Diplomamunka tématerv

Tematikus osztályozást megvalósító alkalmazás klaszterező algoritmusok segítségével

Készítette:

Gégény Benjamin, programtervező matematikus

GEBNAAT.SZE

Témavezető:

Alexin Zoltán, PhD.

A csoportosítás, vagy más néven osztályozás elengedhetetlen az életünkben. Egy rendszerben egyszerű kiigazodni, könnyű keresni. A tematikus osztályozás vagy más néven téma szerinti osztályozás a dokumentumok rendszerezésének leggyakrabban alkalmazott módszere. A tematikus osztályozásnak gyors keresés mellett az átláthatóság egy másik nagy előnye, ezért is mai világban szinte mindenütt jelen van. Pl. az újságok is témák szerint csoportosítják a híreket és a könyvesboltokban/könyvtárakban is témák szerint vannak rendezve a könyvek.

Diplomamunkám célja, egy olyan alkalmazás elkészítése, amely megvalósítja a tematikus osztályozást. Az program klaszterező algoritmusok segítségével fogom megvalósítani. A klaszterező algoritmusoknak két nagy csoportját különböztetjük meg, az egyik a partícionáló a másik a hierarchikus klaszterezés. Néhány ismert algoritmus (Direct k-way clustering, Agglomerative clustering, Repeated Bisections for by k-way refinement) alapján képes lesz a program osztályozni a korpuszban szereplő anyagot.

A alkalmazás C# nyelven készül. A megvalósításhoz a Visual Studio 2008-as verzióját fogom használni. Az input szövegeken előfeldolgozást végzek a könnyebb kezelhetőség érdekében. Össze fogok állítani, egy körül-belül 400 szavas stopszólistát. Ez azokat a szavakat jelenti, amelyeket nem hordoznak érdemi információt a dokumentum jelentésével kapcsolatban, ezért eltávolíthatók. Ez csökkenti számításigényt, gyorsítja feladatmegoldást.

Természetesen ezt a listát a felhasználó majd szabadon módosíthatja. (hozzáírhat vagy éppen kitörölhet belőle)

A programban lesz egy egyszerű kezelőfelület, aminek segítségével a felhasználó kiválaszthatja majd, hogy mely klaszterező algoritmust szeretné használni. Az alkalmazás része lesz egy olyan súgó is, amely segít kiválasztani a megfelelő rendezési módszert az algoritmusokban nem jártas felhasználók számára. Az eredmények a programból grafikusan is megtekinthetők lesznek, de egy szöveges fájlba is lementi a program, ahol a dokumentumok neve az osztályok szerint lesz felsorolva, valamint a művelet végén a feldolgozás pontosságáról is készül egy összefoglaló statisztika.

Az alkalmazás működését egy körülbelül 500 meséből álló korpuszon mutatom be. A meséket korábban már előre meghatározott kategóriákba csoportosították, mint pl. állatmesék, tündérmesék, tréfamesék, stb. Az alkalmazás a mesék automatikus csoportosítását végzi el és összevetné az emberi osztályozással. Továbbá a bemeneten majd szótövesítést fogok végrehajtani, és az eredményeket összehasonlítom az eredeti osztályozás eredményeivel.

Szeged, 2011. szeptember 7.