Towards Content Aggregation on Knowledge Bases through Graph Clustering

Christoph Schmitz

Knowledge and Data Engineering Group, Universität Kassel
http://www.kde.cs.uni-kassel.de/schmitz

Abstract

Recently, several research projects such as PADLR and SWAP have developed tools like Edutella or Bibster, which have been targeted at establishing peer-to-peer knowledge management (P2PKM) systems. In such a system, it is necessary for participants to provide brief descriptions of themselves, so that routing algorithms or matchmaking processes can make decisions about which communities peers should belong to, or to which peer a given query should be forwarded. In this talk, I propose the use of graph clustering techniques on knowledge bases for that purpose. After a brief round-trip over an ontology-based P2P knowledge management scenario, I will demonstrate the automatic generation of self-descriptions of peers' knowledge bases through the use of graph clustering. Viewing the knowledge base of a peer as a graph consisting of concepts and instances, one can employ clustering techniques to partition it into clusters of similar entities. From each cluster, the centroid can then be selected as a representative. This yields a list of entities giving an aggregated self description of the peer.

1 Ontology-Based P2P Knowledge Management

Recently, a lot of effort has been spent at integrating the upcoming research areas of peer-to-peer systems and the semantic web vision [12, 3, 1, 8], based on a notion of peer-to-peer, personal knowledge management (P2PKM for short). In such a scenario, users will model their knowledge – e.g., metadata on research papers they store on their computers – in personal knowledge bases, which can then be shared with other users via a peer-to-peer network.

One crucial point in such a P2P network is that in order to find relevant material to match a user's query, query messages need to be *routed* to peers which will be able to answer the query without flooding the network with unnecessary traffic.

Several proposals have been made recently as to how the network can self-organize into a topology beneficial for routing, and how messages can be routed in a P2PKM network based on the abovementioned scenario [10, 11, 6, 13]. All of these are based on the idea of routing indices as proposed in [2], adapted to the P2PKM scenario.

1.1 P2P Network Model

Following [10], we thus make the following assumptions about peers in a P2PKM network:

- Each peer stores a set of *content items*. On these content items, there exists a *similarity function* called *sim*. We assume $sim(i, j) \in [0, 1]$ for all items i, j, and the corresponding *distance function* $d := 1 - sim$ shall be a metric. For the purpose of this paper, we assume *content items* to be entities from a knowledge base (cf. Section 2.1), and the metric to be defined in terms of the ontology as described in section 2.2.
- 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 indices. A method of obtaining this expertise is outlined in Section 3.

- There is a relation *knows* on the set of peers. Each peer knows about a certain set of other peers, i. e., it knows their expertises and network address (IP, JXTA ID). This corresponds to the routing index as proposed in [2]. In order to account for the limited amount of memory and processing power, the size of the routing index at each peer is limited.
- 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*. Using the *sim* function mentioned above, queries can thus be compared to content items and to peers' expertises.

1.2 Use Cases

Several use cases for P2PKM as sketched above have been implemented recently. In the PADLR and ELENA projects[1], a P2P infrastructure is established for the exchange of learning material among teachers and students; Bibster[2] is a tool for sharing BibTeX entries between researchers; the SCAM tool[3] for knowledge repositories can act as a peer in a P2P network.

These tools assume that peers agree beforehand on an ontology for describing their contents (e. g. the LOM standard for learning objects or the ACM Computing Classification System), and each peer builds a knowledge base on top of this ontology.

2 Ontologies and Metrics on Knowledge Base Entities

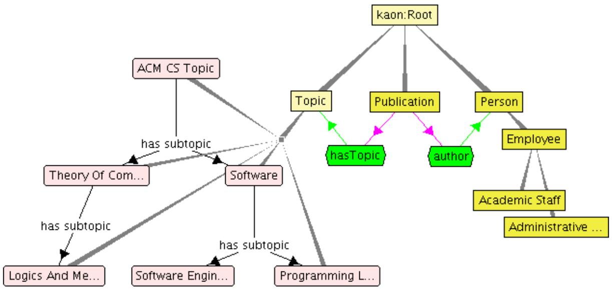

Figure 1: Example Ontology

2.1 Ontology Model

In short, an ontology is a formal conceptualization a group of stake-holders has agreed upon [5]. For the purpose of this paper, we use the view on ontologies proposed by the KAON[4] framework. Basically, an *ontology* consists of concepts with an is-a partial order, and relations between concepts[5]. A *knowledge base* or *OIModel* consists of an ontology and instantiations of concepts and relations. Concepts and instances are both called *entities* (for details cf. [4]).

Two additional important features of KAON OIModels are the attachment of lexical information (labels, descriptions, synonyms, etc.) to entities, and the possibility of nesting OIModels, so that the including model can refer to entities of the included one.

2.2 A Family of Ontology-Based Metrics

An ontology of the kind described above can be viewed as a graph: the set of node comprises the entities, and the relations, relation instances, the is-a and instance-of relationships make up the set of edges. An edge between entities in this graph expresses relatedness in some sense: the instance `phdstud1` may be related to the concepts `PhDStudent` (by an instance-of

[1] `http://www.l3s.de` [2] `http://bibster.semanticweb.org` [3] `http://scam.sourceforge.net/`
[4] `http://kaon.semanticweb.org` [5] We are following the simplified nomenclature of [4].

edge), PhDStudent and Professor would be connected by an edge due to the supervises relation, etc.

On this kind of semantic structure, [9] has proposed to use the distance in the graph-theoretic sense (lengths of shortest paths) as a semantic distance measure. We follow this suggestion and apply it to the abovementioned graph as follows:

- To each edge, a length is assigned; in order to account for the different kinds of relationships, taxonomic edges (is-a, instance-of) get smaller lengths than non-taxonomic edges. This reflects the fact that is-a(PhDStudent,Person) would be considered a closer link between these concepts than, say, rides(Person,Bicycle).
- Edge lengths are divided by the average distance from the root concept of the incident nodes. This reflects the intuition that top-level concepts such as Person and Project would be considered less similar than, e.g., Graduate Student and Undergraduate farther from the root.
- The lengths are normalized such that the longest distance in the graph equals 1, so that $1 - d = sim$ holds.

2.3 Caveats and Pitfalls on Real-World Ontologies

While these strategies of deriving metrics from semantic structures seem straightforward, applying them to ontologies used in real-world applications can turn out to be non-trivial:

Noise and Technical Artifacts Not all of the content of a knowledge base may be genuinely taking part in the user's view of a certain domain; e. g., in KAON lexical information is represented as first-class entities in the knowledge base. This leads to a large number of entities which are not relevant for the semantic distance computation. Similarly, there may a root class which every entity is an instance of, which would render our approach to calculating distances useless.

Modeling Idiosyncrasies Engineering an ontology implies making design decisions, e. g. whether to model something as an instance or as a concept [14]. These decisions carry implications for the weighting of edges, e. g. if a taxonomic relationship is expressed by a special relation which is not one of instance-of, is-a.

To overcome these problems, we have implemented extensive entity filtering and weighting customization strategies which are applied prior to the metric computation itself.

3 Graph Clustering for Content Aggregation

As mentioned above, a peer needs to provide an expertise in order to be found as an information provider in a P2PKM network. From the discussion above, the following requirements for an expertise can be derived:

- The expertise should be provide an aggregated account of what is contained in the knowledge base of the peer, meaning that using the similarity function, a routing algorithm can make good a-priori guesses of what can or cannot be found in the knowledge base.
- The expertise should be orders of magnitude smaller than the knowledge base itself, because it will be used in routing indices.

We propose the use of a version of *k-modes clustering* [7] for this purpose.

3.1 k-Modes Clustering

In short, k-modes clustering works for partitioning a set S of items into k clusters works as follows:

1. Given k, choose k elements of the S as *centroids*
2. Assign each $s \in S$ to the centroid C_i minimizing $d(C_i, s)$

3. For $i = 1 \dots k$, recompute C_i as such that $\sum_{s \text{ assigned to } C_i} d(C_i, s)$ is minimized.
4. Repeat steps 2 and 3 until centroids converge.

This algorithms yields (locally) optimal centroids which minimize the average distance of each centroid to its cluster members. A variation we will use is *bi-section k-modes clustering*, which produces k clusters by starting from an initial cluster containing all elements, and then recursively splitting the cluster with the largest variance with k-modes until k clusters have been reached.

In order to apply this algorithm in our scenario, some changes need to be made:

- The set S to be clustered consists only of those parts of the knowledge base which are not shared between peers; otherwise, the structure of the shared part (which may be comparatively large) will shadow the interesting structures of the private part.
- The centroids will not be chosen from S, but only from the shared part of the ontology. Otherwise, other peers could not interpret the expertise of the peer.

3.2 Example of Knowledge Base Aggregation

As an example, consider a P2PKM network of researchers with knowledge bases about publications according to the ontology shown in Figure 1. Every publication would be related to its corresponding topics from the ACM Computing Classification System[6].

We consider a researcher from DBLP[7] with a sufficient number of papers available on-line, in this case Gruia-Catalin Roman from Washington University. 13 of his papers available with classification at the ACM web site were modeled in a knowledge base, and the clustering algorithm described was run. Table 1 shows some examples of the centroids which were extracted. Note that the k-modes algorithm is non-deterministic because of the random initialization.

k	*Centroids*
2	Network Architecture And Design, Software/Program Verification
2	Requirements/Specifications, Computer-Communication Networks
3	Network Architecture And Design, Requirements/Specifications, Operating Systems
3	Network Architecture And Design, Programming Techniques, Software/Program Verification

Table 1: Centroids for different values of k

A brief examination of the papers and Prof. Roman's home page[8] shows that these centroids indeed reflect his interests of software engineering on the one hand and mobile, distributed computing on the other.

4 Discussion and Work in Progress

In the previous sections, a way of extracting expertises from knowledge bases in a P2PKM setting based on graph clustering was proposed. These expertises consist of entities from the shared part of the ontology which are computed as the centroids in a k-modes clustering procedure. This clustering provides a (locally) optimal set of centroids with respect to the average semantic distance of centroids to knowledge base entities.

While this talk provides anecdotical hints of how the clustering procedure extracts suitable expertises, we are currently conducting a thorough evaluation of this method in conjunction with self-organization techniques for P2PKM networks as described in [10]. Note that usually (e. g. in the text summarization community), the value of aggregations or summaries is measured by evaluating it against human judgment. In our case, however, the aggregations

[6] http://www.acm.org/class/1998/ [7] http://dblp.uni-trier.de

[8] http://www.cs.wustl.edu/~roman/

will be evaluated with regard to their contribution to improving the performance of the P2P network.

Other ongoing research questions include the treatment of literals (e. g. looking for an instance of `PhDStudent` with a last name “Schmitz”) in this metric and the self-organization scheme relying on it, and the formation and labeling of topical communities in the P2PKM network.

References

[1] M. Bonifacio, R. Cuel, G. Mameli, et al. A peer-to-peer architecture for distributed knowledge management. In *Proceedings of the 3rd International Symposium on Multi-Agent Systems, Large Complex Systems, and E-Businesses MALCEB'2002*. Erfurt, Germany, 2002.

[2] A. Crespo and H. Garcia-Molina. Routing indices for peer-to-peer systems. In *Proceedings of the International Conference on Distributed Computing Systems (ICDCS)*. Vienna, Austria, 2002.

[3] M. Ehrig, P. Haase, F. van Harmelen, et al. The SWAP data and metadata model for semantics-based peer-to-peer systems. In M. Schillo, M. Klusch, J. P. Müller, et al. (eds.), *Proceedings of MATES-2003. First German Conference on Multiagent Technologies*, vol. 2831 of *LNAI*, pp. 144–155. Springer, Erfurt, Germany, 2003.

[4] M. Ehrig, S. Handschuh, A. Hotho, et al. KAON - towards a large scale Semantic Web. In K. Bauknecht, A. M. Tjoa, and G. Quirchmayr (eds.), *Proc. E-Commerce and Web Technologies, Third International Conference, EC-Web 2002*, no. 2455 in LNCS. Springer, Aix-en-Provence

[5] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In *Proceedings of the International Workshop on Formal Ontology*. Padova, Italy, 1993.

[6] P. Haase and R. Siebes. Peer selection in peer-to-peer networks with semantic topologies. In *Proceedings of the 13th International World Wide Web Conference*. New York City, NY, USA, 2004.

[7] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.*, 2(3):283, 1998.

[8] W. Nejdl, B. Wolf, C. Qu, et al. Edutella: A p2p networking infrastructure based on rdf. In *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu, Hawaii, 2002.

[9] R. Rada, H. Mili, E. Bicknell, et al. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17, 1989.

[10] C. Schmitz. Self-organization of a small world by topic. In *Proc. 1st International Workshop on Peer-to-Peer Knowledge Management*. Boston, MA, 2004.

[11] C. Schmitz, S. Staab, and C. Tempich. Socialisation in peer-to-peer knowledge management. In *Proc. International Conference on Knowledge Management (I-Know 2004)*. Graz, Austria, 2004.

[12] J. Tane, C. Schmitz, and G. Stumme. Semantic resource management for the web: An elearning application. Proc. 13th International World Wide Web Conference (WWW 2004), 2004.

[13] C. Tempich, S. Staab, and A. Wranik. Remindin': Semantic query routing in peer-to-peer networks based on social metaphors. In W3C (ed.), *Proceedings of the 13th International World Wide Web Conference (WWW 2004)*, pp. 640–649. ACM, New York, USA, 2004.

[14] C. A. Welty and D. A. Ferrucci. What's in an instance? Tech. Rep. #94-18, RPI Computer Science Dept., 1994.