Szociális kapcsolatok feltérképezése gráfalapú adatbányászati módszerekkel

Tasnádi Ervin

III. évfolyamos programtervező informatikus alapszakos hallgató

Témavezetők: Berend Gábor, Dr. Farkas Richárd

SZTE-TTIK Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék

Jelen dolgozatban bemutatunk egy gépi tanuláson alapuló megoldást, ami egy online étteremvéleményező portál felhasználóinak javasol potenciális barátokat. A baráti kapcsolatok előre megjósolhatók a felhasználók szokásai alapján, és emberi beavatkozás nélkül tudjuk elősegíteni az online kapcsolatok kialakítását. Mindez hozzájárulhat ahhoz, hogy elősegítse a XXI. században az egyre nagyobb szerephez jutó Internetes kapcsolattartáson- és ügyintézésen alapuló társadalmak fejlődését. Ehhez rendelkezésünkre áll az étteremvéleményező adatbázis (www.yelp.com), ami tartalmazza, hogy mely felhasználó mely éttermeket látogatja, kikkel van „barátsági" kapcsolatban, illetve milyen véleményeket írnak az egyes éttermekről.

A feladat egy bináris osztályozási problémát definiál, ahol csak pozitív, illetve címkézetlen példák állnak rendelkezésre: ha tudjuk, hogy két felhasználó barát, akkor az igen címkét kapják meg. Ha nincs barátság reláció két felhasználó között, akkor nem tudjuk, hogy ők valóban nem barátok, vagy csak nem ismerik egymást, de lehetnének barátok. Az a feladatunk, hogy a tanítópéldák alapján megítéljük, hogy mennyi esélyt látunk arra, hogy ők barátok legyenek. A megoldás során gépi tanuló algoritmusokkal meghatározzuk a barátság valószínűségét. Fontos megemlíteni, hogy a rendelkezésre álló adathalmazban a barátságok egyoldalúak, azaz nem kölcsönösek: egy felhasználó lehet egy másik barátja úgy, hogy a másik fél nem viszonozza azt.

Az egyéneket több szemszögből vizsgáljuk, hogy kialakítsuk a jellemzőteret. Az első jellemző a felhasználók által írt vélemények nyelvezetének a hasonlósága. Emögött az a motiváció, hogy azt gondoljuk, hogyha egy felhasználó barátja egy másiknak, akkor hasonló nyelvezetük van. Ennek meghatározásához a felhasználókból, és az általuk leírt szavakból egy páros gráfot építünk, melyben a perszonalizált PageRank algoritmussal súlyokat rendelünk a csúcsokhoz, így meghatározva azt, hogy egy bizonyos felhasználóhoz melyik másik a leghasonlóbb. A második jellemző az étteremlátogatási preferenciák hasonlósága. Itt az a feltételezés, hogy a barátok hasonló éttermekbe járnak. Ennek megállapításához az előző módszerhez hasonlóan egy felhasználó-étterem páros gráfot építünk, és a csúcsokhoz PageRank értékeket rendelünk. A harmadik fontos jellemző a szociális kapcsolatokat veszi alapul: egy olyan társaságból valószínűbb, hogy barátokat találunk, ahol már eleve sok kapcsolatunk van. Egy gráfot építünk, ahol a csúcsok a felhasználókat reprezentálják, két csúcs között akkor van egy irányított él a megfelelő irányban, ha az egyik barátja a másiknak. A PageRank algoritmus választ ad arra, hogy milyen a kapcsolat minősége: figyelembe veszi, hogy hányféle képpen lehet eljutni egyik felhasználótól a másikig, illetve milyen messze találhatók a gráfban. Így sokkal többet tudunk mondani arról, hogy mennyire hasonló két felhasználó, mintha csak egyszerűen a legrövidebb utak hosszát számolnánk ki.

A dolgozatban bemutatjuk ezeket az algoritmusokat, és empirikus eredményeket közlünk a Yelp adatbázison.