Big Data

Témavezető: Benczúr András

Az elosztott, Big Data technológiák a következő területeken váltak szükségessé:

- Nagyméretű hasonlósági mátrixok közelítő kitöltése; a küszöb feletti hasonlóságok számában alacsony komplexitású közelítő vagy heurisztikus algoritmusok készítése.
- Regressziós, gyakori termékhalmaz, legközelebbi szomszéd problémák kezelése.
- Rangsor kombinációk tanulása, keresések nagyméretű paraméter-térben.

A feladatok a problémák mérete miatt algoritmikusan nehezek, hiszen például a hasonlósági mátrix kvadratikus méretű, amelynek tárolása külső tárban sem megoldható gyakorlati feladatok esetén, tehát a kimenet méretében szublineáris algoritmusokra van szükség.

Cél hatékony Data Streaming algoritmusok tervezése, kialakítása, különös tekintettel a mobil szenzor adatokat elemző és a valós idejű ajánló alkalmazásokra. Olyan ún. Lambda Architektúra alacsony szintű adatkezelési algoritmus-réteg kutatása világszerte fontos feladat, amely képes egy rendszer részeként nagyon nagy háttér számításokat és mindeközben nagyon gyors, data stream frissítéseket is kiszolgálni.

1 Ajánló rendszerek

Az online és offline viselkedés nyomait összekötő, gazdag és részben strukturálatlan metaadatokkal bővített adatokon történő gépi tanulási eljárások (sztochasztikus gradiens, factorization machine) kialakítása. Olyan módszert kívánunk adni, amely össze képes párosítani az online és az offline, vagy két teljesen más forrásból származó eseménysorozatot, például egy offline kérdőívezés eredményeivel képes egy online könyvértékesítési adaton az ajánlások minőségét javítani.

2 Trend elemzés Web és közösségi média adatokban

Nagy adatokon a trend elemzés fő kihívását az jelenti, hogy a szakértői lekérdezéseket valós időben kell kiszolgálni. Egy adott cég, termék, személy nevét keresve sok millió bejegyzés adatait kell leválogatnunk és másodpercek alatt trend görbéket előállítanunk. Ez a feladat közelítő adatstruktúrák használata nélkül biztosan nem oldható meg. Kutatásunkban megvizsgáljuk a Bloom filterek, MinHash fingerprintek használatának lehetőségét.

3 Hálózati információk terjedése

Célunk közösségi hálózatokat felhasználva, illetve annak hatását vizsgálva jobban megismerni a társadalom szerkezetét és működését, elsősorban a terjedési jelenségek dinamikáját, a különböző kommunikációs csatornák és a rájuk alkalmazott stratégiák szerepét. Valós adatok részletes idő-dinamikájának vizsgálatán keresztül meg kívánjuk érteni, hogy az idő előrehaladtával hogyan kerül át információ egy személytől a kapcsolataihoz: például röviddel azután, hogy valaki egy zeneszámot meghallgatott, meghallgatják-e olyan ismerősei is, akik azt az előadót korábban nem hallgatták. A jelenséghez megfelelő információ-terjedési modelleket kívánunk adni.

Budapest, 2015. február 20.