Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Automation and Applied Informatics

Semantic Information Retrieval in Mobile Peer-to-Peer Networks

Main Results of the Ph.D. Thesis

Bertalan Forstner

Advisor:
Dr. Charaf Hassan Ph.D.
Associate Professor

Budapest University of Technology and Economics
Department of Automation and Applied Informatics

Budapest, 2008

Main Results of the Ph.D. Thesis

Bertalan Forstner

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Automation and Applied Informatics

Goldmann György tér 3.
Budapest,
H-1111

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

Advisor:

Dr. Hassan Charaf Ph.D.
Associate Professor

I. Preliminaries and Objectives

The problem of information retrieval has been one of the most serious challenges in the history of information technology. With the growing number of networked computers it becomes more and more difficult to find a specific document or other information or resource. One solution for this problem is the separation of the roles of the computers (for example the client-server architecture), where the storage of the documents or their indices is located on dedicated computers with rather huge resources. However, the Peer-to-Peer (P2P) information retrieval systems also have increasing popularity because of their architectural advantages. A Peer-to-Peer network is such a fully distributed architecture where each node has the same role.

When talking about information retrieval we should pay special attention to the mobile telecommunication networks. Since the computing resources and the increased usability of the smartphones make these devices with increased proliferation a good platform for representing different kinds of information, we found it important to involve them into the P2P information retrieval world. Mobile communication costs (in terms of money, bandwidth or battery capacity) are even higher than that of wired communication, therefore, it is more important to use effective P2P protocols in their case. A P2P protocol designed for mobile systems should also suit an important specialty of that environment, namely, they should tolerate the strong transient characteristics of these P2P clients: because of economic considerations and the limited connectivity of such devices, they do not spend much time connected to the network. The connectivity of these devices is limited, because of the network coverage and the limited battery capabilities. [56], [57]. Mobil P2P systems have other properties that should be taken into account when designing protocols for them.

While the server computers in the case of client-server architecture should have rather great amount of resources (for example storage space or network bandwidth), the nodes in distributed networks can participate in storage and request serving with significantly lower resources. Most P2P protocols suffer from scalability issues: with the growth of the number of nodes the amount of required network traffic (or other resources) also increases notably to achieve reasonable hit rate. The efforts dealing with this issue can be classified between two significantly different approaches: they can be structured or unstructured. The structured P2P protocols (for example [39], [40], [41], [58]) specify strict rules for the location of documents to be stored, or define to which other peers a node can connect. Although these networks have usually good scalability properties, and their performance can be estimated quite accurately, they become disadvantageous in networks with strong transient character: they can handle the frequent changes in the network population with difficulties at great resource expenses. The second approach examines unstructured networks such as the basic Gnutella protocol [42]. In that case there is no rule for the location of the documents to store, and the connections of the nodes are controlled by few simple rules. For that reason, these systems have limited protocol overhead and can tolerate when nodes frequently enter and leave the network.

The advantages of the two approaches can be combined with semantic layers. The semantic protocols are based on a phenomenon that can be observed during the everyday life. In the real life people's human relations are not random as in a standard decentralized P2P network. These relations are organized along common interests as similar job, hobby, taste, and other characteristics. We use the wording "fields of interest" to describe this kind of categories later. From these fields of interests, it follows that these people constitute some kind of groups and the communication on the organizing topic is more frequent inside these groups than with the rest of people [43].

Most of the semantic solutions that try to increase the hit rate on reasonable traffic with overlay networks, or SONs (Semantic Overlay Networks), are built on the existing networks, both structured or unstructured. The greater part of the advantages of the unstructured approach (such as the tolerance of very transient presence) can still be applied with the SONs. However, the comprehensive examination of the semantic layer and its effects on these systems has only slight focus from researchers.

The performance analysis of the different protocols can be done most practically by analytical models. However, we face difficulties when regarding the SON's, because the models describing standard unstructured networks are not convenient for considering the special parameters of these protocols. For example, the model of [44] deals with the low level metrics describing the quality of the network, but it cannot be used to examine high-level metrics such as the average answer ratio or the quantity of generated network traffic. Although [45] studies P2P networks with different architectures, this model can be used only in extreme cases to describe semantic network layers. The solution presented [46] examines the effect of clustering in the query-propagation graph to the network traffic, but the depth of the analysis is not sufficient enough for our goals. Similarly, the work in [47] is dealing with the network traffic only. Thus my goal was to construct a model which enables calculating the theoretical maximum hit rate based on the patterns of behavior observed at P2P networks with semantic layers, and to discover the parameters required by the nodes to make local decisions in order to maximize the hit rate in a semantic network.

At unstructured P2P protocols, each node should know the number and location in the network of the distinct documents in each topic in order to construct the optimal semantic layer. This is not the case for different reasons; the most important ones are the transient characteristic of the nodes and the intolerable amount of required network traffic. Therefore, the efficient solution should follow a strategy where the nodes discover the relevant parameters with local observations and decisions.

In my research, I examined different proposals regarding the protocols with certain semantic layers (the most important ones are [48], [49], [50], [51]). I paid special attention measuring how they approach the theoretical maximum of the hit rate, or fulfill the requirements of mobile environments. I found that these examined algorithms were not efficient enough, or had too many prerequisites that cannot be fulfilled in the intended context. The efficiency of the protocols highly depends on the accuracy they use the network parameters learned from the model. Therefore, I had to look for a solution that is built on unstructured networks, can be used in mobile context in flexible manner, and approaches the theoretical hit rate. The examined solutions deal with the very common case when the users (or nodes of the network) are searching for documents in the P2P network that can be unequivocally identified, for example, by knowing the document or song title, file name. Therefore, these works differ significantly from the protocols that help looking up relevant resources by some other kind of structured metadata.

The wide applicability of an advanced mobile solution needs to enable the use of different metadata schemes and base protocols. Beside well designed software architecture it needs the efficient implementation of a resource-aware protocol and algorithm. Although there is demand for such a fully distributed information sharing system with these properties [52], such applications barely can be found.

Concluding the open issues, I have the following problems to investigate:

- What are the most important characteristics of mobile P2P networks that should be taken into account when designing semantic overlay networks for them?

- Which parameters are required to enable the nodes to make local decisions to maximize hit rate, or decrease network traffic, in a given semantic context?
- What recall (query hit ratio) can be achieved theoretically with the described kinds of semantic extensions?
- What is an appropriate algorithm and protocol that can be used with devices with low storing and computing resources? What is the best way to construct and maintain the necessary semantic information for them?
- How can a good topology be shaped for an unstructured and fully decentralized overlay network?
- How can an efficient mobile software architecture be constructed that enables the quick incorporation of different semantic information into the peer-to-peer information retrieval world?

II. Methodological Summary

The method of my research was determined by the listed objectives. I used probability theoretical methods to give a closed mathematical model to determine the answering probability of the network. The model calculates the expectable hit rate based on the parameters relevant from the aspect of the given query. The parameters are set according to typical network extents and user activities. During the construction of the model, my primary objective was to take the special aspects of the semantic overlay network (such as the clustering in the query propagation graph or the limited searching space) into account. A further objective was to be able to identify the parameters required to shape the optimal semantic layer.

I elaborated a method based on Bayesian inference to approximate the parameters that require global information on the network. My method approximates the hit rate expectable through the different connections for the different fields of interest based on the observations of the individual nodes. The *a priori* information is provided by the members of the network.

I verified my model and the performance of the SemPeer protocol with a network simulator [13]. The GXS is a simulation tool that is a result of a research project performed by our department together with the Nokia Research Center. This tool is capable of simulating networks with generic P2P protocols and to collect static and dynamic data. I designed and implemented a module for the GXS that realizes my SemPeer protocol extension. With the extension I was able to analyze the dynamic behavior of the protocol, and I also compared the simulated results to the theoretical maximum hit rate predicted by the model. The simulator requires the same input parameters as the model. To determine the parameters, I have designed and implemented a *crawling* client, which collected usage statistics from ten thousands of users.

To validate the results of the model, I used the Gnutella for reference protocol, as it is done by the most of the researches in this field [42]. My decision is justified by the relatively small protocol overhead, the number of accessible reference measurements and the analytical models available for the protocol. In case of the semantic protocols I performed simulations to validate the model. Based on the results I improved the model and then validated it again.

My research involved also practical results. The Symella software package, designed with software engineering methods, is applied in mobile environment with wireless technology, in order to exploit the advantages of my research [53]. For the further reuse of my software, I used design patterns. The source code of the application, which is available as an open source project on the Internet, is downloaded tens of thousand times. I used the Performance Investigator software package, which was designed for Symbian OS, to measure the hardware and resource requirements (CPU load, memory- and energy needs) of the application. I performed the measurements 14 times per each handset with typical user sessions. I normalized the results by calculating the average of the measurements, omitting the 2-2 extreme values.

III. Novel Scientific Results

The novel results are divided into four theses. The theses show how an efficient distributed information retrieval system can be constructed that takes the characteristics of the mobile environment into account. The most important characteristics are the limited energy-, computing power- and memory capacity of the mobile devices, the cost of communication in money, energy and bandwidth, and the transient behavior of the users, which derives from the previous aspects. In my research I also took the distributedness and scalability of the protocol and the autonomy of the peers into account.

While developing the new protocol it was important to find out the hit rate that we can expect by transforming the network based on semantic data. Therefore, I give an analytical model in Thesis I., which calculates the average probability that a query gets answered by any node, based on the parameters characterizing the nodes and the network, without running any simulations. To let the model deal with the high clusteredness caused by the SemPeer protocol, I introduce a new measure, the modified clustering coefficient. In the first thesis, I show how a protocol can decrease the number of processing the redundant queries caused by clustering, and how can it maximize the number of nodes reached by a query.

From the model we can deduce such global parameters, the local availability of which are required to maximize the performance of the semantic layer. Since obtaining the value of such global parameters in a transient network is a complex and resource-consuming task, I rather give the semantic profiles in Thesis II. I use a local, Bayesian inference-based algorithm to keep the profiles up-to-date. I prove that my algorithm approximates the unknown parameters with low variance.

Thesis III presents the SemPeer protocol and algorithm. I prove that, in order to achieve the optimal hit rate in the semantic layer, my solution transforms the network in such way that the answering probability increases. I also give a topology to the overlay network that is free of clusteredness. I give algorithm to construct the topology, and I prove that it eliminates the counterproductive edges.

I designed a concrete software package and an implementation of the protocol-extension to illustrate my results. In Thesis IV, I show the low resource requirements of my implementation, focusing on the size of the messages, the CPU- and memory needs and the power consumption.

Thesis I. Modeling unstructured peer-to-peer networks with semantic overlay network

Related publications: [5], [7], [18], [30] , [36]

I have given an analytic model for clustered P2P information retrieval systems with semantic overlay networks. I have introduced the modified clustering coefficient to describe the clusteredness of the query propagation graph numerically. I have given the relaxed IdealSON P2P protocol to determine the achievable hit rate in static networks. I

have proven that my analytic model can give the answering probability for unstructured P2P protocols with semantic overlay network, with different user behavior patterns. I have shown which global parameters are required locally to construct the optimal topology. I have shown how a node can increase the performance of the semantic layer based on local information.

In this thesis, I give a theoretical basis of calculating the answering probability of unstructured Peer-to-Peer networks with semantic overlay networks in different circumstances. I examine the achievable performance gain with usage patterns of increasing complexity. This increasing complexity is owed to the appearance of semantic links, the multiplicity of fields of interest per user, and the variety of topics stored by the nodes. I needed the *modified clustering coefficient* to numerically characterize the global effect of the clusteredness caused by the local decisions of the nodes.

1.1 Definition (*Query propagation graph*): A directed simple graph representing the nodes and links that are affected by a query issued from a certain node. The query propagation graph for the node *v* is denoted as G^v.

1.2 Definition (*Query propagation tree*): The directed tree that spans the query propagation graph. The query propagation tree for the node *v* is denoted as G'^v.

In my thesis, I have introduced a new concept, the *modified clustering coefficient*, which goal is to characterize the clustering in the query propagation graph, and to help comparing different P2P protocols. In a query propagation graph, the clustering is caused by the counterproductive edges, through which a query message can return to an already visited node.

1.3 Definition (*Counterproductive edges*): Edges in the query propagation graph that point to vertices that can be reached through an alternative path from the initiating vertex. The alternative path is not longer than the path that contains the counterproductive edge. The set of counterproductive edges in a given query propagation graph is denoted with E_r.

1.4 Definition (Classification of counterproductive edges): We can differentiate between the three following kinds of counterproductive links.

a. *backward edges*: links backwards in the propagation graph. In this case a node forwards a message back to a node that already propagated it in an earlier time.
b. *sibling edges*: links between the nodes on the same level. These cause that node *A* forwards a query to node *B* which received the query after the same number of hops (hop number) than node *A* did.
c. *skew edges*: link to neighbors of a sibling node. In that case a node receives the same query with the same hop number from different nodes.

The clustering coefficient introduced by Watts and Strogatz [55] characterizes the connectedness of the neighbors of a given node with a value between 0 and 1. The modified clustering coefficient, introduced by me, is a value that matches the following criteria.

1.5 Definition *(Properties of the modified clustering coefficient (C_{mod})):* A clustering coefficient C_{mod} is satisfactory to describe a clustered query-propagation graph if it has the following properties:

a. $C_{mod} \in [0,1]$

b. $\mathrm{G}'=\mathrm{G}\cup \mathrm{e}_{\mathrm{i}}\setminus \mathrm{e}_{\mathrm{j}}, \mathrm{e}_{\mathrm{i}}\in \{\mathrm{E}_{\mathrm{r}}\}, \mathrm{e}_{\mathrm{j}}\notin \{\mathrm{E}_{\mathrm{r}}\} \Leftrightarrow \mathrm{C}_{\mathrm{mod,G}}\leq \mathrm{C}_{\mathrm{mod,G'}}$

c. $\mathrm{C}_{\mathrm{mod}}$=0 if the subgraph of the nodes reached by a query constitute a tree.

I have given the following definitions for modified clustering coefficients.

1.6 Definition *(The C_{mod1} modified clustering coefficient):* For a node in the network, we calculate the $C_{mod1,r}$ value for node *r* as follows.

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

$\mathrm{C}_{\mathrm{mod1}}$ is calculated from the average $\mathrm{C}_{\mathrm{mod1,r}}$ values of the individual nodes.

1.1 Subthesis: I have proven that for the Peer-to-Peer protocols where nodes with a nodal degree of *k* forward messages with a given Time-to-Live *TTL* value in the query propagation graph, $\mathrm{C}_{\mathrm{mod1}}$ is a valid clustering coefficient.

During the model construction it was of used to introduce a modified clustering coefficient the value of which is derivable from the number of nodes reached by a query message.

1.6 Definition *(The C_{mod2} modified clustering coefficient):* For a node *r* in the network, we calculate the $C_{mod2,r}$ value as the root of the following equation in interval [0, 1]:

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

where E_q stands for the number of reached nodes by a query.

1.2 Subthesis: I have proven that for the Peer-to-Peer protocols where nodes with a nodal degree of *k* forward messages with a given Time-to-Live *TTL* value in the query propagation graph, $\mathrm{C}_{\mathrm{mod2}}$ is a valid clustering coefficient.

I have defined a relaxed problem for semantic information retrieval in mobile P2P networks, in which, by ignoring some aspects of the real networks, it is easier to achieve the optimal average hit rate in an easier way. I call the relaxed problem and its protocol the *IdealSON.*

1.6 Definition *(The IdealSON relaxed problem):* The IdealSON protocol (or IdealSON problem) is a semantic layer over an unstructured P2P network that has the following characteristics:

a. The nodes in the network are fixed: there are no joins and leaves.
b. The documents stored by each node never change.
c. The connections of the nodes are fixed.
d. The semantic connections of the nodes point to nodes with similar fields of interest.
e. The graph components formed by the nodes with similar interest are coherent.

In case of the IdealSON protocol the aim of the nodes is to maximize the number of distinct documents achievable through a given semantic connection, in the given topic. Being the network static implies that this maximization of the number of documents results in the highest achievable recall value with unstructured, message flooding protocol, with a given classification of the topics. Therefore, the IdealSON is the steady state of the network that is to be achieved by the real network with the help of the protocol extension.

I designed the analytical model to be able to describe nodes both with single or multiple fields of interest. To achieve this goal I had to separate the hit rate in case of off-topic questions and queries for documents in the field of interest of the initiating node. The analytical SON model is produced by combining these 2 cases. In case of a query for a relevant (that is, not off-topic) document, the hit rate can be approximated with higher precision in the semantic layer than in the random links of the base network. Therefore, it is important to forecast the probability of finding a node with the required field of interest in a level of the query propagation tree. Before giving a proposition for this reason, I explain the used notation.

Let P_i stand for the probability that a node on the level i of the propagation tree has the same topic as the originator node. To be able to approximate this value, we introduce $P_{SemLink}$, which denotes the probability that an outlink points to a node with similar topic. $P_{SameTopic}$ is the probability that a random (off-topic) connection points to a node with the same field of interest as the originator.

1.3 Subthesis: I have shown that the probability of finding such a node on the level i of the query propagation tree that can be characterized with the same topic as the query q can be calculated as follows.

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

The value of P_0 is 1 if the topic of the resulting document can be characterized with any of the fields of interest of the initiating node, otherwise P_0 equals 0.

Let t stand for the number of different fields of interest (topics) in the system. In case of the following subthesis I suppose that these topics occur with uniform distribution in the network. Based on Subthesis 1.3, I have proven the following.

1.4 Subthesis: I have shown that in a network that can be characterized with t topics, if a query is initiated for a document that can be described with the field of interest of the initiator node, the following formula holds for the average hit rate:

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

where D_t is the number of different documents in the topic, D_n is a random variable characterizing the distinct documents stored by the node, and D is the total number of documents in the system.

In case of queries for off-topic documents the following statement holds:

1.5 Subthesis: I have shown that if a query is initiated for an off-topic document by a node, the following formula holds for the average hit rate:

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

where D_n is a random variable characterizing the distinct documents stored by the node, and D is the total number of documents in the system.

Based on Subthesiss 1.1-1.5, I have proven the following statement for the average hit rate in an IdealSON system.

1.6 Subthesis: I have shown that if the pattern of the issued queries is known, and the nodes contain off-topic documents with $P_{doc,offtopic}$, the following formula holds:

$$P_{success,IdealSON} = P_{issued,topic} P_{success,topic} + P_{issued,offtopic} P_{success,offtopic} . \tag{6}$$

I have also shown the most suitable strategy to achieve the highest hit rate based on the locally available data.

1.7 Subthesis: I have proven that in absence of knowledge on the total number of documents in a given topic, the recall in the semantic layer can be increased by replacing a connection with another one that provides more distinct documents in its propagation graph, supposing that we can get more relevant documents through the new connection that the other connections cannot deliver:

$$P_{success,t}^{C_1 \cup ... \cup C_l \cup ... \cup C_k} \leq P_{success,t}^{C_1 \cup ... \cup C_j \cup ... \cup C_k} , \forall i, j : D_t^{C_l} \leq D_t^{Cj} . \tag{7}$$

The detailed results of the first thesis can be found in Chapter 4 of my thesis book.

Thesis II. Semantic profiles to characterize the users and query propagation graphs by observations

Related publications: [3], [9], [10], [17], [31], [32]

I have proposed four different profiles to represent and maintain the probability of occurrence of concepts stored in taxonomies. I have elaborated a methodology based on Bayesian process to approximate the probabilities characterizing the fields of interest of the individual users based on local observations. About the profiles I have shown that the expected value of the probabilities calculated by my method has low variance and low computing requirements. I have also given a method based on local observations that approximates and maintains the random variables that characterize the probability of a hit rate in different topics through the query propagation trees.

Thesis I showed us the parameters required to maximize the number of replies for the queries in an ideal environment. To produce such parameters that are not accessible for a node locally, I have introduced four semantic profiles. In practice and during the simulations I used the music genre tag of the MP3 music files or the commonly used WordNet taxonomy [54] for English words as semantic data. The nodes of the profiles are the *synsets:*

2.1 Definition (*Synset*): A set of one or more concepts that are synonyms.

2.2 Definition (*Semantic profile*): The semantic profile is a weighted taxonomy tree, where the nodes represent the different synsets describing the stored documents, along with the number of occurrence of the given synset (D_t^*).

The probability of the occurrence of the topics that describe the shared documents of a node can be read out from the semantic profile. I have shown that if we suppose that the stored documents characterize the fields of interest of the user, then his semantic profile describes these fields of interest on a specific generalization level as accurate as the extracted concepts cover the content of the documents. I use the semantic profile later as a priori information for describing the fields of interest of the user.

The connection profile is used by the nodes to characterize their individual connections.

2.3 Definition (*Connection profile*). The connection profile is a weighted taxonomy, where, along with the concepts, the expectable hit rate of a query for a document that can be described with the given concept *t* through the given connection *C* is stored. This probability is denoted with P_t^C. The maximum number of distinct documents in that topic that can be reached through the given connection is also stored (D_t^C).

I have shown how a connection profile can be maintained with local observations.

2.1 Subthesis: I have proven the following statement. Let α be the number of cases when the connection *C gave* results for a query for a document in the topic *t*, and *β* be the number of observations when it *did not*. In that case $\mathrm{P}_\mathrm{t}^\mathrm{C}$ is a random variable with beta distribution that represents the *belief* of a node that connection *C* gives positive answers in context *t*, therefore the random variable's expectable hit rate in the topic *t* through the connection *C* can be calculated as

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

with the following variance:

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

The importance of this subthesis is that it requires a relatively low number of observations to obtain the expected value of P_t^C with small variance for each topic and connection. In order to enable the nodes to compare the value of an existing connection to a new one, and to enable the fast convergence of the random variables in the connection profile to the value of P_t^C, I introduced the reply profile.

2.4 Definition (*Reply profile*): The reply profile for a node *c* is a taxonomy where, along with each synset *t*, the maximum number of known documents in that topic and the estimation of the recall value from node *c* are represented. This probability is denoted with R_t, or, when it is important to define the corresponding node, it is denoted as R_t^c.

The aim of the reply profile is to help in gathering information on the expectable hit rate through the query propagation graph of a non-neighbor node in the different topics. The reply profiles are initialized by the connection profiles, and their applicability in transient environment is proven in the following subthesis.

2.2 Subthesis: I have proven that if the used protocol ensures that a node maintains the number of semantic connections in a specific topic by supplementing a leaving neighbor by a node with similar field of interest, the connection profile maintains the semantic knowledge of the nodes in the query propagation graph by observing only the standard P2P messages, without additional network traffic.

The following statement ensures the satisfaction of the condition of the proposition.

2.3 Subthesis: I have shown that the value of R_t^c equals with the minimum hit rate achievable through the given connection: $\mathrm{P}_{\mathrm{success,t}} \geq \mathrm{R}_t^c$.

Finally, in order to describe the fields of interest of a given user, I introduced the query profile.

2.5 Definition (*Query profile*). The query profile of a node is a weighted taxonomy where, along with the synsets, a probability is stored which value represents that a query issued by the given node is pointed for a document that can be described by the synset *t*. This probability is denoted as Q_t.

2.4 Subthesis: I have proven the following statement. Let Q_t be a random variable, representing the belief that a query issued by the given node is pointed for a document that can be described by the synset *t*. If, by observing the replies to the queries issued by the node, we obtain that α times the query was pointed to a document with the topic *t* and β times to documents with other topics, then Q_t will be a variable with beta distribution, therefore, its expected value can be calculated as

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

with the following variance:

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

Due to the updates based on the observations, the *query* and *connection* profiles track the changes in the fields of interests of the user or the neighbors. However, because of the short connected period of the users, it is important that the more recent data have greater weight when calculating the value of the variables. Therefore, I introduced an *f* fading factor, which is a floating point number from the interval [0,1]. With the help of this factor I modified the update method of the profiles as follows.

$$(Q_j)' = Beta(f\alpha_j+1, f\beta_j), \ \forall j : j \in \{t \cup hypernyms(t)\} \tag{12}$$

$$(Q_i)' = Beta(f\alpha_i, f\beta_i+1), \ \forall i : i \in \{t \cup hypernyms(t)\} \tag{13}$$

The symbol *hypernyms(t)* stands for the set of synsets that are hypernyms of the topic *t*.
The detailed results of the second thesis can be found in Chapter 5 of my thesis.

Thesis III. Semantic Peer-to-Peer Information Retrieval Protocol and Algorithm

Related publications: [6], [12], [13], [14], [16], [20], [25], [27]

I have given algorithms and protocol extensions to support semantic information retrieval in mobile environment, and I have proven that my solution approaches the performance achieved by the IdealSON protocol. I have given the Disjoint Rings topology and I have given a distributed algorithm to shape that topology. I have shown that this topology is free of counterproductive edges. I have provided measurements that prove that the highest hit rate can be achieved by proper combination of semantic and random links as described in Thesis I. I have proven by measurements that the performance of my protocol cannot be influenced significantly by the transient behavior of mobile peers. I have shown that decrease in the network traffic can be achieved with my protocol extension.

The commonly used unstructured Peer-to-Peer protocols and extensions build on the *ad hoc* standard Gnutella protocol, or on its five message types. The construction of the network is performed in a random manner; therefore the queries reach random nodes in the network. It is reasonable to reorganize the connections of the nodes in such way that the new neighbors can answer to the query with higher probability. This reorganization can happen in a simple way based on statistics on messages or on some kind of description of the fields of interests of the neighboring nodes.

Based on the advantages of different protocol extensions I have elaborated a new protocol layer that takes the characteristics of the mobile networks into account. Since my protocol acts as an additional, semantic layer on the basic protocol, it can be implemented in such a way that the advanced nodes supporting this protocol extension can collaborate seamlessly with the standard clients that know the basic protocol only.

I have elaborated extensions to unstructured P2P protocols and to the operation of the participating nodes.

1. *When connecting, nodes send their reply profiles to each other. (Definition 2.2)*
2. *When a query reaches a hit at a non-neighbor node, the reached node attaches its reply profile to the answer (QueryHit) message.*
3. *The nodes can indicate the actual abstraction level they use with the taxonomy in their messages.*

The operation of the SemPeer-enabled nodes differs in the following rules from the clients that support only the basic Gnutella protocol.

1. *Each node constructs its semantic profile based on the stored documents as described in (Definition 2.2). The most relevant keywords should be extracted.*
2. *Each node maintains its query profile based on the files received for their queries (Definition 2.5). Based on the query profile the nodes identify one or more important fields of interest of the user, and reserve a few outgoing link slots for semantic connections for them. The ratio of the standard links and the semantic links in each topic should be defined according to their weight in the query profile.*
3. *Whenever a node receives a query hit through one or more of its connections, it modifies its connection profiles (Definition 2.3).*
4. *When a node receives a reply profile from another node as part of a QueryHit message, it decides whether to insert that node in its connections list, based on Algorithm 3.1.*

5. *When the candidate node accepts the connection request, and the number of semantic links is less than the predefined maximum for the given topic, the new link can be established without any further considerations. If, however, the maximum number of semantic connections has already been reached, the new link must replace one of the existing connections, which has the lowest connection value (that is, the less similar one). In both situations, the received semantic data should be stored for further comparisons.*

According to rule 4, the query profile should be compared to the reply profile of the new node. I have given the following algorithm to make the comparison.

3.1 Algorithm: SEMPEER_QUERYHIT_HANDLE (Message *m*)

```
if neighbor(m.IP) OR NOT Topology_Check(m.ip) return;
for all identified fields of interest t do
  NewHitRate = GetHitRate(m.profile, t);
  if ConnectionCount(t) < MaxConnectionCount(t) OR NewHitRate >
     MinHitRate(t) then
      The node that sent the profile should be notified about the connection request
endif
```

First the algorithm excludes the nodes from the further examination that are already neighbors or which will violate the topology policy when connected. Then it checks whether the number of connections in the given topic is less than the predefined maximum, or the new connection could be of higher value than the existing connections for that topic. In case the new node suits the criteria, then the originator contacts it with a connection message. With the help of Subthesis 2.2, I have proven the following statement.

3.1 Subthesis: I have proven that Algorithm 3.1 does not decrease the expectable hit rate in the given topic ($P_{success,t}$) for the connecting node.

The protocol should contain parts that ensure the construction of an appropriate topology of the network. In case of SemPeer protocol we have three expectations of the topology:

1. *The advantages offered by the semantic connections should be utilized.*
2. *Clustering in the query propagation path should be eliminated.*
3. *There should not be nodes with special roles to construct the topology. Also, the algorithm that shapes the topology should not use considerable resources.*

I have elaborated the Disjoint Rings topology that matches the above described criteria.

3.2 Subthesis: I have given the following algorithm to construct the topology, and then I have proven the following statements in connection with the topology.

3.2 Algorithm: TOPOLOGY_CHECK(IP PeerIP)

```
if PeerIP[4] mod MinLoopSize + 1 = self.ip[4] mod MinLoopSize return true;
return false;
```

I have proven, that if $TTL < MinLoopSize < 256$ holds, then the use of Algorithm 3.2 ensures the followings:

i. the number of *backward* links is zero,
ii. number of *sibling* links is zero,
iii. and the maximum number of *skew* links equals $n*(k^{(TTL-1)} - 1)$.

Furthermore, I have proven that if the nodes apply Algorithm 3.2 and they recognize and drop incoming connections that send duplicate queries, then the clustering in the propagation path will be eliminated. As a corollary, I have shown that the Clustering Coefficient of Watts and Strogatz [55] equals zero in a network with Disjoint Rings topology.

I have made measurements with the simulating environment described in [13] on the performance of the network using the SemPeer protocol and Disjoint Rings topology to determine the increase in performance. Furthermore, I compared the hit rates of the SemPeer to those of IdealSON protocols, using the analytic model with the parameters of the simulator. The results of the analysis and the measurements can be concluded as follows.

3.3 Subthesis: I have proven by measurements the increase in performance with the application of my SemPeer protocol extension and Disjoint Rings topology. Specifically, I have shown the following.
I have proven by measurements that in case of rare topics and nodes having a specific field of interest the SemPeer protocol extension ensures better hit rate than protocols without using semantic knowledge.
I have also proven by measurements that the semantic layer has poor performance with the off-topic queries. This implies the usage of the standard, random links and the semantic links in the appropriate ratio. This results in significantly better performance in case of using the SemPeer protocol over the use of other unstructured ones.
I have also proven by measurements that my *Disjoint Rings topology* in fact eliminates the counterproductive edges in the query propagation graph. I have shown that, without applying my topology, the performance of the network decreases quickly because of the search space getting narrower.
With the comparison of the results of the analytic model and the simulator, I have shown that the performance of the SemPeer protocol extension approaches that of the network defined by IdealSON.
I have proven by measurements that the number of network messages required to achieve a given hit rate is significantly lower when applying the SemPeer protocol. I have shown by simulations that the relevant document can be reached within fewer hops distance from the originating node in my semantic overlay network than in the base network layer. The SemPeer protocol extends the basic protocol messages with a lower amount of data in average than the other similar semantic protocol extensions.
I have proven by measurements that the transient characteristic of the nodes does not influence the performance of the SemPeer protocol extension.

The detailed results of the third thesis can be found in Chapter 6 of my thesis book.

Thesis IV. The properties of the implementation of the results

Related publications: [1], [4], [8], [9], [10], [21], [22], [23] [24], [26], [28]

I have contributed a specific implementation of my protocol as a generic Gnutella extension about which I have shown that it generates small overhead. I have designed and implemented a modular software package for my extension, which can be used with different base protocols and which is independent from the used taxonomy. I have shown that the implementation of the SemPeer protocol extension has small memory and computing requirements. I have also shown that the implementation does not require significant extra power from the mobile device.

The P2P protocols for mobile environments have such characteristics that depend not only from the used algorithm but from the appropriate implementation. Namely, besides keeping the number of network messages low, the size of the individual messages is also an important aspect. When talking about protocol extensions that, based on theoretical calculations, have low computing resource and memory requirements, it should be examined whether a general implementation can be given that utilizes the resources (CPU, memory, battery power) of the concrete devices well.

Keeping these aspects in mind I have designed and implemented a software package for the most popular unstructured P2P protocol, the Gnutella [42], which package is still the most popular Gnutella implementation for the Symbian OS smartphone platform. We made this implementation, the **Symella**, publicly available as an open source project. The more than 50.000 active user of Symella makes it possible to conduct further measurements on the behavior of mobile P2P users.

In the software package, the code parts for the user interface and business logic are clearly separated, enabling simple porting of the application to different software platforms. At the time of writing, my solution is implemented for the S60 and Series80 platforms.

The business logic implementing the **semantic layer** is also separated from the code of the **base protocol**. This makes it possible to replace the Gnutella protocol with another extendible basic P2P protocol. Furthermore, the business logic of the **SemPeer extension** is also separated from the software modules dealing with the **concept taxonomy and metadata extraction.**

I have implemented the protocol extension with the standard General Gnutella Protocol Extension (GGEP).

4.1 Subthesis: I have proven by measurements that the protocol extension causes less than 2 percent increase in power needs on average smartphone devices (Nokia N80 and N95) compared to the basic Gnutella client, with typical user behavior. Increase in the computing power requirements can hardly be detected.

4.2 Subthesis: I have shown that, in case of Ping and Query messages, my extension increases the message size by a constant 11 bytes. In case of the reply messages (Pong, QueryHit) the increase in bytes can be calculated as

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

where l is the abstraction level used by the node, N_t is the number of fields of interest sent with the message, t is the identifier of a given topic, D and R are the number of documents and the probability value as defined by the *reply profile*. In this calculation, only the size of the message elements is important, which is denoted with the operator | | .

I have shown that the memory requirement of my implementation depends on the used taxonomy, and my package fits very well to the memory size of the smartphones. Specifically, I have shown that the utilization of the taxonomy of music genres requires 1.138 bytes, and the utilization of the WordNet taxonomy requires 162.608 bytes of memory from the device.

The detailed results of the fourth thesis can be found in Chapter 7 of my thesis book.

IV. Application of the Novel Scientific Results

From the results in my theses, it can be seen that they are built on each other, making it possible to design a concrete implementation. However, the results of the individual theses can be applied at different problems from the area of information retrieval. My analytic model can be used to compare the performance of different distributed protocols, or to predict their performance without performing simulations. The model can be applied also when optimizing or instrumenting such systems, for example, to find the optimal parameters for a system with a given minimal hit rate or maximum number of network messages.

The profiles presented in Thesis II are applicable in such type of problems where users should be modeled with combined content-based and collaborating learning methods. User modeling, besides the information retrieval, is the basis of content recommendation systems because they enable to compare the interests of a single user to whole groups of users efficiently, together with the capability to describe a new user with initial profile. Another very important practical result is the ability of quickly constructing social networks, that is, identifying groups of users in a given set of users that match according to given classifying properties.

The protocol presented in Thesis III is one of the most important results of my work, since it forms the basis of distributed information retrieval in mobile environment. Besides the concrete implementation, my results give means to the efficient organization of topology construction on social networks.

The Symella software package presented in Thesis IV is a proof of the practical applicability of my results. The package has been downloaded more than fifty thousand times since its publication and, according to the statistics and feedbacks, most of the users regularly use it. Imre Kelényi and Gergely Csúcs helped me a lot in the implementation of the results; therefore, they are my co-authors in many of the articles about the software package.

V. Related Publications

Books, book parts

[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.

Book chapters

[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

Journals

[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

Publications in International Conference Proceedings:

[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.

Publication in Hungarian Conference Proceedings in English:

[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

Publications in Hungarian Conference Proceedings:

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

Lecture notes published online

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

VI. Citations

[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.