Towards Self-Organizing Communities in Peer-to-Peer Knowledge Management

Christoph Schmitz

Knowledge and Data Engineering Group, FB 17, Universität Kassel, D-34121 Kassel

Abstract. Recently, many researchers and practitioners have been working at the intersection of knowledge management, peer-to-peer, and online social networks. Results include concrete applications such as P2P knowledge sharing or collaboration platforms on the web (wikis, folksonomy tools). On the other hand, a lot of analyses and models are available that aim at explaining the processes that govern the emergence and self-organization of social networks and their resulting properties, which have proven to be effective in spreading information among the members of the network. From these different perspectives, this position paper outlines the vision of self-organized P2P knowledge management. We point out building blocks which are already available, some of the challenges, and possible solutions.

1 Introduction: Vision

Currently, we are seeing a convergence of research areas such as knowledge management, social networks, peer-to-peer, and the Semantic Web [3]. The vision is that users will be able to have integrated models of their knowledge (documents, e-mails, blogs, etc.) available on their desktop, which can then be linked and exchanged in a peer-to-peer fashion to create online social networks of collaborating peers. The goal of this paper is to outline some important steps towards this kind of peer-to-peer knowledge management in the light of the abovementioned trends and movements.

In [3], Decker and Frank outline their vision of a *Social Semantic Desktop* (SSD). The SSD will combine the ideas of the Semantic Web, P2P, desktop applications such as e-mail, file managers and PIMs, and social networking; it will provide a platform for collaborative work in knowledge-intensive fields. Some of the main features of the SSD are

- wrapping and semantic integration of various knowledge sources using standardized desktop ontologies
- querying and browsing of metadata
- distributed querying and content-based routing
- community building

We will complement this vision with some more detailed aspects of the abovementioned points. As a running example, we will point out some points needed to build a P2P version of a personal KM tool such as Haystack [13].

1.1 Self-Organization and Autonomy

Social networking on the internet is by its very nature a non-deterministic, decentralized process: users decide where they contribute to wiki pages, take part in online communities, or if they maintain their profiles in Orkut.

Possible implementations of the SSD will not only support these processes, e. g. by providing adequate metadata, but may also mimic this social behavior on the technical side. By making use of semantic similarity measures on profile information, the SSD will be able to infer automatically which other peers in the network are likely to be interesting neighbors in the future. Thus, the SSD can proactively create a network topology which enables querying and browsing strategies to exploit locality with respect to topical communities. Furthermore, if the underlying similarity measure reflects the human judgement of similarity, this self-organized community structure can be presented to the user, recommending peers in the network that share the same interests.

Note that a crucial point of an SSD is *autonomy* [10]: while users are willing to share their knowledge, they will like to maintain control over it, just like one keeps in control of one's web pages: others may read, use, and even replicate them, but the owner can control their contents, shut them down, and generally maintain ownership of them. Thus, a self-organized topology in which peers offer their own respective contents and describe it to their neighbors is more suitable in this setting than a Distributed Hash Table approach, in which a homogeneous pool of resources is spread over peers as dictated by the underlying data structures.

1.2 Metadata for Communities and Peers

In order to create and make use of this self-organized community structure, the underlying algorithms need to have a semantic description of the peer's contents and interests. Gathering this kind of profile information is a knowledge acquisition problem, since a user may be unwilling or unable to provide an exact description of his knowledge and his interests such as "this user deals with articles from journals and conferences regarding artificial intelligence". Thus, the SSD will have to augment or replace manual acquisition of profile information by automatically extracting it from the knowledge base or from the user's behavior in the past.

The same holds for communities: having rich metadata about others will enable peers to form communities and to have a notion of what a community is about. Having community descriptions will provide an aggregated view of the network, allowing users to choose which communities to join, detect new trends, or find useful information they did not realize was available.

1.3 Semantic Similarity

A crucial building block in the abovementioned aspects is the notion of semantic similarity. At several points, the similarity of profiles, queries, knowledge items etc. must be assessed. Thus, the SSD application will have to implement metrics for semantic similarity, e. g. for ranking query results or making routing decisions.

2 Towards a Self-Organizing P2PKM system

2.1 Pieces already in Place: Related Work

As described in [3], many of the foundations of a Social Semantic Desktop are available today:

Personal KM Tools Systems such as Haystack or Gnowsis have been built which enable the integration of data such as e-mails, appointments, documents etc. in a semantic desktop environment by adding relevant metadata and thus linking these pieces of information.

Tools for Collaboration, Social Networking, and Folksonomies Today, wikis are ubiquitous, as are social networking projects such as Orkut or FOAF. Newer developments in that area include so-called "folksonomy" projects: del.icio.us, CiteULike, and many others, which allow users to freely create tags describing contents, so that a shared understanding of the respective domains can evolve over time.

Semantic P2P The potential of enriching P2P applications with semantic metadata has been recognized for several years. Since around 2002, projects such as Edutella [12], SWAP [17], or Edamok [1] have pursued the combination of P2P and Semantic Web ideas.

2.2 Missing Links

The abovementioned building blocks as such are not sufficient, however, to build a Social Semantic Desktop. Personal KM tools lack the P2P connectivity, semantic P2P applications are limited to very specific use cases (Bibster) or rather serve as platforms upon which end-user applications need to be built (Edutella), and collaboration platforms are mostly based on centralized servers.

In this chapter, we point out some missing links that need to be filled in order to build personal P2P KM applications which can self-organize into communities of peers with shared interests: Self-organization of peers needs to be further explored; it relies on peers describing themselves semantically, and on metrics on these self descriptions.

Self-Organized Semantic Topologies The P2P network topology of connected P2PKM applications can be structured along a common ontology, putting peers of similar description close to each other, thus enabling community building and semantic routing. Semantic topologies along these lines have been evaluated in [16, 6, 18]. They can be built using only knowledge locally available at the peers with greedy strategies for rewiring into topical clusters, and they have been shown to improve query performance. Some questions regarding self-organized semantic topologies remain open, though:

Parameters of Self Organization In [16], several parameters of the topical clustering process and their influence on query performance have been explored. More insight needs to be gained in the effects of those parameters on the query performance, the proper parameters for a given network, and possible ways of automatically determining appropriate parameters.

Treatment of Literals Using a semantic topology as described above, one can efficiently answer queries related to concepts and relations. It does not, however, account for queries on literal values such as "return all publications by an author called 'Mayer' ", or even "[...] called 'M.*er' ". While fuzzy and range queries in DHTs have been discussed [5], we need to be able to combining conceptual and literal queries.

Scalability While the DHT community has proven the feasibility of DHTs for networks of millions of nodes [8], all implementations in the self-organized semantic P2P area such as presented in [16, 6, 18] are several orders of magnitude smaller than that. It needs to be shown that they can scale to realistic sizes.

Self-Descriptions of Peers To retrieve relevant knowledge in a P2PKM system, one needs to find one or more peers which are able to provide it. For that purpose, there need to be content-based *routing indices.* The question remains, however, of how to obtain the aggregated descriptions of peers needed to establish these routing indices. Such a self-description would outline in a few relevant concepts what kind of knowledge the peer contains. We are currently evaluating the use of graph clustering techniques applied to knowledge bases for this purpose.

Metrics on Knowledge Base Entries In order for a P2P topology to self-organize, and to route queries semantically, a similarity measure on entities from knowledge bases is needed, in order to determine the relevance of results for a user query, the similarity of a query to a peer's self description for routing purposes, or the similarity of peers' self descriptions to each other in order to build the semantic topology. Linguistics and psychology have a history of research on semantic metrics on conceptual knowledge [19, 14]. The knowledge management community has picked up these results and adapted them to its needs [9, 4]. In order for the abovementioned results on metrics to be applied in a self-organized P2PKM setting, several questions remain:

Applicability of these Approaches Many of the existing approaches make assumptions which do not apply to the SSD. For example, works such as [19] take into account detailed linguistic properties of concepts; information theoretic metrics such as [15] assume the availability of full text. Neither of these will always be available in our setting. Thus, existing semantic metrics will have to be adapted.

Similarity of Unaccounted Entities In a distributed P2P setting, only a subset of conceptual entities will be shared by all participants; on this subset, a metric can be agreed upon. The knowledge bases of each peer, however, will contain entities which are only known to that one peer. Hence, one needs to devise methods for computing the similarity of those entities to others.

Meaning of Similarity The abovementioned metrics try to approximate the human understanding of similarity. For the SSD, this goal is only partially valid: while one wants to deliver results which the user regards as "similar" to the intended outcome of a query, the question remains if this notion of similarity is optimal for the self-organization process. Possibly, a different kind of metric could prove to be better for routing and clustering purposes.

Usage of Multiple Metrics It is worth discussing whether the same metric should be used for all of the abovementioned purposes (ranking, routing, self-organization). For focused web crawlers, it has been shown [2] that different metrics can be employed for different purposes to increase the overall performance of the crawler.

Detecting and Labeling Communities There have been efforts aimed at detecting implicit communities in P2P systems [7]. These are based on attributes (keywords) with which the peers describe themselves. Using semantic self descriptions as outlined at the beginning of this section instead would allow for a much richer description of peers and communities. Thus, community detection algorithms will have to be extended or new ones designed which make use of the added possibilities of a conceptual description of peers.

3 Conclusion and Discussion

In order to build an SSD application in which users can share their personal knowledge and form communities around topics, a Haystack-type personal KM tool will have to be equipped with the capabilities necessary to participate in a social network of SSDs. If we get the missing pieces from Section 2.2 into place, the SSD will be able to make decisions about which other peers to recommend to the user, where to forward queries, and how to find other peers, based on self-descriptions of others and a semantic similarity measure on these. Based upon these decisions, it will be able to help building communities of peers (and of the people they represent).

On the other hand, there are other tools available which have quite a head start: in order to offer advantages over the omni-present Google, centralized collaboration tools, and DHT or super-peer based P2P applications for file sharing or instant messaging which have been around for a while, the SSD will have to make good use of a rich, semantically interlinked personal knowledge base which the user can maintain to his own taste, in order to provide better search results, less information overload, less maintenance overhead, and discovery of knowledge sources which the user would not have learned about otherwise.

References

1. Matteo Bonifacio, Roberta Cuel, Gianluca Mameli et al. A peer-to-peer architecture for distributed knowledge management. In *Proc. 3rd International Symposium on Multi-Agent Systems, Large Complex Systems, and E-Businesses MAL-CEB'2002*, Erfurt, Germany, October 2002.
2. Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11–16):1623–1640, 1999.
3. Stefan Decker and Martin R. Frank. The networked semantic desktop. In *Proc. WWW Workshop on Application Design, Development and Implementation Issues in the Semantic Web*, Net York City, NY, May 2004.

4. Marc Ehrig, Peter Haase, Nenad Stojanovic, and Mark Hefke. Similarity for ontologies - a comprehensive framework. In *Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, at PAKM 2004*, DEC 2004.
5. Jun Gao and Peter Steenkiste. An adaptive protocol for efficient support of range queries in dht-based systems. In *Proc. 12th IEEE International Conference on Network Protocols*, Berlin, October 2004.
6. Peter Haase and Ronny Siebes. Peer selection in peer-to-peer networks with semantic topologies. In *Proc. 13th International World Wide Web Conference*, New York City, NY, USA, May 2004.
7. Mujtaba Khambatti, Kyung Dong Ryu, and Partha Dasgupta. Structuring peer-to-peer networks using interest-based communities. In *Proc. International Workshop On Databases, Information Systems and Peer-to-Peer Computing (P2PDBIS)*, Humboldt University, Berlin, Germany, September 2003.
8. Ben Leong and Ji Li. Achieving one-hop DHT lookup and strong stabilization by passing tokens. In *Proc. 12th International Conference on Networks (ICON 2004)*, Singapore, November 2004.
9. Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In *Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002. Madrid, Spain, October 1-4, 2002*, volume 2473 of *LNCS/LNAI*. Springer, 2002.
10. Oscar Mangisengi and Wolfgang Essmayr. P2P knowledge management: an investigation of the technical architecture and main processes. In *Proc. 14th International Workshop on Database and Expert Systems Applications (DEXA 2003)*, Prague, September 2003.
11. Inderjeet Mani and Mark T. Maybury, editors. *Advances in Automatic Text Summarization.* MIT Press, 1999.
12. Wolfgang Nejdl, Boris Wolf, Steffen Staab, and Julien Tane. EDUTELLA: Searching and annotating resources within an RDF-based P2P network. Submitted to IPTPS 2002, December 2001.
13. Dennis Quan and David R. Karger. How to make a semantic web browser. In *Proc. 13th International World Wide Web Conference*, New York City, NY, USA, May 2004.
14. Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, January/February 1989.
15. Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*, Montreal, Canada, 1995.
16. 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.
17. Christoph Tempich, Marc Ehrig, Steffen Staab, et al. SWAP: Ontology-based knowledge management with peer-to-peer. In U. Reimer, A. Abecker, S. Staab, and G. Stumme, editors, *Workshop ontologiebasiertes Wissensmanagement, WM 2003, Lucern*, volume P-28 of *Lecture Notes in Informatics (LNI)*, pages 17–20, Bonn, 2003. Gesellschaft für Informatik.
18. Christoph Tempich, Steffen Staab, and A. Wranik. Remindin': Semantic query routing in peer-to-peer networks based on social metaphors. In W3C, editor, *Proceedings of the 13th International World Wide Web Conference (WWW 2004)*, pages 640–649, New York, USA, MAY 2004. ACM.
19. A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.