AZ ASSZOCIATÍV ADATFELDOLGOZÁS ÉS ALKALMAZÁSAI

NAGY ISTVÁN♣

Mottó: Ha tudjuk, miért tesszük,
és azt is, hogyan,
tegyük meg,
hogy jobb legyen a világ.
(Ogana Yan)

BEVEZETÉS

Az asszociatív adatfeldolgozás (ADP; Associative Data Processing) számunkra egyaránt jelenti a numerikus adatok hasonlósági feldolgozását, vagyis az adattér, és az attribútum (tulajdonság) tér redukcióját, és ez utóbbi révén egy lényeg-kiemelési, és hatásmechanizmus-felismerési eszközt, valamint a szekvenciális struktúrák (szövegek, fonéma-szekvenciák, vagy akár DNS-láncok) hasonlítását.

A L0GANA asszociatív szövegkereső (bemutatását lásd a www.logana.com honlapon) egy olyan program, mely valamely böngészőbe (mint például az Internet Explorer, az Opera, a FireFox, stb.) beépülve lehetővé teszi a hibásan keresett, vagy hibásan tárolt, illetve többalakú szöveges adatok hatékony keresését is szövegfájlokban, adatbázisban, vagy az Interneten.

A keresés során a felhasználó két adatot ad meg. Az egyik egy kereső-kifejezés, mely a legegyszerűbb esetben egy keresőszó, a másik pedig az úgynevezett határhasonlóság. Az asszociatív szövegkeresés lényege, hogy a keresőszótól nem várjuk el a teljes pontosságot (ráadásul mindezt nyelv- és karakterkészlet-független módon!). A határhasonlóság, egy százalékban megadott érték, mely azt fejezi ki, hogy egy talált szót csak akkor fogad el az asszociatív szövegkereső, ha a keresőszóhoz viszonyított, úgynevezett szöveg-hasonlóságának mértéke meghaladja ezt a határhasonlóságot.

Általánosságban elmondható, hogy egy keresőeszköz legfontosabb paraméterei a ***teljesség*** és a ***sebesség***. A *teljesség* azt jelenti, hogy a felhasználó minden olyan objektumot megtaláljon, mely rendelkezik az általa elvárt tulajdonságokkal, a *sebesség* pedig azt, hogy mindezt a számítógépektől manapság megszokott reakcióidőn (néhány másodpercen) belül.

Például a hagyományos internetes kereső eszközök (Google, Yahoo, Bing, stb.) sebessége megfelel a gyakorlati igényeknek, a teljességi feltételnek azonban csak rendkívül szűk körben tesznek eleget, ugyanis nem képesek sem a felhasználói keresések, sem a tárolt dokumentumok szöveghibáit (az esetleges többalakúságot) megfelelően kezelni.

Ennek igazolására próbáljuk megkeresni a "Veiszäcker" szót bármelyik hagyományos szövegkeresőben... (Richard Karl Freiherr von Weizsäcker, 1984 és 1994 között a Német Szövetségi Köztársaság elnöke volt.)

A legérdekesebb ebben az esetben a Google, ez ugyanis (a honlapunk éppen e példájára való hivatkozáson kívül) kilenc találatot is ad; például egy svéd csevegő oldal hét hozzászólását, de mindegyik esetben éppen ebben az alakban (Richard von Veiszäcker) hivatkoznak a német politikusra. Ennek a sikernek azonban két sajnálatos mellékhatása van. Az egyik, hogy e hibásan beírt hozzászólások a helyesen beírt név keresésekor már nem találhatók meg, a másik pedig, hogy e találatok mellett éppen a helyesen írt névre való hivatkozások nem találhatók meg...

A Bing mindössze kettőt talált meg a svéd csevegő oldal hozzászólásai közül, és mást se, a Yahoo pedig három találatot ad, köztük két svéd csevegőt... Nézzük, mit mond az Ask! "Your search for Veiszäcker did not match with any Answers results", és amit ajánl: "Make sure all words are spelled correctly", vagy "Try different keywords", és még van néhány hasonló javaslata...

Hát ezekkel a jó tanácsokkal bizony nem sokra megyünk, noha a helyesen írt "Weizsäcker" szó (melyre a Google több, mint félmillió találatot ad!) az előbbitől csak egy karakterben ("W" helyett "V"), és egy szomszédcserében ("Z" és "S") tér el, vagyis igen hasonlóak (hasonlóságuk 80% (!) a L0GANA asszociatív szöveghasonlító eljárása szerint...).

♣ Logana Információ-Kutatási Team, www.logana.com
"Nagy István" <analog@logana.com>, 06-30/547-65-94

Az asszociatív szövegkeresés tehát a L0GANA világviszonylatban egyedülálló képessége, a jelenleg létező kereső eszközök (Google, Bing, Yahoo, Ask, stb.) ehhez még csak hasonló funkcióval sem rendelkeznek. E képességet a saját fejlesztésű, nyelv- és karakterkészlet-független, rendkívül hatékony asszociatív szöveghasonlító algoritmusunk biztosítja.

Asszociatív Adatfeldolgozási Alkalmazások

(Kutatás-fejlesztési feladatcsoportok)

Az alábbiakban feladatokként fogalmazunk meg alkalmazás-orientált módszereket, melyek hátterében a *szekvenciális struktúrák hasonlítása*, vagy még általánosabban a *hasonlósági adatelemzés* áll. A feladatként való megfogalmazás a hasonlósági szemlélet sokoldalúságára utal, de persze egyúttal az innovációs lehetőségekre is fel akarja hívni a figyelmet.

A bemutatott feladatok részfeladatai között lehetnek átfedések, ez azonban nem baj, hisz ilyen a világ! Egy távcső lencséjének készítésekor, és egy szép üvegváza csiszolásakor is sok az azonos technológiai elem, az pedig igazán nem baj, ha a kutató-fejlesztő munkacsoportok időnként együttműködnek...

Alapfogalmak

A szekvenciális struktúrák hasonlóságát a mennyiségi és a sorrendi hasonlóság segítségével határozzuk meg. Ezek mindegyike 0 és 1 közötti törtszám, ahol a 0 a teljes különbözőséget, az 1 pedig a teljes azonosságot reprezentálja.

A *mennyiségi hasonlítás* során az összehasonlítandó szekvenciák azonos karaktereinek mennyisége határozza meg (sorrend-független módon) a mennyiségi hasonlóság mértékét (természetesen a többszörös kölcsönös előfordulást többszörösen figyelembe véve). A *sorrendi hasonlítás* során a (nem feltétlenül közvetlen) rákövetkezések mennyisége határozza meg a sorrendi hasonlóság mértékét. Végül a mennyiségi és a sorrendi hasonlóság szorzata adja magát a hasonlóságot. Megjegyezzük, hogy a mennyiségi hasonlóság mindig szimmetrikus, míg a sorrendi hasonlóság nem mindig. A találati sikeresség érdekében ezért általában mindkét irányban elvégezzük a sorrendi hasonlóságot, és a nagyobb értéket véglegesítjük.

A *határhasonlóság* egy 0 és 1 közötti törtszám (a felhasználó felé százalék) melyet két szekvencia hasonlóság értékének legalább el kell érnie az *asszociatív hasonlítás* esetén, és ennél kisebbnek kell lennie a *disszociatív hasonlítás* esetén az elfogadáshoz.

Az *asszociatív keretrelatív hasonlítás* során a vizsgált szekvencián egy, a keresett szekvencia hosszával egyező méretű keretet (az ebben lévő szekvenciát nevezzük keretszónak) mozgatunk karakterenként, és a keresett szekvenciát minden egyes keretszóval összehasonlítjuk. Az illeszkedést ekkor a karakterenkénti hasonlóságértékekből álló *hasonlósági függvény* olyan lokális maximumainál vizsgáljuk, ahol e maximum eléri, vagy meghaladja az előre megadott *határhasonlóság-értéket*. Az ismételt vizsgálatok során előfordulhat, hogy a keret méretét csökkentjük, és e csökkentett méretű keretet a keresett szekvencián is mozgatjuk. A genetikai vizsgálat esetén egy karakter egy triplet, a beszédfeldolgozás esetén pedig egy karakter egy fonéma.

Az *asszociatív hasonlítás* során tehát egy vizsgált szekvenciát valamely referenciának tekintett szekvencia halmazra és egy határhasonlóság értékre vonatkozóan elfogadunk, ha a vizsgált szekvencia a referencia szekvenciák *legalább egyikéhez legalább határhasonlóság mértékben* hasonlít.
A *disszociatív hasonlítás* során viszont akkor fogadunk el egy vizsgált szekvenciát valamely referenciának tekintett szekvencia halmazra és egy határhasonlóság értékre vonatkozóan, ha a vizsgált szekvencia a referencia szekvenciák *mindegyikéhez a határhasonlóság értékénél kevésbé* hasonlít.

1.) Szekvenciális struktúrák hasonlítása

N1. Asszociatív szövegkeresés dokumentumokban, táblázatokban és adatbázisokban

Előfordul időnként, hogy egy szöveges dokumentumban éppen a számunkra fontos szavak vannak hibásan, vagy legalábbis más helyesírással írva, és persze az is, hogy mi nem ismerjük egy számunkra fontos szó pontos alakját (tehát a keresőszó hibás alakú). Ilyenkor igen nagy szükségünk lenne egy hatékony hibatoleráns szövegkeresőre...
Feladat: A szakirodalomban áttekinteni az ilyen jellegű kutatásokat, valamint létrehozni egy ezzel a hasonlítási képességgel rendelkező lekérdező nyelvet, és e nyelvet használó alkalmazást készíteni az asszociatív szöveghasonlítás technikájával.

N2. Asszociatív internetes keresőrobot

A különböző létező internetes kereső eszközök (Google, Bing, Yahoo, stb.) igen korlátozott mértékben alkalmasak a hibásan keresett, a hibásan tárolt (de megfelelő karaktersorozattal keresett), vagy egyszerűen csak némileg más alakú szöveges adatok megtalálására, azonban az asszociatív szöveghasonlítási technika lehetőséget ad erre.
Feladat: Áttekinteni a jelenleg elterjedt keresőrobot megoldások elvét és megvalósítását a szakirodalom alapján, és létrehozni egy, az asszociatív szöveghasonlítási technikát alkalmazó internetes kereső eszközt.

N3. Automatikus kulcsszó és tárgyszó kigyűjtés szöveges dokumentumok esetén asszociatív és disszociatív szöveghasonlítással

A kereső kulcsszavak (vagy legalábbis az azokhoz "eléggé" hasonló szavak) előfordulásainak, vagy egy szöveg szignifikáns szavainak (a tárgyszavainak) automatikus kigyűjtése a könyvtártudomány egyik legizgalmasabb területe. E feladat még ismeretlen nyelvű szöveg esetén is megoldható az asszociatív, illetve a disszociatív szöveghasonlítás technikájával!
Feladat: A szakirodalomban áttekinteni az ilyen jellegű kutatásokat, valamint egy ezzel a képességgel rendelkező, referencia-szótárt használó, valamint referencia-szótár nélküli alkalmazást készíteni az asszociatív, illetve a disszociatív szöveghasonlítás technikájával.

N4. Dokumentumok hasonlóságának vizsgálata

Két dokumentum hasonlóságát értelmezhetjük a következőképpen: 1.) Állítsuk elő az egyik dokumentumból mindazon szavakat, melyek a másik dokumentum valamelyik szavához egy megadott mértéknél (S_1) hasonlóbbak (asszociatív kigyűjtés). 2.) Az így kigyűjtött szavak közül tartsuk meg azokat, melyek egy megadott mértéknél (S_2) kevésbé hasonlítanak az általánosan használt szavak referencia szótárának minden szavához (disszociatív szűrés). 3.) Az így kapott szavak közül csak azokat tartsuk meg, melyek egy adott hasonlósági mértéknél (S_3) kevésbé hasonlítanak egymáshoz (disszociatív térredukció).
Feladat: A szakirodalomban áttekinteni az ilyen jellegű kutatásokat, valamint elkészíteni egy, a fenti módszert alkalmazó dokumentum-hasonlító alkalmazást.

N5. Szekvenciális struktúrák keretrelatív asszociatív felismerésének módszere

Kiemelt hírnek számít a két- és háromdimenziós alakzatok felismerésére alkalmas eszközök megjelenése, pedig az egydimenziós szekvenciális struktúrák hasonlósági felismerése sem triviális. (Gondoljunk például egy számítógépes vírus keresésére, egy népi motívum-szekvencia, vagy egy ismeretlen írás szerkezetének felismerésére, szóra bontására, stb.)
Feladat: A szakirodalomban áttekinteni az ilyen jellegű kutatásokat, valamint egy ezzel a képességgel rendelkező alkalmazást készíteni az asszociatív keretrelatív szöveghasonlítás technikájával.

N6. Mutáns génszekvencia asszociatív keresése DNS-láncokban

A DNS-lánc 64-féle, úgynevezett szervesbázis-láncból (tripletből) épül fel, és kb. 1 milliárd tripletből áll. Mivel pedig egy gén kb. 100 tripletet tartalmaz, így egy "DNS-mondat" kb. 10 millió "GénSzóból" épül fel. Az orvoskémiai módszerekkel szemben a génszekvenciák (génrészek) felismerése (akár a mutáns génszekvenciáké is!) elvégezhető az asszociatív keretrelatív szöveghasonlí-

tás technikájával.
Feladat: A szakirodalomban áttekinteni a DNS-láncokkal kapcsolatos strukturális ismereteket, és egy nagy erőforrású számítástechnikai környezetben alkalmazást készíteni a génszekvenciák felismerésére az asszociatív keretrelatív szöveghasonlítás technikájával.

N7. A fonéma alapú asszociatív beszédfelismerés

A fonémák a kiejtett hangok, vagyis a beszélt nyelv „betűi". Az ezeken alapuló beszédfelismerés jobban bővíthető, rugalmasabb rendszert biztosít, mint a hagyományos teljes szavas beszédfelismerés.
Feladat: A szakirodalomban áttekinteni az ilyen jellegű kutatásokat, valamint elkészíteni egy ilyen jellegű alkalmazást az asszociatív keretrelatív szöveghasonlítás technikájával. A cél tehát egy olyan alkalmazás készítése, amely a folyamatos vagy diszkrét-szavas beszédfelismerés eredményeként keletkező szöveget helyesírásilag korrekt szöveggé alakítja.

N8. A hibatoleráns információ-átvitel asszociatív szöveghasonlítással

Az információ-elmélet a spektrális analízis, és a korreláció-elemzés segítségével megfelelő eszközöket ad a zajos csatornán áthaladó információ hibatoleráns átviteléhez, de hogyan javíthatók ki a forráshibák (pl. a selypítés, a pöszeség, vagy a dadogás)? Ha mi megértjük (márpedig általában megértjük) a beszédhibás, vagy a tájszólásban elhangzó beszédet, akkor elvárható, hogy megértse a gép is! Véges szókészlet (például diszpécseri parancsnyelvi vezérlés) esetén az asszociatív keretrelatív szöveghasonlítással e feladat hatékonyan megoldható!
Feladat: A szakirodalomban áttekinteni az ilyen jellegű kutatásokat, valamint egy ezzel a képességgel rendelkező alkalmazást készíteni az asszociatív keretrelatív szöveghasonlítás technikájával.

N9. Szövegjavítás asszociatív szöveghasonlítással

Egy szöveg különböző okokból tartalmazhat hibásan írt szavakat. Ennek oka lehet a szöveg jellege, mely lehet például egy beszédfelismerés eredményeként keletkező fonéma-szekvencia, egy szkennelést követő OCR eredmény, egy SMS-ben küldött parancsnyelvi üzenet, de lehet hogy valamilyen szaknyelvi (például orvosi) szövegről van szó, vagy egyszerűen csak nem tud valaki helyesen írni. Ilyenkor az intelligens szövegszerkesztőkben aktivizálhatók a helyesírás ellenőrző eszközök, ám ha maga a szógyök hibás (főként már a szógyök eleje!), akkor bizony ezek sem tudnak segíteni.
Feladat: Egy olyan alkalmazás készítése, mely képes az asszociatív szöveghasonlítás technikájával egy szöveg hibásan írt szavait kijavítani, illetve javaslatot tenni a lehetséges javításokra. Értelmezze a hibák jellegét, és mértékét (hasonlóságként), és adjon lehetőséget az alkalmazás felhasználójának arra, hogy a javítást különböző (határhasonlósággal korlátozott) hasonlósági tartományokban végezze. A megvalósítás részeként tekintse át a szakirodalmat az ilyen jellegű kutatásokkal, illetve forgalmazott termékekkel kapcsolatban.

N10. Szekvenciális struktúrák asszociatív keresésének hardveres előszűrése

A szekvenciális struktúrák asszociatív keresése (például dokumentumban, táblázatban, vagy adatbázisban való szövegkeresés) során bizonyos statisztikai feltételeknek nem megfelelő szekvenciák (jellemzően szavak) kiszűrését egy hardveres előszűrés lényegesen gyorsíthatja.
Feladat: A szakirodalomban áttekinteni az FPGA és a GPGPU eszközök hardver-környezetbe való illesztését, az ezeket kezelő programozási nyelvek használatát, valamint egy ezekre alapuló alkalmazás elkészítésének keretében a megfelelő mennyiségi hasonlítást elvégző hardverstruktúra kialakításával az asszociatív szöveghasonlítási feladatokra egy előszűrési megoldást adni. (Ennek eredményeként a csak szoftveresen végezhető, és így erőforrás-igényesebb sorrendi hasonlítást már egy lényegesen kisebb szekvencia-halmazra vonatkozóan kell csupán elvégezni.) Kiegészítő feladat a létrehozott eszköz feldolgozási hatékonyságot növelő hatását bemutatni valamely szekvenciális struktúra asszociatív hasonlítását végző alkalmazás esetén (együttműködve egy másik alkalmazás megvalósítóival).

2.) Numerikus asszociatív adatfeldolgozás

Az alábbi feladatok lényegesen eltérnek az előzőktől, mivel a struktúrálatlan (vagy legalábbis annak tekintett) objektumokon végeznek adatelemzést. Ezúttal a *hasonlósági* feldolgozás helyett célszerűbb az *asszociatív* jelző használata, mivel az algebrailag egzakt hasonlóság fogalmat (hasonlóság egy reflexív és szimmetrikus reláció) sajnos a geométerek és (ennek következtében) a statisztikusok "némileg" eltérő módon értelmezik (lásd hasonló háromszögek, vagy a korlátlanul használt hasonlósági függvények), vagyis triviális ekvivalencia-relációként (amelyre tehát a hasonlósági reláció fenti tulajdonságain kívül még a tranzitivitás is teljesül!). Ennek pedig alapvető következménye, hogy az így keletkező osztályok mereven elhatárolódnak (diszjunktak, azaz nem átfedők), vagyis az objektumok már egészen kis eltérések esetén is más osztályba kerülhetnek. Ennek pedig már egyszerű (például adózó) állampolgárként is érezzük a hatását...

N11. Az asszociatív adatfeldolgozás módszere

A (numerikus) asszociatív adatfeldolgozás során a vizsgált objektumteret az objektumok figyelembe vett tulajdonságai (az attribútumok) alapján először egy többdimenziós numerikus adattérré alakítjuk. Ezután a numerikus értékeken valamilyen hasonlósági függvényt értelmezve, egy, hasonlóság-értékekkel súlyozott teljes gráfot kapunk, melyen a határhasonlóságnál kisebb hasonlósági értékű éleket töröljük. A következő lépésben a valamilyen szempontból legmegfelelőbb objektumok (ezeket nevezzük reprezentatív objektumoknak) határhasonlóság sugarú környezeteit határozzuk meg, és ezek az alterek összességében alkotják a hasonlósági teret. (A reprezentatív objektumok kijelölésének egyik feltétele lehet például az, ha az objektumba befutó élek hasonlóságainak összege meghalad egy előre megadott értéket.) Könnyen belátható, hogy az egyes alterek lehetnek átfedőek, tehát ez a hasonlósági tér egy valódi hasonlósági relációt alkot, ezért ezeket az altereket hasonlósági osztályoknak is nevezhetjük. Az így létrejött hasonlósági téren elvégezhető adatfeldolgozások főbb eredményei:
1.) A reprezentatív objektumok halmazát tekinthetjük egy *hasonlósági térredukció* eredményének. Az így redukált téren az attribútumokra vonatkozó adatelemzések lényegesen hatékonyabban végezhetők el, ami különösen hasznos az igen nagy objektumterek (például adattárházak) esetén.
2.) A reprezentatív objektumokhoz tartozó alterekben (a hasonlósági osztályokban) elvégezhetjük azon attribútumok kiválasztását, amelyek szignifikánsan megkülönböztetik az azonos altérbeli objektumokat a más altérbeli objektumoktól. (Tehát a különböző hasonlósági osztályokban különböző attribútumok (vagy legalábbis különböző attribútum-értékek) "tartják össze" az objektumokat.) Ezáltal alterenkénti *hasonlósági attribútum-redukciót* végeztünk el, melyet *lényeg-kiemelésnek* is nevezhetünk.
3.) A hasonlósági osztályok létrejötte, és az egyes osztályokon elvégzett attribútum-redukció lehetőséget ad *új összefüggések, hatásmechanizmusok felismerésére*.
Feladat: A szakirodalomban áttekinteni a hasonlósági osztályozásokat, és egy alkalmazás elkészítésén keresztül bemutatni az asszociatív adatfeldolgozást. Kiegészítő feladat annak megvizsgálása, hogy a hasonlósági attribútum-redukció milyen lehetőséget ad a lényeg-kiemelésre, új összefüggések, hatásmechanizmusok felismerésére.

N12. Adattárházak építése és asszociatív elemzése

A korszerű döntéstámogató rendszerek a különböző adatforrásból kinyert nagyon nagymennyiségű adatot valamely jól megfogalmazott feladatosztály számára először előkészítik (szűrés, aggregálás), megfelelően hatékony struktúrába rendezik (adattárház létrehozás), végül különböző idősor-elemzési, statisztikai és egyéb vizsgálatokat, feldolgozásokat végeznek rajtuk (ezt nevezik adatbányászatnak). Az asszociatív adatfeldolgozás révén létrehozott redukált téren hatékonyabb elemzések végezhetők, az attribútum-redukció eredményeként pedig új összefüggések ismerhetők fel.
Feladat: A szakirodalomban áttekinteni az adattárház-építési és adatbányászati technikákat, valamint egy meghatározott nagyméretű adattéren, egy meghatározott elemzési céloknak megfelelő adattárház létrehozása, és ezt valamely alkalmazás keretén belül feldolgozni az asszociatív adatfeldolgozás technikájával.