Budapesti Műszaki és Gazdaságtudomány Egyetem
Informatikai Tudományok Doktori Iskola
Híradástechnikai Tanszék

Elosztott gépi intelligencia módszerek néhány vonatkozása

Tézisfüzet

Benkő Borbála Katalin
Konzulens: Imre Sándor, DSc

2012

Tartalomjegyzék

Tartalomjegyzék 1

Bevezetés 2

I. OLAKO: Egy nyílt kollektív-adaptív rendszer 3
I.1. Háttér és motiváció 3
I.2. Célok 4
I.3. Új tudományos eredmények 4
I.4. Eredmények alkalmazása 5

II. Javított önszerveződés Keverő ODC algoritmussal 6
II.1. Háttér és motiváció 6
II.2. Célok 7
II.3. Új tudományos eredmények 7
II.4. Eredmények alkalmazása 8

III. Vox Populi: Névelem alapú szentiment bányászat természetes nyelvi szövegekből 8
III.1. Háttér és motiváció 8
III.2. Célkitűzések 9
III.3. Új tudományos eredmények 10
III.4. Eredmények alkalmazása 11

Publikációs lista 12

Bevezetés

Disszertációm új, elosztott, autonóm, kollektív gépi intelligencia módszereken alapuló megoldásokkal–például autonóm entitások közötti tudásmegosztás, emergens önszerveződés, tudás önrevízió–foglalkozik, bemutatva hogy ezen módszerek hogyan alkalmazhatók sikeresen néhány konkrét problémakörre, a rendszer adaptivitását illetve teljesítményét növelendő. A szakirodalomból ismert algoritmusokat kiterjesztettem, új problématerületekre alkalmaztam, illetve gyakran az elosztottságból adódó emergens hatásokat is kihasználtam.

A disszertáció három alterületre vonatkozó kutatási eredményeimet tartalmazza. Az összekötő kapocs az alterületek között az új, elosztott, autonóm gépi intelligencián alapuló megközelítés.

Az **első téziscsoport** a nyílt, kollektív-adaptív multiágens rendszerek területén elért eredményeimet tárgyalja. Egy új megközelítést (OLAKO modell) ajánlok az erősen dinamikus környezeti elvárások problémájára, mely autonóm ágensek online tanulásán és az ágenstársadalom belüli páronkénti tudásmegosztáson, illetve a tudás ágens szintű optimalizációján és tisztításán alapul. Megmutattam, hogy a model magasfokú adaptivitást tesz lehetővé, a tudásmegosztás lerövidíti tanuláshoz szükséges időt, illetve a tudás ön-optimalizáció job tanulási görbét eredményez.

A **második tézicsoport** az önszerveződő mobil hálózatok igény szeriti klaszterezéséhez kapcsolódó eredményeimet összegzi. Egy új, igény szerinti klaszeterező algoritmust javasolok (Keverő ODC), mely a szakirodalomból ismert igény szerinti klaszterező algoritmus (ODC, On Demand Clustering / Igény Szerinti Klaszterezés algoritmus) kiterjesztése. Megmutattam, hogy a Keverő ODC segít az ODC algoritmusnál fennálló topologikus éhezési probléma elkerülésében, nagyobb klasztereket és optimálisabb klaszter alapú terhelésmegosztást tesz lehetővé a lokalitási elv megsértése nélkül, továbbá a kommunikációs költség jelentős növelése nélkül.

A **harmadik téziscsoport** természetes nyelvű szövegekből történő szentiment bányászat területén elért eredményeimet mutatja be. Definiáltam a névelem alapú szentiment bányászat problémáját, és egy általános, nyelvfüggetlen modellt javasoltam hozzá. Kísérleteken keresztül megmutattam, hogy a javasolt model magas pontossági és visszaadási értékeket képes nyújtani a gyakorlatban is. Ezen felül egy felügyelet nélküli ön-revíziós módszert javasoltam a tudás karbantartására illetve a tudás egészségének monitorozására, mely felügyelet nélküli, azaz annotált korpuszt nem igénylő, szótár ön-ellenőrzésén és a (nem annotált) teszt korpusz automatikus teljességvizsgálatán alapul.

Közreműködők a mögöttes kutatások során

A tézisek mögött húzódó kutatás gyakran más kutatókkal és hallgatókkal közösen folyt (a részleteket lásd a disszertáció előszavában). A megfogalmazott tézisek azonban én saját eredményekeimen alapulnak.

I. OLAKO: Egy nyílt kollektív-adaptív rendszer

I.1. Háttér és motiváció

Az elmúlt évtizedben egyre nagyobb tudományos figyelem vetült az *autonóm ICT*[1] és *adaptív rendszerek* témakörére. Definíció szerint ezek a rendszerek emberi beavatkozás nélkül is képesek a környezetben beálló bizonyos fokú változásokhoz alkalmazkodni. Az egyik szokásos alapmodell az adaptivitáshoz egy visszacsatolási kör alkalmazása: a rendszer megfigyeli a környezetét, kiválasztja a legmegfelelőbb cselekvést/beavatkozást, kivitelezi azt, majd újrakezdi a kört a hatás megfigyelésével. A hagyományos irányító rendszerekkel szemben a fő különbség, hogy egy adaptív rendszerben az elérendő cél menet közben megváltozhat, és a rendszernek dinamikusan követnie kell a célban beálló változásokat (statikus vs dinamikus optimum).

Napjaink adaptív rendszerei többnyire explicit, előre elkészített (pre-injektált) modelleket használnak–világ modell, önmodell–, melyek leírják, hogy a könyezet illetve maga a rendszer hogyan működik. A rendszer ezeket a modelleket használja mind normál működése során a következő cselekvés kiválasztásához, mind a nem várt helyzetek kezelésére. A modell alapú működést számos járulékos eszköz (következtetés, ontológiák) teszi hatékonnyá. A megközelítés fő hátulütője azonban, hogy túlságosan is a modellek és a modellezett hatások teljességére, helyességére és időbeli állandóságára épít. Mindaddig, amíg a modellek összehangban vannak a valósággal, a rendszer helyesen és hatékonyan tud működni és adaptálódni, azonban nem modellezett helyzetekben vagy a valóság és a modell közötti eltérés esetén mindez azonnal szertefoszlik, sőt, a helytelen modellek akár teljesen helytelen cselekvési spirálba is sorodhatják a rendszert. Azok az erős prekoncepciók tehát, melyek a rendszer működését és alkalmazkodását egyes esetekben olyan hatékonyá teszik, másszor épphogy teljesen ellehetetlenítik azt.

Kutatásom egyik célja annak a vizsgálata, hogy építhető-e *adaptív rendszer erős, explicit prekoncepciók nélkül*, vagy legalábbis azok lazításával, például online tanuláson alapulva. Azaz lehetséges-e az adaptációs folyamatot bizonyos szinten szabadjára engedni, és hagyni, hogy a rendszer maga találja meg a helyes utat az előre beprogramozott instrukciók követése helyett. Egy ilyen rendszerrel szemben a nyitottság számos szinten is elvárás: nyílt világmodell, mely a megfigyelt tények alapján folyamatosan finomodik és alkalmazkodik a valóságban beálló esetleges változásokhoz; nyílt adaptációs startégiák melyek során a rendszer egyszerű elemekből építkezik, nem pedig kész, komplex stratégiákat használ; szociális nyitottság, ahol az ágensek megoszthatják egymással a tudásukat még akkor is, ha a másik ágens nem biztos, hogy ugyanazzal a problémával találkozik; végül nyitottság a tudás folyamatos optimalizációjára (tömörítés, tisztítás, elavultatás).

Az adaptív és autonóm rendszerekkel kapcsolatban gyakran merül fel a *kollektív viselkedés*, csoport intelligencia és emergens hatások témaköre. Ezek az állat- és növényvilágban megfigyelhető módszereken alapuló, és ott már bizonyítottan működő algoritmusok sokszor jól alkalmazhatók számítástechnikai problémákra is. Ha nagyon

[1] ICT: information and communication technology / információs és kommunikációs technológia

sok entitás végzi ugyanazt az egyszerű műveletsort, a csoportosan elért hatás emergens lehet, azaz az eredeti műveletsornál alapvetően magasabb szintű. A szakirodalom itt főként néhány konkrét alkalmazási területre, például az emergens önszerveződésre, koncentrál.

Kutatási célom ezen a területen annak a megértése volt, hogy a kollektív tanulásnak milyen előnyei és hátányai lehetnek a nyílt adaptív rendszerek területén: a tudás megosztása az egyedek között hogyan befolyásolja a tanulási görbét, illetve milyen emergens hatásokkal kell számolnunk, ha a művelet nem csupán elenyésző számban, hanem a teljes társadalom szintjén–korlátolt megfigyelőképesség és változó környezeti elvárások mellett–rutinszerűen megy végbe. Tudomásom szerint ez a problémakör a szakirodalomban ezelőtt még nem került a figyelem középpontjába.

Új megközelítést javaslok a változó és nem teljesen megfigyelhető környezeti elvárások problémájára, mely erősen adaptív online tanuláson, tudás ön-optimalizáción és a tanuló egyedek közötti kollektív tudásmegosztáson alapul. A javasolt OLAKO modell a szakirodalomból ismert módszerektől két alapvető pontban különbözik: (i) nagymértékben online tanuláson és ön-* tulajdonságokon alapuló autonóm adaptív rendszer, (ii) lehetővé teszi az egyedek önállóan kifejlesztett és optimalizált-adaptálódott tudásának intenzív, akár a teljes társadalomra kiterjedő megosztását.

I.2. Célok

A kutatás célja egy olyan megoldási irány kidolgozása és vizsgálata volt, mely a dinamikusan változó környezeti kihívások problémájára explicit, pre-injektált tudás nélkül, kollektív-adaptív módon ad választ. A központi kérdések a következők:

- Alkalmazható-e online tanulás egy kollektív-adaptív rendszerben? Lehetséges e így a rendszerbe előzetesen betáplálandó explicit tudás mértékét csökkenteni, vagy azt teljesen elhagyni?
- Hogyan egyeztethető össze a tudásmegosztás elve azzal, ha az ágensek egymástól függetlenül feljesztik és optimalizálják egyéni tudásukat, és így akár drasztikusan eltérő irányokba fejlődhetnek?
- Milyen hatásai lehetnek a tudásmegosztásnak egyéni illetve közösségi szinten?
- Lehet-e az dinamikusan épített, esetleg részben importált tudást futási időben, igény szerint ön-optimalizálni? Hogyan tudja egy ágens futási időben a világról való absztrakcióit (például a megfigyelt világleírók körét) felülvizsgálni, illetve dinamikusan újraszervezni vagy akár részben elvetni?
- Az online tanulás lehetővé teheti-e, hogy a rendszer beavatkozás nélkül alkalmazkodjon az elvárásokban a futás során dinamikusan bekövetkező komoly változásokhoz?

I.3. Új tudományos eredmények

I-1. tézis *(OLAKO modell) A dinamikusan változó környezeti elvárások problémakörére egy komplex modellt, egy online tanulás alapú kollektív-adaptív rendszert javasoltam. Az OLAKO modell az online tanulást az ágensek közötti*

emergens tudásmegosztással és egy autonóm tudás ön-optimalizációs mechanizmussal kombinálja. A modell vizsgálatára egy formális keretrendszert javasoltam, a formális leírásból következő néhány alapvető állítással egyetemben.

Kapcsolódó publikációk: J3, J8u, C9

I-2. tézis *(OLAKO tudás ön-optimalizáció) A szakirodalomból ismert LSTD (legkisebb hibanényzetek időkülönbsége) tanulási algoritmust kiterjesztettem az online, dinamikusan beszúrt és levetett világleírók (feature-ök) elvével. A világleírók újraszervezésére egy hierarchikus csoportbafoglaláson és egyesítésen alapuló modellt javasoltam. A tudás ön-optimalizációjához konkrét algoritmust javasoltam világleírók futási idejű, dinamikus bevezetésére és elvetésére. Algoritmust javasoltam a tudás de-optimalizációjához, melynek segítségével az optimalizált tudás visszaalakítható egy, az eredeti világleírókat tartalmazó közelítő formára. Szimulációval megmutattam, hogy a tudás ön-optimalizáció 11-18%-kal gyorsabb tanulást eredményez a mintaproblémán.*

Related publications: J3, J8u, C9

I-3. tézis *(OLAKO tudásmegosztás) Modellt és algoritmust javasoltam a függetlenül kifejlesztett LSTD tanuláson alapuló tudásblokkok más ágensek tudásbázisába való integrációjára. Tudásimportálási stratégiákat javasoltam, beleértve az ágensek párosítási elvét, az importálási döntés meghozatalát és magát a tudásintegrációs mechanizmust. Szimulációval megmutattam, hogy a tudásimport javítani képes az ágens teljesítményét akár még gyenge donortudás mellett is (+46-84% vesztéselkerülési arány és +2% sikerarány a mintaproblémán a tudásimport nélküli esethez képest). Leírtam a tudásimportnak az ágenstársadalom szintjén keletkező emergens hatását. Szimulációval megmutattam, hogy a teljes rásadalom szintjén történő random tudásimportnak emergens negatív hatása lehet, a mintaproblémán az ágensek tudása ezen stratégia mellett teljesen ellehetetlenült, és az idő előrehaladtával sem javult. Szimulációval megmutattam, hogy a társadalmi szintű "okos" tudásimportálási stratégiáknak, ahol az importálási döntés a donortudás haszonsságbecslésén alapul, emergens pozitív hatása lehet, a mintaproblémán ez a stratégia 21%-kal magasabb sikerarányt és 30%-kal magasabb vesztéselkerülést eredményezett a tudásimport nélküli változathoz képest.*

Kapcsolódó publikációk: J3, J8u, C9

I.4. Eredmények alkalmazása

Az OLAKO modell egy általános megközelítés. Tetszőleges problémakörre alkalmazható, melyben a probléma világállapotok sorozatával, cselekvések véges halmazával és esetenkénti visszajelzéssel modellezhető. Tudásmegosztási szempontból előnyös, ha sok tanuló ágens találkozik hasonló problémával.

Egy konkrét alkalmazási terület az intelligens kütyük (gadget) és szolgáltatások világa közvetlen humán használatra, például egy adaptív grafikus felület, egy proaktív gyorsítótár vagy egy proaktív intelligens épület vezérlőrendszer, mely az emberi felhasználó mindennapi életét megkönnyíti azáltal, hogy rátanul mind az emberek általános szokásaira, mind a konkrét felhasználó egyedi elvárásaira. Egy másik alkalmazási terület az adaptív hálózatok világa, például egy adaptív felhasználói élmény (QoE) alapú média adó, mely a felhasználó visszajelzése alapján finomítja a mögöttes videófolyam beállításait. Egy harmadik konkrét terület az online játékok és virtuális valóság rendszerekhez kapcsolódik, itt az OLAKO alapú ágensek a virtuális térben hasonlóképp tanulnak rá és alkalmazkodnak az ember és a játék elvárásaihoz, mint a disszertációban bemutatott mintavilág játékosai.

II. Javított önszerveződés Keverő ODC algoritmussal

II.1. Háttér és motiváció

A hálózati önszerveződés a 3GPP LTE rendszerektől a pervazív számítástechnikán át az ad-hoc mobil hálózati alkalmazásokig számos területen kap egyre nagyobb szerepet. Ezen a forgatókönyvek szerint a hálózati elemek erőforrásokat osztanak meg egymással, kommunikálnak, terheléskiegyenlítést végeznek vagy szolgáltatásokat nyújtanak egymásnak anélkül, hogy ennek vezérléséhez központi menedzsment entitásra lenne szükség.

Míg az emberi hálózatirányító hozzáférhet bizonyos, a hálózat egészére vonatkozó információkhoz, önszerveződési esetben az egyes elemeknek a döntéseiket kizárólag lokális információikra kell alapozniuk–még ha ismert is, hogy ez se nem teljes, se nem objektív–, illetve az eszközölt beavatkozások is tisztán lokálisak lehetnek. Ezt a szakirodalom gyakran a *lokalitás elvének* is nevezni. Az önszerveződés gyakran alapul egyszerű, de emergens hatásokkal rendelkező algoritmusokon, ahol egyszerű lépések ismételt, párhuzamos futtatása "intelligens" magasszintű viselkedést hoz létre, azaz egy magasabb szintű minőséget eredényez a teljes rendszer szintjén. Az emergens algoritmusok gyakran a természetből merítenek az inspirációt, biológiai, kémiai vagy fizikai folyamatokból, például madarak vagy hangyák csoportkommunikációjából, az emberi agyban lezajló folyamatokból vagy éppen az emberi immunrendszerből.

Önszerveződő hálózatokban *klaszterezés* alatt azt értjük, hogy a hálózat elemei olyan más elemeket keresnek a hálózaton belül, amelyek megfelelnek a klaszterezési feltételnek (önmagukhoz hasonló elemeket normál klaszterezés esetén, illetve kiegészítő párt inverz klaszterezés esetén), majd egy felettes hálózati szinten kapcsolatot építenek ki velük. Az így kialakuló felettes hálózat (overlay network) számos célra használható. Az egyik szokásos felhasználási mód a *terheléskiegyenlítés*, azaz a hálózati elemek a bejövő terhelést megosztják a nem túlterhelt klaszterszomszédokkal. A hatékony klaszterképzés a hatékony terheléskiegyenlítésnak szükséges de nem elégséges feltétele, a terheléselosztáshoz még meg kell válaszolni, hogy melyik elem, mikor és mely másik elemnek ad át a saját terheléséből, illetve ezt a döntést milyen információk alapján hozza meg.

Munkám a biológiailag inpirált, teljesen elosztott önszerveződésre és kifejezetten a nagy hálózatok esetére koncentrál, külön figyelemmel a terheléselosztási célú klaszterezére. Kiinduló problémának az igény szerinti klaszterezést választottam, mely esetben a hálózati elemek akkor kezdenek klasztereződni vagy klasztert bővíteni, ha erre igény van, azaz például magasabb a bejövő terhelés, mint amivel a klaszter jelenleg elbír. Ez az igény szerinti megközelítés az általános esettől két fontos pontban tér el: (i) az előállt klaszterek mérete a terheléstől függ, azaz a klaszterméret nem fogja elérni az elméleti maximumot, ha a bejövő terhelés ezt nem indokolja, (ii) a klasztereződési sebessége különösen fontos, hisz minden bővítési kezdeményezés mögött azonnali igény áll. Egy gyors, de kisebb mértékű bővülés könnyen hasznosabb lehet, mint egy lassú, de nagyarányú bővülés.

Munkám újdonsága a javított klaszterezési képesség (gyorsabban és nagyobb klaszterek jönnek létre), melyet egy szakirodalomból ismert alap algoritmus kiterjesztésével értem el, és terheléskiegyelítési szcenáriókon keresztül bizonyítottam.

II.2. Célok

A kutatás célja a British Telecom eredeti ODC (On Demand Clustering, igény szerinti klaszterezés) algoritmusának javítása volt a lokalitási elv megsértése és jelentős kommunikációs többlet okozása nélkül. A központi kérdések a következők voltak:

- Mi az ODC algoritmus gyenge pontja vagy szűk keretmetszete?
- Hogyan javítható az ODC teljesítménye?
- A módosított algoritmusnak milyen hatásai vannak? Vannak-e váratlan, esetleg emergens mellékhatások?
- A módosított algoritmus hogyan viszonyul a lokalitási elvhez, illetve milyen kommunikációs többletköltséget von maga után?

II.3. Új tudományos eredmények

II-1. tézis *(Topologikus éhezés) Azonosítottam és leírtam az topologikus éhezés jelenségét, mely egy emergens negatív mellékhatás a közvetlen nyereség alapú link rekonfigurációra épülő önszerveződési mechanizmusokban. Megmutattam, hogy az erdeti ODC algoritmus elősegíti a topologikus éhezés kialakulását.*

Kapcsolódó publikációk: J4, J5, C8

II-2. tézis *(Keverő ODC) Az eredeti ODC algoritmust kiterjesztettem a közvetlen nyereség nélküli link rekonfiguráció elvével. Az így előálló algoritmus (Keverő ODC) apró változtatásokat eszközöl a közeli overlay topológiában abban az esetben, ha a klaszterbővítési kísérlet sikertelen volt. Szimulációkkal megmutattam, hogy a Keverő ODC a topologikus éhezés elkerülése által 25-30%-kal több klaszterezett csomópontot, körönként 17-23%-kal több feldolgozott terhelést, 20-30%-kal hatékonyabb munkamegosztást és 30-50%-kal alacsonyabb össz feldolgozási időt eredményez az eredeti ODC algoritmushoz képest.*

Kapcsolódó publikációk: J4, J5, C8

II-3. tézis *(Keverő ODC kommunikációs költsége) Megmutattam, hogy a Keverő ODC algoritmus kommunikációs költsége maximum lineárisan arányos az eredeti ODC algoritmus kommunikációs költségével. Megmutattam, hogy ha a klaszterbővítési kísérlet mindig sikeres, akkor a két algoritmus kommunikációs költsége megegyezik.*

Kapcsolódó pubikációk: J4, J5, C8

II-4. tézis *(Lokalitás megőrzése Keverő ODC esetén) Szimulációkon keresztül megmutattam, hogy a terheléskiegyenlítésre használt Keverő ODC nem töri meg a lokalitás elvét. A megosztott terhelés által megtett átlagtávolság 56-63%-kal kisebb, mint a klaszterméretben bekövetkező növekmény.*

Kapcsolódó publikációk: J4, J5, C8

II.4. Eredmények alkalmazása

A Keverő ODC minden olyan területen alkalmazható, ahol igény szerinti klaszterezésre van szükség.

Az egyik fontos alkalmazási terület a peer-to-peer hálózatok és adat gridek világa. Ebben az esetben a klaszterezés tipikusan vagy terheléselosztási, vagy adatreplikációs célt szolgál.

A másik jelentős alkalmazási terület a szolgáltatás ökoszisztémák, például a szolgáltatásra igényt tartó csomópont inverz klaszterezéssel kereshet egy közeli szolgáltatót, vagy a szolgáltatást nyúltó csomópontok terheléseloszlási céllal normál klaszterezést is végezhetnek.

III. Vox Populi: Névelem alapú szentiment bányászat természetes nyelvi szövegekből

III.1. Háttér és motiváció

A közösségi média előretörésével az, amit mások egy-egy témáról gondolnak vagy főleg írnak, már nem csak egy érdekes járulékos információ, hanem egyre inkább az üzleti illetve személyes döntések meghatározó eleme. A szentiment elemzés vagy más néven *vélemény bányászat* számottevő figyelmelmet kapott az utóbbi időben főleg a hírek, termék véleményező oldalak, blogok és a közösségi háló bejegyzések terén.

A vélemény bányászat a szövegben kifejezett szemantikus irány meghatározásával foglalkozik. Ezt akár egy speciális dokumentumkategorizációs feladatnak is felfolghatjuk, ahol a meghatározandó kategória a dokumentumban jelen levő szentiment, nem pedig a szöveg témája. Néhány fontos alfeladat: polaritás elemzés az

érzelmi irányultság irányának meghatározására (pozitív, negatív, semleges), erősség vizsgálat a kifejezett vélemény intenzitásának leírására, valamint szubjektivitási vizsgálat az objektív és szubjektív faktor közötti különbségtételre. Munkám során az alfeladatok közül az első kettővel foglalkoztam.

Szintement elemzéshez számos eszköz és modell ismert a szakirodalomból az egyszerű szótár alapú megközelítéstől a komplex nyelvtani elemzést és gépi tanulást is használó módszerekig. A *szótár alapú* modell, melyet gyakran neveznek szóhalmaz modellnek is, egy olyan listát használ, melyben minden érzelmet kifejező szó mellett a szó által kifejezett polaritás és intenzitás érték szerepel; a szöveg érzelmi összértéke pedig a benne szereplő szavak szentiment értékének egyszerű összegeként áll elő. A szótár lehet manuális vagy automatikusan előállított. Automatikus szótárkészítésre egy példa a SentiWordNet kezdeményezés, mely a WordNet nevű szemnatikus szóháló felhasználásával néhány egyértelműen erős polaritású szóból kiindulva következtette ki az érzelmet hordozó szavak listáját és szentiment értékét. A szakirodalomban ismertetett bonyolultabb modellek a szótárat nyelvtani analízissel, témastruktúra analízissel és gépi tanulással is kombinálják.

A véleménybányászati feladat több különböző szinten is elképzelhető, például a teljes szöveg szintjén, bekezdés szinten, mondatonként.

Kutatásom fókusza némiképp eltér a sztenderd véleménybányászati feladattól, ahol az elemzés a teljes dokumentumra vagy annak egy strukturális elemére (pl. bekezdésre) vonatkozik. Ehelyett én egy-egy a szövegben előforduló *névelemekre vonatkozólag* vizsgálom a szentiment információt, tehát azt elemzem, hogy a szöveg milyen véleményt fogalmaz meg egy adott termékekre, cégre, személyre vagy épp földrajzi helyre vonatkozóan. Az érzelmi információt névelem előfordulásonként nyerem ki, majd névelemenként az egész szövegre vonatkozóan aggregálom.

A szakirodalomban előforduló modellek az emberi nyelvek egyfajta statikusságát tételezik fel, azaz nem számolnak azzal a lehetőséggel, hogy egyes szavak érzelmi töltete az idő előrehaldtával változhat.

Az *adaptivitás* ezzel szemben az általam javasolt modell központi eleme. Egy olyan automatikus, felügyelet nélküli szótár ön-elemző módszert javaslok, melynek segítségével a szótárban fennálló dinamikus hibák felderíthetők, illetve akár ki is javíthatók. Az algoritmus felügyelet nélküli olyan értelemben, hogy nem igényel előzetesen annotált referencia korpuszt. A módszer lehetőséget ad arra, hogy a szótár az emberi nyelvhasználat–főleg az informális szókészlet, szleng–változásait lekövesse. A másik fő előny, hogy a tisztítási képesség révén egy kissé szennyezett–például gépi fordítással előállított–szótárból kiindulva is készíthetünk tiszta adathalmazt.

A munka újdonságértéke egyrészt az újszerűen megfogalmazott feladat, azaz hogy a bányászat a névelemekre vonatkozik, illetve a felügyelet nélküli szótár ön-revíziós algoritmuson alapuló nagyfokú rugalmasság és adaptivitás.

III.2. Célkitűzések

A kutatás célutűzése annak vizsgálata volt, hogy milyen módszerekkel érhető el és támogatható a névelem alapú szentiment bányászat különböző természetes nyelvű szövegekre vonatkozóan. A központo kérdések a következők voltak:

- Milyen architektúra szükséges a névelem alapú vélemény bányászathoz?
- Hogyan lehet a meglevő algoritmusokat kiterjeszteni vagy módosítani ahhoz, hogy a névelemre vonatkozó érzelmi töltetet azonosítsák ahelyett, hogy a teljes szövegrészre vonatkozóan hoznának ítéletet?
- Lehetséges-e a rendszert ön-tudatossá, adaptivvá tenni, azaz felvértezni felügyelet nélküli tanuláson alapuló ön-revíziós és ön-elemző képességgel?
- Lehetséges-e a szentiment szótár helyességét annotált referncia korpusz nélkül is megbecsülni? Lehetséges-e hasonló felügyelet nélküli tanuláson alapuló algoritmusokat a rendszer más egységeiben is alkalmazni?

III.3. Új tudományos eredmények

III-1. tézis *(Vox Populi) Modellt javasoltam az adaptív, névelem alapú szentiment bányászati feladatra. Definiáltam a rendszer alapelemeit és az adatutakat. Kísérletekkel igazoltam, hogy a modell alkalmazható magyar nyelvű szövegekből való szentiment bányászatra.*

Kapcsolódó publikáció: J10u

III-2. tézis *(Névelem központú szentiment bányászati algoritmus) Algoritmust javasoltam a szövegben előforduló névelemekhez kapcsolódó szentiment értékek meghatározására. Az algoritmus az alapvető nyelvtani elemzést (szótövesítés, negáció) kombinálja egy kiterjesztett szóhalmaz modellel, mely a névelem körüli környeztetre (súlyozott kontextuskeretek) épít. Módszert javasoltam az algoritmus futási paramétereinek, például a kontextuskeretek hosszának, dinamikus finomhangolására. Annotált referencia korpuszon végzett mérésekkel megmutattam, hogy az algoritmus 68-89%-os pontosság és 67-87% visszaadás értéket nyújt.*

Kapcsolódó publikáció: J10u

III-3. tézis *(Szótár önrevíziós algoritmus) Algoritmust javasoltam a szótár referencia korpusz nélküli önellenőrzésére. Fordmális definíciót adtam az algoritmus lépéseihez és a kapcsolódó előfeltevésekhez. Bebizonyítottam, hogy ha az előfeltevések fennálnak, akkor az algoritmus annotált referencia korpusz használata nélkül is azonosítja a szótában megbúvó hibákat. Adatstruktúrát (elvárás-megfigyelés mátrix) javasoltam az algoritmus eredményeinek reprezentálására.*

Kapcsolódó publikációk: J10u

III-4. tézis *(Ponttosság és visszaadás becslés bizonyíték alapú szótárkiterjesztésre) Algoritmust javasoltam a szótár bizonyíték alapú kiterjesztésének pontosság és visszaadás becslésére. Mérésekkel megmutattam, hogy a szótár bizonyíték alapú kiterjesztése a Vox Populi modellel magyar nyelvű szövegekre csak gyenge eredményt (43-44% pontosság és 43-44% visszaadást) hozna, ezért nem javasolt.*

Kapcsolódó publikációk: J10u

III.4. Eredmények alkalmazása

A Vox Populi modell és a névelem alapú vélemény bányászat az üzleti döntéshozók, illetve akár az átlagember számára is közvetlen használható eredményt nyújt. A modell elhetővé teszi, hogy egy termékhez vagy más névelemhez kapcsolódó véleményeket akár különböző forrásokat intergrálva és/vagy időtengely mentén összegezve jelenítsük meg. A számos felhasználási lehetőség közül néhány példa a használatra: (i) Jelenlegi helyzet lekérdezése. Mi a termékről kialakult általános vélemény? (ii) Idősor analízis. Hogyan változott a termék megítélése az idők során? (iii) Forrás szerinti vizsgálat. Hogyan értékelik a terméket a vezető hírekben és a blogokon, személyes bejegyzésekben? (iv) Páronkénti összevetés. A két márka/termék közül melyiknek jobb a megítélése? Melyikkel kapcsolatban változott jobban az emberek véleménye az elmúlt időszakban?

A Vox Populi rendszer egy kísérleti verziója a Magyar Telekom Nyrt-nél 2010 óta próbaüzemben működik.

Publikációs lista

Könyvfejezetek

B1 Borbála Katalin Benkő
Increasing the syntactical parse efficiency using "strong rules".
In: Eloína Miyares Bermúdez, Leonel Ruiz Miyares (eds)
Linguistics in the Twenty First Century.
Cambridge: Cambridge Scholars Publishing, 2006. pp. 289-298.
(ISBN: 1904303862)

B2 Tamás Katona, Borbála Katalin Benkő
Information retrieval in homogeneous document sets using syntactical parse information.
In: Eloína Miyares Bermúdez, Leonel Ruiz Miyares (eds)
Linguistics in the Twenty First Century.
Cambridge: Cambridge Scholars Publishing, 2006. pp. 307-316.
(ISBN: 1904303862)

B3 Endre Sándor Varga, Bernát Wiandt, Borbála Katalin Benkő, Vilmos Simon
Autonomously Evolving Communication Protocols: The Case of Multi-Hop Broadcast.
In: Dinesh Verma, Pietro Lio' (eds)
Biologically Inspired Networking and Sensing: Algorithms and Architectures.
New York: IGI Global, Information Science Reference, 2011.
(In Press)

Folyóiratcikkek

J1 Borbála Katalin Benkő, Tamás Katona, Róbert Schulcz
CASCADAS – Autonomic communication and pervasive situated services.
HÍRADÁSTECHNIKA LXII:(1) pp. 41-46. (2007)

J2 Borbala Katalin Benkő, Vilmos Simon
Nature-Inspired Adaptivity in Communication and Learning.
LECTURE NOTES IN COMPUTER SCIENCE 6209: pp. 323-325. Paper 10.1007/978-3-642-14547-6_31. (2010)

J3 Borbála Katalin Benkő, Dávid Lányi
An Open Learning System.
PERIODICA POLYTECHNICA-ELECTRICAL ENGINEERING. (2011)
(In Press)

J4 Márton Legény, Borbála Katalin Benkő
Topology Cache Aided Self-organization Directions for Ad-hoc Mobile Networks.
CYBER JOURNALS-JOURNAL OF SELECTED AREAS IN TELECOMMUNICATIONS 2011:(3) pp. 56-63. Paper 08. (2011)

J5 Borbala Katalin Benkő, Marton Legény
Improved On Demand Clustering on Scale-Free Topologies
Infocommunications Journal (2011)
(In Press)

J6 Laszlo Muka, Borbala Katalin Benkő

Meta-level performance management of simulation: The problem context retrieval approach
PERIODICA POLYTECHNICA-ELECTRICAL ENGINEERING. (2011)
(In Press)

J7u Borbala Katalin Benkő, Nermin Brgulja, Antonio DiFerdinando, Edzard Höfig, Rico Kusber, Marco Mamei
Autonomic Communication Elements: Design,Evaluation and Application
Under review for ACM Transactions on Autonomous and Adaptive Systems
Submitted: December 2008
Revision submitted: January 2011

J8u Borbala Katalin Benkő, David Lányi, Karoly Farkas
OLAKO: An Approach for an Open Collective-Adaptive System
Under review for ACM Transactions on Autonomous and Adaptive Systems
Submitted: Marc 04 2011

J9u Endre Sándor Varga, Bernát Wiandt, Borbála Katalin Benkő, Vilmos Simon
Autonomous Online Evolution of Communication Protocols
Under review for: Elsevier Journals, Journal of Systems and Software, Special Issue on State-of-the-Art in Self-Adaptive Systems
Submitted: April 2011

J10u Borbala Katalin Benkő, Zsolt Kardkovacs
Sentiment Mining for Named Entities from Dynamic Web Contents with Self-Supervision
Under review for IEEE Transactions on Knowledge and Data Engineering
Submitted: May 2011

Konferenciacikkek

C1 Borbála Katalin Benkő
On increasing the syntactical parse efficienc.
In: Proceedings of UkrObraz2004. Kyiv, Ukrajna, 2004-2004. Kyiv:

C2 Borbála Katalin Benkő
Increasing the syntactical parse efficiency using "strong rules".
In: Proceedings of the 9th International Symposium for Social Communication. Cuba, 2005.01-2005.01. Santiago de Cuba: pp. 562-566.(ISBN: 959-7174-04-9)

C3 Borbála Katalin Benkő, Tamás Katona
On the Efficient Indexing of Grammatical Parse Trees for Information Retrieval.
In: Proceedings of INISTA2005. Istanbul, Turkey, 2005-2005. Istanbul: pp. 366-369.

C4 Tamás Katona, Borbála Katalin Benkő
Information retrieval in homogeneous document sets using syntactical parse information.
In: Proceedings of the 9th International Symposium for Social Communication. Cuba, 2005.01-2005.01. Santiago de Cuba (ISBN: 959-7174-04-9)

C5 Edzard Höfig, Björn Wüst, Borbála Katalin Benkő, Antonietta Mannella, Marco Mamei, Elizabetta Di Nitto
On Concepts for Autonomic Communication Elements.
In: Proceedings of IEEE International Workshop on Modelling Autonomic Communications Environments. Dublin, Ireland, 2006 Dublin: pp. 1-5.

C6 Borbala Katalin Benkő, Robert Schulcz
Mapping external functionalities into autonomic services.
In: Bio-Inspired Models of Network, Information and Computing Systems, 2007. Bionetics 2007. 2nd. Budapest, Hungary, 2007.12.10-2007.12.12. Budapest: pp. 270-276. (ISBN: 978-963-9799-05-9)
DOI: 10.1109/BIMNICS.2007.4610125

C7 Borbála Katalin Benkő, Nermin Brgulja, Edzard Höfig, Rico Kusber
Adaptive Services in a Distributed Environment.
In: Proceedings of the Eighth International Workshop on Applications and Services in Wireless Networks. Kassel, Germany, 2008-2008. Washington: IEEE Computer Society Press, pp. 66-75. (ISBN: 978-0-7695-3389-6)

C8 Márton Legény, Borbála Katalin Benkő
Design of Novel Self-Organization Mechanisms Through Simulation.
In: ISC'2010. Budapest, Magyarország, 2010.06.07-2010.06.09. pp. 10-14. (ISBN: 978-90-77381-5-57)

C9 Dávid Lányi, Borbála Katalin Benkő
A Nontraditional Approach for a Highly Interactive Collective-Adaptive System.
In: NexTech 2010. Firenze, Italy, 2010.10.25-2010.10.30. Firenze: pp. 1-6. Paper 19. (ISBN: 978-1-61208-000-0)

C10 Peter H. Deussen, Edzard Höfig, Borbála Katalin Benkő
Body Aura - A New Approach Towards Ambient Intelligence: Vision Paper.
In: NexTech 2010. Firenze, Italy, 2010.10.25-2010.10.25. Firenze: pp. 1-6. Paper 23. (ISBN: 978-1-61208-000-0)

Könyvefejezetek magyar nyelven

B4 Benkő Borbála Katalin
Futballnyelv az interneten.
In: Balázs Géza, Grétsy László (ed)
Sportnyelvünk a 21. század elején.
Nemzeti Kulturális Örökség Minisztériuma Budapest:2003. pp. 55-66.
(ISBN: 963-210-122-7)

Folyóiratcikkek magyar nyelven

J11 Borbála Katalin Benkő, Zsolt Pándi
Dokumentumkategorizálás távközlési alkalmazásban.
HÍRADÁSTECHNIKA LX:(11) pp. 45-49. (2005)

J12 Borbála Katalin Benkő, Tamás Katona, Róbert Schulcz
CASCADAS – Autonóm kommunikáció és pervazív, helyzetfüggô szolgáltatások.
HÍRADÁSTECHNIKA LXI:(12) pp. 23-28. (2006)