classe d'equivalència i s'han encerclat els clans $X = \{A, B\}$ i $Y = \{C, D\}$. Són clans perquè la resta de nodes del graf els "veuen" de la mateixa manera. Això vol dir que les arestes que uneixen els clans amb la resta d'elements del graf són del mateix color o classe d'equivalència. El subconjunt $\{A, C\}$ no és un clan del graf ja que el node E el distingeix: $\text{negre}(E, A) \neq \text{verd}(E, C)$.

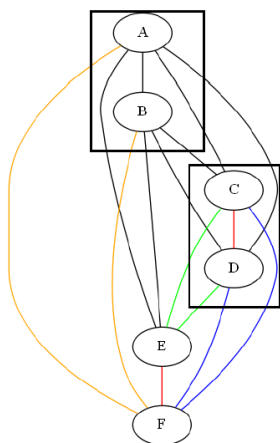


Figura 5. Representació dels clans X i Y .

Els clans es poden diferenciar en:

- Clans trivials

Els subconjunts de longitud u i els que contenen tots els nodes del graf són clans trivials. Els clans trivials sempre es consideren clans primers.

- Clans primers

Un clan és primer si no es superposa amb altres clans del graf. A i B es superposen si $A \cap B \neq \emptyset$, $(A \cap B) \subset A$ i $(A \cap B) \subset B$. Per tant, no són clans primers. El procediment redueix el temps de la cerca dels clans, que normalment és elevat en grafs molt grans. Així que quan es vol descomposar un graf amb una grandària considerable és recomanable utilitzar els clans més freqüents.

- Clans més freqüents

Els clans més freqüents d'un graf els genera el programari Apriori [R]. El programari mitjançant els subconjunts més freqüents retorna els clans primers dels grafs.

Estructura

El treball finalment desenvolupa la teoria de la 2-estructures, i en particular es demostra que cada 2-estructura pot ser construïda a partir de la descomposició d'un graf en clans primers. En el procés de descomposició anterior s'obté una 2-estructura a través d'una representació jeràrquica en forma d'arbre.

Una 2-estructura és una seqüència de valors agrupats, formada per un subconjunt finit D , anomenat domini, i una relació d'equivalència $R \subseteq E(D) \times E(D)$ en el conjunt $E(D) = \{(u, v) \mid u \in D, v \in D, u \neq v\}$ de les seves arestes. Per això una 2-estructura es pot definir com un graf $G = (D, R)$ dirigit i complet, on el domini D representaria els tots els nodes i les classes d'equivalències R , les arestes acolorides d'aquest.

Exemple 5. A la figura 6 es pot veure la representació gràfica d'una 2-estructura. Cada estructura rectangular, anomenada clúster, representa un clan primer no trivial del graf lineal de la figura 5. Les arestes internes dels clústers serveixen per a diferenciar les classes d'equivalències a les que pertany cada element del clan primer corresponent. Les arestes externes entre els clústers serveixen per indicar els clans primers que

formen part d'un altre clúster (o clan primer). Els clans trivials de longitud igual a u es troben a les fulles del subarbre al que queda connectat cada clan primer representat.

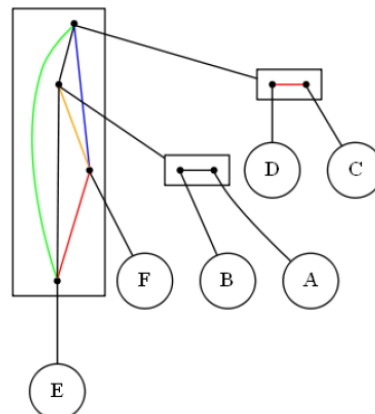


Figura 6. 2-estructura equivalent al graf de la figura 5.

Resultats

La investigació va néixer amb la finalitat de crear una eina per a l'estudi de les 2-estructures i establir resultats prometedors que relacionessin les 2-estructures amb l'anàlisi de dades relacionals; objectius que s'han complert.

Les 2-estructures resulten un mètode visual molt potent per a relacionar, classificar i analitzar les dades relacionals. Difícilment es pot extreure tanta informació d'un fitxer o de la taula d'una base de dades relacional.

Treball futur

Part de l'objectiu era la construcció d'una eina base per a investigacions futures en relació a la teoria de les 2-estructures. Tot i així, és impossible avançar-se i enumerar totes les possibles utilitats que podria tenir. Per això, serà necessari seguir completant i adaptant aquest treball per ajustar-se a les necessitats dels treballs futurs.

Com a millores es proposen:

- La ampliació del nombre de colors per a poder diferenciar moltes més classes d'equivalències.
- Un altre punt interessant seria l'anàlisi en profunditat de la creació d'una 2-estructura a partir de diferents taules SQLite. I el seu anàlisi corresponent.
- Per últim, es proposa com a següent pas natural, la millora dels algorismes per augmentar la capacitat del tractament de les bases de dades relacionals amb major nombre de dades.

Referències

- [1] A. Ehrenfeucht i G. Rozenberg. "The Theory of 2-Structures: A Framework for Decomposition and Transformation of Graphs". World Scientific, 1999.
- [2] SQLite. <https://www.sqlite.org/>
- [3] L. Libkin. "Elements of Finite Model Theory". Springer, 2004.
- [4] Programari Apriori. <http://borgelt.net/apriori.html>