



M Ű E G Y E T E M 1 7 8 2

Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Mészáros Tamás

Ágensalapú információbeszerzés és -megosztás

Ph.D. értekezés tézisei

Témavezető

Dr. habil. Dobrowiecki P. Tadeusz

egyetemi docens

Budapest, 2015. május

1 Bevezetés

Világunkban egyre nagyobb szerephez jut az információ: munkánk és szabadidős tevékenységeink során korábban sosem látott mennyiségben áll rendelkezésünkre, beszerzésére és megosztására egyre többféle lehetőségünk van. Az információk elmúlt évtizedekben elterjedt elektronikus elérési és megosztási módszerei új dimenziókat nyitottak tudásunk bővítésére, eredményeink kidolgozására és közzétételére.

Figyelembe véve az elektronikusan elérhető adatok hatalmas mennyiségét az igényelt információk beszerzése elképzelhetetlen megfelelő informatikai támogatás nélkül. Az adatok felkutatásának, beszerzésének, illetve értelmezésének támogatása számos egymástól igen különbözően felépülő és működő informatikai rendszer célkitűzése. A mindenki által használt „egyszerű” webböngészőktől kezdve a felhasználókat támogató információbeszerző rendszereken, a webes és a telefonokba integrált (hangalapú) keresőkön át az ipari adattisztító, „*big data*” adatelemző és az internetes forrásokból származó információkat az üzleti folyamatokba integráló rendszerekig találunk példákat. Ezek számos különböző módszer alkalmazásával igyekeznek az információbeszerzés és -elemzés tevékenységének minél nagyobb részét átvállalni a felhasználójuktól.

Ezek az informatikai rendszerek a társadalom széles rétegével állnak kapcsolatban a gyerekektől kezdve az idősekig, így alkalmazásuk sikere nemcsak feladatuk megoldásán, hanem felhasználói felületük megvalósítását is múlik. Ahogy egyre több funkciót biztosítanak számunkra ezek a rendszerek, és egyre kisebb eszközökön képesek működni, úgy válik egyre nehezebbé e feladat megoldása. A felhasználói felületük kialakítása sok esetben nehezebb, kritikusabb probléma, mint funkcióik megvalósítása.

Az információbeszerző és -szolgáltató rendszerek sikere megfelelő tárgyterületi és rendszerkezelői tudással rendelkező emberi szakértők bevonása nélkül nagyon sok területen elképzelhetetlen. Mindez alapvetően korlátozza a rendszerek teljesítményét és növeli alkalmazásuk költségeit. Számos kutatás és gyakorlati rendszer tűzte ki célul az emberi intelligencia átemelését a számítógépes információbeszerzés folyamatába, például az adatelemzés, a szövegbányászat, a természetes nyelvfeldolgozás, a szakértői rendszerek és a felhasználó-modellezés módszereinek alkalmazásával. Az ilyen **tudásintenzív információbeszerző és -szolgáltató rendszerek** képezik kutatási tevékenységem tárgyát.

2 Célkitűzések

Kutatásaim során alapvetően azt vizsgáltam, hogy a **mesterséges intelligencia** különböző módszerei hogyan alkalmazhatók az információbeszerzés és -megosztás területén. Alapvető hipotézisem az volt, hogy az **intelligens ágens** megközelítésmód megfelelő alapot teremt egy olyan általános modell megalkotására, amely képes integrálni a területen alkalmazott igen sokféle módszert a statisztikai elemzésektől („big data”) kezdve a tartalom gépi megértését célzó megoldásokig („right data”), valamint rugalmas, robusztus és bővíthető architektúrát kínál a gyakorlati megvalósításukhoz is.

Külön hangsúlyt fektettem az információs igény minél pontosabb megfogalmazására. Azt kívántam megvizsgálni, hogy a felhasználók információs igénye hogyan fogalmazható meg úgy természetes nyelven, hogy az minél pontosabb leírása legyen a feladatnak, ugyanakkor számítógépes rendszerekkel elemezhető és értelmezhető legyen.

Munkám harmadik pillérét a strukturált dokumentumkezelés jelentette, amelynek módszereit és eszközeit a tartomelemzés és -integráció megvalósítására kívántam felhasználni. Olyan megoldást kívántam kialakítani, amely egy egységes, ugyanakkor rugalmas, bővíthető tartalomreprezentációt kínál, és illeszkedik az ágensalapú megvalósítás módszereihez és eszközeihez.

Célul tűztem ki a terület részletesebb megismerését, az ágensmodellek és -rendszerek alkalmazhatóságának vizsgálatát, majd ezekre építve egy ágensalapú információbeszerző és -megosztó rendszer elvi felépítésének és működésének kidolgozását. Emellett vizsgálni kívántam a természetes nyelvű interfészek kialakításának lehetőségeit, az ehhez szükséges elemzési és értelmezési módszereket, technológiákat és eszközöket. A tartalomintegráció területén az XML technológia különböző részeinek alkalmazhatóságát kutattam, mint a bővíthető deklarációs rendszerek, a strukturált tartalmak feldolgozása és transzformációja, valamint XML dokumentum táruk létrehozása és lekérdezése.

Kutatási célkitűzéseim mellett fontosnak tartottam az elért eredmények mérnöki gyakorlatba ültetését is, sok esetben gyakorlati problémák ösztönözték az új megoldások kialakítását. Munkámat egyaránt tekintem tudományos és mérnöki tevékenységnek: a felmerült problémák, a megoldásukra kidolgozott új módszerek, elméleti eredmények és az ezek alapján elkészített implementációk mindvégig szoros kapcsolatban álltak.

3 Az információbeszerzés és -szolgáltatás áttekintése

Az információbeszerzés célvezérelt feladat. A felhasználó megfogalmazza információs igényét egy adott témakörben, amelyre válaszul arra releváns tartalmakat vár. Az információbeszerző rendszer értelmezi az igényt, felkutatja a forrásokat, beszerzi és feldolgozza a releváns tartalmakat, majd továbbítja őket a felhasználó felé. A siker sok tényezőtől függ: a felhasználó elvárásainak minél pontosabb meghatározásától, a források felkutathatóságától és az azokból beszerzett tartalmak értelmezhetőségétől.

Az információs igény megfogalmazása, reprezentációja, felhasználása a különböző rendszerekben igen eltérő lehet. Például adatbázis-kezelők esetében egy SQL kifejezés, webes keresés esetén a hiányzó információ jellemzésére, körülírására használt kulcsszavak felsorolását. Bármilyen módszert is alkalmazunk, a felhasználótól nem várható el sem az alkalmazott technológiák sem az adatmodellek ismerete. Ennek érdekében a rendszerek igyekeznek az információs igény megfogalmazásának lehető legegyszerűbb módját biztosítani. Ez azonban jellemzően az igény megfogalmazási pontosságának rovására történik: egy szabad szöveges mezőben felsorolt kulcsszavakkal nehezen lehet pontosan meghatározni az elvárásokat.

A felhasználók által keresett információk nem mindig jól definiált szerkezetű és elérhető tartalomtárakban vannak, hanem alig strukturált, gépileg nehezen értelmezhető dokumentumokban találhatók szétszórva a világháló rendszereiben. Ezt a környezetet nagyfokú komplexitás, térbeli és időbeli elosztottság és változékonyság, valamint technológiai sokszínűség jellemzi. A konkrét alkalmazás, a forráskörnyezet tulajdonságai, a keresett adatról rendelkezésre álló információk, a hozzáférési, megbízhatósági és konzisztencia problémák miatt a feladat sikeres megoldására olyan rendszer lehet képes, amely egyszerre többféle módszert futásidőben adaptálva alkalmaz az igényelt információ beszerzésére.

A forrásokból jellemzően nem az információs igénynek megfelelő szerkezetű és formájú tartalmakat tudjuk beszerezni, hanem egy olyan adathalmazt, amelyből a kívánt információ előállítható lehet. Ennek megvalósítása az információkinyerési lépés feladata. Még jól ismert szerkezetű források esetében is szükség lehet feldolgozási lépések végrehajtására, ismeretlen források esetében jellemzően összetett elemzések végrehajtására van szükség az igényelt információ felismerésére és kinyerésére. Az információkinyerés egy összetett feladat, amelynek nincs egyetlen, eredményre vezető

módja. Számos megoldást dolgoztak ki az adat- és szövegbányászat, a természetes nyelvű szövegek gépi megértése, stb. terén. Sok sikeres alkalmazás esetében jellemző az is, hogy az adott terület jellemzőinek megfelelő, egyedi módszerek eredményezték a leghatékonyabb megoldást.

Az információbeszerzés lépései mellett külön megemlítendő a felhasználók szerepe a beszerzés folyamatában, amely alapvetően meghatározza a rendszerek teljesítőképességét. A legszélesebb körben használt webes keresés a felhasználó intenzív közreműködését igényli: a keresési igény megfelelő kulcsszavakra bontása, a visszakapott találati lista értékelése és a keresési minta esetleges módosítása, a weboldalak letöltése és értelmezése, végül a kívánt információk kinyerése. Még egy ilyen egyszerűnek tűnő alkalmazás is szakértelmet kíván: hogyan válasszuk ki a legmegfelelőbb kulcsszavakat, hogyan találjuk meg a legjobb forrásokat, értelmezzük azok tartalmát, stb.. Egy vállalati környezetben működő rendszer esetében az üzemeltetés tárgyterületi és rendszerkezelői tudást is megkíván. Egy hatékony, nagy adatmennyiséget feldolgozni képes rendszer azonban nem képzelhető el humán szakértők folyamatos közreműködésével, az információelérési tudás minél nagyobb részét kell a rendszerbe integrálni, szükséges a minél inkább autonóm megvalósítás.

A web talán a legmeghatározóbb terület, ahol az információbeszerzés előbbieken részletezett problémáival szembesülünk. A web alaptermotechnológiáit fejlesztő és szabványait gondozó szervezetet (World Wide Web Consortium, W3C) a szemantikus web koncepció megalkotásával alapjaiban kívánja megreformálni a tartalomtárolást és -elérést: a HTML megjelenésorientált tárolási módszeréről az XML szemantikus jelöléseket alkalmazó rendszerére kíván áttérni a gépi feldolgozás megkönnyítése érdekében. Bár ez egy nagyon fontos mérföldkő a web történetében, a megoldás elterjedése még várat magára.

A szemantikus web technológiái ugyanakkor már kialakultak és elterjedtek, széles körben elérhetők az információfeldolgozó rendszerekben. Az XML adatformátum alkalmazási és implementációs környezettől független, önleíró adatformátum. A hozzá kapcsolódó feldolgozási eljárások és programozói eszközök ma már minden platformon használhatók. Az adattároló rendszerek (adatbázisok és dokumentum táruk) is kiegészültek a strukturált dokumentumok tárolására és lekérdezésére alkalmas módszerekkel és eszközökkel.

4 Elért eredmények

4.1 A megvalósított ágensalapú rendszer felépítése

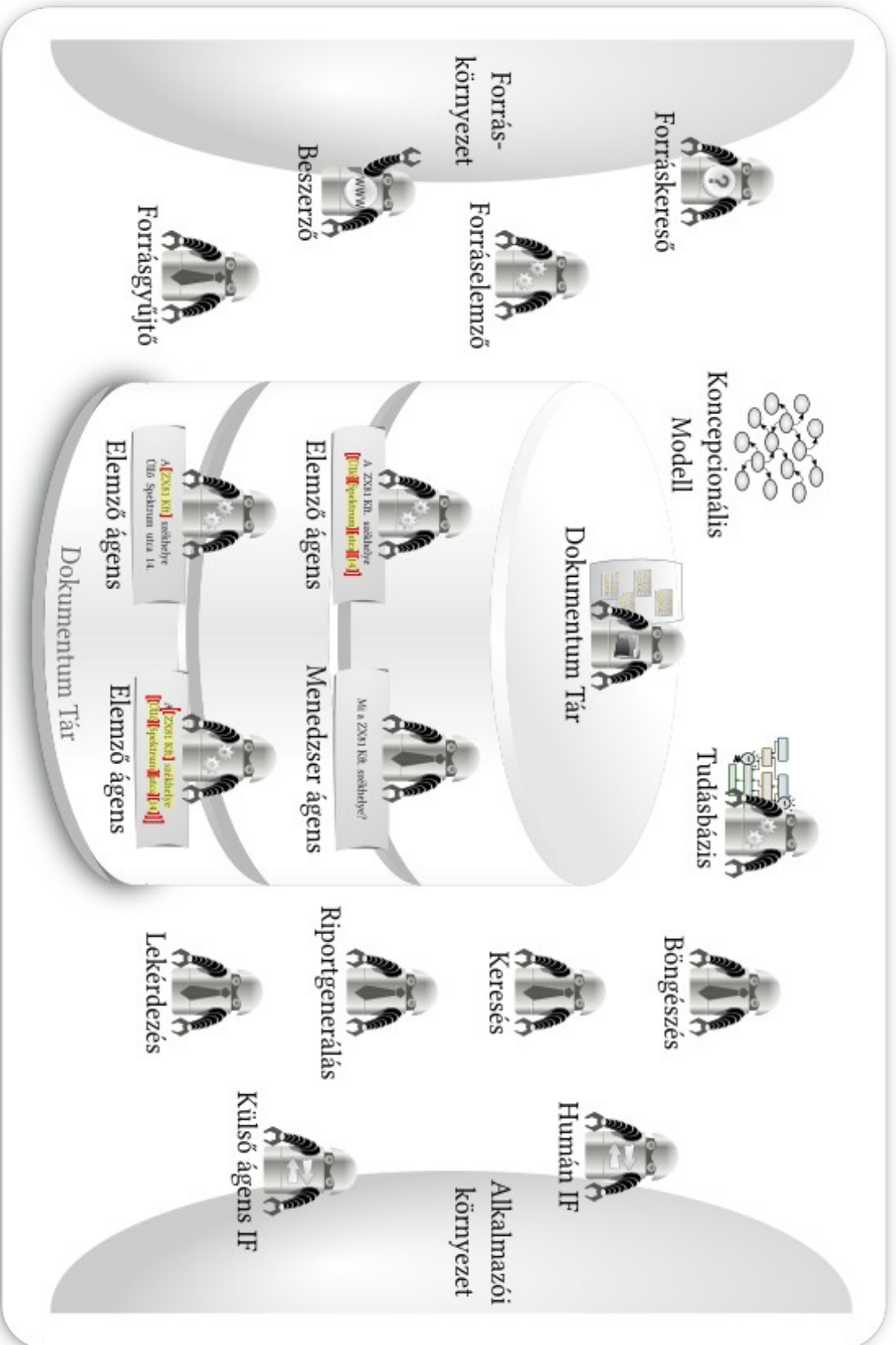
Az információbeszerzés és -szolgáltatás feladatának megoldására egy összetett, ágensalapú rendszert hoztam létre, amely három fő alrendszerből áll:

- Az **Információbeszerző Alrendszer (IBA)** felkutatja az igényelt információk forrásait, majd beszerzi és elemzi a forrástartalmakat.
- A **Tudásfuzionáló Alrendszer (TFA)** egy strukturált dokumentum tárban helyezi el a beszerzett tartalmakat, amelyeket egymással és tárgyterületi tudásbázisokkal egy közös koncepcionális modellre alapozva képes integrálni.
- Az **Információlekérdező Alrendszer (ILA)** interfész fogadja és elemzi a felhasználók információs igényeit, majd előállítja a keresett információkat a TFA információira és szolgáltatásaira támaszkodva.

A három alrendszerben az intelligens ágensek sokféle szerepet töltenek be (1. ábra):

- **Menedzser ágensek** irányítják a rendszer különböző feladatainak végrehajtását, úgymint a források felkutatását, az elemzőrendszer működését, és a lekérdező, böngésző, kereső és riportgeneráló funkciók megvalósítását.
- **Vállalkozó ágensek** jelentkezhetnek és megbízást kaphatnak részfeladatok megoldására, például források felkutatására és tartalmak beszerzésére, elemzésre, a felhasználói kontextus modellezésére, stb..
- **Interfész ágensek** teremtenek kapcsolatot külső rendszerekhez, felhasználókhoz és a rendszerbe integrált tudásalapú komponensekhez is.
- **Tudásalapú ágensek** leképezik a tárgyterület egyes aspektusait a rendszeren belülre, a beszerzett tartalmakkal bővíthetik, frissíthetik tudásbázisukat.

A rendszer a forráskörnyezetből beszerzett tartalmakat egy speciális, alkalmazási igények szerint bővíthető **XML formátumban** reprezentálja. Ez biztosítja a forrástartalmak különféle elemzéseinek integrációját, támogatja az alkalmazói környezet által elvárt kimeneti formátumok előállítását, és biztosítja az ágens kommunikációs protokoll tartalmi szintjének adatformátumát is. A **Koncepcionális Modell** feladata a tartalommodellek, a kommunikáció és a rendszerkomponensek közös fogalmi alapjának biztosítása. Az információs igények egyszerű, de pontos leírására, valamint humán szakértők integrálására **kontrollált természetes nyelvű interfészek**, valamint a mögöttük meghúzódó elemző és értelmező komponensek szolgálnak.



1. ábra: ágensek az információbeszerző és -szolgáltató rendszerben

4.2 A megfogalmazott tézisek

Az alábbiakban ismertetem téziseimet és az azokat alátámasztó hivatkozásokat.

1. Téziscsoport: ágensalapú architektúra

Kutatásaim kiindulópontja az ágensalapú információbeszerzés vizsgálata volt. Alapvető eredményeit az F030 763 OTKA projekt keretében értem el [1].

1.1 Megállapítottam, hogy az intelligens ágens megközelítésmód több szempontból is alkalmas információbeszerző rendszerek felépítésének és működésének leírására és megvalósítására. Az intelligens ágensek tulajdonságai (környezetbe ágyazottság, autonómia, reaktív, proaktív, konkurens, célvezérelt működés), valamint a több-ágens rendszerek kommunikációs és kooperációs megoldásai jól illeszkednek az ilyen rendszerek működéséhez és a velük szemben támasztott követelményekhez. [1], [2]

1.2 Kidolgoztam egy több-ágens architektúrát az információforrások felkutatására. A vállalkozó hálók protokoll alkalmazásával egy menedzser ágens fogadja az információs igény leírását, majd forráskereső, tartalombeszerző és -elemző vállalkozó ágenseket bíz meg a dekomponált részfeladatok végrehajtásával. [2], [3], [4], [5]

1.3 Kidolgoztam egy több-ágens architektúrát és működési algoritmust a tartalomelemzés 2.2 tézisben megfogalmazott elvi működésének megvalósítására. Meghatároztam az elemző ágensek cselekvéseit és a több-ágens rendszer működését. Értékelő ágensek bevezetésével kooperációs sémákat alakítottam ki az elemzésekkel és forrástartalmakkal kapcsolatos futásidejű problémák kezelésére. A célvezérelt működést vállalkozási hálók protokoll alkalmazásával valósítottam meg. [1], [2], [6], [7]

1.4 Kialakítottam egy megoldást a tudásalapú komponensek, a Dokumentum Tár és a Koncepcionális Modell integrációjára. A fuzionált tudástár az architekturális integrációt interfész ágensek alkalmazásával valósítja meg. Az egységes ágens kommunikációs protokoll mellett tartalmi szinten a közös koncepcionális modellre és az XML tartalom formátumra épít. Részletesen a Bayes-hálók integrációját ismertettem. [8]

1.5 Kidolgoztam az Információlekérdező Alrendszer ágensalapú megvalósítását, amelynek kereső, lekérdező, böngésző és riportgeneráló funkcióit menedzser ágensek valósítják meg a fuzionált tudástár ágenseire támaszkodva, míg az alkalmazói környezet rendszereivel és felhasználóival interfész ágensek kötik össze és kontextus modellező ágensek segítenek a külső felkérések pontosabb értelmezésében. [8], [9], [10]

2. Téziscsoport: elemzőrendszer és strukturált tartalom reprezentáció

Kutatásaim másik fő területe egy olyan univerzális elemzőrendszer és dokumentum formátum kialakítását célozta, amely képes kiszolgálni a rendszer és az alkalmazók igényeit. Ezeket a kutatásaimat alapvetően az IKF projekt keretében végeztem. [2], [5]

2.1 Kidolgoztam egy XML-alapú formátumot, amely egy dokumentumon belül képes a forrástartalmak és azok különbözőképpen elemzett (különböző struktúrájú) változatainak tárolására. Az implementációs eszközöktől független formátum két részből áll: az elemzett tartalomból (amely az alkalmazás igényei szerint tetszőleges módon kialakítható), valamint a beszerzési és az elemzési metaadatokból. [2], [3], [5], [11]

2.2 A dokumentum formátumra alapozva kialakítottam a tartalomelemző-rendszer elvi működését, amelyet elemzőkből felépülő modelltranszformátor-hálózatként határoztam meg. Ez a megoldás dekomponálja a feladatot az ágensalapú megvalósítás számára, valamint lehetőséget teremt az elemzés formális leírására, elemzési hálózatok tervezésére és verifikálására is. XML eszközökre támaszkodva kidolgoztam az elemzők működési algoritmusát, amely meghatározza a releváns bemeneti tartalomelemeket, előállítja az elemzés eredményét, majd azt a kívánt formába önti. [5], [6], [7], [11], [12], [13]

2.3 Kidolgoztam egy módszert az elemzések által előállított tartalomváltozatok összetartozó részeinek (attribútumok, elemek és azokon belüli szövegrészletek) összekapcsolására. A kidolgozott megoldás egy XLink hivatkozáshálózat alkalmazásával írja le a strukturált dokumentumrészek között kapcsolatokat. Az így kialakított nézethálózat alapjául szolgálhat magyarázatgeneráló funkciók megvalósításának is. [13]

3. Téziscsoport: Kontrollált természetes nyelvű felhasználói felület

Munkám harmadik nagyobb része a felhasználói felületek kialakításához kapcsolódik. Ezen a területen kontrollált természetes nyelvű interfészek kialakításának módszereit dolgoztam ki, illetve valósítottam meg több gyakorlati alkalmazásban.

3.1 Kidolgoztam egy kontrollált természetes nyelvű megoldást az alkalmazói környezet humán felhasználói által megfogalmazott információs igények leírására. Segítségével lehetővé válik egy könnyen használható, rugalmas, és nagy kifejezőerejű felhasználói felület kialakítása. A nyelvtani szabályok, a szókincs és a szemantika rögzítésével valósítottam meg a természetes nyelvű mondatok feldolgozását és átalakítását (lefordítását) az információs igények belső reprezentációjára. [10], [14]

3.2 Meghatároztam az elégséges kifejezőerejű nyelv megalkotásának rendszeren belüli forrásait. Kialakítottam egy eljárást környezetfüggetlen kontrollált nyelvek modellalapú elkészítésére, amely a tárgykörnyezet koncepcionális modelljének nyelvi réteggel történő kiegészítésére épít. [14], [15]

3.3 Kidolgoztam a kontrollált nyelvű felhasználói felület működési módját. Bemutattam, hogy javaslattevő funkció megvalósításával hogyan alakítható ki egy prediktív szövegbeviteli rendszer, amely segíti a felhasználókat a nyelv korlátos szabályainak és szókincsének alkalmazásában. Kidolgoztam egy nyelvi elemzésre és generálásra is alkalmas reprezentációt, és részletesen bemutattam az elemzés és javaslattétel algoritmusának működését környezetfüggetlen nyelvtanokra. [14], [15]

5 Gyakorlati eredmények

5.1 OTKA F030763: ágensek és információkeresés (témavezető)

Az OTKA kutatásom alapvető célkitűzése annak vizsgálata és feltárása volt, hogy az intelligens ágensek hogyan alkalmazhatóak elosztott információs rendszerekben az információkeresés és -szolgáltatás, valamint az ember-gép kapcsolat területein.

Az általam összeállított kutatási terv kiterjedt az ágens módszertanok és az ágensalapú szoftver architektúrák vizsgálatára, az ágensek alkalmazásának lehetséges módjaira, valamint a felhasználó modellezés területére. A projekt során egy demonstrációs célú prototípus rendszert is kifejlesztettem („DocMan”), amely dokumentumok tárolására, keresésére és lekérdezésére alkalmas, Letöltő, Analízis és Kereső Ágenseket tartalmazott, és alapvetően különféle statisztikai módszereket alkalmazott a beszerzett dokumentumok elemzésére. A rendszer különféle szótárak formájában integrálta a tárgyterületi tudást is, valamint itt jelent meg a természetes nyelvű felhasználói felület első, kezdetleges változata is (a keresőfelület egyszerű kiegészítésén keresztül).

A rendszert integráltuk egy Bayes-hálón alapuló modellező eszközzel (SEBANN) is, amely a dokumentum tárolás és a tudásalapú komponensek összekapcsolásának első változatát jelentette a már bemutatott fogalmi annotációk segítségével.

5.2 IKF: Információ és Tudás Tárház projekt (résztevő)

Az IKF (Information and Knowledge Fusion – Információ és Tudás Tárház) projektben az információbeszerzés és szolgáltatás területének általános, elvi kérdéseit, felépítését és működését vizsgáltuk. Feladatom alapvetően a rendszer architektúrájának megtervezése, valamint a dokumentumbeszerzés és -elemzés kialakítása volt.

A pénzügyi szektor számára fejlesztettünk tudásalapú információkeresési megoldást. A rendszer különböző információ forrásokból (webrendszerek, külső és belső adattárházak, stb.) merít információt az alkalmazási környezet igényeinek megfelelően, majd ezt strukturált formában tárja a felhasználó elé. Az IKF projektben – kollégáimmal közösen – elért elméleti eredmények alapozták meg az architektúrával foglalkozó fejezet alapvetéseit: ott jelentek meg a tárgy, forrás és alkalmazói környezetek fogalmai, ezek főbb tulajdonságai és viszonyai, valamint egy információbeszerző és -szolgáltató rendszer felépítésének váza és alapvető folyamatai. A projektben elért eredményekre támaszkodva fejlesztettem tovább a DocMan rendszer ágensalapú felépítését és működését is.

Az IKF projektben végzett kutatásaim másik jelentős területe a dokumentumok elemzéséhez kapcsolódik. Ennek során a DocMan egyszerű, statisztikai elemzőit egy jóval összetettebb rendszerrel váltottam fel, amely sokféle elemzési módszer együttes alkalmazását tette lehetővé. A projektben jelent meg az egységes, XML-alapú dokumentum formátum, valamint a nézetek és azok hálózata, amelyeket alapvetően saját eredménynek tekintek, és továbbfejlesztett változatukat, valamint az ágens megközelítésmódon alapuló megvalósításukat tézisek formájában rögzítettem a dolgozatomban.

Az elemzőrendszer sokféle prototípus implementációja készült el az IKF projekthez kapcsolódó fejlesztések keretében. Ezek közül kiemelkedik az egyik hallgatóm által készített Doctopus rendszer [6], amely a koncepció teljes implementációját megvalósította. A Doctopus egy moduláris elemzőrendszer és egy hozzá kapcsolódó tervekészítő és futtatórendszer, alapelveiben és tartalomreprezentációjában hasonlít az általam kidolgozott elemzőrendszerre, eltérő – nem ágensalapú – működési móddal rendelkezik. Disszertációmban továbbfejlesztettem az elemzőrendszer működéséhez szükséges XML tartalom modelleket, különösen azok nézethálózatának részleteit, emellett a koncepcionális modellezéssel megteremttem a tartalom modellek egységes fogalomrendszerét.

A harmadik terület, ahol az IKF projekt és doktori munkám kapcsolódik a dokumentum beszerzés működése. Az IKF projektben az általam kidolgozott tervek és egy korábbi implementációm (DocuHunter) alapján egy új keresőrendszer készült el (Spyder, [3]). Ennek terveit és működését továbbgondolva alakítottam ki doktori dolgozatomban a vállalalkozási hálók protokoll szerint működő több-ágens rendszert.

5.3 Főnévlista: a főnévi vonzattár webes felülete (témavezető)

A projekt során a Magyar Főnévi Vonzattár adatbázisához készítettem egy webes felületet, amely az adatok lekérdezésére és különböző elemzések futtatására szolgált [16]. Az adatbázis körülbelül 30 000 magyar főnév vonzatosságáról tartalmaz adatokat. Az egyes főneveknél mintegy harmincféle vonzat adatát, valamint több vonzat esetén ezek kapcsolódásait, összefüggéseit is tárolja. A felhasználói felület készítése során a nagyméretű, komplex adattár lekérdezésének bonyolultságából fakadó problémákkal szembesültem.

A Főnévlista webes rendszerében a klasszikus, űrlalapú lekérdező felület mellett – a felmerült felhasználói problémák kezelésére – egy természetes nyelvű felhasználói interfészt is megvalósítottam. Ennek tervezése során foglalkoztam a szintaktikai és szemantikai elemzés, valamint a fordítás problémáival, és megoldásként szűkített nyelvtani szabályok és szókincs alkalmazását választottam.

Ezen első implementáció után a kontrollált természetes nyelvű lekérdező és tudásbeviteli rendszert több más fejlesztés során is alkalmaztam. A természetes nyelvű webes interfészek kialakítására PHP nyelven létrehoztam az EasyLingua rendszeremet, Android környezetben pedig kisebb otthonautomatizálási feladatok megoldására használtam egy Java-alapú kontrollált nyelvű beviteli rendszert.

5.4 Kontrollált nyelvű önéletrajz-készítő

A kontrollált természetes nyelvű tudásbeviteli beviteli módszer alapjául szolgáló megoldást elsőként a Képesség és Kompetenciák Európai Szótárához készített önéletrajz-készítő alkalmazásában alakítottam ki. A szótár egy olyan többnyelvű teaurusz, amely Europass formátumú önéletrajzok készítéséhez használható fel [17].

Az Europass CV szerkezete kötött, egyes részein (pl. a szak- és nyelvtudás leírása) a mondszerkezetekre is léteznek ajánlások. Ezekre támaszkodva létrehoztam egy olyan alkalmazást (CVBuilder), amely az önéletrajzok bizonyos mezőinek kitöltését támogatta egy kontrollált természetes nyelvű szerkesztőfelületen. A nyelvtani szabályok kiterjeszhető szimbólumok segítségével hivatkoztak a teaurusz megfelelő elemeire .

5.5 Tudásbevitel kontrollált természetes nyelven (témavezető)

A tudásbevitel megvalósításának kérdéseivel a **kontrollált nyelvű absztraktok** (controlled language abstract, CLA) készítése kapcsán is foglalkoztam. Ezek természetes nyelvű írások olyan összefoglalói, amelyek formális gépi (például logikai) reprezentációra alakíthatók át.

Egy hallgatómmal közösen kialakított rendszer a Zotero webes publikációtárba integrál egy CLA szerkesztő felületet [18]. A rögzített szemantikus kivonatok, a Zotero által gyűjtött metaadatok és a szövegek statisztikai jellemzői egy tudástárba kerülnek. A szemantikus kivonatok a rendszer Prolog állításokra fordítja le, és egy tudásbázist épít belőlük.

6 A disszertáció témájához kapcsolódó publikációk

- [1] Mészáros Tamás, "Intelligens ágensek alkalmazása elosztott információs rendszerekben," OTKA szakmai zárójelentés OTKA-F030 763, 2002.
- [2] T. Mészáros, Z. Barczikay, F. Bodon, T. P. Dobrowiecki, and G. Strausz, "Building an Information and Knowledge Fusion System," in *Proc. of IEA/AIE-2001, Springer Lecture Notes in Computer Science*, 2001.
- [3] P. Varga, T. Mészáros, C. Dezsényi, and T. P. Dobrowiecki, "An Ontology-Based Information Retrieval System," in *Developments in Applied Artificial Intelligence*, Loughborough, UK, 2003, vol. 2718, pp. 217–228.
- [4] C. Dezsényi and T. Mészáros, "Domain Knowledge Based Document Retrieval," presented at the IEEE-TTTC-International Conference on Automation, Quality and Testing, Robotics, A&QT-R 2002 (THETA 13), Kolozsvár, Romania, 2002.
- [5] Dezsényi Cs., Varga P., Mészáros T., Strausz Gy., "Tudásalapú információkinyerés: az IKF projekt," vol. 51, no. 5, pp. 193–208, 2004.
- [6] C. Dezsényi, T. P. Dobrowiecki, and T. Meszaros, "Adaptive information extraction from unstructured documents," *International Journal of Intelligent Information and Database Systems*, vol. 1, no. 2, pp. 156–180, 2007.
- [7] C. Dezsényi, T. P. Dobrowiecki, and T. Mészáros, "Adaptive Document Analysis with Planning," in *Multi-Agent Systems and Applications IV*, Budapest, Hungary, 2005, vol. 3690, pp. 620–623.
- [8] P. Antal, T. Mészáros, B. De Moor, and T. Dobrowiecki, "Annotated Bayesian networks: a tool to integrate textual and probabilistic medical knowledge," in *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on*, 2001, pp. 177–182.
- [9] P. Antal, B. De Moor, D. Timmerman, T. Mészáros, and T. Dobrowiecki, "Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain," in *Computer-Based Medical Systems, 2002. (CBMS 2002). Proceedings of the 15th IEEE Symposium on*, 2002, pp. 213–218.
- [10] T. Mészáros and T. P. Dobrowiecki, "Controlled Natural Languages for Interface Agents," in *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Budapest, Hungary, 2009, pp. 1173–1174.
- [11] C. Dezsényi, T. Dobrowiecki, and T. Mészáros, "Information Extraction with AI Planning," *Relation*, vol. 10, no. 1.12, p. 5349, 2008.
- [12] Dezsényi Cs., Mészáros T., Dobrowiecki T., "Adaptív dokumentumelemzés információkinyeréshez," *Híradástechnika*, vol. 60, no. 5, pp. 32–34, 2005.
- [13] C. Dezsényi, T. Mészáros, and T. Dobrowiecki, "Parser Framework for Information Extraction," in *Proc. of EUROFUSE-2004 Workshop on Data and Knowledge Engineering*, Warsaw, Poland, 2004.
- [14] T. Meszaros and T. Dobrowiecki, "Rapid prototyping of application-oriented natural language interfaces," in *2010 32nd International Conference on Information Technology Interfaces (ITI)*, 2010, pp. 97–102.
- [15] T. Mészáros and T. Dobrowiecki, "Rapid development of application-oriented natural language interfaces," in *Intelligent Engineering Systems (INES), 2014 18th International Conference on*, 2014, pp. 55–59.
- [16] Mészáros T., Dobrowiecki T., Kiss M., "Kontrollált természetes nyelvű lekérdezés webes adatbázisokhoz," in *Networkshop 2007*, Eger, Hungary, 2007.
- [17] Förhécz A., Mészáros T., "Többnyelvű tezaurusz építése és szolgáltatása webes környezetben," in *Networkshop 2007*, Eger, Hungary, 2007.
- [18] Hornyák Zsuzsanna Éva, Mészáros Tamás, "Publikációgyűjtemény tudásbázisának építése természetes nyelven," in *Networkshop 2014*, Pécs, 2014, pp. 65–69.

