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) a számunkra egyaránt jelenti az adatok hasonlósági feldolgozását, 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.

Az ANAL0G asszociatív szövegkereső (bemutatását lásd a www.logana.com web-helyen) egy olyan program, mely a szokásosan használt böngészőkbe (Internet Explorer, Opera, 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 másik adat, 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 három találatot is ad; egy svéd csevegő oldal három hozzászólását, ahol persze éppen ebben az alakban (Richard von Veiszäcker) hivatkoznak a német politikusra. Ennek a sikernek itt 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 azt üzeni, hogy "Veiszäcker keresése nem adott találatot.", a tippje pedig: "Ellenőrizze a helyesírást.". Aranyos, nem? Nézzük, mit mond a Yahoo? " We did not find results for: Veiszäcker. Try the suggestions below or type a new query above.", és amit ajánl: "Check your spelling."

Hát ezekkel a jó tanácsokkal bizony nem sokra megyünk, noha a helyesen írt "Weizsäcker" szó (melyre a Google több, mint egymillió találatot ad!) az előbbitől csak egy karakterben ("W" helyett "V"), és egy szomszédcserében ("Z" és "S") tér el, vagyis igen hasonlóak (hasonlóságuk 80% (!) az ANAL0G asszociatív szöveghasonlító eljárása szerint...).

♣ Óbudai Egyetem, Neumann János Informatikai Kar,
Oracle Kompetencia Központ,
Logana Információ-Kutatási Team
[Nagy.Istvan@Nik.Uni-Obuda.hu]

Az asszociatív szövegkeresés tehát az ANAL0G világviszonylatban egyedülálló képessége, a jelenleg létező kereső eszközök (Google, Yahoo, Bing, stb.) ehhez még csak hasonló funkcióval sem rendelkeznek. E képességet a saját fejlesztésű, nyelv- és karakterkészlet-független, rendkívül hatékony asszociatív szöveghasonlító algoritmusunk biztosítja.

Az Asszociatív AdatFeldolgozás Alkalmazásai

(Kutatás-fejlesztési feladatcsoportok)

Az alábbi feladatcsoportok részfeladatai között lehetnek átfedések, hiszen az alapfeladat mindegyikben hasonló; *a szekvenciális struktúrák hasonlítása*, illetve esetenként egy ennél még általánosabb feladat, *a hasonlósági adatelemzés*. Ez azonban nem baj, hisz ilyen a világ! Egy távcső lencséjének készítésekor, és egy szép üvegváza csiszolásakor is sok az azonos technológiai elem, és hát az igazán nem ártalmas, ha a kutató-fejlesztő munkacsoportok időnként együttműködnek...

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. 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.

N3. 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.

N4. 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 szöveghasonlítás 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.

N5. 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.

N6. 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.

N7. Szövegjavítás asszociatív szöveghasonlítással

Egy szöveg különböző okokból tartalmazhat hibásan írt szavakat. Ennek oka lehet a szöveg jellege, mely lehet például egy beszédfelismerés eredményeként keletkező fonémaszekvencia, 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.

Tekintse át a szakirodalmat az ilyen jellegű kutatásokkal, illetve forgalmazott termékekkel kapcsolatban.

N8. 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.

N9. Asszociatív internetes keresőrobot

A különböző létező internetes kereső eszközök (Google, Bing, Yahoo, stb.) igen korlátozott mértékben alkalmasak a hibásan keresett, vagy hibásan tárolt (de megfelelő karaktersorozattal keresett) 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 létrehozni. Az alkalmazás mutassa be a megfelelő kiegészítő hardver eszközök (FPGA, GPGPU) használatának jelentőségét.

N10. Szekvenciális struktúrák keretrelatív asszociatív felismerése

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.

Az alábbi két feladat "kilóg" a sorból, mivel a szövegnél általánosabb objektumon végez feldolgozást. Ezekben az esetekben is használjuk a *hasonlósági feldolgozás* kifejezést, az *asszociatív* jelző azonban azért célszerű, mivel az algebrailag egzakt hasonlóság fogalmat (hasonlóság egy olyan reláció, mely reflexív és szimmetrikus) 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 osztályozások diszjunktak (azaz nem átfedők), aminek már egyszerű (például adózó) állampolgárként is érezzük a hatását...

N11. Asszociatív adatfeldolgozás

Az asszociatív adatfeldolgozás során a vizsgált adatteret először többdimenziós numerikus adattérré, majd hasonlósági adattérré alakítjuk, majd a reprezentáns elemek kijelölésével redukáljuk az adatteret, és az így keletkező, részben átfedő hasonlósági osztályok mindegyikén még a reprezentáns attribútumok kiválasztását is elvégezzük. Összességében tehát az adattéren hasonlósági térredukciót, és hasonlósági attribútum-redukciót is végzünk.

Feladat: A szakirodalomban áttekinteni a hasonlósági osztályozásokat, és egy alkalmazás elkészítésén keresztül megvalósítani az asszociatív adatfeldolgozást. Kiegészítő feladat annak megvizsgálása, hogy a hasonlósági attribútum-redukció milyen lehetőséget ad a lényegkiemelésre, 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 idősorelemzési, statisztikai és egyéb vizsgálatokat, feldolgozásokat végeznek rajtuk (ezt nevezik adatbányászatnak).

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.