Introducció

Breu introducció de mi mateixa.

Durant els propers 15 minuts us parlaré sobre les estructures involucrades en la teoria de les 2-estructures i de com aplicar-les a l'anàlisi de dades relacionals. I a partir d'alguns exemples veurem el seu disseny en profunditat.

Objectius

Els objectius principals han estat:

[Llegir transparència]

[Figura] Aquí podem observar la transformació a la que finalment s'arriba. Un conjunt de dades relacionals es transforma en una 2-estructura. Noteu que les 2-estructures resulten un mètode visual molt potent per a relacionar, classificar i analitzar dades relacionals. En les properes transparències desglossaré aquesta transformació.

Diagrama de flux

El procés de transformació es pot dividir en 4 fases:

- L'obtenció i l'adequació de les dades relacionals.
- La creació del graf a partir d'aquestes dades.
- La descomposició del graf.
- La creació de la 2-estructura a partir dels elements amb els que s'ha dividit el graf.

Dades relacionals

El format dels fitxers que inicialment contenen les dades relacionals són fitxers amb extensió ARFF, TXT o DB. Cada fitxer es converteix en una taula SQLite. S'han utilitzat aquest tipus de taules ja que alhora d'iniciar una línia d'investigació és molt important la senzillesa i la generalització que aporten.

Graf complet

A continuació explicaré la creació de grafs utilitzant les dades Titanic. Que són un exemple típic (particularment senzill) en la docència de l'anàlisi de dades.

Una vegada s'han col·locat les dades Titatic en una taula SQLite, amb ella es crea un graf complet. Un graf és complet quan entre totes les parelles de nodes del graf existeix una aresta. En el procés de creació del graf complet es recórrer la taula d'aquesta manera [indicar el recorregut]. Cada valor diferent, sense comptar els valors repetits, representarà un node del graf. I cada node estarà enllaçat amb la resta de nodes (arestes discontínues). Així serà un graf complet. Per exemple, el node Yes està connectat amb [resseguir les arestes amb el dit].

Graf de Gaifman

Després es torna a recórrer la mateixa taula per a identificar les parelles de nodes que apareixen a la taula [indicar el recorregut]. Per a cada parella de valors existirà un enllaç o aresta contínua en el graf. Per exemple, entre els elements 3rd i Female existeix una aresta [resseguir l'aresta amb el dit].

En aquest segon recorregut també es compten el nombre d'aparicions de cada parella de valors de la taula, que s'anomena nombre d'equivalències; i s'etiqueten les arestes amb aquest nombre entre els nodes equivalents. El graf resultant s'anomena graf de Gaifman o graf primer.

Existeixen 4 tipus de grafs de Gaifman: pla, pla amb llindar, lineal i exponencial. Definiré un per un a continuació.

- **Un graf pla** és un graf de Gaifman que conté dues classes d'equivalències. Les classes d'equivalències representen les arestes acolorides del graf. En aquest cas, 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 discontínues).

- **Un graf pla amb llindar** és un graf pla que conté dues classes d'equivalències. Una classe d'equivalència representa les arestes amb el nombre d'equivalències per sota el llindar (arestes discontínues). El llindar és un valor introduït per l'usuari. L'altre classe d'equivalència representa les arestes on el nombre d'equivalències és superior o igual a el llindar (arestes contínues). Per exemple, en aquest graf, el llindar és 1000. Per tant, les arestes etiquetades amb un valor igual o superior a 1000 es diferencien de la resta [anècdota divertida].

- **Un graf lineal** és un graf pla que conté més de dues classes d'equivalències. Cada classe d'equivalència li correspon un color diferent. Per exemple, les arestes amb el nombre d'equivalències 4 o 6 [indicar colors].

- **Un graf exponencial** és un graf lineal [es crea a partir del 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. Cada interval li correspon un color diferent. Per exemple, el color turquesa comprèn l'interval exponencial (2, 3). L'interval exponencial (4, 7) el representa el color verd.

Descomposició

Després el graf es descomposa.

[Llegir transparència]

El subconjunt {'A', 'B'} forma un clan perquè la resta de nodes es relacionen de la mateixa manera amb el subconjunt [assenyalar les arestes].

El subconjunt {'A', 'B'} no forma un clan perquè per exemple el node E el distingeix [assenyalar les arestes].

Clans

Existeixen 3 tipus de clans.

[Llegir transparència].

Superposició

Com acabo de comentar, un clan és primer si no es superposa amb cap altre clan. Per tant, un clan primer no complirà el principi de superposició. Aquest principi diu que existeix superposició entre dos clans si es compleixen aquestes tres condicions.

Per exemple, considerant els clans A i B, existeix superposició entre ells si:

- La intersecció no és buida.
- La intersecció no està continguda totalment en A.
- La intersecció no està continguda totalment en B.

Clans primers

A continuació veiem un exemple de la descomposició del graf en clans. Tots els clans del graf són: [llegir transparència i comprovar un d'exemple].

Els clans trivials del graf són: [llegir transparència]. Ens fixem que són els clans de longitud u i el clan que conté tots els elements del graf. Els clans primers són: tots els clans trivials més [llegir transparència].

Perquè quedi una mica més clara la definició de clan primer. Un exemple de clan primer. El subconjunt {'b', 'f'} és un clan primer ja que no es superposa amb altres clans: {'e', 'd'} i {'a', 'e', 'd'}. A més, el clan {'a', 'd', 'e', 'f', 'b'} el conté. La intersecció de {'b', 'f'} amb {'a', 'd', 'e', 'f', 'b'} és {'b', 'f'} (ell mateix) i {'b', 'f'} és subconjunt del clan que conte tots els nodes del graf, {'a', 'd', 'e', 'f', 'c', 'b'}.

Els clans primers contenen altres clans primers [indicar]. Aquesta característica és molt útil per la factorització de la 2-estructura. Que divideix les 2-estructures en altres 2-estructures [indicar].

Estructura

Finalment es desenvolupa la teoria de la 2-estructures, i en particular es demostra que una 2-estructura es pot construir a partir de la descomposició d'un graf en clans primers.

Mitjançant el procés de descomposició del graf anterior on hem obtingut els clans primers [indicar] s'obté una 2-estructura a través d'una representació jeràrquica en forma d'arbre. Com es pot observar hem acolorit els clans perquè es vegi més bé la seva equivalència a la 2-estructura.

Com veieu, cada estructura rectangular, anomenada clúster, representa un clan primer (amb longitud igual o superior 2). 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.

La factorització de les 2-estructures [indicar].

Resultats

[Llegir transparència].

Treball futur

[Llegir transparència].