Ontology-Aware Classification and Association Rule Mining for Interest and Link Prediction in Social Networks

Waleed Aljandal waleed@ksu.edu
Vikas Bahirwani vikas@ksu.edu
Doina Caragea dcaragea@ksu.edu
William H. Hsu bhsu@ksu.edu

Department of Computing and Information Sciences, Kansas State University
234 Nichols Hall, Manhattan, KS 66506-2302

Abstract

Previous work on analysis of friendship networks has identified ways in which graph features can be used for prediction of link existence and persistence, and shown that features of user pairs such as shared interests can marginally improve the precision and recall of link prediction. This marginal improvement has, to date, been severely limited by the flat representation used for interest taxonomies. We present an approach towards integration of such graph features with ontology-enriched numerical and nominal features (based on interest hierarchies) and on itemset size-sensitive associations found using interest data. A test bed previously developed using the social network and weblogging service *LiveJournal* is extended using this integrative approach. Our results show how this semantically integrative approach to link mining yields a boost in precision and recall of known friendships when applied to this test bed. We conclude with a discussion of link-dependent features and how an integrative constructive induction framework can be extended to incorporate temporal fluents for link prediction, interest prediction, and annotation in social networks.

Keywords: ontology-aware classification, association rule mining, link prediction, interest prediction, social networks, constructive induction, annotation

1 INTRODUCTION

This paper presents an integrative, ontology-enriched framework for link prediction in social networks. The framework combines previously-developed approaches for feature construction and classification – namely, computing topological graph features [HKP+06], shared membership counts [BCA+08], and aggregates across all shared memberships [AHB+08]. It augments them with a ontology extraction mechanism based on partitioning and agglomerative hierarchical clustering. [BCA+08] This mechanism extends our feature construction task to a more general one of feature extraction, while enabling it to handle *diverse* memberships, such as: interests that a user can hold, communities he or she can belong to, *etc*. Previous work has focused more on scaling up the algorithms for graph feature construction from hundreds of users to thousands [HLP+07] and applying association rule mining to construct features useful in estimating link probability and strength [AHB+08].

Meanwhile, the ontology-aware classification approach used by Bahirwani *et al.* [BCA+08] has incorporated generic glossaries and other definitional data sources such as *WordNet-Online*, the *Internet Movie Database (IMDB)*, and *Amazon Associates' Web Service (AWS)*. These methods laid the groundwork for an integrative feature extraction system, considerably improving the precision and recall of link prediction. [BCA+08, AHB+08] There are, however, several directions where further technical advances are needed in order to make the classification-based prediction approach effective within a recommender system. These include:

1. Extension of the framework to include interest prediction, based on an itemset size-adaptive measure of interestingness for association rules generated in frequent pattern mining [AHB+08]
2. Extension of the framework to incorporate technical glosses
3. Application of the framework to incorporate semantic metadata regarding user profiles: specifically, schemas describing eligible interests and memberships that a pair of candidate users can have in common
4. Feature selection, extraction, and discovery methods that are sensitive to recommendation context and able to leverage the above metadata
5. Development of data description languages using description logics that can capture fluents such as set identity over time

We first define our link prediction framework, then the holistic framework for ontology-enriched classification. Next, we describe our social network test bed in brief, and report new positive results after extending the framework along the first aspect above. Finally, we discuss the data integration and modeling operations needed to implement the other four using present-day social networks and Semantic Web representations such as OWL.

2 BACKGROUND

2.1 Friendship Networks in Social Networks

Most social networking services include friend-listing mechanisms that allow users to link to others, indicating friends and associates. Friendship networks do not necessarily entail that these users know one another, but are means of expressing and controlling trust, particularly access to private content. In blogging services such as SUP's *LiveJournal* or *Xanga*, this content centers on text but comprises several media, including: interactive quizzes, voice posts, embedded images, and video hosted by other services such as *YouTube*. In personal photograph-centric social networks such as News Corporation's *MySpace*, *Facebook*, Google's *Orkut*, and Yahoo's *Flickr*, links can be annotated ("How do you know this person?") and friends can be prioritized ("top friends" lists) or granted privileges as shown in Figure 1.

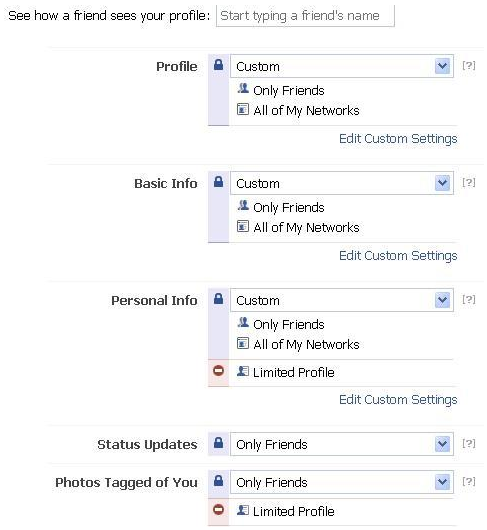

Figure 1. An excerpt of *Facebook*'s access control lists for user profile components. © 2008 Facebook, Inc.

Some vertical social networks such as *LinkedIn*, *Classmates.com*, and *MyFamily.com* specialize in certain types of links, such as those between colleagues, past employers and employees, classmates, and relatives. As in vertical search and vertical portal applications, this specialization determines many aspects of the data model, data integration, and user knowledge elicitation tasks. For example, *LinkedIn*'s friend invitation process requires users to specify their relationship to the invited friend, an optional or post-hoc step in many other social networks.

Friendship links can be undirected, as in *Facebook* and *LinkedIn* (requiring reciprocation, also known as confirmation, to confer access privileges) or directed, as in *LiveJournal* (not necessarily requiring reciprocation).

2.2 Prediction Tasks: Link Existence and Persistence

Link analysis techniques, such as supervised learning of classification functions for predicting link existence [HKP+06, HLP+07] and persistence [HWP08], have been applied to social networks such as *LiveJournal*. This approach is based on inductive generalization over three types of features:

1. **node-dependent**: specific to a user *u* to whom a friend is being recommended, or to a recommended user *v*
2. **pair-dependent**: based on co-membership of *u* and *v* in a domain-specific set (see below)
3. **link-dependent:** .based on annotation of known relationships, or aggregation between them in the entity-relational data modeling sense

Examples of pair-dependent attributes include measures of overlap among common:

- interests
- communities, forums, groups
- fandoms (*fan of*), endorsements (*supporter of*)
- institutions (schools, colleges and universities, companies, etc.)

Measures of overlap depend on the abstract data type of the attributes. For interests, communities, fandoms, and endorsements, they are most often simple counts – that is, the size of the intersection of two users' membership sets, computed by string comparison. Overlap can, however, be a weighted sum of similarity measures between concepts; our focus in this paper is the development and application of concept hierarchies based on such measures. For institutions, the base types for computing overlap can be intervals – typically, the time periods that two people were both at a university or company.

Most features for link prediction are node-dependent or pair-dependent. For example, Hsu *et al.* derived seven topological graph features and five interest-related features of potential relevance to link existence prediction in *LiveJournal*'s directed friendship network. [HKP+06] They then used supervised inductive learning over pairs of candidate features known to be within two degrees of separation to find discriminators between direct friends and "friend of a friend" (FOAF) pairs within a limited *LiveJournal* friendship graph, initially containing 1000 users [HKP+06] that was later extended to 4000 users [HLP+07]. In later work [HWP08], they extended the "friend vs. FOAF" task to predicting day-by-day link persistence in a time series of repeated web crawls.

Computation of topological graph features, such as the degree of separation (shortest path length) between a pair of users, can yield information such as alternative paths as a side effect. Figure 2 illustrates one use of such information in the professional social network *LinkedIn*.

In this paper, we focus on link existence prediction between users and between interests. The task is formulated as follows: given a graph consisting of *all other extant links*, specify for a given pair (*u*, *v*) that are either friend (distance 1) or FOAF (distance 2), the true

distance. For user-to-user links, our experiments are conducted using this "friend vs. FOAF" task over a 1000-node *LiveJournal* data set created by Hsu *et al.* We seek to improve the precision and recall of link existence prediction beyond that achieved using node-dependent and pair-dependent features on flat interest hierarchies.

Figure 2. Minimal-length paths for a third-degree connected pair in *LinkedIn.* © 2008 LinkedIn, Inc.

2.3 Link Mining

Link mining refers to the problem of finding and analyzing associations between entities in order to infer and annotate relationships. It may therefore require data modeling, integration, and mining by means of machine learning from known or putative links. The links can be user-specified, as for the social networks discussed earlier in this section, or latent for text information extraction tasks such as that of McCallum *et al.* [MWC07], who used the Enron e-mail corpus to infer roles and topic categories. For a much more complete survey of link mining approaches that emphasizes statistical relational learning approaches and graphical models, we refer the interested reader to Getoor and Diehl [GD05].

Ketkar *et al.* [KHC05] compare data mining techniques over graph-based representations of links to first-order and relational representations and learning techniques that are based upon inductive logic programming (ILP). Sarkar and Moore [SM05] extend the analysis of social networks into the temporal dimension by modeling change in link structure across discrete time steps, using latent space models and multidimensional scaling. One of the challenges in collecting time series data from LiveJournal is the slow rate of data acquisition, just as spatial annotation data (such as that found in LJ maps and the "plot your friends on a map" meme) is sparse.

Popescul and Ungar [PU03] learn an entity-relational model from data in order to predict links. Hill [Hi03] and Bhattacharya and Getoor [BG04] similarly use statistical relational learning from data in order to resolve identity uncertainty, particularly coreferences and other redundancies (also called deduplication). Resig *et al.* [RDHT04] use a large (200000-user) crawl of *LiveJournal* to annotate a social network of instant messaging users, and explore the approach of predicting online times as a function of friends graph degree.

3 METHODOLOGY

3.1 Ontology-Aware Link Mining

Hsu *et al.* reported a near-baseline accuracy of 88.5% and very low precision of 4.5 – 5.4% for the *LiveJournal* link existence prediction task using shared interests alone. [HPL+07] They report that adding shared interests to graph features yielded an incremental improvement of 6.5% in precision (from 83.0% to 89.5%) for decision trees, which achieved the best baseline and final precision on cross-validation data. This illustrates that using literal string equality to compute "similarity of interest sets" between two users does not result in effective features for predicting link existence in the friends network of *LiveJournal*. We hypothesized that this was due to the semantically limited similarity measure and that a measure based on an ontology, such as a concept hierarchy of interests as depicted in Figure 3, would yield further improvement. [BCA+08]

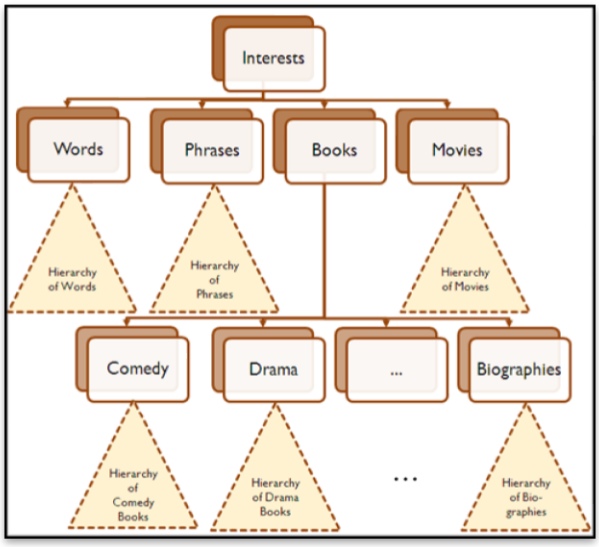

Figure 3. Concept hierarchy of interests. [BCA+08]

By applying unsupervised learning to a complete lexicon of interest terms, with reference dictionaries as sources of knowledge about term similarity, we constructed two types of interest-based features:

- **nominal**: measured for grouped relationships for a candidate pair of entities by name (e.g., are *u* and *v* both interested in topics under the category of mobile computing?)
- **numerical**: interestingness measures that are computed across these grouped relationships (e.g., *how many* interests that *u* is interested does *v* share, and *how rare* are these interests?)

All features in these two categories are examples of pair-dependent co-membership features as discussed in Section 2.2, and can be computed using the ontology.

3.2 Nominal Features: Abstraction using Ontologies

As in many social networks, the number of distinct interests in *LiveJournal* ranges from thousands to hundreds of thousands as the number of users grows. The bit vector for all "shared interests" between users becomes so large and sparse that for nominal interests it is *only* feasible to use an ontology. Rather than continue to use literal string equality, which results in this overly stringent and sparse representation, we clustered interests to form a concept hierarchy and used the aggregate distance measure between user interests to more accurately determine their degree of interest overlap.

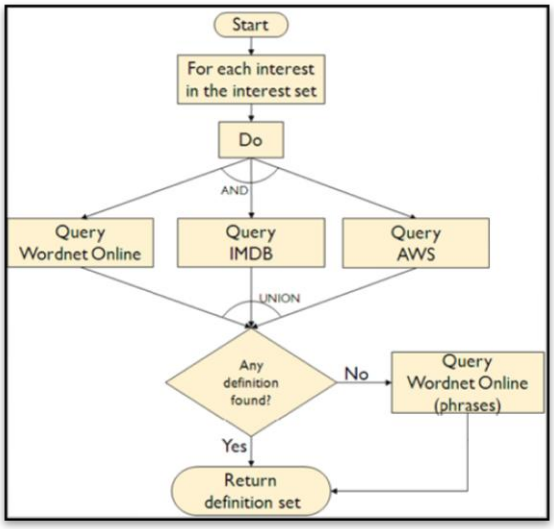

Figure 4. Procedure for consulting definitional data sources prior to constructing interest ontology. [Ba08]

The actual hierarchy consisted of single-word concepts formed from individual terms; *LiveJournal* allows up to four 15-character terms per interest. The similarity metric used for clustering was the number of matching terms in a unified definition set obtained as shown in Figure 4: each term of an interest was looked up in *WordNet-Online*, the *Internet Movie Database (IMDB)*, and *Amazon Associates' Web Service (AWS)*. Hierarchical Agglomerative and Divisive (HAD) clustering, a hybrid bottom-up linkage-based and divisive (partitional) algorithm, was used to generate the hierarchy. The output, consisting of 19 clusters, is summarized in Figure 5; note that the level of abstraction can be manually set, as we do in our experiments. We refer the interested reader to Bahirwani [Ba08] for additional details of the clustering algorithm and documentation on the data sources consulted.

Our link existence prediction system uses, as a baseline, the computed graph features specified by Hsu *et al.* [HKP+06]. To this we add nominal features: one Boolean value for each pair of interests in the Cartesian product of those for a user *u* and a user *v*. This is computed by first clustering single interest keywords to build a concept hierarchy, then mapping each interest of *u* and *v* to its abstract ancestor in the concept hierarchy before computing the nominal features (a bit vector).

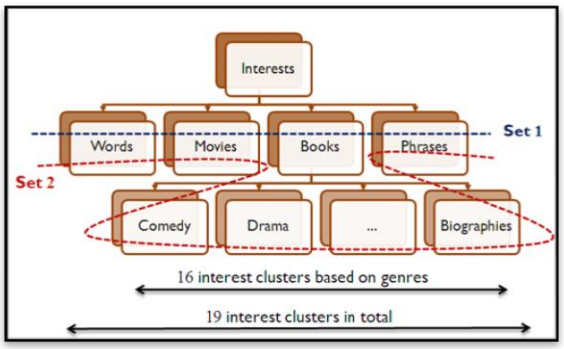

Figure 5. Example of clusters found using Hierarchical Agglomerative and Divisive (HAD) algorithm.

3.3 Numerical Features: Estimation by Association Rule (AR) Mining

Interestingness measures are descriptive statistics computed over rules of the form $u \rightarrow v$, which in our application denotes that "when *u* holds an interest, then *v* also holds that interest". This allows us to apply algorithms for association rule (AR) mining based on calculation of frequent itemsets, which by analogy with market basket analysis denote sets of users who are all interested in one topic. Each interestingness measure captures one or more desiderata of a data mining system: novelty (surprisingness), validity (precision, recall, accuracy), expected utility, and comprehensibility (semantic value).

We use the count of common interests, plus eight normalized AR interestingness measures over common interests, as numerical friendship prediction features. Each measure is a statistic over the set common interests of *u* and *v*, and expressed as a function of the rule $u \rightarrow v$.

1. The **number of common interests**:
   $|\ \text{Itemsets}(u) \cap \text{Itemsets}(v)\ |$
2. **Support ($u \rightarrow v$) = Support ($v \rightarrow u$)** = $\mathrm{P}(u, v)$
3. **Confidence ($u \rightarrow v$)** = $\mathrm{P}(v|u)$
4. **Confidence ($v \rightarrow u$)** = $\mathrm{P}(u|v)$
5. **Lift ($u \rightarrow v$)** = $\frac{\mathrm{P}(v|u)}{\mathrm{P}(v)}$
6. **Conviction ($u \rightarrow v$)** = $\frac{\mathrm{P}(u)\mathrm{P}(v)}{\mathrm{P}(u,\neg v)}$
7. **Match ($u \rightarrow v$)** = $\frac{\mathrm{P}(u,v)-\mathrm{P}(u)*\mathrm{P}(v)}{\mathrm{P}(u)*(1-\mathrm{P}(u))}$
8. **Accuracy ($u \rightarrow v$)** = $\mathrm{P}(u, v) + \mathrm{P}(\neg u, \neg v)$
9. **Leverage ($u \rightarrow v$)** = $\mathrm{P}(v|u) - \mathrm{P}(u)\mathrm{P}(v)$

A normalization step is used to sensitize the AR mining algorithm to the popularity of interests, which is measured by the sizes of itemsets. Intuitively, it is more significant for two candidate users to share rare interests than popular

ones, a property which gives itemset size a particular semantic significance in this application domain. For the derivation of a parametric normalization function, we refer the interested reader to Aljandal *et al.* [AHP+08]

3.4 Combining AR Mining and Interest Ontologies

The same ontology used is also applied to concrete (literal) interests to generate numerical features for abstract interests: that is, interests are first generalized as in Section 3.2; interestingness measures are then computed over the abstract interest categories; finally, the resultant measures are normalized using the size of each abstract itemset (list of interest-holders).

4 EXPERIMENT DESIGN

4.1 Link Prediction: 1000-user *LiveJournal* Data Set

We used the 1000-user data set developed by Hsu *et al.* [HLP+07], which includes about 22000 unique interests that are shared by at least two users. (Interests held by only one user are of no interest for link prediction, so singleton itemsets are pruned as is often done in frequent itemset mining.) As mentioned in Section 3.2, these are clustered using the HAD algorithm to form 19 clusters, resulting in 19 + 19 = 38 nominal features for every candidate pair (*u*, *v*). To these we add the original 7 graph features and the 9 numerical features. This integrated data set incorporates all of the ontology-enhanced relational features described in Section 3. It is sampled at a ratio of 50% negative (non-friends) to 50% positive (friends) for training, while the holdout validation data set has a natural ratio of about 90% to 10%. This follows the approach used by Kubat, Matwin, and Holte to learn classifiers for anomaly detection where the naturally observed rate of negative examples was much greater than that of positive examples. [KM97, KHM98]

4.2 Interest Prediction: *LiveJournal* Data Set

We also use the integrated, ontology-enhanced data set to predict whether an individual user *u* lists a member of one of the 19 abstract interest categories, given the fraction of their friends in the network that also list that category.

5 RESULTS

5.1 Link Existence Prediction

Table 1. Results without ontology-based features.

Inducer	Accuracy (%)	Precision	Recall	F-Score	AUC
Random Forest	65.3	0.014	0.556	0.028	0.605
Logistic	85.5	0.034	0.556	0.065	0.68
ADTree	78.8	0.023	0.556	0.045	0.694

Table 1 shows the precision, recall, F-score and area under the specificity-sensitivity curve (ROC-AUC) for the three inductive learning algorithms with the highest ROC-AUC. Each was trained using graph, nominal, and numerical features computed *without* the ontology. Table 2 lists the same results *with* the ontology.

Table 2. Results *with* ontology-based features

Inducer	Accuracy (%)	Precision	Recall	F-Score	AUC
Random Forest	70.0	0.020	0.857	0.038	0.829
Logistic	89.7	0.056	0.857	0.104	0.894
ADTree	82.7	0.034	0.857	0.065	0.925

5.2 Interest Prediction

We evaluated the nominal and numerical features using five classifier models and inductive learning algorithms: support vector machines (SVM), Logistic Classification, Random Forests, decision trees (J48), and decision stumps (OneR). Table 3 and Table 4 list the results for SVM and Logistic Classification, which achieved the highest ROC-AUC score using all available features. [Ba08] The overall highest AUC was achieved using numerical features along with Logistic Classification, although the precision is still improved by the inclusion of nominal features.

Table 3. Results using Support Vector Machines.

Nom	Num	Precision	Recall	F-Score	AUC
*		0.617	0.693	0.601	0.558
	*	0.829	0.826	0.817	0.918
*	*	0.833	0.838	0.829	0.921

Table 4. Results using Logistic Classification.

Nom	Num	Precision	Recall	F-Score	AUC
*		0.618	0.684	0.611	0.570
	*	0.838	0.846	0.839	0.924
*	*	0.845	0.844	0.843	0.919

6 CONCLUSIONS AND FUTURE WORK

We have shown how concept hierarchies learned from interest terms, generic dictionaries, and topical dictionaries can result in ontology-aware classifiers. These have greater precision and recall on link existence and interest prediction than those based only on graph features, nominal features, and numerical interestingness measures for association rules.

In future work, we will examine how to extend the framework to incorporate multi-word interests and technical definitions. Other memberships listed in Section 2.2 may also benefit from ontology discovery – especially fandoms and communities, which have their own description pages and metadata in most social networks. The association rule mining approach *and* the semantics

of itemset size extend naturally to these domains, making these a promising area for exploration of ontology-aware classification. To be able to account for the relationship between membership popularity and significance towards link existence, however, it will be important for our feature discovery methods to capture some domain-specific semantics of links and itemset membership. For example, we do not expect that itemset size normalization methods will apply in all market basket analysis domains, even though they seem to be effective in some social networks. Finally, returning to the *LinkedIn* example in Figure 2, an ontology that includes temporal fluents such as part-of ("Blogger became part of Google in 2004") and use them to infer relational fluents ("*u* and *v* have been Google employees since 2004") will allow us to construct semantically richer feature sets that we believe will be more useful for link existence and persistence prediction.

7 ACKNOWLEDGEMENTS

We thank Tim Weninger for assistance with implementations and useful discussions on graph feature discovery.

8 REFERENCES

[AHB+08] Aljandal, W., Hsu, W. H., Bahirwani, V., Caragea, D., & Weninger, T. (2008). Validation-based normalization and selection of interestingness measures for association rules. In *Proceedings of the 18th International Conference on Artificial Neural Networks in Engineering (ANNIE 2008)*, to appear. St. Louis, MO, 09 - 12 Nov 2008.

[Ba08] Bahirwani, V. (2008). Ontology engineering and feature construction for predicting friendship links and users' interests in the Live Journal social network. M.S. thesis, Kansas State University.

[BCA+08] Bahirwani, V., Caragea, D., Aljandal, W. & Hsu, W. H. (2008). Ontology engineering for social network data mining. In *Proceedings of the 2nd ACM SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD 2008)*, held in conjunction with the *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*. Las Vegas, NV, 24 - 27 Aug 2008.

[BG04] Bhattacharya, I. & Getoor. L. (2004). Deduplication and group detection using links. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Link Analysis and Group Detection (LinkKDD2004)*, Seattle, WA, USA, August 22-25, 2004.

[GD05] Getoor, L. & Diehl, C. P. (2005). Link mining: a survey. *SIGKDD Explorations, Special Issue on Link Mining*, 7(2):3-12.

[Hi03] Hill, S. (2003). Social network relational vectors for anonymous identity matching In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Statistical Learning of Relational Models (SRL)*, Acapulco, MEXICO, August, 2003.

[HWP08] Hsu, W. H., Weninger, T., & Paradesi, M. S. R. (2008). Predicting links and link change in friends networks: supervised time series learning with imbalanced data. In *Proceedings of the 18th International Conference on Artificial Neural Networks in Engineering (ANNIE 2008)*, to appear. St. Louis, MO, 09 - 12 Nov 2008.

[HLP+07] Hsu, W. H., Lancaster, J. P., Paradesi, M. S. R., & Weninger, T. (2007). Structural link analysis from user profiles and friends networks: a feature construction approach, In *Proceedings of the 1st International Conference on Weblogs and Social Media*, (pp. 75-80). Boulder, CO, 26 - 28 Mar 2007.

[HKP+06] Hsu, W. H., King, A. L., Paradesi, M., Pydimarri, T., & Weninger, T. (2006). Collaborative and structural recommendation of friends using weblog-based social network analysis. In Nicolov, N., Salvetti, F., Liberman, M., & Martin, J. H. (*Eds.*), *Computational Approaches to Analyzing Weblogs - Papers from the 2006 Spring Symposium*, pp. 24-31. AAAI Press Technical Report SS-06-03. Stanford, CA, 27 - 29 Mar 2006.

[KHC05] Ketkar, N. S., Holder, L. B., & Cook, D. J. (2005). Comparison of graph-based and logic-based multi-relational data mining. *SIGKDD Explorations, Special Issue on Link Mining*, 7(2):64-71.

[KHM98] Kubat M., Holte R., & Matwin S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2/3):195-215.

[KM97] Kubat, M., & Matwin S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the 14th International Conference on Machine Learning* (ICML 1997), pp. 179-186.

[MWC07] McCallum, A, Wang, X., & Corrada-Emmanuel, A. (2007). *Journal of Artificial Intelligence Research (JAIR)*, 30:249-272.

[PU03] Popescul, A. & Ungar, L. H. (2003). Statistical relational learning for link prediction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Statistical Learning of Relational Models (SRL 2003)*, Acapulco, MEXICO, August, 2003.

[RDHT04] Resig, J., Dawara, S, Homan, C. M., & Teredesai, A. (2004) Extracting social networks from instant messaging populations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Link Analysis and Group Detection (LinkKDD2004)*, Seattle, WA, USA, August 22-25, 2004.

[SM05] Sarkar, P. & Moore, A. (2005). Dynamic social network analysis using latent space models. *SIGKDD Explorations, Special Issue on Link Mining*, 7(2):31-40.