

Az Infomap algoritmus világa - Kivonat

Urbanics András

A biológiai és társadalmi rendszerek hálózati struktúrába rendezett adatai, mára már óriási mértékben felhalmozódtak. Azért, hogy ezekből az adatokból hasznos információkat tudjunk kinyerni, először egyszerűsíteniük kell a struktúrát olyan módon, hogy az ábrázolások könnyen értelmezhetőek legyenek, ugyanakkor megtartsák a hálózat összetettségét, mint ahogyan a kartográfusnak is egyensúlyt kell találnia a térkép készítése során a felesleges részletek kidolgozása és a fontos struktúrák túlzott egyszerűsítése között. Hasznos lehet, ha a hálózat gráfjában csúcsokat és éleket modulokká alakítunk úgy, hogy azok a legfontosabb információkat és kapcsolatokat tükrözzék. Az ilyen jó ábrázolások egyszerre fogják kiemelni és egyszerűsíteni a mögöttes struktúrát és a hálózat kapcsolatait, ahogy a térképek is jól tömörítik a valóságot, miközben a lehető legtöbb információt igyekeznek szolgáltatni.

A dolgozat első fejezete a matematikai alapapparátust építi fel. Bevezeti az információelméleti alapmennyiségeket és többek között említésre kerül Shannon zajtalan forráskódolási tétele, valamint a Markov-láncok Ergod-tétele.

A második rész ezekre építve írja le az Infomap algoritmus működését: bolyongjunk a gráfon memória nélkül, vagyis ugorjunk csúcsról csúcsra úgy, hogy bárhol is vagyunk aktuálisan, a kimenő élek súlyaival arányos valószínűséggel válasszuk a következő állomást a múlttól teljesen függetlenül. Minden klaszterezéssel természetes módon lehet tömöríteni az utunkat. Az Infomap algoritmus kimenete az a klaszterezés, amivel ez a tömörítés a leghatékonyabb.

A harmadik rész ismerteti a Lancichinetti-Fortunato-Radicchi szabványgráfot, amely a hatványeloszlás szerint generált csúcsoknak és modulméreteknek köszönhetően hasonló a valós hálózatokhoz. Az így keletkező ismert struktúrájú hálózatok segítségével lehetőség adódik az Infomap algoritmus tesztelésére. Bevezetésre kerül két hasonlósági mérték, így az ismert klasztereket már egyszerűen összevethetjük az algoritmus által adott klaszterekkel. Végül a leírtakat alkalmazva futtatási eredményeket közöl a dolgozat.