How to model Semantic Peer-to-Peer Overlays?

Christoph Schmitz
FG Wissensverarbeitung, University of Kassel, Germany

Alexander Löser
IBM Almaden Research Center, San José, CA

Abstract: Simulation studies are frequently used to evaluate new semantic peer-to-peer searching techniques. However, one major problem is the lack of a common evaluation standard and of common data sets. Therefore, comparing results and routing approaches from different research groups often is not possible. Based on our recent experiments and simulations we present a common model, evaluation metrics, and data sets for semantic peer-to-peer networks. Ideally, we like to encourage the community to work towards a common evaluation standard.

1 Introduction

Peer-to-peer based knowledge management (P2PKM) applications have received a lot of attention over the past years. Projects such as Edutella, SWAP, or Edamok [NWQ+02, EHvH+03, BBB+04] have combined ideas from the knowledge management, semantic web, and peer-to-peer (P2P) areas, creating tools that facilitate the exchange of ontology-based knowledge modeled using semantic web formalisms.

One particular concern within the area of P2P information management is the efficiency with which a given piece of information can be retrieved from the network. For many application scenarios the retrieval of pieces of data labeled with some binary identifier is sufficient. For those cases, index structures such as distributed hash tables (DHT) have been the most successful approach (see [ATS04] for an overview), as they can give deterministic or stochastic guarantees concerning retrieval costs.

For the scenarios considered in the aforementioned P2PKM applications, however, several of the assumptions used in the DHT world do not hold. Foremost, the knowledge needs of a user may not be easily specified in terms of binary identifiers. Users may not be willing to store information in arbitrary locations in the network, and they will wish to explore the network not only by querying, but also using other modes of interaction which do not fit very well into the idea of data spread across the network according to a hashing scheme.

Thus, most of the work done in the P2PKM area relies on the idea of local routing indices [CGM02a], in which each peers stores metadata about the contents of its neighbors, and a semantic overlay network based on these indices which may be built in a more or less deterministic fashion.

The rest of this paper is structured as follows: in Section 2, we introduce our model for

a P2PKM network. Section 3 formalizes the evaluation of P2PKM networks. Section 4 discusses properties of data sets for P2PKM evaluation.

2 Semantic Peer-to-Peer Overlay Networks

Our view on a peer in a P2PKM network is the following [Sch04, LTQ+05, TSW04]:

- Each peer stores a set of *content items*. On these content items, there exists a *similarity function* sim which can be used to determine the similarity of content items to each other. We assume $1 - sim$ to be a metric.
- Each peer provides a self-description of what it contains, in the following referred to as *expertise*. Expertises need to be much smaller than the knowledge bases they describe, as they are transmitted over the network and used in other peers' routing tables. In our case, the expertise consists content items selected as representatives for the peer, but in general, the expertise could also include peer metadata like query languages supported, additional capabilities of the peer etc. Peer expertises can be compared to each other and to queries using the sim function.
- Each peer knows about a certain set of other peers, i. e., it knows their network address (IP, JXTA ID) and possibly additional information. This relation is said to maintain *shortcuts* to other peers which can be on different levels: a peer may be known to contain information about a certain topic (content shortcut), or to know other peers knowledgeable about a certain topic (recommender shortcut), or to be well connected in the network (bootstrapping shortcut). This routing information is a routing index as proposed by Crespo et al. [CGM02a]. The size of the routing index at each peer is limited to account for the limited amount of memory and processing power.
- Peers query for content items on other peers by sending query messages to some or all of their neighbors; these queries are forwarded by peers according to some *query routing strategy*. It makes use of the information available in the routing index. Furthermore, different routing strategies can be combined in a meta strategy, e. g. in a chain of responsibility (taking the most promising strategies first), or by picking one randomly according to a given probability distribution.
- Peers form new shortcuts based on a *shortcut strategy*. According to this strategy, it is determined which shortcuts and which kind of shortcuts is established, and how the limited space of the routing index is shared among the different kinds of shortcuts.

3 Evaluation Functions and Parameters

Following [ESS+03], this section defines evaluation as a function, which can be instantiated using concrete input parameters and output figures.

3.1 Evaluation as a Function

The intended behavior of a P2PKM system can be described in terms of certain criteria which have to be met in order to satisfy the requirements. On the other hand, the performance of the system will be determined by certain input parameters that govern its behavior and thus the output criteria.

$$(i_1, i_2, \ldots, i_n) \xrightarrow{F} (o_1, o_2, \ldots, o_m)$$

This can be modelled as a function. The function (F) describes the setting and the basic algorithms used, that is, the interior of our system. Different parameters are used as input (i_k) e.g. the routing table size. Specific output figures (o_k) result from it, e. g. the average network load.

3.2 Modeling the Evaluation Function

Modeling a peer to peer network has been investigated and discussed widely in the past within the community. Typically and in particular in our simulation framework, the peer-to-peer network is described using the following sub-models:

- **Content model.** A naive model would distribute the content equally among the peers, which is not true in a real world setting [IRF04]. To model the distribution of files among the peers in a more realistic way, authors either use 1.) synthetic distributions that base on a synthetic model or 2.) crawl/extract data from a real world model.

  Section 4 discusses possible data sets in more detail.

- **Query model.** Very little information is available about which peer issued which query. [SGG03, CFB04] observe that the queries follow a Zipfan distribution. However both studies give little information about which peer issues which query. To our best knowledge we are not aware of a study where queries are mapped to individual peers. In our simulation in [LTQ+05], we choose a uniform query distribution instead of a Zipf distribution, thus the shortcut algorithm does not take advantage of repeated queries for popular topics. Furthermore, there first was a 'learning phase' where the peer network was confronted with half of the possible queries. Then, there was an explicit 'test phase', in which one could observe how the peer network would re-adjust to a second set of queries disjoint from the first. Although studies

of interest shift in file sharing networks show a smoother transition between interests, this drastic shift is better suited to exemplify the behavior of the algorithms [LKXR05].

In the rare case that usage logs for the particular evaluation dataset are available – e.g. if the data stem from a web portal – these logs can be used, of course, to get a query distribution.

- **Gnutella style network model.** The simulation in [LTQ+05] is initialized with a network topology which resembles the small-world properties of file sharing networks as found in Gnutella [IRF04]. We simulated 1024 peers and connect each peer with 5 remote peers on the network layer.
- **Volatile network characteristics.** In [LTQ+05], a first set of simulations assumes a static network, i.e. all peers are always online. In the second round of simulations, we use a dynamic network model observed for Gnutella networks by [SGG03]: 60% of the peers have an availability of less than 20%, while 20% of the peers are available between 20 and 60% and 20 % are available more then 60%. Hence only a small fraction of peers is available more than half of the simulation time, while the majority of the peers is only online a fraction of the simulation time. The real online and offline times of peers are determined randomly at simulation runtime so that their cumulative online times resemble the distribution found in [SGG03].

3.3 Input Parameters

A list of possible input parameters that can be entered into the system will follow:

Number of peers. The size of the peer-to-peer network affects the results of the system. The scalability of the system is represented by this number.

Number of documents or statements. Another type of scalability is checked with this parameter. Whereas peers are physical locations, this parameter describes content objects. They represent the smallest entities in the system.

Network topology. Most of the decisions around topology directly influence the function. But depending on the chosen topology different parameters can be used for further adjustment. When using indexes an important figure is the *index size*. How much content will eventually be stored in the network and how detailed is the knowledge about other peers. Slightly different is the *level of connectivity* or *the size of the routing table*. These are figures representing the characteristics of the network.

3.4 Evaluation Metrics

The evaluation criteria for peer-to-peer (P2P) routing algorithms depend on the requirements of the application scenario. In our setting we are interested in retrieving as many as

possible documents that exactly match the query with a given number of messages. Therefore we limit our hard evaluation criteria to two metrics which are highly dependent on each other:

Recall describes the proportion between all relevant answers in peer network and the retrieved ones. Hereby, we define 'relevant' as ' matches the query exactly'. We imply that a high recall corresponds to a high quality of service. We do not rank the answers and assume that all of them are equally relevant to the query.

Messages per query represent the required search costs. This criteria is used to determine the efficiency of the routing algorithm. The less messages the algorithm produces for one query the more efficient it is.

Note that the precision as used in Information Retrieval does not apply here, as we assume all matches to be relevant. The messages per query, though, can be seen as a similar counterpart to the recall measure, as it expresses the overhead one has to spend in order to retrieve a given number of relevant results.

In [LTQ+05, Sch04] we consider further metrics, such as clustering co-efficient and average network diameter. Although we think these metrics are helpful in understanding the behavior of the system, we believe that they only have an implicit effect on how a user will experience the system.

4 Data Sets

Work in the P2PKM area usually relies on data sets lifted from other contexts, e.g. the Open Directory [LTQ+05] or the ACM Digital Library [SHJS06]. Depending on the origin of the data and their translation into Semantic Web languages such as RDF or OWL, the data sets may exhibit very different characteristics:

Tree or Graph. Many current Semantic Web ontologies are dominated by one or more large tree-shaped taxonomies and contain few or no non-hierarchical relationships. On the other hand, ontologies may evolve which contain more relations across the hierarchy.

A-Box or T-Box emphasis. Real-world knowledge bases often contain either large numbers of concepts and few instances, or a shallow conceptualization and many instances. Depending on which side a usage scenario lies on, certain aspects of the routing strategy, such as the similarity metric, may perform very differently and need to be adjusted.

For our simulation we used different data sets that are available for further experiments: [1]

1. **Bibster.** This data set bases on real query data captured from the P2P bibliography network Bibster [HBM+04]. We use data from observations in a four month period. In this time 520 peers were online. As we logged only the queries and the number of retrieved answers not all shared BibTEX items are available for the simulation setup.

[1] http://ontoware.org/projects/swapsim/, http://www.kde.cs.uni-kassel.de/schmitz/acmdata

From the answers and queries we constructed a data set containing in total 26.173 distinct BibTEX items and 37 distinct classes. The items are classified against the different topics available from an ACM-topic hierarchy.

2. **DMOZ.** Participants of the open directory project (DMOZ) manually categorize Web pages of general interest into a topic hierarchy. Editors contribute links to Web pages, defined subtopics and associate related topics to the DMOZ topic pages. The DMOZ data is available as an RDF dump comprising a small schema and many instances. Our subset consists of 1657 topics and 1024 peers.

3. **ACM.** The ACM Digital Library (ACM DL) [2] contains metadata about papers, their topics, authors, etc. We obtained the metadata about those papers from ACM DL which are present in the DBLP bibliography. This yields information about 39,067 papers about 1232 out of the 1474 ACM topics, written by 53,074 authors. This information was combined into per-author knowledge bases where each author contained the information about all his papers. Both the distribution of papers per author, as well as papers per topic, show a power law distribution.

4. **Synthetic data set.** The data in the Bibster as well as the DMOZ data set is distributed manually by humans working with the respective systems. In order to explore the influence of data distribution parameters on our approach several synthetic data sets are created. Such a dataset comprises an instantiated ontology formalizing the complete knowledge available in the network, and an assignment of knowledge to peers in the network. The number of classes, the number of properties and the number of sub-class relationships together with their respective distributions determine the schema of the ontology. The number of instances and the number of relations between instances determine the distribution of the instance data. The distributions are modelled as Zipf distributions with parameter settings according to observations from real world data sets. The parameter settings for the schema generation are based on [TV03] while the parameter settings for instance generation are based on observation of [CFB04]. Data distribution on the peers follow the model presented in [CGM02b].

Further details about the data sets are described in [LÖ5, Tem06].

As in Section 3.2, other sources of information about the actual distribution of data on peers can be used, if available, to distribute contents on peers. For example, usage logs of the ACM DL website could be used to obtain a realistic distribution of contents over users (i.e. peers).

5 Conclusion

Accurate simulations of semantic peer-to-peer techniques require both data sets and evaluation methods. We have surveyed our work in content query and network modeling, de-

[2] http://www.acm.org/dl

scribed four data sets and our evaluation metrics. However, a common modeling standard for the behavior of peer-to-peer networks and in particular semantic peer-to-peer networks is still an unresolved problem.

The P2PKM area is relatively new and has not settled yet for standard metrics and datasets as in, e. g. the Information Retrieval field. Still, we believe that many P2PKM settings fit into the framework described here if the community agrees on common data sets and an evaluation test bed, and would like to encourage other projects to publish their data sets as well.

Acknowledgement. We thank Christoph Tempich (AIFB Karlsruhe) for his valuable input and the Bibster and DMOZ data sets.

References

[ATS04] Stephanos Androutsellis-Theotokis and Diomidis Spinellis. A survey of peer-to-peer content distribution technologies. *ACM Comput. Surv.*, 36(4), 2004.

[BBB+04] Matteo Bonifacio, Paolo Bouquet, Paolo Busetta, Alberto Danieli, Antonia Donè, Gianluca Mameli, and Michele Nori. KEEx: A Peer-to-Peer Solution for Distributed Knowledge Management. In *Proc. MobiQuitous Workshop on Peer-to-Peer Knowledge Management (P2PKM 2004)*, Boston, MA, USA, August 2004.

[CFB04] Vicent Cholvi, Pascal Felber, and Ernst Biersack. Efficient search in unstructured peer-to-peer networks. *European Transactions on Telecommunications: Special Issue on P2P Networking and P2P Services*, (15):535–548, November 2004.

[CGM02a] Arturo Crespo and Hector Garcia-Molina. Routing Indices For Peer-to-Peer Systems. In *Proc. International Conference on Distributed Computing Systems (ICDCS)*, Vienna, Austria, July 2002.

[CGM02b] Arturo Crespo and Hector Garcia-Molina. Semantic Overlay Networks for P2P Systems. Technical report, Comp. Science Dep., Stanford University, 2002.

[EHvH+03] Marc Ehrig, Peter Haase, Frank van Harmelen, Ronny Siebes, Steffen Staab, Heiner Stuckenschmidt, Rudi Studer, and Christoph Tempich. The SWAP data and metadata model for semantics-based peer-to-peer systems. In *Proc. MATES-2003.*, Erfurt, Germany, September 2003.

[ESS+03] Marc Ehrig, Christoph Schmitz, Steffen Staab, Julien Tane, and Christoph Tempich. Towards Evaluation of Peer-to-Peer-based Distributed Information Management Systems. In *Proceedings of the AAAI Spring Symposium on Agent-Mediated Knowledge Management (AMKM-03)*, Stanford, March 2003.

[HBM+04] P. Haase, J. Broekstra, M.Ehrig, et al. Bibster - A Semantics-Based Bibliographic Peer-to-Peer System. In *3rd. Int. Semantic Web Conference (ISWC)*, 2004.

[IRF04] Adriana Iamnitchi, Matei Ripeanu, and Ian Foster. Small-World File-Sharing Communities. In *23th. IEEE InfoCom HongKong*, 2004.

[LÖ5] Alexander Löser. *Adaptive Overlays in Peer-to-Peer Netzwerken*. PhD thesis, Technische Universität Berlin, 2005.

[LKXR05] J. Liang, R. Kumar, Y. Xi, and K. Ross. Pollution in P2P file sharing systems. In *IEEE INFOCOM*, 2005.

[LTQ+05] Alexander Löser, Christoph Tempich, Bastian Quilitz, Steffen Staab, Wolf Tilo Balke, and Wolfgang Nejdl. Searching Dynamic Communities with Personal Indexes. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, *Proc. 4th International Semantic Web Conference, ISWC 2005*, volume 3729 of *LNCS*, pages 491 – 505, Galway, Ireland, NOV 2005. Springer-Verlag GmbH.

[NWQ+02] Wolfgang Nejdl, Boris Wolf, Changtao Qu, Stefan Decker, Ambjrn Naeve Michael Sintek, Mikael Nilsson, and Tore Risch Matthias Palmr. EDUTELLA: A P2P Networking Infrastructure Based on RDF. In *Proc. 11th International World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, May 2002.

[Sch04] Christoph Schmitz. Self-Organization of a Small World by Topic. In *Proc. 1st International Workshop on Peer-to-Peer Knowledge Management*, Boston, MA, August 2004.

[SGG03] S. Saroiu, P. K. Gummadi, and S. D. Gribble. A measurement study of peer-to-peer file sharing systems. *Multimedia Systems*, 9(2), 2003.

[SHJS06] Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Content Aggregation on Knowledge Bases using Graph Clustering. In *Proceedings of the 3rd European Semantic Web Conference*, Budva, Montenegro, June 2006.

[Tem06] Christoph Tempich. *Ontology Engineering and Routing in Distributed Knowledge Management Applications*. PhD thesis, Universität Karlsruhe (TH), 2006.

[TSW04] Christoph Tempich, Steffen Staab, and Adrian Wranik. REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphers. In *13th World Wide Web Conference (WWW)*, 2004.

[TV03] Christoph Tempich and Raphael Volz. Towards a benchmark for semantic web reasoners - an analysis of the DAML ontology library. In *2nd Workshop on Evaluation of Ontology-based Tools (EON2003) at the 2nd International Semantic Web Conference (ISWC 2003), volume 87. CEUR-Workshop proceedings http://ceur-ws.org.*, 2003.