Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Automatizálási és Alkalmazott Informatikai Tanszék

Szemantikus információ-visszakeresés mobil Peer-to-Peer hálózatokban

Ph.D. értekezés tézisei

Forstner Bertalan

Tudományos vezető:
Dr. Charaf Hassan Ph.D.
egyetemi docens

Budapesti Műszaki és Gazdaságtudományi Egyetem
Automatizálási és Alkalmazott Informatikai Tanszék

Budapest, 2008

Ph.D. értekezés tézisei

Forstner Bertalan

Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Automatizálási és Alkalmazott Informatikai Tanszék

1111 Budapest, Goldmann György tér 3.

E-mail: forstner.bertalan@aut.bme.hu
Telefon: 463-1668
Fax: 463-3478

Tudományos vezető:

Dr. Charaf Hassan Ph.D.
egyetemi docens

I. Bevezetés és célkitűzések

Az információ-visszakeresés feladata a kezdetektől a számítógépes technológiák egyik legnagyobb kihívása. A hálózatba kapcsolt számítógépek számának növekedésével egy-egy dokumentum megtalálása egyre nehezebbé válik. Megoldást jelenthet erre, ha az egyes számítógépek szerepeit szétválasztjuk (pl. kliens-szerver architektúra), ahol a dokumentumok, vagy a felkutatásukhoz szükséges adatok (indexek) tárolását dedikált, nagy erőforrásokkal rendelkező számítógépekre bízzuk. Másik, főbb megoldásként egyre inkább a Peer-to-Peer (P2P) információ-visszakereső rendszereket részesítik előnyben annak felépítéséből adódó előnyei miatt. A P2P hálózat egy olyan teljesen elosztott architektúra, ahol az egyes csomópontok egyenrangú félként vesznek részt a kommunikációban.

Az információ-visszakereső rendszerekről beszélve különös figyelmet kell fordítanunk a mobil távközlési rendszerekre. Mivel az okostelefonok, úgynevezett smartphone-ok számítási teljesítménye és növekvő képessége egyre több, különböző típusú információ megjelenítésére teszi alkalmassá ezeket a platformokat, fontossá válik a bevonásuk a P2P információ-visszakeresés világába. A mobil kommunikáció költségei (mind pénz, mind sávszélesség vagy elemkapacitás tekintetében) felülmúlják a vezetékes kommunikációét, ezért esetükben még fontosabb a hatékony P2P protokollok használata. A mobil rendszerekre tervezett P2P protokollnak figyelembe kell venni ezen környezetek egy jelentős tulajdonságát, azok erősen tranziens jellegét. Ez annak köszönhető, hogy gazdasági megfontolásokból, vagy korlátozott kapcsolódási képességük miatt a mobil eszközök felhasználói nem töltenek jelentős időt a hálózaton. A készülékek kapcsolódási képessége leginkább a hálózati lefedettség és az energiakapacitás miatt korlátozott [56], [57]. A mobil P2P rendszerek még további sajátosságokkal rendelkeznek, amelyeket figyelembe kell venni a protokollok kifejlesztésekor.

Míg a kliens-szerver architektúra esetén a kiszolgálóoldalnak jelentős erőforrásokkal kell rendelkeznie (például háttértár, vagy kommunikációs sávszélesség szempontjából), addig az elosztott hálózatok esetében az egyes csomópontok jelentősen kisebb erőforrás-ráfordítással is képesek hatékonyan részt venni a tárolás és kiszolgálás feladatában. A legtöbb P2P protokoll ugyanakkor skálázhatósági problémákkal küszködik: a csomópontok számának növekedésével a megfelelő találati arány csak jelentős hálózati forgalom mellett érhető el. Az ennek a problémának a hatékony megoldását célul kitűző erőfeszítések két nagy csoportra oszthatóak. A strukturált P2P rendszerek (például [39], [40], [41], [58]) közös jellemzője, hogy a csomópontok kapcsolódását, illetve az egyes dokumentumok letárolásának helyét szigorú szabályok határozzák meg. Ezek a hálózatok jó skálázhatósági paraméterekkel rendelkeznek, és működésbeli hatékonyságuk is viszonylag jól becsülhető, de nehezen, vagy igen jelentős erőforrás-ráfordítással kezelik a P2P hálózatok populációjában a gyakori belépések és kilépések következtében megfigyelhető erős tranziens jelleget. A strukturálatlan rendszerek protokolljai esetében, mint az alap Gnutella [42] is, a dokumentumok tárolási helyére nincs előírás, és a csomópontok kapcsolódását is csak néhány egyszerű szabály írja elő. Éppen ezért ezek a rendszerek viszonylag kis protokollrezsivel (overhead) rendelkeznek, és jól viselik a csomópontok hálózatba való gyakori ki- és belépését.

A kétfajta megközelítés előnyeit szemantikus hálózati réteg felhasználásával lehet ötvözni. Az ezt célzó kutatások abból a mindennapi élet több területén is megfigyelhető jelenségből indulnak ki, hogy egy adott személy kapcsolatai nem véletlenszerűek: a kapcsolatok valamilyen közös szál (pl. hasonló foglalkozás, hobbi, zenei ízlés; a továbbiakban érdeklődési terület) mentén alakulnak ki. Ebből következik, hogy a közös érdeklődési terület

mentén ezek az emberek csoportokat képeznek, és a kommunikáció a csoporton belül gyakoribb, mint a csoporton kívül [43].

P2P protokollok szemantikus megközelítéssel történő kifejlesztésére több kutató iskola is javaslatot tett. Ezek az elképzelések mind strukturálatlan, mind strukturált hálózatokra általában új rétegként ráépülve próbálják elérni a találati arány növelését elfogadható forgalom mellett (SON - *Semantic Overlay Networks*). A strukturálatlan megközelítésnek a felhasználás szempontjából való előnyei a SON-ok esetében is megmaradnak. Ugyanakkor a szemantikus réteg és a belőle származó jelenségek vizsgálatára mindeddig kevés figyelmet fordítottak. Létjogosultságukat azonban a rájuk épülő alkalmazások iránti növekvő igény (mint amilyen például a mobil P2P információ-visszakereső rendszerek is) igazolják.

A teljesítmény javítását célzó szemantikus hálózati réteg megfelelő kialakításához szükséges paraméterek megismeréséhez, illetve a réteggel elérhető válaszadási valószínűség kiszámításának céljából matematikai modelleket célszerű alkalmazni. A SON-ok esetében ugyanakkor nehézségekbe ütközünk, hiszen a hagyományos strukturálatlan hálózatokat leíró modellek nem alkalmasak az általam vizsgált rendszerek speciális tulajdonságainak figyelembe vételére. A céljaimhoz, vagyis a szemantikus réteg különböző paraméterek melletti tulajdonságainak, illetve az elérhető válaszadási valószínűségnek kiszámításához legközelebb álló megoldások közül [44] a hálózat minőségét leíró alacsony szintű mértékekkel foglalkozik, de nem használható olyan magasabb szintű értékek, mint az átlagos válaszadási valószínűség, vagy generált hálózati forgalom vizsgálatára. [45] különböző felépítésű P2P hálózatokat vizsgál ugyan, modelljét azonban csak extrém esetekben lehet felhasználni szemantikus hálózati rétegek leírására. A [46] által bemutatott megoldás a kérés-továbbítási fában történő csoportképződés hálózati forgalomra való hatását vizsgálja, ám nem elég mély szinten. Hasonlóan [47] is csak a hálózati forgalommal foglalkozik. Célom tehát egy olyan modell elkészítése volt, amellyel kiszámítható a szemantikus réteggel rendelkező P2P hálózatokban megfigyelt viselkedésminták alapján a lehető legmagasabb elméleti találati arány.

Ahhoz, hogy a strukturálatlan P2P protokoll kialakíthassa az optimális szemantikus réteget, a hálózatban található, különböző témájú dokumentumok számáról, illetve elhelyezkedéséről pontos ismeretekkel kellene rendelkeznie minden egyes csomópontnak. Ennek több akadálya is van, melyek közül a legfontosabbak a mobil csomópontoknál tapasztalt erősen tranziens viselkedés, illetve az elérhetetlenül magas hálózati forgalom-igény. Ebből adódik, hogy a hatékony megoldás sokkal inkább egy olyan stratégiát kell, hogy kövessen, amellyel az egyes csomópontok lokális megfigyelések és döntések sorozatával térképezhetik fel a számukra releváns paramétereket.

Kutatásaim során különböző javaslatokat vizsgáltam meg a szemantikus adatokon alapuló réteget tartalmazó protokollok, illetve protokoll-kiterjesztések tekintetében (pl. [48], [49], [50], [51]) különösen arra koncentrálva, hogy az elméleti találati arányt milyen mértékben közelítik meg. A megvizsgált algoritmusok azonban vagy nem elég hatékonyak, vagy túl sok előfeltételezéssel rendelkeznek, vagy egyáltalán nem alkalmazhatóak a mobil P2P rendszerekben. A használható protokollok hatékonysága nagyban függ attól, hogy – a modellben megismert – hálózati paramétereket milyen mennyiségben illetve pontossággal használják fel. Egy olyan megoldásra van tehát szükség, amely strukturálatlan hálózatokra építhető, rugalmasan alkalmazható a mobil hálózatokban megismert körülmények között is, és jól közelíti az elméleti válaszadási valószínűséget. Az általam vizsgált megoldás arra az esetre koncentrál, amikor a felhasználók (vagyis az egyes csomópontok) egyértelműen azonosítható dokumentumokat keresnek a P2P hálózaton, vagyis a keresett dokumentumnak valamilyen azonosítóját ismerik (pl. dokumentum címe, zeneszám, album címe).

Egy fejlett megoldás széleskörű alkalmazhatóságát az biztosíthatja, ha ezen megoldás segítségével mobil környezetben különböző metaadatsémákra és alapprotokollokra építve magas válaszadási valószínűséget sikerül elérni. Ehhez nemcsak megfelelően tervezett szoftver architektúra szükséges, hanem egy alacsony erőforrás-igényű protokoll és algoritmus hatékony megvalósítása is. Bár mutatkozik fogyasztói igény ilyen tulajdonságokkal rendelkező, alacsony energiaigényű, teljesen elosztott mobil információmegosztó és -kereső rendszerre [52], hatékony megvalósításuk még várat magára.
Mindezek alapján téziseimben a következő kérdésekre kerestem választ:

- Melyek a mobil P2P hálózatok azon sajátosságai, amiket figyelembe kell venni, amikor szemantikus P2P réteget tervezünk hozzájuk?
- Mely paraméterek ismerete szükséges ahhoz, hogy az egyes csomópontok lokális döntésekkel maximalizálják találati arányukat, vagy csökkentsék az általuk generált hálózati forgalmat, egy adott szemantikus környezetben?
- Milyen válaszadási valószínűség (recall value) érhető el elméletben a vizsgálandó, strukturálatlan szemantikus kiterjesztésekkel?
- Milyen algoritmus és protokoll a megfelelő az alacsony számítási képességekkel és háttértárral rendelkező eszközök esetében? Mi a legjobb módja esetükben a szemantikus információk tárolásának és karbantartásának?
- Hogyan lehet megfelelő topológiát kialakítani egy strukturálatlan és teljesen elosztott hálózati rétegben?
- Hogy néz ki egy megfelelő szoftverarchitektúra mobil eszközökre, amely architektúra a készülékek tulajdonságainak figyelembe vétele mellett különböző típusú szemantikus adatok alkalmazására is képes?

II. A kutatás módszertana

A célkitűzésben felvázolt problémák határozták meg a kutatásom irányát. A P2P hálózat válaszadási valószínűségének, vagyis a sikeresen megválaszolt lekérdezések arányának megállapításához először valószínűségszámítási módszerekkel egy zárt matematikai modellt állítottam fel. A modell adott kérdés szempontjából releváns paraméterek alapján származtatja a várható valószínűséget, míg a paraméterek beállítása tipikus hálózati méretek és felhasználói aktivitások alapján történik. A modell megalkotásánál elsődleges célom volt, hogy a szemantikus adatok alapján átalakított hálózat különleges jellemzőit (például csoportképződés a kéréstovábbítási gráfban, szűkített keresési tér) a modell figyelembe vegye. További kitűzött célom volt, hogy a modellből következtetni lehessen az optimális szemantikus réteg megalkotásához szükséges paraméterekre.

Azon paraméterek közelítésére, amelyek a hálózatról való globális tudást előfeltételeznek, Bayes-i eljárást dolgoztam ki. Az eljárás a hálózat tagjai által biztosított *a priori* információt az egyes csomópontok megfigyelései alapján pontosítva közelíti a különböző klienseken keresztül elérhető válaszadási valószínűséget az egyes témakörökben.

Az elkészített modell helyességét, illetve a SemPeer protokoll teljesítményét szimulátor segítségével verifikáltam [13]. A GXS a tanszéki csoport által a Nokia Research Centerrel együtt indított kutatási projekt eredményeképpen született eszköz, amely általános P2P protokollokat alkalmazó hálózatok szimulációjára, statikus és dinamikus adatok gyűjtésére alkalmas. A GXS Peer-to-Peer szimulátort az általam kifejlesztett SemPeer protokollt megvalósító modullal bővítve az alap- és az új protokollok dinamikus viselkedését is vizsgálni tudtam, illetve az eredményeknek a modell által jósolt beállt állapothoz (elméleti

maximumhoz) való viszonyát is ellenőrizhettem. A szimulátor a modellel azonos paraméterekkel dolgozik. A paraméterek meghatározásához egy úgynevezett *crawling* klienst készítettem, amely valós, mobil P2P felhasználók tízezreinek használati statisztikáit gyűjtötte össze.

A modell eredményeinek ellenőrzésére alapvető referenciaként a legtöbb kutatás által is használt Gnutella protokollt [42] választottam. Döntésemet a viszonylag alacsony protokollköltség és a nagyszámú elérhető referenciaanyag mellett a Gnutellát leírni képes matematikai modellek megléte is indokolta. A szemantikus protokollok esetében a szimulációs eredményekkel való összevetéssel tudtam a modell ellenőrzését elvégezni. Az ellenőrzések alapján a modellt tökéletesítettem és visszaellenőriztem.

Az általam bemutatott új eredményeket a gyakorlatban is hasznosítottam. A szoftvermérnöki módszerekkel megtervezett, Symellának nevezett szoftvercsomag az előnyök kihasználása céljából mobil környezetben, vezeték nélküli kommunikációs technológiával működik [53]. A tervek további felhasználhatósága céljából a szoftver architektúrájának megalkotásakor tervezési mintákat alkalmaztam. Az alkalmazás forráskódját, amely Open Source formában elérhető, azóta több tízezren töltötték le. A szoftvercsomagnak a protokoll-kiterjesztésből adódó erőforrásigényeit – mint a CPU-használat, memória-, illetve energiaigény –, a SymbianOS alapú készülékekre kidolgozott Performance Investigator szoftvercsomaggal mértem fel. A mérések során a megfigyelt átlagos mobil munkamenetekkel végeztem többször 14 mérést. A méréseknél a mért adatokat a 2-2 szélsőértéket elhagyva, a maradék 10 értékből átlagot számítva normalizáltam.

III. Új eredmények összefoglalása

Kutatómunkám eredményét négy tézisben foglaltam össze. Ezen tézisek arra adnak a választ, hogy miként lehet egy, a mobil környezet sajátosságait figyelembe vevő, hatékony elosztott információ-visszakereső rendszert építeni. A sajátosságok közül a legfontosabb szempontok a mobil eszközök limitált energia-, számítási- és memóriakapacitása, a kommunikáció anyagi-, energia- és sávszélesség költsége, és az ezekből adódó jellegzetességek – mint pl. a kliensek erősen tranziens jellege – voltak. Szintén szem előtt tartott szempont volt a protokoll teljesen elosztott jellege, a skálázhatóság, illetve a csomópontok autonomitásának megőrzése.

A protokoll kutatása során fontos volt megismerni, hogy milyen válaszadási valószínűség várható a hálózat szemantikus adatokra támaszkodó optimális átalakításából. Ehhez az *első tézisben* matematikai modellt alkottam, amely a hálózatot és a csomópontokat leíró releváns paraméterek alapján konkrét szimulációk futtatása nélkül megadja annak valószínűségét, hogy egy csomópont az általa elindított keresésre (feltett kérdésre) pozitív választ kap-e (ezt az értéket a továbbiakban a P2P hálózat *válaszadási valószínűségének* nevezem). Annak érdekében, hogy a modell hatékonyan kezelje a SemPeer protokoll által okozott nagymértékű csoportképződést, egy új mértéket vezetek be, a *módosított csoportképződési együtthatót*. Ennek ismeretében ugyancsak ebben a tézisben mutatom meg, hogy egy protokoll hogyan érheti el a csoportképződésből adódó redundáns üzenet-feldolgozások számának csökkentését, illetve a kérés által elért csomópontok számának maximalizálását.

A modellből kikövetkeztethetőek azok a globális paraméterek, amelyek lokális ismerete a szemantikus réteg teljesítményének maximalizálásához szükséges. Mivel ezen globális paraméterek megismerése a tranziens hálózatban nehézkes és erőforrás-igényes feladat

lenne, a *második tézisben* szemantikus profilokat alkotok, amelyeknek naprakészen tartásához egy lokális, Bayes-i eljáráson alapuló algoritmust készítettem. Erről belátom, hogy az ismeretlen paramétereket alacsony szórású valószínűségi változókkal közelíti.

A *harmadik tézis* a SemPeer protokollt és algoritmust mutatja be. Bebizonyítom, hogy a megoldásom a szemantikus rétegben az optimális találati arány eléréséhez folyamatosan úgy alakítja át a hálózatot, hogy a válaszadási valószínűség növekedjen. Emellett a kéréstovábbítási gráfban megfigyelhető csoportképződéstől mentes topológiát adok a szemantikus réteg formálására. Algoritmust adok a topológia kialakítására, amelyről bebizonyítom, hogy valóban kiküszöböli a kontraproduktív éleket.

Az eredményeim illusztrálására konkrét protokollkiterjesztés-megvalósítást, illetve szoftvercsomagot terveztem. A *negyedik tézisben* megmutatom a megvalósítások alacsony erőforrásigényét az üzenetek méretére, a számítási- és memóriaigényre, illetve az energiafogyasztásra koncentrálva.

I. Tézis. Szemantikus réteggel rendelkező strukturálatlan Peer-to-Peer hálózatok válaszadási valószínűségének matematikai modellezése

Kapcsolódó publikációk: [5], [7], [36], [18], [30]

Matematikai modellt adtam a szemantikus réteggel, valamint különböző tulajdonságokkal rendelkező P2P protokollok válaszadási valószínűségének kiszámítására. A kéréstovábbítási gráfban megfigyelhető klasztereződés számszerű leírására bevezettem a módosított csoportképződési együtthatót. A statikus hálózatban elérhető válaszadási valószínűség meghatározásához bevezettem az absztrakt IdealSON P2P protokollt. Az általam megadott matematikai modellről bebizonyítottam, hogy különböző viselkedési minták mellett alkalmas a szemantikus réteggel rendelkező strukturálatlan P2P protokollok által elérhető válaszadási valószínűség kiszámítására. Beláttam, hogy mely globális paraméterek lokális ismerete szükséges az optimális topológia kialakításához. Megmutattam, hogy lokális információk birtokában mely döntésekkel tud egy csomópont a szemantikus réteg teljesítményén javítani.

Ebben a tézisben elméleti alapot adtam arra, hogy strukturálatlan Peer-to-Peer hálózatok esetén különböző peremfeltételek mellett szemantikus réteggel milyen válaszadási valószínűséget lehet elérni. A modellel egyre komplexebb viselkedésminták mentén vizsgáltam az optimális esetben elérhető teljesítmény-növekedést. A komplexitást a véletlen kapcsolatok melletti szemantikus linkek megjelenése, a csomópontokat jellemző érdeklődési területek multiplicitása, illetve az általuk tárolt dokumentumok tematikája jelentette. A *módosított csoportképződési együttható* bevezetésére a lokális döntésekkel operáló strukturálatlan P2P protokollok esetében megjelenő globális hatás jellemzése és számszerűsítése érdekében volt szükség.

1.1 Definíció (*Kéréstovábbítási gráf*): A P2P hálózat egy adott csomópontjából kiinduló keresés által bejárt csomópontok és kapcsolatok hálózatát ábrázoló gráf. Jelölése: G^v.

1.2 Definíció (*Kéréstovábbítási fa*): A kéréstovábbítási gráfból elhagyva azokat az éleket, amelyek egy, már meglátogatott csomópontra mutatnak, kapjuk a kéréstovábbítási fát. A kéréstovábbítási fa egyes szintjeihez a kiindulási csomóponttól azonos távolságra található csomópontok tartoznak. Jelölése: G'^v.

A tézishez bevezettem egy új fogalmat, a *módosított csoportképződési együtthatót*, amelynek célja, hogy jellemezze a kéréstovábbítási gráfban jelentkező csoportképződést, és segítségével összevethetőek legyenek különböző P2P protokollok. Egy kéréstovábbítási gráfban a csoportképződést az úgynevezett kontraproduktív élek okozzák, amelyeken keresztül a keresőkérdés visszajuthat egy már meglátogatott csomóponthoz.

1.3 Definíció (*Kontraproduktív élek*): A kéréstovábbítási gráfban szereplő olyan élek, amelyek olyan csomópontra mutatnak, amelyeket a kiinduló csomópontból egy alternatív útvonalon el lehet érni. Az alternatív útvonal nem hosszabb, mint az, amelyik a kontraproduktív élet tartalmazza. A kontraproduktív élek egy adott kéréstovábbítási gráfban megjelenő halmazát E_r jelöli.

1.4 Definíció (A kontraproduktív élek osztályozása): A kontraproduktív éleket a következő három típusba sorolhatjuk:

a. hátramutató (*backward) élek*: élek, amelyek a kéréstovábbítási gráfban hátrafelé mutatnak. Ebben az esetben az él kiinduló csomópontja olyan csomópontnak továbbít egy üzenetet, amely már korábban megkapta és továbbította azt.
b. oldalra mutató (*sibling*) élek: A kéréstovábbítási fa azonos szintjein elhelyezkedő csomópontok (*testvér csomópontok*) közötti élek halmaza. Ezek az élek okozzák azt, hogy egy *A* csomópont egy olyan *B* csomópontnak továbbít egy üzenetet, amely ugyanannyi lépés (*hop)* után az *A* csomóponthoz hasonlóan megkapta már egyszer azt.
c. ferde *(skew)* élek: a testvér csomópontok szomszédos csomópontjaira mutató élek halmaza. Ez esetben a mutatott csomópont több útvonalon, azonos lépésszám után többször kapja meg ugyanazt az üzenetet.

A Watts és Strogatz által bevezetett csoportképződési együttható [55] egy adott csomópont szomszédainak egymás közötti kapcsolatainak mennyiségét jellemezte egy 0 és 1 közötti értékkel. Az általam bevezetett módosított csoportképződési együttható egy olyan érték, amely megfelel a következő tulajdonságoknak.

1.5 Definíció *(A módosított csoportképződési együttható (C_{mod}) tulajdonságai):* A módosított csoportképződési együttható a Peer-to-Peer hálózatban egy adott keresőkérdés által elért csomópontok halmazát jellemzi a következő módon:

a. $C_{mod} \in [0,1]$
b. $G' = G \cup e_i \setminus e_j, e_i \in \{E_r\}, e_j \notin \{E_r\} \Leftrightarrow C_{mod,G} \leq C_{mod,G'}$
c. C_{mod}=0 akkor és csak akkor áll fenn, ha a kéréstovábbítási gráf egy fa.

A következő definíciókat adtam módosított csoportképződési együtthatókra.

1.6 Definíció *(A C_{mod1} módosított csoportképződési együttható):* A hálózat egy *r* csomópontja esetén a $C_{mod1,r}$ értéke a következőképpen számítandó:

$$C_{\mathrm{mod1,r}} = \frac{\left|\left\{E_r^*\right\}\right|}{\sum_{m=1}^{TTL}\left[k^m\left(\sum_{n=0}^{m-1}\left(k^n\right)+k^m-1\right)\right]+\sum_{m=1}^{TTL-1}k^m\left(k^{m+1}-k\right)}. \qquad (1)$$

A C_{mod1} értéket az egyes csomópontoknál számolt $C_{mod1,r}$ értékek átlagából számíthatjuk.

1.1 Altézis: Beláttam, hogy azon P2P protokollok esetén, ahol az egyes csomópontok *k* kapcsolaton keresztül, *TTL* kezdeti érvényességi paraméterrel (Time-to-Live, az üzenet továbbításokban mért élettartama) küldenek üzeneteket, C_{mod1} egy megfelelő módosított csoportképződési együttható.

A modellalkotás során hasznosnak bizonyult egy olyan *módosított csoportképződési együttható* bevezetése, amelynek értéke a keresőkérdés által elért csomópontok számából levezethető.

1.6 Definíció *(A C_{mod2} módosított csoportképződési együttható):* A hálózat egy *r* csomópontja esetén a $C_{mod2,r}$ értéke a következő egyenlet [0, 1] közötti gyökeként számolható:

$$E_q = \sum_{i=1}^{TTL} \left[(1 - C_{\mathrm{mod}\,2,r})k \right]^i , \qquad (2)$$

ahol E_q az egy adott kérdés által elért csomópontok száma.

A továbbiakhoz bebizonyítottam a következő altézist.

1.2 Altézis: Megmutattam, hogy azon P2P protokollok esetén, ahol az egyes csomópontok *k* kapcsolaton keresztül, *TTL* kezdeti érvényességi paraméterrel küldenek üzeneteket, C_{mod2} egy megfelelő módosított csoportképződési együttható.

Definiáltam egy egyszerűsített (*relaxed*) feladatot a mobil környezetben történő szemantikus P2P információ-visszakeresés feladatára, amelyben a valós környezet egyes tulajdonságainak figyelmen kívül hagyásával egyszerűbb módon lehet a hálózat átlagos válaszadási valószínűsége tekintetében optimális eredményt elérni. Az egyszerűsített problémát, és az abban használt protokollt IdealSON-nak neveztem el.

1.7 Definíció *(Az IdealSON egyszerűsített feladat):* Az IdealSON protokoll (vagy IdealSON feladat) egy olyan, P2P hálózat fölött elhelyezkedő réteg, amely az alábbi tulajdonságokkal rendelkezik:

a. A hálózat csomópontjai adottak: senki nem csatlakozik vagy hagyja el a hálót.
b. Az egyes csomópontok által tárolt dokumentumok sosem változnak.
c. A csomópontok kapcsolatai rögzítettek.
d. A csomópontok szemantikus kapcsolatai hasonló érdeklődésű területtel rendelkező csomópontokra mutatnak.
e. A hasonló érdeklődési területekkel rendelkező csomópontok és kapcsolataik által definiált részgráf összefüggő.

Az IdealSON protokoll esetén a csomópontok célja tehát a szemantikus kapcsolatukon keresztül elérhető, adott témakörű, különböző dokumentumok számának maximalizálása. Adott strukturálatlan, üzenetárasztáson alapuló protokollal nem lehet elérni magasabb átlagos találati arányt az érdeklődési területek adott finomságú felosztása esetén, mint amekkorát az IdealSON protokoll a kéréstovábbítási gráfban található dokumentumok számának maximalizálásával elérhet, mivel a kapcsolatok feltételeknek megfelelő kialakítása minimálisra csökkenti a keresési teret. Az IdealSON tulajdonképpen a hálózatnak az az

állandósult állapota, amit majd a valós környezetben alkalmazott protokoll-kiterjesztésünk minden pillanatban el akar érni.

A matematikai modellt úgy alkottam meg, hogy nulla vagy több érdeklődési területtel rendelkező csomópontok esetén is alkalmazható legyen. Ehhez szükséges volt szétválasztanom az adott érdeklődési területnek megfelelő, illetve az azon kívül eső (*offtopic*) kérdésekre érkező válaszok valószínűségét. Egy adott SON protokollnak megfelelő végső modell ezek kombinálásával áll elő. Az érdeklődési területnek megfelelő keresőkérdés esetén SON protokollt használva a várható találati arányt pontosabban lehet közelíteni, mint az alattas réteg normál kapcsolatain keresztül érkező válaszok esetén, ezért fontos megbecsülni, hogy a kéréstovábbítási fa adott szintjén milyen valószínűséggel találunk megfelelő érdeklődési területtel rendelkező csomópontot. Ezt a becslést adja meg a következő altézis, amihez bevezetek néhány jelölést.

Álljon P_i annak valószínűségeként, hogy egy, a kéréstovábbítási fa *i*-edik szintjén található csomópont rendelkezik a *q* keresőkérdés témakörének megfelelő érdeklődési területtel. Ennek az értéknek a közelítő számításához bevezettem a $P_{SemLink}$ jelölést, amely annak valószínűségét adja meg, hogy egy adott, a keresőkérdés érdeklődési területének megfelelő csomópont kimeneti kapcsolata szintén az adott érdeklődési területű csomópontra mutat. $P_{SameTopic}$ pedig azt a valószínűséget számszerűsíti, hogy egy, az alapprotokollban meglévő kapcsolat által mutatott csomópont érdeklődési területei között a keresett dokumentum témaköre is megtalálható.

1.3 Altézis: Beláttam, hogy a *q* keresőkérdés által keresett dokumentum témakörének megfelelő érdeklődési területtel rendelkező csomópont találatának valószínűsége a kéréstovábbítási fa *i-dik* szintjén a következő:

$$\begin{aligned} P_i &= P_{i-1}P_{SemLink,q} + P_{i-1}\overline{P}_{SemLink,q}P_{SameTopic,q} + \\ &+ \overline{P}_{i-1}\overline{P}_{SemLink,q}P_{SameTopic,q} = P_{i-1}P_{SemLink,q} + P_{i-1}\overline{P}_{SemLink,q}P_{SameTopic,q} \end{aligned} \qquad (3)$$

P_0 értéke 1, amennyiben a kiindulási csomópont érdeklődési területei közé a keresett dokumentum témaköre beletartozik. Ellenkező esetben értéke 0.

Legyen *t* a rendszerben megkülönböztetett érdeklődési területek (*topic*) száma. A következőkben feltételezem, hogy a különböző érdeklődési területek egyenletes eloszlással szerepelnek a hálózatban. A fenti altézisek segítségével a következő megállapításokat tettem.

1.4 Altézis: Megmutattam, hogy a *t* darab érdeklődési területtel jellemezhető hálózatban a csomópont érdeklődési területének megfelelő keresőkérés esetén az átlagos találati valószínűségre fennáll az alábbi összefüggés:

$$P_{success,topic} \le 1 - \prod_{i=1}^{TTL}\left(1-\frac{1}{D_t}\right)^{D_n \overline{P}_{doc,offtopic} P_i [(1-C)k]^i} * \prod_{i=1}^{TTL}\left(1-\frac{1}{D}\right)^{D_n P_{document,offtopic} \overline{P_i}((1-C)k)^i}, \qquad (4)$$

ahol D_t az adott témakörökhöz tartozó egyedi dokumentumszám, D_n véletlen változó a csomópont által tartalmazott egyedi dokumentumok számának jellemzésére, D a rendszerben található összes egyedi dokumentum száma.

Továbbá a csomópont érdeklődési területen kívül eső keresőkérdésre a következő állítás igaz:

1.5 Altézis: Bebizonyítottam, hogy egy adott csomópont érdeklődési területén kívül eső keresőkérdésre az átlagos válaszadási valószínűségre fennáll az alábbi összefüggés:

$$P_{success,offtopic} \le 1-\prod_{i=1}^{TTL}\left(1-\frac{1}{D}\right)^{D_n \overline{P}_i((1-C)k)^i} * \prod_{i=1}^{TTL}\left(1-\frac{1}{D}\right)^{D_n P_{document,offtopic} P_i((1-C)k)^i}, \quad (5)$$

ahol D_n véletlen változó a csomópont által tartalmazott egyedi dokumentumok számának jellemzésére, D a rendszerben található összes egyedi dokumentum száma.

Az 1.1-1.5 altézisekre támaszkodva beláttam az IdealSON rendszerben a csomópontok által várható válaszadási valószínűségre a következőt:

1.6 Altézis: Megmutattam, hogy amennyiben a kibocsátott keresőkérdések témakörök közti eloszlásának valószínűsége ismert, és a csomópontok $P_{doc,offtopic}$ valószínűséggel tárolnak az érdeklődési körükön kívül eső dokumentumokat, az alábbi állítás igaz:

$$P_{success,IdealSON} = P_{issued,topic} P_{success,topic} + P_{issued,offtopic} P_{success,offtopic}. \quad (6)$$

A modellalkotás folyamán bebizonyítottam, melyik a legmegfelelőbb stratégia a csomópontok által elérhető találati arány maximalizálására a lokálisan elérhető adatok alapján.

1.7 Altézis: Beláttam, hogy adott témakörű dokumentumok számának ismeretének hiányában a szemantikus rétegben elérhető válaszadási valószínűség úgy növelhető, ha egy csomópont egy kapcsolatát egy olyannal helyettesíti, amelyen keresztül több, különböző dokumentumot lehet elérni annak kéréstovábbítási gráfján keresztül, feltéve, hogy az új kapcsolaton keresztül több olyan, az érdeklődési területbe eső dokumentumot kapunk, amelyet a többi kapcsolaton keresztül nem kapunk meg:

$$P_{success,t}^{C_1 \cup ... \cup C_l \cup ... \cup C_k} \le P_{success,t}^{C_1 \cup ... \cup C_j \cup ... \cup C_k}, \forall i,j : D_t^{C_l} \le D_t^{C_j}. \quad (7)$$

Az első téziscsoport eredményeit a disszertáció 4. fejezetében fejtem ki részletesen.

II. Tézis. Szemantikus profil a felhasználók és kéréstovábbítási gráfok tanulás alapján történő jellemzésére

Kapcsolódó publikációk: [3], [17], [31], [32]

Kialakítottam négy különböző profilt taxonómiákban tárolt fogalmak előfordulási valószínűségeinek hatékony reprezentálására és karbantartására. Kidolgoztam egy Bayes-i eljáráson alapuló módszert az egyes felhasználók érdeklődési területeit jellemző valószínűségi változók helyi megfigyeléseken alapuló közelítésére. A profilokról megmutattam, hogy az eljárásommal kapott valószínűségi változók várható értéke egyszerű számítással és alacsony szórással áll elő. Szintén lokális megfigyelésen alapuló módszert adtam a csomópontok kapcsolatain keresztül elérhető kéréstovábbítási gráfokon az egyes témakörökben várható válaszok valószínűségeinek hatékony inicializálására és karbantartására.

Az előző tézisben bemutatott modell feltárta, hogy – a vizsgált, ideális környezetben – mely adatok ismerete szükséges ahhoz, hogy egy csomópont minél több keresőkérdésére választ kapjon. Az ezek közt előforduló, nem hozzáférhető adatok értékeinek minél pontosabb előállítására bevezettem négy szemantikus profilt. A szemantikus profil egy súlyozott

taxonómia, amely az adott csomópont által tárolt, vagy rajta keresztül elérhető dokumentumokat leíró kulcsszavakat tartalmazza, a kulcsszó előfordulási számával együtt. A gyakorlatban és a szimulációk során az MP3 zenei fájlok metaadataiból kinyerhető zenei stílust, vagy pedig az angol szavakon alapuló, általánosan elfogadott és naprakész WordNet taxonómiát használom a tárolt kulcsszavak közötti relációk kinyerésére [54]. A profilok csomópontjai az úgynevezett *synset-ek*.

2.1 Definíció (*Synset, szinonim halmaz*): Egy vagy több szinonim fogalomból álló halmaz.

2.2 Definíció (*Semantic* vagy *szemantikus* profil): Egy csomópont szemantikus profilja egy olyan súlyozott taxonómia, amelynek csomópontjai a tárolt dokumentumokat leíró szinonim halmazokból és a halmazok elemeinek a dokumentumok leírásában történő előfordulási számából (D_t^*) állnak.

A szemantikus profilból megállapítható az egyes témák előfordulásának valószínűsége a megosztott dokumentumok között. Megmutattam, hogy ha feltételezzük, hogy a tárolt dokumentumok jellemzik a felhasználó érdeklődési területeit, akkor a szemantikus profil olyan pontossággal írja le ezeket az érdeklődési területeket valamely absztrakciós szinten, amilyen pontosan a dokumentumokból kinyert kulcsszavak jellemzik magukat a dokumentumokat. A szemantikus profilt a felhasználó érdeklődési területeinek egy kezdeti közelítéseként használom fel.

Az egyes csomópontok kapcsolataik jellemzésére használják a *connection* profilt.

2.3 Definíció (*Connection* vagy *kapcsolati* profil). A *connection profil* egy olyan, *C* kimenő kapcsolathoz rendelt súlyozott taxonómia, amelyben a szinonim halmazok mellett az adott kapcsolaton keresztül az egyes *t* szinonim halmazokkal leírható dokumentumokra irányuló keresőkérdésre a kapcsolathoz tartozó kéréstovábbítási gráfon keresztül várható válaszadási valószínűség is szerepel. Ezt a válaszadási valószínűséget P_t^C-vel jelöljük. Az adott kapcsolaton keresztül elérhető, az egyes témakörökhöz tartozó különböző dokumentumok maximális száma (D_t^C) is eltárolásra kerül.

Megmutattam, hogy lokális megfigyelések alapján hogyan lehet a *connection* profilt naprakészen tartani.

2.1 Altézis: Bebizonyítottam a következő állítást. Legyen α azon esetek száma, amikor egy csomópont a *C* kapcsolaton keresztül *t* témakörrel leírható dokumentumokra indított keresések esetén találatot ért el, β pedig annak száma, amikor nem ért el találatot. Megmutattam, hogy P_t^C beta eloszlású valószínűségi változó, amely a csomópont abbeli hiedelmét reprezentálja, hogy a *C* kapcsolaton keresztül a *t* témakörben találatot érhetünk el. Ennek várható értéke a következőképpen számítható:

$$\mathbf{E}(P_t^C) = \frac{\alpha}{\alpha + \beta}, \tag{8}$$

az alábbi szórásnégyzet mellett:

$$Var(P_t^C) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{9}$$

Az altézis jelentősége, hogy viszonylag kevés megfigyelésből alacsony szórással kapjuk meg P_t^C várható értékét minden témakörre és kapcsolatra. Annak érdekében, hogy a csomópont összevethesse egy elérhető kapcsolat értékét egy meglévő kapcsolatával, valamint abból a célból, hogy a *connection* profil értékei gyorsabban konvergáljanak P_t^C-hez, bevezettem a *reply* profilt.

2.4 Definíció (*Reply profil*): Egy *c* csomópont *reply* profilja egy olyan taxonómia, ahol az egyes *t* szinonim halmazokkal együtt eltároljuk a szinonim halmazzal jellemezhető ismert dokumentumok maximális számát, illetve a *c* csomóponton keresztül a témakörben kiküldött keresőkérdés válaszadási valószínűségét. Ezt R_t^c-vel jelöljük.

A *reply* profil tehát segít a nem szomszédos csomópontokon keresztül elérhető kéréstovábbítási gráfok egyes témakörökre jellemző válaszadási valószínűségek megismerésében. A *reply* profillal inicializált *connection* profilok tranziens környezetben való alkalmazhatóságát bizonyítottam a következő altézisben.

2.2 Altézis: Beláttam, hogy amennyiben egy szemantikus protokoll olymódon tartja karban egy csomópont szemantikus kapcsolatait, hogy egy leváló szomszédot olyan, másik csomóponttal helyettesít, amelynek megegyező érdeklődési területe van, akkor a *connection* profil csak a standard P2P üzenetforgalom megfigyelésével, további hálózati forgalom generálása nélkül képes a teljes kéréstovábbítási gráf csomópontjairól alkotott szemantikus ismereteket karbantartani.

E tézis feltételének kielégítését biztosítja a következő megállapításom.

2.3 Altézis: Megmutattam, hogy R_t^c értéke egyenlő az adott csomóponton keresztül minimálisan elérhető válaszadási valószínűséggel az adott témakörben, vagyis $P_{success,t} \geq R_t^c$.

Végül egy adott felhasználó érdeklődési területeinek leírásához bevezettem a *query* profilt.

2.5 Definíció (*Query* vagy *lekérdezési profil*). Egy csomópont query profilja egy súlyozott taxonómia, amelyben az egyes szinonim halmazokkal együtt azt is eltároljuk, hogy a profilhoz tartozó csomópontból indított keresőkérdés milyen valószínűséggel írható le az adott témakörrel. Ezt a valószínűséget Q_t-vel jelöljük.

2.4 Altézis: Bebizonyítottam a következő állítást. Legyen Q_t egy véletlen változó, amely azt a hipotézist reprezentálja, hogy a csomópont által kibocsátott keresőkérdésre válaszul kapott dokumentum a *t* témakörrel írható le. Amennyiben a csomópont múltbeli kérdéseire kapott válaszokat megvizsgálva azt kapjuk, hogy α alkalommal a *t* témakörrel leírható, míg β alkalommal t-től eltérő témakörrel leírható dokumentumot kaptunk, Q_t egy beta eloszlású valószínűségi változó, melynek várható értéke a következőképpen számítható:

$$\mathbf{E}(Q_t) = \frac{\alpha}{\alpha + \beta}, \tag{10}$$

az alábbi szórásnégyzet mellett:

$$Var(Q_t) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{11}$$

A *query* és a *connection* profilok a megfigyelésekkel történő frissítés következtében követik a felhasználó, illetve a kapcsolatok érdeklődési területeiben történő változásokat. A felhasználók rövid csatlakoztatott állapotban töltött ideje miatt azonban szükségessé válhat, hogy a frissebb adatok nagyobb súllyal essenek számításba a valószínűségi változók meghatározásánál. Ebből a célból egy *f* fakulási faktort (*fading factor*) vezettem be, melynek értékkészlete a [0,1] halmaz, és segítségével a megfigyelések alapján történő profilfrissítést a következőképpen módosítottam.

$$(Q_j)' = Beta(fa_j+1, fb_j), \ \forall j : j \in \{t \cup hypernyms(t)\} \quad (12)$$
$$(Q_i)' = Beta(fa_i, fb_i+1), \ \forall i : i \in \{t \cup hypernyms(t)\} \quad (13)$$

A *hypernyms(t)* halmaza a taxonómiában a *t* témakör általánosított fogalmait jelenti. A frissítés a *connection* profil esetében is hasonlóan alakul.

A második téziscsoport eredményeit a kapcsolódó disszertáció 5. fejezetében fejtem ki részletesen.

III. Tézis. Szemantikus Peer-to-Peer információ-visszakereső protokoll és algoritmus

Kapcsolódó publikációk: [6], [12], [13], [14], [16], [20], [25], [27]

Algoritmust és protokollt adtam a mobil környezetben történő szemantikus információ-visszakeresés támogatására, ezekről igazoltam, hogy a megoldásom megközelíti az IdealSON elméleti protokoll által elérhető teljesítményt. Megterveztem a Disjoint Rings topológiát, elosztott algoritmust adtam az előállítására. Az így előállt topológiáról bebizonyítottam, hogy mentes a kontraproduktív élektől. Mérésekkel igazoltam, hogy a csomópontok a szemantikus és véletlen kapcsolatoknak az I. tézisben leírt arányú alkalmazásával tudják az elérhető válaszadási valószínűséget maximalizálni. Mérésekkel igazoltam, hogy a mobil P2P hálózatokban tapasztalható tranziens viselkedés nem befolyásolja jelentősen a protokollom teljesítményét. Igazoltam, hogy a protokollom használata jelentős hálózati forgalomcsökkenést eredményez.

Az általánosan használt strukturálatlan P2P protokollok valamilyen formában az *ad hoc* szabványra, a Gnutella alapprotokollra, vagyis annak öt üzenettípusára épülnek. A hálózat kiépülése véletlenszerűnek tekinthető, így az egyes keresések gyakorlatilag a hálózat véletlenszerűen meghatározott csomópontjait érik el. Ésszerűnek tűnik az egyes csomópontok kapcsolatainak átszervezése olymódon, hogy a szomszédok minél nagyobb valószínűséggel tudjanak eredményesen válaszolni a keresőkérdésre. Az átszervezés történhet egyszerűen, az üzeneteken alapuló statisztikák figyelembe vételével, megbízható, kiemelt szerepű csomópontok segítségével, vagy érdeklődési kategóriákon alapuló üzenettovábbítással.

Az egyes protokollbővítések előnyeire építve kidolgoztam egy olyan új protokollréteget, amely figyelembe veszi a mobil környezetnek a bevezetésben leírt sajátosságait. Mivel ez egy réteg az alapprotokoll fölött, ezért megfelelő implementációval az új protokollt támogató csomópontok együtt tudnak működni a csak az alapprotokollt ismerő számítógépekkel.

Kiterjesztéseket dolgoztam ki a strukturálatlan P2P protokollokra, illetve a résztvevő csomópontjaik működésére. A protokoll kiegészítése a következő pontokban nyilvánul meg:

1. *A csomópontok kapcsolódásakor a résztvevők elküldik egymásnak szemantikus profiljukat (2.2 Definíció)*
2. *Amennyiben egy nem szomszédos csomópont egy keresés esetén találatot jelez, a válaszüzenethez csatolja a saját reply profilját*
3. *A csomópontok az üzenetekhez csatolhatják a taxonómia általuk használt absztrakciós szintjét.*

Az egyes SemPeer csomópontok a következő szabályok szerint bővítik az alap csomópontok funkcionalitását.

1. *Az egyes csomópontok az általuk tárolt dokumentumok alapján felépítik a szemantikus profiljukat (2.2 definíció). Egy dokumentumból a leginkább jellemző konstans számú metaadatot vagy kulcsszavakat kell kinyerni.*
2. *A keresőkérdésekre érkező válaszok alapján a csomópontok karbantartják a saját query profiljukat (2.5 definíció). Az egyes témakörök súlya alapján meghatározzák a preferált szemantikus kapcsolataikat.*
   3. Amennyiben egy keresőkérdésre válasz(ok) érkeznek a csomópont egy vagy több kapcsolatán keresztül, akkor a csomópont ezek alapján felfrissíti a connection profiljait (2.3 definíció).
4. *Amikor az adott csomópont egy nem szomszédos csomóponttól megkapja annak szemantikus profilját, a 3.1 algoritmus alapján eljárva dönt az adott csomópont kapcsolatok közé való felvételéről.*
5. *Amennyiben a másik csomópont engedélyezi a kívánt kapcsolat létrehozását, és a szemantikus kapcsolatainak száma nem érte még el a paraméterként beállított maximális értéket, az újabb link nehézség nélkül létrehozható. Ha viszont az összes kapcsolati helye foglalt, akkor az új link a meglévő szemantikus szomszédok közül a legkisebb kapcsolati értékkel rendelkező helyére vehető fel. A felvett csomópont szemantikus adatait mindkét esetben el kell tárolni a későbbi összehasonlítások céljából.*

A 4. szabály értelmében a *query* profilt az ismeretlen csomópont *reply* profiljával kell összevetni, amihez a következő algoritmust adtam.

3.1 Algoritmus: SEMPEER_QUERYHIT_HANDLE (Message *m*)

```
1 if neighbor(m.IP) OR NOT Topology_Check(m.ip) return;
2 for all a csomópont t érdeklődési köreire do
3    NewHitRate = GetHitRate(m.profile, t);
4    if ConnectionCount(t) < MaxConnectionCount(t) OR NewHitRate >
       MinHitRate(t) then
5       A választ küldő csomópontot értesítjük a kapcsolatfelvételi igényről
6 endif
```

Az algoritmus először kizárja a már jelenleg is szomszédos csomópontokat a vizsgálatból, valamint megvizsgálja, hogy az új csomóponthoz való kapcsolódás nem sértené-e meg a hálózati topológiát. Ezután megvizsgálja, hogy az új kapcsolat számára van-e még szabad hely a kimenő kapcsolatok listájában, illetve ha nincs, akkor értékesebb lehet-e az új kapcsolat, mint a már meglévőek. Amennyiben az új kapcsolat a kritériumoknak megfelel, a

választ küldő csomópontot értesíti a kapcsolat felvételi igényünkről. A 2.2 altézissel összhangban megmutattam a következőt.

3.1 Altézis: Bebizonyítottam, hogy a 3.1 algoritmus használata nem csökkenti a kapcsolódó csomópontnak az adott témakört érintő találati valószínűségét ($P_{success,t}$).

A protokollnak olyan elemeket is tartalmaznia kell, amelyek a megfelelő hálózati topológia kiépítését biztosítják. A SemPeer esetében a topológiának három szempontot kell figyelembe vennie:
1. *Az érdeklődési területeknek megfelelő kapcsolódásokat ki kell használni*
2. *A kéréstovábbítási gráfban megjelenő csoportképződést ki kell küszöbölni*
3. *Kitűntetett csomópont nélkül, a lehető legkisebb erőforrás-ráfordítással kell felépíteni a topológiát.*

Kidolgoztam a fenti követelményeknek megfelelő *Disjoint Rings* topológiát. A topológia lényege, hogy a kérdéstovábbítási gráf klaszterződéséért felelős kapcsolatokat kiküszöböli, amellett, hogy a hasonló érdeklődési területtel rendelkező csomópontok felé való kapcsolódást nem zárja ki.

3.2 Altézis: A topológia kialakítására a következő algoritmust adtam, majd a topológiával kapcsolatban a következő állításokat igazoltam.

3.2 Algoritmus: TOPOLOGY_CHECK(IP PeerIP)
1 **if** *PeerIP*[4] mod *MinLoopSize* + 1 = *self.ip*[4] mod *MinLoopSize* return true;
2 return false;

Bebizonyítottam, hogy ha $TTL < MinLoopSize < 256$ fennáll, a 3.2 algoritmus használata biztosítja a következőket:
i. a *backward* kapcsolatok száma 0,
ii. a *sibling* kapcsolatok száma 0,
iii. A *skew* kapcsolatok maximális száma a kéréstovábbítási gráfban $n*(k^{(TTL-1)} - 1)$.

Igazoltam továbbá, hogy ha a 3.2 algoritmus használata mellett a csomópontok felismernek és elvetnek minden olyan bejövő kapcsolatot, amelyek duplikált üzeneteket küldenek, akkor ez megszünteti a kéréstovábbítási gráfban a csoportképződést. Ennek következményeként beláttam, hogy a *Disjoint Rings* topológiában a Watts és Strogatz által definiált csoportképződési együttható [55] értéke 0.

A SemPeer protokollt és a Disjoint Rings topológiát alkalmazó hálózat teljesítményét a [13] szimulációs környezetben alkalmazva méréseket végeztem az elért hatékonyság megállapítására. Emellett az analitikus modellt a szimuláció paramétereivel alkalmazva összehasonlítottam a SemPeer és az IdealSON protokollok válaszadási valószínűségét. Az elemzés eredményét a következőkben foglalom össze.

3.3 Altézis: Mérésekkel igazoltam az általam javasolt Sempeer protokoll, illetve Disjoint Rings topológia használatával elérhető teljesítménynövekedést. Konkrétan megmutattam a következőket.

Mérésekkel igazoltam, hogy specializált csomópontok és ritka témakörök esetén a SemPeer protokoll jelentősen jobb válaszadási valószínűséget biztosít, mint a szemantikus információkkal nem rendelkező protokollok.
Szintén mérésekkel igazoltam, hogy a szemantikus réteg gyenge teljesítményt nyújt az érdeklődési területen kívül indított keresőkérdésekre. Emiatt a standard, véletlen linkek és a szemantikus linkek megfelelő arányát kell alkalmazni inhomogén hálózatokban. Ily módon a SemPeer protokoll teljesítménye határozottan magasabb, mint strukturálatlan társaié.
Mérésekkel igazoltam továbbá, hogy az általam javasolt *Disjoint Rings* topológia valóban kiküszöböli a kontraproduktív éleket a kéréstovábbítási gráfban. Szintén megmutattam, hogy a topológia használata nélkül a keresési tér szűkülése miatt gyorsan visszaesik a hálózat teljesítménye.
Az analitikus modell eredményeit mérésekkel összevetve igazoltam, hogy a SemPeer protokoll jól megközelíti az IdealSON által definiált optimális hálózati réteg teljesítményét.
Mérésekkel igazoltam, hogy egy adott válaszadási valószínűség eléréséhez szükséges üzenetek száma jelentősen csökken a SemPeer protokoll használata esetén. Szimulációkkal megmutattam, hogy az általam javasolt szemantikus hálózati rétegben átlagosan a kiinduló csomóponttól kisebb távolságra találjuk meg a keresett dokumentumot, mint az alattas rétegben. A SemPeer protokoll kisebb adatmennyiséggel egészíti ki az üzeneteket, mint más szemantikus protokoll-kiterjesztések.
Szintén mérésekkel igazoltam, hogy a SemPeer protokoll teljesítményére a csomópontok tranziens viselkedése nem gyakorol jelentős hatást.

A harmadik téziscsoport eredményeit a kapcsolódó disszertáció 6. fejezetében fejtem ki részletesen.

IV. Tézis. Az eredmények gyakorlati megvalósításának tulajdonságai

Kapcsolódó publikációk: [4], [1], [8], [21], [22], [23] [24], [26], [28]

Megadtam a III. tézisben kifejtett protokoll-kiterjesztésemnek egy Gnutella-protokollra épülő konkrét megvalósítását, amelyről beláttam, hogy alacsony hálózati többletköltséggel rendelkezik. Megterveztem és megvalósítottam egy moduláris szoftvercsomagot, amelyről megmutattam, hogy különböző alapprotokollokkal lehet használni. Szintén megmutattam, hogy az általam tervezett szoftvercsomag a használt taxonómiától függetlenül alkalmazható. Mérésekkel igazoltam, hogy a SemPeer protokoll kevés memóriakapacitást és számítási teljesítményt igényel a kliens készülékektől. Szintén mérésekkel igazoltam, hogy a protokollom alkalmazása csak alacsony fogyasztási többletet jelent a készülékek oldalán.

A mobil környezetben alkalmazandó P2P protokollnak vannak olyan tulajdonságai, amelyek nemcsak a használt algoritmustól, hanem azok megfelelő megvalósításától is függenek. Nevezetesen a hálózati üzenetek számának alacsony szinten tartása mellett az egyes üzenetek mérete is fontos paraméter. Az elméleti megfontolások alapján alacsony számítási és memóriakapacitást igénylő kiegészítés esetében azt szükséges megvizsgálni, hogy adható-e olyan általános megvalósítás, amely az elterjedt kliens készülékekkel szemben alacsony erőforrásigényt (pl. CPU, memória, felhasznált többlet energia) támaszt.

Ezek fényében terveztem és valósítottam meg a mindmáig legnépszerűbb strukturálatlan P2P protokollra, a Gnutellára [42] épülő **Symella** környezetet, amely

megvalósítása óta az egyetlen natív strukturálatlan P2P kliens megvalósítás a legelterjedtebb smartphone platformra, a Symbian operációs rendszerre. A megvalósítást nyílt forráskódú projektként elérhetővé tettük. Az alkalmazás több mint 50,000 aktív felhasználója lehetővé teszi, hogy további méréseket végezzünk a mobil P2P hálózatok sajátosságaival kapcsolatban.

A szoftvercsomagban élesen elválik az üzleti logika és a felhasználói interfész, lehetővé téve a különböző platformokra való egyszerű hordozhatóságot. Jelenleg az S60 és Series80 felhasználói felületű kliens készülékekre implementáltam a megoldásomat.

A **szemantikus réteget** megvalósító üzleti logikát a szoftvercsomagban **elkülönítettem az alapprotokolltól**, ezáltal lehetővé tettem, hogy a Gnutellát lecserélve más alapprotokollra is alkalmazni lehessen a kiterjesztésemet. Szintén **szétválasztottam a SemPeer kiterjesztés üzleti logikáját és a dokumentumok metaadatait és a belőlük képzett taxonómiát felhasználó modulokat**, lehetővé téve bármilyen metaadat-kinyerő és kezelő algoritmus alkalmazását.

A protokoll-kiterjesztés gyakorlati megvalósítását a Gnutella protokoll szabványos kiterjesztési protokolljával, a GGEP -vel készítettem el.

4.1 Altézis: Mérésekkel igazoltam, hogy a protokoll-kiterjesztés a jelenleg elterjedt smartphone készülékeken (Nokia N80, N95) tipikus felhasználási viszonyok között **kevesebb, mint 2% energiaigény-növekedést okoz** az alapprotokollt megvalósító klienshez képest, mérésekkel alig kimutatható CPU felhasználási növekmény mellett.

4.2 Altézis: Megmutattam, hogy a kérdés jellegű üzenetek (Ping, Query) esetében a kiterjesztésem konstans 11 byte-tal növeli meg az üzenetek méretét. A válasz jellegű üzenetek (Pong, Queryhit) esetében a növekmény mértéke (byte)

$$11 + |l| + N_t * (|t| + |D| + |R|), \qquad (14)$$

ahol l a csomópont által használt absztrakciós szint, N_t az üzenettel küldött érdeklődési területek száma, t egy adott érdeklődési terület, D és R pedig a *reply* profilban szereplő dokumentumszám és valószínűség. A konkrét, különböző profilbeli elemek a számítás tekintetében nem fontosak, csak az adatelemek mérete, amelyet az | | operátorral jeleztem.
Megmutattam, hogy a megvalósítás memóriaigénye a felhasznált taxonómiától függ, és **a smartphone-ok tipikus operatív memóriáját csak nagyon kis mértékben használják**. Konkrétan megmutattam, hogy a zenei stílusok taxonómiáját alkalmazva 1,138 byte, míg a WordNet taxonómiájára építve 162,608 byte memóriaigényre számíthatunk az általam készített megvalósítással.

A negyedik téziscsoport eredményeit a disszertáció 7. fejezetében fejtem ki.

IV. Gyakorlati alkalmazhatóság

A téziseimben végigvezetett eredményekből kitűnik, hogy azok egymásra épülnek, ilymódon lehetővé téve a konkrét implementáció elkészítését. Az egyes téziseimben bemutatott eredmények külön-külön is alkalmazhatóak különféle, az információ-visszakeresést célul kitűző feladatok során. A matematikai modellem alkalmas különböző elosztott protokollok teljesítményének összehasonlítására, illetve a várható eredmények előrejelzésére szimulációk elvégzése nélkül. A modell méretezési és optimalizálási problémákra is megoldást nyújt, amikor például egy adott válaszadási valószínűség vagy maximális üzenetszám mellett kell egy adott rendszer paramétereit meghatározni.

A második tézisben bemutatott profilok olyan problémák esetén alkalmazhatóak, amikor felhasználókat kell modellezni kombinált, tartalmon és együttműködésen alapuló technológiákkal. A felhasználói modellezés az információ-visszakeresés mellett a tartalomajánló rendszerek legfontosabb alapja. Ebből nyilvánvalóan következik az eredményeim szociális hálón alapuló kereskedelmi megoldások környezetében történő alkalmazhatósága.

A harmadik tézisben bemutatott protokoll munkám egyik legfontosabb eredménye, a mobil környezetben történő elosztott információ-visszakeresés alapvetése. A konkrét megvalósításon túl eredményeim eszközt adnak a szociális hálók feletti topológiák hatékony szervezésére is.

A negyedik tézisben bemutatott Symella szoftvercsomag bizonyítja, hogy eredményeim a gyakorlatba is átültethetőek. A szoftvercsomagot publikálása óta több, mint 50.000 felhasználó töltötte le, és a statisztikák, valamint visszajelzések alapján túlnyomó többségük rendszeresen használja is. A szoftvercsomag elkészítésében többek között Kelényi Imre és Csúcs Gergely voltak segítségemre, akik ezért számos cikkben társszerzőim is egyben.

V. Saját publikációk jegyzéke

Könyv, könyvrész

[1] Csúcs G., **Forstner B.,** Charaf H., Marossy K., *Symbian alapú szoftverfejlesztés*, Budapest, Szak Kiadó, 2004, pp. 11-18, 97-157

[2] **B. Forstner,** P. Ekler, I. Kelényi, Bevezetés a mobilprogramozásba. Budapest, Szak Kiadó, 2008. ISBN 978-963-9863-01-9. pp. 33-118.

Könyvfejezet

[3] **B. Forstner**, I. Kelényi, G. Csúcs, *Peer-to-Peer Information Retrieval Based on Fields of Interest.* In Frank H. P. Fitzek (editor), Towards Cognitive and Cooperative Wireless Networking: Techniques, Methodologies and Prospects, Springer Verlag, 2007, ISBN 978-1-4020-5968-1 pp. 235-249

[4] I. Kelényi, G. Csúcs, **B. Forstner**, *Peer-to-Peer file sharing for mobile devices,* In: Frank H. P. Fitzek (editor), Mobile Phone Programming and its Application to Wireless Networking, Springer Verlag, 2007, ISBN 978-1-4020-5968-1 pp. 311-325

Folyóiratcikk

[5] **B. Forstner,** Dr. H. Charaf, Modeling Peer-to-Peer Networks with Interest-Based Clusters, *Transactions on Enformatika, Systems Sciences and Engineering, pp. 38-43, Volume 8, October 2005, ISBN 975-98458-7-3* LR(google scholar)

[6] **B. Forstner,** H. Charaf, Neighbor Selection in Peer-to-Peer Networks Using Semantic Relations, *WSEAS Transactions on Information Science and Applications*, Volume 2, Issue 2, pp. 239-244, February 2005, ISSN 1790-0832

[7] **B. Forstner,** H. Charaf, Probabilistic Model for Semantic Peer-to-Peer Overlay Networks, *WSEAS Transactions on Information Science and Applications*, Volume 3, Issue 4, pp. 691-696, April 2006, ISSN 1709-0832

[8] **Forstner B.,** Kelényi I.: Szemantikus protokollt tartalmazó mobil Peer-to-Peer kliensszoftver. Híradástechnika 2006/9

Nemzetközi konferencia kiadványa

[9] **B. Forstner**, H. Charaf, Applying User Profiles in Mobile Peer-to-Peer Environment, *1st IEEE International Peer-to-Peer for Handheld Devices Workshop at Fifth Annual IEEE Consumer Communications and Networking Conference (CCNC2008)*, 10-12 January, 2008, Las Vegas, USA

[10] **B. Forstner**, I. Kelényi, H. Charaf, Applying User Profiles in Transient Peer-to-Peer Environment, *IEEE Cognitive and Cooperative Wireless NetworksWorkshop (CoCoNet)at IEEE ICC 2008, 19-23 May. 2008, Beijing, China*

[11] **B. Forstner,** H. Charaf, cPEED: A Rapid Web Application Development Framework, *Proc. of Parallel And Distributed Computing And Networks*, Innsbruck, Austria, February 17-19, 2004

[12] **B. Forstner,** H. Charaf, Semantic Peer-to-Peer Information Retrieval, *MicroCAD 2004 International Scientific Conference*, University of Miskolc, Hungary, March 18-19 2004

[13] **B. Forstner,** G. Csúcs, K. Marossy, H. Charaf, Evaluating Performance of Peer-To-Peer Protocols with an Advanced Simulator, *Conference on Parallel And Distributed Computing And Networks*, Innsbruck, Austria, Feb. 15-17, 2005

[14] **B. Forstner,** H. Charaf, Adaptive Peer-to-Peer Network Using Semantic Relations, *IEEE 3rd International Conference on Computational Cybernetics* (ICCC 2005), Mauritius, April 13-16, 2005

[15] **B. Forstner,** H. Charaf, General-purpose Module-based Web Development Environment, *MicroCAD 2005 International Scientific Conference*, University of Miskolc, Hungary, 10-11 March 2005

[16] G. Csúcs, K. Marossy, **B. Forstner,** H. Charaf, An Advanced Simulator for Peer-to-Peer Protocol Analysis, *MicroCAD 2005 International Scientific Conference*, University of Miskolc, Hungary, 10-11 March 2005

[17] **B. Forstner,** H. Charaf, Semantic Profile-based Neighbor Selection in Peer-to-Peer Networks, *MicroCAD 2005 International Scientific Conference*, University of Miskolc, Hungary, 10-11 March 2005

[18] **B. Forstner,** H. Charaf, Modelling Clustered Peer-to-Peer Networks, *IASTED International Conference on Communication Systems and Applications* (CSA 2005), July 19-21, 2005, Banff, Alberta, Canada

[19] L. Lengyel, T. Levendovszky, G. Mezei, **B. Forstner,** H. Charaf, Metamodel-Based Model Transformation with Aspect-Oriented Constraint, *In Proc. of the International Workshop on Graph and Model Transformation (GraMoT 2005)*

[20] **B. Forstner,** Dr. H. Charaf, The Parallel Rings Topology in Semantic Peer-to-Peer Networks, *6th International Symposium of Hungarian Researchers on Computational Intelligence, November 18-19, 2005, Budapest, Hungary*

[21] R. Kereskényi, **B. Forstner,** H. Charaf, Universal communication component on Symbian Series60 platform, *6th International Symposium of Hungarian Researchers on Computational Intelligence, Nov. 18-19, 2005, Budapest, Hungary*

[22] L. Lengyel, T. Levendovszky, G. Mezei, **B. Forstner** and H. Charaf, Towards a Model-Based Unification of Mobile Platforms, *ACS/IEEE International Conference on Computer Systems and Applications*, 3/8/2006 - 3/11/2006, Dubai/Sharjah

[23] **B. Forstner,** L. Lengyel, T. Levendovszky, G. Mezei, I. Kelényi and Dr. H. Charaf, Model-Based System Development for Embedded Mobile Platforms, *Proc. of 13th Annual IEEE International Conference and Workshop on the Engineering of Computer Based Systems (ECBS), March 27th-30th, 2006, Potsdam, Germany*

[24] B. Forstner, L. Lengyel, I. Kelényi, T. Levendovszky and Dr. H. Charaf, Supporting Rapid Application Development on Symbian Platform, *IEEE EUROCON 2005 The*

International Conference on "Computer as a tool". November 21-24, 2005, Belgrade, Serbia & Montenegro

[25] **B. Forstner,** R. Kereskényi, H. Charaf, Optimization of Semantic Peer-To-Peer Network Topology for Mobile Environment, *MicroCAD 2006 International Scientific Conference*, *University of Miskolc, Hungary, 16-17 March 2006*

[26] R. Kereskényi, **B. Forstner,** H. Charaf, Designing a Universal Communication Framework on Different Mobile Platforms, *MicroCAD 2006 International Scientific Conference*, *University of Miskolc, Hungary, 16-17 March 2006*

[27] **B. Forstner,** R. Kereskényi, H. Charaf, Eliminating Clustering in the Propagation Tree of Semantic Peer-to-Peer Networks, *IASTED Conference on Parallel And Distributed Computing And Networks*, Innsbruck, Austria, February 14-16, 2006

[28] R. Kereskényi, **B. Forstner,** H. Charaf, Using Design Patterns in Mobile Communication Software Development, *IASTED Conference on Parallel And Distributed Computing And Networks*, Innsbruck, Austria, Feb. 14-16, 2006

[29] **B. Forstner,** I. Kelényi, G. Csúcs, H. Charaf, *Hybrid Web- and Mobile-based E-learning with Rich Media Support,* In Proc. of Methods, Materials and Tools for Programming Education (MMT2006), April 4-5, 2006

[30] **B. Forstner,** Dr. H. Charaf, Analytical Model for Semantic Overlay Networks in Peer-to-Peer Systems, *4th WSEAS International Conference on Software Engineering, Parallel & Distributed Systems,* Feb. 15-17, 2006, Madrid, Spain

[31] **B. Forstner,** H. Charaf, Bayesian Approach to Improve the Performance of Transient Peer-to-Peer Networks, In Proc. of European Computing Conference, Athens, Greece, September 25-27, 2007.

[32] **B. Forstner,** Imre Kelényi and H. Charaf: Applying User Profiles in Transient Peer-to-Peer Environment. IEEE Cognitive and Cooperative Wireless Networks Workshop 2008. Beijing, China 19-23 May 2008 (submitted)

[33] I. Kelényi, **B. Forstner.** Distributed Hash Table on Mobile Phones. 1st IEEE International Peer-to-Peer for Handheld Devices Workshop. 10-12 January 2008, Las Vegas, Nevada.

[34] I. Kelényi, P. Ekler, **B. Forstner,** A Comparison of Mobile Peer-to-peer File-sharing Clients. *MicroCAD 2008 International Scientific Conference*, University of Miskolc, Hungary, 20-21 March 2008 (submitted)

[35] I. Kelényi, **B. Forstner,** Deploying BitTorrent Into Mobile Environments. In Proc. of European Computing Conference, Athens, Greece, September 25-27, 2007.

Helyi részvételű rendezvény kiadványában megjelent idegen nyelvű előadás

[36] **B. Forstner**: An Analytic Model for Peer-to-Peer Systems with Semantic Overlay Network *In Proc. of AACS'06 Workshop*, Budapest, Hungary June 30, 2006. pp. 11-28

Magyar nyelvű konferenciaelőadás

[37] **Forstner B.,** Elterjedt technológiákra épülő hatékony webfejlesztő keretrendszer és kódgenerátor, *Második Magyarországi PHP Konferencia*, 2004. március 27.

Nyomtatott/elektronikusan közzétett egyetemi jegyzet

[38] **Forstner B.**, Integrált Információs Rendszerek. BME Automatizálási Tanszék, 2003.

VI. Irodalomjegyzék

[39] A. Rowstron and P. Druschel, *Storage management and caching in past, a large-scale, persistent peer-to-peer storage utility.* In Proceedings of SOSP'01, 2001.

[40] Ion Stoica, Robert Morris, David Karger, Frans Kaashoek, and Hari Balakrishnan, *Chord: A scalable peer-topeer lookup service for internet applications.* In Proceedings of SIGCOMM'2001, August 2001.

[41] Ben Y. Zhao, John Kubiatowicz, and Anthony Joseph, *Tapestry: An infrastructure for fault-tolerant wide-area location and routing.* Technical Report UCB/CSD-01-1141, University of California at Berkeley, Computer Science Department, 2001.

[42] The Gnutella homepage, http://rfc-gnutella.sourceforge.net/index.html

[43] Barabási, A. L. (2002). Linked. The New Science of Networks. Cambridge MA: Perseus Publishing.

[44] Jovanovic, M. A., Annexstein, F. S., and Berman, K. A. (2001). Modeling peer-to-peer network topologies through "small-world" models and power laws. In Proc. of the IX. Telecommunications Forum (TELEFOR).

[45] Ge, Z., Figueiredo, D. R., Jaiswal, S., Kurose, J., and Towsley, D. (2003). Modeling peer-peer file sharing systems. In Proc. of INFOCOM 2003, San Francisco, USA.

[46] Yang, B., Garcia-Molina, H.: Efficient search in peer-to-peer networks. In: Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS), Vienna, Austria,. (2002)

[47] S. Sen, Jia Wang, *Analyzing peer-to-peer traffic across large networks*, Networking, IEEE/ACM Transactions on Volume 12, Issue 2, April 2004

[48] K. Sripanidkulchai, B. Maggs, H.Zhang, *Efficient content location using interest-based locality in peer-to-peer systems,* Infocom, 2003

[49] S. S. Shashidhar Merugu and E. Zegura. *Adding structure to unstructured peer-to-peer networks: the use of small-world graphs*, Journal of Parallel and Distributed Computing, 65(2):142-153, Feb, 2005.

[50] Hanhua Chen, Hai Jin, and Xiaomin Ning, *Semantic Peer-to-Peer Overlay for Efficient Content locating*, Proceedings of MEGA'06, Harbin, China, Jan.16-18, 2006.

[51] Yang, B., Garcia-Molina, H.: *Efficient search in peer-to-peer networks.* Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS), Vienna, Austria,. (2002)

[52] Chesnais, S. (2007). The netsize guide. Technical report, Netsize Group, http://www.netsize.com

[53] The Symella homepage, http://www.symella.aut.bme.hu

[54] The WordNet project homepage, http://www.cogsci.princeton.edu/~wn/

[55] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. In Nature, volume 393, pages 440–442.

[56] Creus, G. B. and Kuulusa, M. (2007). Mobile Phone Programming and its Application to Wireless Networking, chapter *Optimizing Mobile Software with Built-in Power*, pages 449–462. Number ISBN 978-1-4020- 5968-1. Springer Verlag.

[57] Flinn, J. and Satyanarayanan, M. (1999). *Energy-aware adaptation for mobile applications*. In Symposium on Operating Systems Principles, pp. 48–63.

[58] Sylvia Ratnasamy, Paul Francis, Mark Handley, RichardKarp, and Scott Shenker, *A scalable contentaddressable network.* In Proceedings of SIGCOMM'2001, August 2001.