Ontology Extraction using Social Network *

Masahiro Hamasaki[1], **Yutaka Matsuo**[1], **Takuichi Nishimura**[1], **Hideaki Takeda**[2]
(1) National Institute of Advanced Industrial Science and Technology (AIST)
(2) National Institute of Informatics (NII)
{masahiro.hamasaki,y.matsuo,takuichi.nishimura}@aist.go.jp, takeda@nii.ac.jp

Abstract

This paper proposes integration of a social network with the tripartite model of ontologies by P. Mika. That model is based on three dimensions, i.e. actors, concepts and instances, and illustrates ontology emergence using actor-concept and concept-instance relations. However, another important ingredient is the actor-actor relation. For example, a vocabulary is sometimes shared within a community, which consists of dense relations among persons. Through considering of who knows whom (as described in FOAF) and who collaborates with whom, the extracted ontology might be improved. We propose an advanced model based on Mika's work, and describe a case study using the model. We show an application of an extracted ontology for information recommendation for academic conferences.

1 Introduction

Social networks play important roles in our daily lives. They overwhelmingly influence our lives without our knowledge of their implications. Many applications use social networks [Staab *et al.*, 2005]. In the context of the Semantic Web, social networks are crucial to achieve a web of trust, which enables the estimation of information's credibility and a souce's trustworthiness [Golbeck and Hendler, 2005; Massa and Avesani, 2005].

The relation between social networks to ontology emergence is described by P. Mika [Mika, 2005b]. That model takes three dimensions into consideration: actors, concepts, and instances. Two networks are considered in the model. One is the affiliation network of people and concepts. The other is the affiliation network of concepts and instances. The understanding and use of terms might change, reflected in the set of associations between concepts and instances created by users. It provides ontology from two viewpoints: From the affiliation network of concepts and instances, we can extract a classification hierarchy. From the affiliation network of people and concepts, we can extract a hierarchy based on sub-community relationships.

Several researchers have suggested *emergent semantics*. A community might evolve and their commitments might change because members continually leave and enter the group. Towards the dynamic change of ontologies, it is important to develop a method to extract an ontology. The expectation is that numerous individual interactions would engender global effects that are observable as semantics. This approach can support realization of a scalable and easily maintainable Semantic Web. Mika's model provides methods to grasp emergence of semantics through introduction of a structure of members in a community.

Although Mika's study proposes an elegant model of actors-concepts-instances, we can advance it further. We call Mika's model as the Mika model in this paper. One important factor that we explore in our analysis is actors-actors relation. Although Mika mentions *social networks*, the network that was conceptualized in that study is an *affiliation network* (or a two-mode network) that shows the relation between actors and concepts. The affiliation network can be folded to generate an association of concepts in terms of overlapping instances or concepts. We integrate the Mika model with another kind of social network, called an *adjacent network*. It is the social network to which we refer in a more usual senses, e.g., when mentioning social networking in SNSs or a social network by friend-of-a-friend (FOAF) aggregation.

An adjacent network, in which each tie represents a relation between actors such as *knows*, collaborates, and being friends with, can enhance the Mika model. The major advantages of using the Mika model along with the adjacent networks are to give solutions for two problems: the *data sparsity* problem and *word-sense disambiguation*.

In this paper, we propose an advanced model for ontology extraction based on the Mika model. Moreover, we show a case study using the model: emergent ontology at academic conferences. By considering who knows whom (as described in FOAF) and who collaborates with whom, the extracted ontology can be improved. The contribution of this paper is summarized as follows:

- We propose a model called the *HAMA* model, which integrates the Mika model with an adjacent social network dimension.

*This research was partially supported by a Grant-in-Aid for Scientific Research. No. 18700163 (Young-B) provided by Japan Society for the Promotion of Science.

- Our approach underscores the potential importance of the integration of FOAF data with metadata through an example of social bookmarking.

The next section explains the Mika model as a previous work, and our HAMA model. Section 3 shows a case study using the model. It presents an ontology extraction at an academic conference. We used a social network obtained in the community support system for academic conferences called the *POLYPHONET Conference*. We discuss and show related works in section 4, and conclude this paper in Section 5.

2 Tripartite Model with a Social Network

2.1 Mika model: tripartite model of ontologies

Mika proposed a tripartite model of ontologies using a hypergraph. The set of vertices is partitioned into three disjoint sets $A = \{a_1, \ldots, a_k\}$, $C = \{c_1, \ldots, c_l\}$, and $I = \{i_1, \ldots, i_m\}$ corresponding to the set of actors (users), the set of concepts (tags, keywords), and the set of annotated (bookmarks, photographs etc). According to the model, *folksonomies* are regarded as follows: users tag objects with concepts, creating ternary associations between the user, the concept and the object. Consequently, the folksonomy is defined by a set of annotations $T \subseteq A \times C \times I$, consisting of a hyperedge in the whole hypergraph.

This tripartite hypergraph can be reduced into three (two-mode) graphs for each pair among three:

Concept-instance graph (CI) represents the relation of concepts by measuring instance overlaps. It therefore consists a lightweight ontology O_{ci}. We also obtain a network of instances by measuring overlaps of concepts.

Actor-instance graph (AI) shows how actors are related by overlapping instances. We also obtain a network of instances.

Actor-concept graph (AC) shows how actors are related by sharing the same concepts. It therefore consists of a lightweight ontology O_{ac}. It also provides a social network of users based on sets of objects.

Two ontologies O_{ci} and O_{ac} are formulized as follows: We denote a matrix representing the relation of actors and concepts as $\mathbf{B}_{ac} = \{b_{ij}\}$, where $b_{ij} = 1$ if actor a_i is affiliated with concept c_j. Then it can be folded into two graphs: a social network of users based on overlapping sets of objects ($\mathbf{S}_{ac} = \mathbf{B}_{ac}\mathbf{B}_{ac}^T$) and a lightweight ontology of concepts based on overlapping sets of communities ($\mathbf{O}_{ac} = \mathbf{B}_{ac}^T\mathbf{B}_{ac}$). In addition for matrix $\mathbf{B}_{ci}$ where $b_{ij} = 1$ if concept c_i is affiliated with instance i_j, a lightweight ontology of concepts based on overlapping sets of instances ($\mathbf{O}_{ci} = \mathbf{B}_{ci}\mathbf{B}_{ci}^T$) is obtained. The two ontologies are shown and compared in Mika's paper, and a preliminary evaluation was made by questionnaire was made to show the effectiveness of O_{ac}.

2.2 Integration of Adjacent Networks

We expand the Mika model by including relations among actors. In a social tagging system, some users are mutual friends and neighbors that are detectable by aggregating FOAF documents or SNS data. Considering that a Web ontology carries minimal commitment among users in their local interaction, who talks to whom is an important source of information to refine emergent ontologies.

Assume that a researcher is a colleague of another researcher. One might annotate a Web page as "Semantic Web, folksonomy." Then, it is probable that the other will also annotate the page similarly, if not, the annotation is useful for the other. This can provide a solution for a data sparsity problem; the tagging of a user might cover tagging of that user's neighbors. In another example, a set of persons use a tag "network." Some people (in Semantic Web or social science) use the word meaning a social network, but other people (in communication engineering) use the word to denote network infrastructure. This problem is known typically as a word disambiguation problem: a word might have multiple senses, e.g., a *polysemy* (such as "bank as a river edge" and "bank as a financial institution"). But why does this not cause a critical inefficiency in the real world? The reason is that we can use the context of the utterance: surrounding words, previous sentences, situation of the conversation, and so on. Especially, who talks to whom is an important key to facilitate disambiguation. Therefore, if we examine who talks to whom, it is possible to infer more reliably in which sense a user is using the tag: a person who uses "network" as a social network might have neighbors who also use the term in that meaning.

The data sparsity problem can be stated more clearly. Consider a social tagging system. A user tags items, such as Web pages, pictures, academic papers, and so on. The Mika model can obtain the emergent ontology. Nevertheless, a problem arises when each user tags few items, or equivalently, when many more items items must be tagged than the number of users. Using an adjacent network, tags annotated by one can also be used for friends. In other words, concepts (or instances) used by persons in a close social relation can be considered as relevant. This is natural because (lightweight) ontology is seen as a mutual understanding of the lowest necessary for communication [Mika, 2005b]; the adjacent network represents the communication channel. An adjacent social network contributes to the data sparsity problem: we can obtain a more accurate ontology using the network even when tagging data are few.

A concept (represented by a tag or a keyword) might have multiple meanings. It might degrade the correctness of ontology, performance of retrieval, and efficiency of information sharing. Some solutions are available for this: a concept might be inferred as having multiple senses when its instances are not mutually relevant. We must make clusters of instances, judging the content relevancy of clusters. Thereby, we can discern whether there might be multiple meanings or not, which requires complicated algorithms. However, if we bring a social network, it might be solved in a quite straightforward way: if a concept is annotated to different instances by different groups of persons, the concept has different meanings. Otherwise, the concept has the same meaning. This approach is based on the simple fact that people use different labels so that they can discriminate concepts when communicating with their friends and colleagues. Therefore, it is plausible that the neighbors on the adjacent network will come to use a similar label for the same concept, but it does

not necessarily coincide with the label in another community. Such is often the case in academic communities; several well-known examples exist in which the same topic is studied in different communities using different terms; contrarily, where the same word many times represents different meanings in different academic communities.

2.3 The HAMA model: tripartite model with friends and neighbors

Let us formulate our concept following the Mika model. We denote $\mathbf{B}_{ac}$ as a $k \times l$ matrix for O_{ac}, and $\mathbf{B}_{ci}$ as a $l \times m$ matrix for O_{ci}.

We consider the effects imparted by neighbors. We denote a $k \times k$ matrix as $\mathbf{A}_{aa} = \{a_{ij}\}$, which represents relations among actors. It is called an *adjacent matrix*, and each element a_{ij} is binary. Note that $\mathbf{A}_{aa}$ and $\mathbf{S}_{ac}(= \mathbf{B}_{ac}\mathbf{B}_{ac}^T)$ are different although both are a $k \times k$ matrix: the former represents a direct relation among users, whereas the latter represents the similarity of concepts among users.

The resultant matrix $\mathbf{A}_{aa}\mathbf{B}_{ac}$ is a $k \times l$ matrix similarly as $\mathbf{B}_{ac}$ if we take a product of two matrices $\mathbf{A}_{aa}$ and $\mathbf{B}_{ac}$. This matrix sums up the columns of $\mathbf{B}_{ac}$ corresponding to the neighbors of each actor. Then, concept-concept relations are obtained by substituting $\mathbf{B}_{ac}$ with $\mathbf{A}_{aa}\mathbf{B}_{ac}$, i.e., $(\mathbf{A}_{aa}\mathbf{B}_{ac})^T\mathbf{A}_{aa}\mathbf{B}_{ac}$. It represents the relation among concepts measured using common actors considering the neighbors on a social network.

The preceding explanation is straightforward using matrices, but several other operations are also possible for consideration of the actor relation into the Mika model. We can state our model, called the *Homophilic Actor-Mediated Activation model (HAMA model)*, more generally as follows.

The HAMA model

Depending on some distance measures of two actors, we define the matrix $\mathbf{A}_{aa}$; then we modify $\mathbf{B}_{ac}$ considering $\mathbf{A}_{aa}$.

We might apply clustering as a preprocess to the network to discern the forms of communities more clearly.

The HAMA model can mitigate the data sparsity problem. Neighbors on a social network will alleviate the data sparsity problem because tagging information of the neighbors is used.

We argue that the HAMA model is effective under the following assumptions: the neighbors of a person have high similarity of concept usage to the person on average. If this does not hold, the resultant matrix might be inferior to the original. However, when we take the matrix for friends' or colleagues' data, this seems to work well because of the very nature of emergent ontology: lightweight ontology is a mutual understanding among members.

2.4 Concept Gathering

We propose an algorithm for word-sense disambiguation based on the HAMA model, called *concept gathering*.

First, the Mika model can be interpreted as the following procedure.

1. Separate concepts appropriately, and
2. Relate two if they share (more than a certain threshold number of) the same actors/instances.

The resultant relations show the related network of concepts. In other words, we first consider each concept as different and isolated. Then we relate them sequentially.

We can expand the procedure to an extreme. The idea is: to consider *each concept by each actor* as different and isolated, then relate them one by one. We denote the primitive of each concept by each actor as a pre-concept. The procedure is the following.

Concept Gathering algorithm

1. Separate pre-concepts appropriately. There can be $l \times k$ pre-concepts at maximum.
2. Merge two if
   - the labels are the same, and
   - they share the same actors/instances.
3. Regard the cluster of pre-concepts as a concept.

This gathering model will show concepts as overly separated. However, if we integrate the social network of actors, it will be done as the following.

2' merge two if
   - the labels are the same, and
   - they share the same actors/instances or neighboring actors

A concept might have multiple meanings. For that reason, it is important to recognize the sense of a concept by a user.

3 Emergent Ontology at an Academic Conference

In this section, we describe a case study of applying the HAMA model and concept gathering to a research community. We develop an academic conference system called Polyphonet, in which a paper is regarded as an instance, a keyword as a concept, and an author as an actor. Moreover, because Polyphonet is equipped with an SNS function, it provides information related to actor-to-actor relations.

3.1 Polyphonet Conference

Polyphonet Conference (hereafter, Polyphonet) is a community support system for an academic conference. Polyphonet has functions of social networking and an online program of a conference. A user can find research papers and related persons. In the online program part, a user can bookmark interesting presentations (papers, demonstrations and posters) and explore recommended presentations and recommended researchers.

Although Polyphonet does not support tagging to each presentation[1], we regard it as social tagging by the following approximation; we regard each presentation as tagged by the

[1] Actually it does have a tagging function, but not many users used the function.

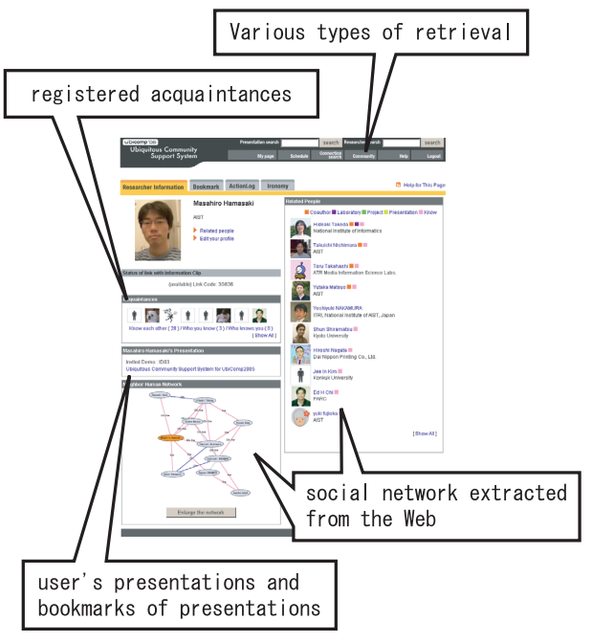

Figure 1: Start Page of *Polyphonet Conference*.

keywords of the presentation, and if a user bookmarks the presentation, it is interpreted that a user has tagged the presentation by the keywords.

In summary, in Polyphonet we obtain the following data, which makes the HAMA model and concept gathering feasible:

- A user can register their friends and acquaintances (as an SNS).
- A user can bookmark presentations, and annotate them by their keywords.

Polyphonet was operated at five conferences to promote participants' communication: 17th, 18th and 19th Annual Conferences of the Japan Society of Artificial Intelligence (JSAI2003, JSAI2004, and JSAI2005) and at The International Conferences on Ubiquitous Computing (UbiComp2005 and UbiComp2006). More than 500 participants attended each conference; about 200 people at each actually used the system. We use the data of JSAI2005 in this paper because those users were the most active among the five conferences.

Similar data are obtainable using SNS + social bookmarking sites. Alternatively, we can aggregate FOAF documents for users of social bookmarking sites. For situations in which Semantic Web proceeds in the future, there might be sufficient tagging data with explicit user profiles obtained using FOAF. Our model would work well in such a situation.

3.2 Ontology Emergence at an Academic Conference

We describe some examples for emergent ontologies for an academic conference. The ontologies represent the relation

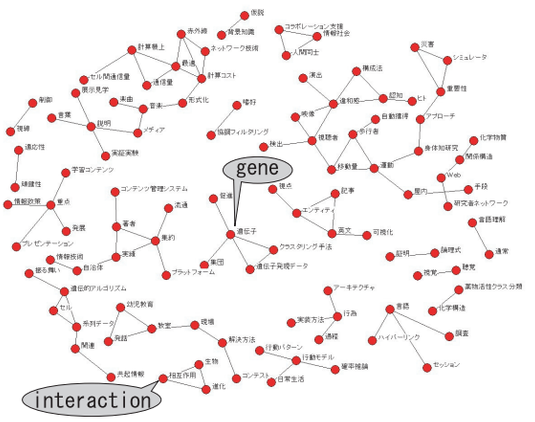

Figure 2: Keyword network extraction with coocurrence in articles (O_{ci}).

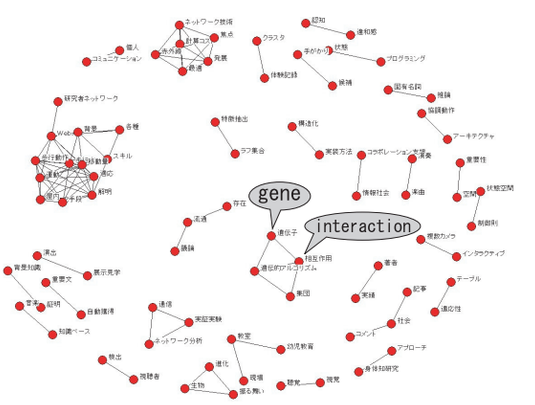

Figure 3: Keyword network extraction with coocurrence in articles (O_{ac}).

among keywords in a paper. Two ontologies of the Mika model applied to JSAI2005 are shown in Figs. 2 and 3. The thresholds are determined so that the numbers of edges are equal (100 edges). Although comparison of two ontologies is certainly difficult, some differences between them are readily apparent. We can infer little connection from words such as "gene" or "interaction" in O_{ci}, but they are more connected in O_{ac}, meaning that the hidden relevance is clarified by considering common actors.

The ontology obtained by the HAMA model is shown in Fig. 4. It is more evenly distributed. By considering friend and acquaintance information, the concepts can be related more closely, demonstrating the model's effectiveness against data sparsity problems.

Table 1 shows network centralization. It takes Bonacich eigenvector centrality [Bonacich, 1972] as a measurement. The centrality of O_{ci} is higher than O_{ac}. General or ambiguous words appear in various contexts in documents. Therefore, such words have many co-occurrences with other words.

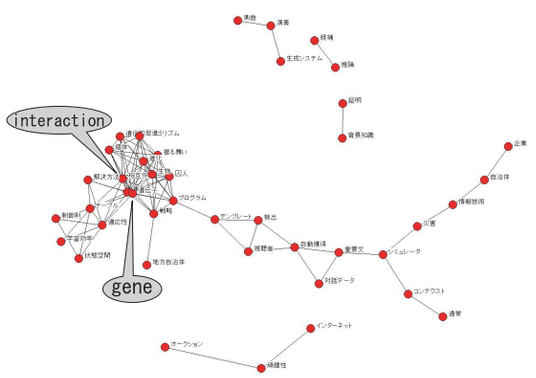

Figure 4: An O_{ac} network with a social network (where the relationship is 'acquaintance').

Table 1: Network centralization based on eigenvector centrality.

	Network Centralization
O_{ci}	103.04
O_{ac}	51.11
Hama model	45.8

The centrality of our HAMA model is less than O_{ac}. The HAMA model forms clusters of words, meaning that words are merged into a cluster based on the social network.

3.3 Evaluation of Concept Gathering

We have 353 keywords for 297 papers overall. There are 4063 pre-concepts, representing tags by different users. The average user has 54.2 pre-concepts. The user with the most has 199 pre-concepts.

There are 1530 concepts if we merge pre-concepts by neighbors. There are 584 concepts if we merge them according to common instances. There are 470 concepts if we merge them according to both rules.

Let us examine the pre-concept "agent". In all, 20 users have the pre-concept "agent" and their pre-concepts are merged into three concepts. The biggest one has 17 users, the second largest has two users, and the smallest has only one user. The biggest indicates "agent" as a program with a function to interact with users. The second indicates "agent" in the context of machine learning. Next, if we examine "game". Six users have the pre-concept "game" and theirs merged into two concepts. On the one hand, the concept relates to "game theory". On the other hand, the concept indicates "game" as a kind of entertainment.

3.4 Application of Lightweight Ontology for Recommendation

It is difficult to compare superiority among ontologies. Therefore, we conducted a task-based evaluation. Extracted lightweight ontology is useful for information recommendation. For users who tag one paper, we can recommend other papers based on an ontology. It will produce a higher rate of acceptance than that of the other if an ontology is superior to another. The assumed hypothesis is: a user will accept with higher probability if the user sees recommendation by instances with ontological proximity.

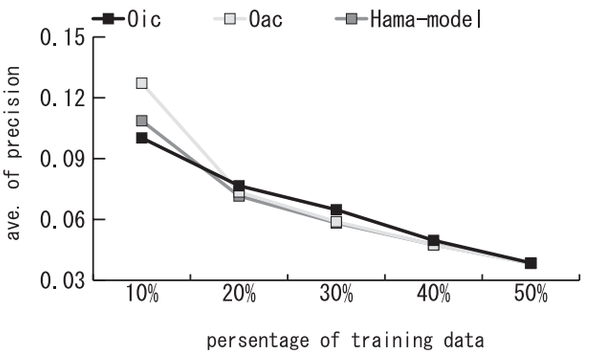

Figure 5: Precision of information recommendation.

We gathered the tagging data and simulated whether a user will accept a recommended instance according to ontological similarity. As a kind of cross-validation, a training set and a test set are used separately. We measure the ontological similarity from the training set, and then provide recommended instances to each user. If a user checked the instance in the test set, it is correct; otherwise it is incorrect. Although users did not tag all the instances in which they might be interested, this figure shows an approximation of the effectiveness of ontologies.

Figure 5 displays the average precision of the recommendation to each user by O_{ci}, O_{ac} and the HAMA model. We must tune up some parameters and thresholds when we adapt these methods to recommendation services. In this case, the system recommends all instances that have an ounce of ontological proximity. For that reason, all precisions are low. The graph indicates that, with increased training data, the precision decreases because the ratio of correct answers (except training data) decreases. All precisions are almost equal among both ontologies. The O_{ac} and O_{ac} with the HAMA model are better than O_{ai} when fewer training data are used.

Figure **??** depicts the average recall of the recommendation against the amount of the training data. It shows that recall of O_{ac} and O_{ac} with the HAMA model is higher than O_{ai} as a whole. Especially, O_{ac} with the HAMA model is better than others when few training data are used. This result indicates the effectiveness of the HAMA model against the data sparsity problem.

4 Discussion and Related Works

Our work is generally based on the work by P. Mika [Mika, 2005b], which we described in section 2. Several investigations have examined extracting word association and general knowledge from the Web. Cimiano proposed a pattern-based approach to categorize instances with regard to an ontology [Cimiano *et al.*, 2005]. It uses statistical information on the Web to annotate Web resources.

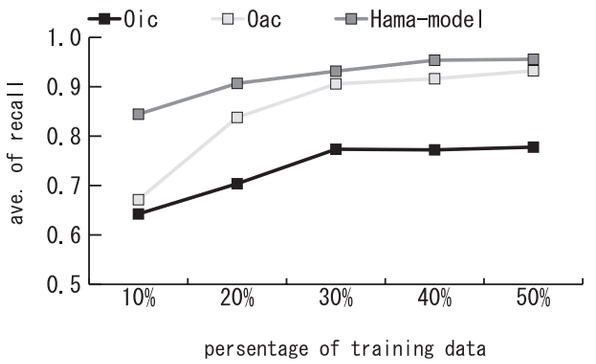

Figure 6: Recall of information recommendation.

In natural language processing research, taxonomy extraction from a corpus has been studied for over a decade. Recently, folksonomy on social tagging is widely examined. Social tagging systems have become increasingly popular because they can improve many information services [Marlow *et al.*, 2006]. Scott analyzed dynamics of social tagging and discovered that stable patterns emerge in tag proportions [Golder and Huberman, 2006]. He pointed out a problem of tag polysemy. Our work has tackled that problem from a social network point of view.

In our work, we use a social network: Various methods exist to obtain social networks. Automatic detection of relations is also possible from various sources of online information such as e-mail archives, schedule data, and Web citation information [Adamic and Adar, 2003; Tyler *et al.*, 2003; Miki *et al.*, 2005]. We can collect FOAF files and obtain a FOAF network [Finin *et al.*, 2005; Mika, 2005a].

5 Conclusion

In this paper, we proposed integration of a social network using Mika's tripartite model of ontologies. That model is based on three dimensions, i.e., actors, concepts and instances, and illustrates ontology emergence by actor-concept and concept-instance relation. We add another important dimension, actor-actor relation, because the HAMA model provides a solution for data sparsity and polysemy problems.

Our case study, an ontology extraction at an academic conference, emphasizes the effectiveness of the HAMA model. We will further investigate the applicability of the HAMA model to different data sets.

References

[Adamic and Adar, 2003] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[Bonacich, 1972] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2:130–120, 1972.

[Cimiano *et al.*, 2005] Philipp Cimiano, Siegfried Handschuh, and Steffen Staab. Towards the self-annotatin web. In *Proceedings of WWW2004*, 5 2005.

[Finin *et al.*, 2005] Tim Finin, Li Ding, and Lina Zou. Social networking on the semantic web. *The Learning Organization*, 2005.

[Golbeck and Hendler, 2005] Jennifer Golbeck and James Hendler. Inferring trust relationships in web-based social networks. *ACM Transactions on Internet Technology*, 2005.

[Golder and Huberman, 2006] Scott Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[Marlow *et al.*, 2006] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Tagging papers, taxonomy, frickr, academic article, toread. In *Proceedings of Hypertext2006*, 2006.

[Massa and Avesani, 2005] Paolo Massa and Paolo Avesani. Controversial users demand local trust metrics: an experimental study on epinions.com community. In *Proceedings of AAAI-05*, 2005.

[Mika, 2005a] Peter Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3, 2005.

[Mika, 2005b] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of ISWC2005*, 2005.

[Miki *et al.*, 2005] Takeru Miki, Saeko Nomura, and Toru Ishida. Semantic web link analysis to discover social relationship in academic communities. In *Proccedings of SAINT2005*, 2005.

[Staab *et al.*, 2005] Steffen Staab, Pedro Dmingos, Tim Finin, Peter Mika, Anupam Joshi, Jennifer Golbeck, Andrzej Nowak, Li Ding, and Robin R. Vallecher. Social network applied. *IEEE Intelligent systems*, pages 80–93, 2005.

[Tyler *et al.*, 2003] Josh Tyler, Dennis Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. *Communities and technologies*, pages 81–96, 2003.