Predicting Friendship Links in Social Networks Using a Topic Modeling Approach

Rohit Parimi and Doina Caragea

Computing and Information Sciences,
Kansas State University, Manhattan, KS, USA 66506
{rohitp,dcaragea}@ksu.edu

Abstract. In the recent years, the number of social network users has increased dramatically. The resulting amount of data associated with users of social networks has created great opportunities for data mining problems. One data mining problem of interest for social networks is the friendship link prediction problem. Intuitively, a friendship link between two users can be predicted based on their common friends and interests. However, using user interests directly can be challenging, given the large number of possible interests. In the past, approaches that make use of an *explicit* user interest ontology have been proposed to tackle this problem, but the construction of the ontology proved to be computationally expensive and the resulting ontology was not very useful. As an alternative, we propose a topic modeling approach to the problem of predicting new friendships based on interests and existing friendships. Specifically, we use Latent Dirichlet Allocation (LDA) to model user interests and, thus, we create an *implicit* interest ontology. We construct features for the link prediction problem based on the resulting topic distributions. Experimental results on several *LiveJournal* data sets of varying sizes show the usefulness of the LDA features for predicting friendships.

Keywords: Link Mining, Topic Modeling, Social Networks, Learning.

1 Introduction

Social network such as *MySpace, Facebook, Orkut, LiveJournal* and *Bebo* have attracted millions of users [1], some of these networks growing at a rate of more than 50 percent during the past year [2]. Recent statistics have suggested that social networks have overtaken search engines in terms of usage [3]. This shows how Internet users have integrated social networks into their daily practices.

Many social networks, including *LiveJournal* online services [4] are focused on user interactions. Users in *LiveJournal* can tag other users as their friends. In addition to tagging friends, users can also specify their demographics and interests in this social network. We can see *LiveJournal* as a graph structure with users (along with their specific information, e.g. user interests) corresponding to nodes in the graph and edges corresponding to friendship links between the users. In general, the graph corresponding to a social network is undirected. However,

J.Z. Huang, L. Cao, and J. Srivastava (Eds.): PAKDD 2011, Part II, LNAI 6635, pp. 75–86, 2011.

in *LiveJournal*, the edges are directed i.e., if a user 'A' specifies another user 'B' as its friend, then it is not necessary for user 'A' to be the friend of user 'B'. One desirable feature of an online social network is to be able to suggest potential friends to its users [8]. This task is known as the link prediction problem, where the goal is to predict the existence of a friendship link from user 'A' to user 'B'. The large amounts of social network data accumulated in the recent years have made the link prediction problem possible, although very challenging.

In this work, we aim at using the ability of machine learning algorithms to take advantage of the content (data from user profiles) and graph structure of social network sites, e.g., *LiveJournal*, to predict friendship links. User profiles in such social networks consist of data that can be processed into useful information. For example, interests specified by users of *LiveJournal* act as good indicators to whether two users can be friends or not. Thus, if two users 'A' and 'B' have similar interests, then there is a good chance that they can be friends. However, the number of interests specified by users can be very large and similar interests need to be grouped semantically. To achieve this, we use a topic modeling approach. Topic models provide an easy and efficient way of capturing semantics of user interests by grouping them into categories, also known as topics, and thus reducing the dimensionality of the problem. In addition to using user interests, we also take advantage of the graph structure of the *LiveJournal* network and extract graph information (e.g., mutual friends of two users) that is helpful for predicting friendship links [9]. The contributions of this paper are as follows: (i) an approach for applying topic modeling techniques, specifically LDA, on user profile data in a social network; and (ii) experimental results on *LiveJournal* datasets showing that a) the best performance results are obtained when information from interest topic modeling is combined with information from the network graph of the social network b) the performance of the proposed approach improves as the number of users in the social network increases.

The rest of the paper is organized as follows: We discuss related work in Section 2. In Section 3, we review topic modeling techniques and Latent Dirichlet Allocation (LDA). We provide a detailed description of our system's architecture in Section 4 and present the experimental design and results in Section 5. We conclude the paper with a summary and discussion in Section 6.

2 Related Work

Over the past decade, social network sites have attracted many researches as sources of interesting data mining problems. Among such problems, the link prediction problem has received a lot of attention in the social network domain and also in other graph structured domains.

Hsu et al. [9] have considered the problems of predicting, classifying, and annotating friendship relations in a social network, based on the network structure and user profile data. Their experimental results suggest that features constructed from the network graph and user profiles of *LiveJournal* can be effectively used for predicting friendships. However, the interest features proposed in [9] (specifically, counts of individual interests and the common interests

of two users) do not capture the semantics of the interests. As opposed to that, in this work, we create an implicit interest ontology to identify the similarity between interests specified by users and use this information to predict unknown links.

A framework for modeling link distributions, taking into account object features and link features is also proposed in [5]. Link distributions describe the neighborhood of links around an object and can capture correlations among links. In this context, the authors have proposed an Iterative Classification Algorithm (ICA) for link-based classification. This algorithm uses logistic regression models over both links and content to capture the joint distributions of the links. The authors have applied this approach on web and citation collections and reported that using link distribution improved accuracy in both cases.

Taskar et al. [8] have studied the use of a relational Markov network (RMN) framework for the task of link prediction. The RMN framework is used to define a joint probabilistic model over the entire link graph, which includes the attributes of the entities in the network as well as the links. This method is applied to two relational datasets, one involving university web pages, and the other a social network. The authors have reported that the RMN approach significantly improves the accuracy of the classification task as compared to a flat model.

Castillo et al. [7] have also shown the importance of combining features computed using the content of web documents and features extracted from the corresponding hyperlink graph, for web spam detection. In their approach, several link-based features (such as degree related measures) and various ranking schemes are used together with content-based features such as *corpus precision* and *recall*, *query precision*, etc. Experimental results on large public datasets of web pages have shown that the system was accurate in detecting spam pages.

Caragea et al. [10], [11] have studied the usefulness of a user interest ontology for predicting friendships, under the assumption that ontologies can provide a crisp semantic organization of the user information available in social networks. The authors have proposed several approaches to construct interest ontologies over interests of *LiveJournal* users. They have reported that organizing user interests in a hierarchy is indeed helpful for predicting links, but computationally expensive in terms of both time and memory. Furthermore, the resulting ontologies are large, making it difficult to use concepts directly to construct features.

With the growth of data on the web, as new articles, web documents, social networking sites and users are added daily, there is an increased need to accurately process this data for extracting hidden patterns. Topic modeling techniques are generative probabilistic models that have been successfully used to identify inherent topics in collections of data. They have shown good performance when used to predict word associations, or the effects of semantic associations on a variety of language-processing tasks [12], [13]. Latent Dirichlet Allocation (LDA) [15] is one such generative probabilistic model used over discrete data such as text corpora. LDA has been applied to many tasks such as word sense disambiguation [16], named entity recognition [17], tag recommendation [18], community recommendation [19], etc. In this work, we apply LDA

on user profile data with the goal of producing a reduced set of features that capture user interests and improve the accuracy of the link prediction task in social networks. To the best of our knowledge, LDA had not been used for this problem before.

3 Topic Modeling and Latent Dirichlet Allocation (LDA)

Topic models [12], [13] provide a simple way to analyze and organize large volumes of unlabeled text. They express semantic properties of words and documents in terms of probabilistic topics, which can be seen as latent structures that capture semantic associations among words/documents in a corpus. Topic models treat each document in a corpus as a distribution over topics and each topic as a distribution over words. A topic model, in general, is a generative model, i.e. it specifies a probabilistic way in which documents can be generated.

One such generative model is Latent Dirichlet Allocation, introduced by Blei et al. [15]. LDA models a collection of discrete data such as text corpora. Figure 1 (adapted from [15]) illustrates a simplified graphical model representing LDA. We assume that the corpus consists of M documents denoted by $\mathrm{D} = \{\mathbf{d_1}, \mathbf{d_2} \cdots \mathbf{d}_M\}$. Each document $\mathbf{d_i}$ in the corpus is defined as a sequence of N_i words denoted by $\mathbf{d_i} = (w_{i1}, w_{i2} \cdots w_{iN_i})$, where each word w_{ij}belongs to a vocabulary V. A word in a document $\mathbf{d_i}$ is generated by first choosing a topic z_{ij} according to a multinomial distribution and then choosing a word w_{ij} according to another multinomial distribution, conditioned on the topic z_{ij}. Formally, the generative process of the LDA model can be described as follows [15]:

1. *Choose the topic distribution* $\theta_i \sim Dirichlet(\alpha)$.
2. *For each of the* N_i *words* w_{ij}*:*
   (a) *Choose a topic* $z_{ij} \sim Multinomial(\theta_i)$.
   (b) *Choose a word* w_{ij} *from* $p(w_{ij}|z_{ij}, \beta)$ *(multinomial conditioned on* z_{ij}*).*

From Figure 1, we can see that the LDA model has a three level representation. The parameters α and β are corpus level parameters, in the sense that they are assumed to be sampled once in the process of generating a corpus. The variables θ_i are document-level variables sampled once per document and the

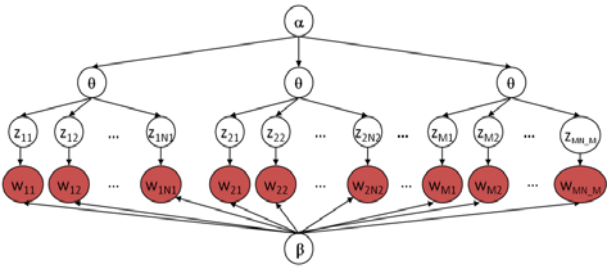

Fig. 1. Graphical representation of the LDA model

variables z_{ij} and w_{ij} are at the word level. These variables will be sampled once for each word in each document. For the work in this paper, we have used the LDA implementation available in *MALLET, A Machine Learning for Language Toolkit* [20]. MALLET uses Gibbs sampling for parameter estimation.

4 System Architecture

As can be seen in Figure 2, the architecture of the system that we have designed is divided into two modules. The first module of the system is focused on identifying and extracting features from the interests expressed by each user of the *LiveJournal*. These features are referred to as *interest based features*. The second module uses the graph network (formed as a result of users tagging other users in the network as 'friends') to calculate certain features which have been shown to be helpful at the task of predicting friendship links in *LiveJournal* [9]. We call these features, *graph based features*. We use both types of features as input to learning algorithms (as shown in Section 5). Sections 4.1 and 4.2 describe in detail the construction of interest based and graph based features, respectively.

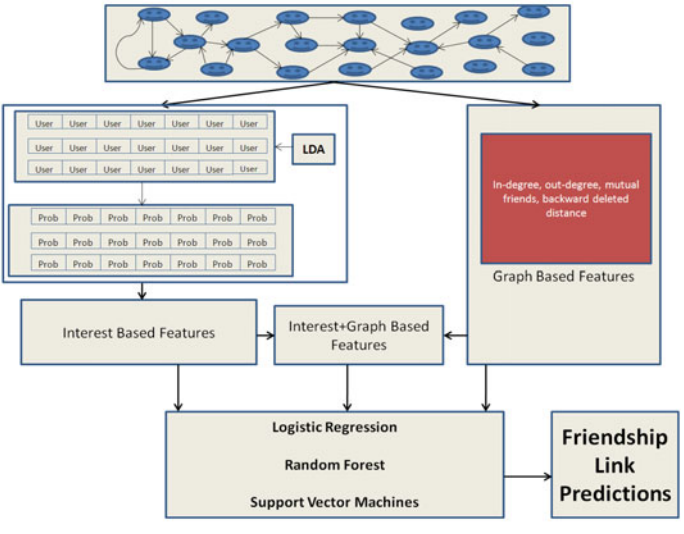

Fig. 2. Architecture of the system used for link prediction

4.1 Interest Based Features

Each user in a social network has a profile that contains information characteristic to himself or herself. Users most often tend to describe themselves, their likes, dislikes, interests/hobbies in their profiles. For example, users of *LiveJournal* can specify their demographics and interests along with tagging other users of the social network as friends. Data from the user profiles can be processed into

useful information for predicting/recommending potential friends to the users. In this work, we use a topic modeling technique to capture semantic information associated with the user profiles, in particular, with interests of *LiveJournal* users. Interests of the users act as good indicators to whether they can be friends or not. The intuition behind interest based features is that two users 'A' and 'B' might be friends if 'A' and 'B' have some similar interests. We try to capture this intuition through the feature set that we construct using the user interests.

Our goal is to organize interests into "topics". To do that, we model user interests in *LiveJournal* using LDA by treating *LiveJournal* as a document corpus, with each user in the social network representing a "document". Thus, interests specified by each user form the content of the "user document". We then run the MALLET implementation of LDA on the collection of such user documents. LDA allows us to input the number of inherent topics to be identified in the collection used. In this work, we vary the number of topics from 20 to 200. In general, the smaller the number of topics, the more abstract will be the inherent topics identified. Similarly, the larger the number of topics, the more specific the topics identified will be. Thus, by varying the number of topics, we are implicitly simulating a hierarchical ontology: a particular number of topics can be seen as a cut through the ontology. The topic probabilities obtained as a result of modeling user interests with LDA provide an explicit representation of each user and are used to construct the interest based features for the friendship prediction task, as described in what follows: suppose that $A\,[1 \cdots n]$ represents the topic distribution for user 'A' and $B\,[1 \cdots n]$ represents the topic distribution for user 'B' at a particular topic level n. The feature vector, $F(A, B)$ for the user pair (A, B) is constructed as: $F(A, B) = (|A\,[1] - B\,[1]\,|, |A\,[2] - B\,[2]\,|, \cdots, |A\,[n] - B\,[n]\,|)$. This feature vector is meant to capture the intuition that the smaller the difference between the topic distributions, the more semantically related the interests are.

4.2 Graph Based Features

Previous work by Hsu et al. [9] and Caragea et al. [10], [11], among others, have shown that the graph structure of the *LiveJournal* social network acts as a good source of information for predicting friendship links. In this work, we follow the method described in [9] to construct graph-based features. For each user pair (A, B) in the network graph, we calculate *in-degree* of 'A', *in-degree* of 'B', *out-degree* of 'A', *out-degree* of 'B', *mutual friends* of 'A' and 'B', *backward deleted distance* from 'B' to 'A' (see [9] for detailed descriptions of these features).

5 Experimental Design and Results

This section describes the dataset used in this work and the experiments designed to evaluate our approach of using LDA for the link prediction task. We have conducted various experiments with several classifiers to investigate their performance at predicting friendship links between the users of *LiveJournal*.

5.1 Dataset Description and Preprocessing

We used three subsets of the *LiveJournal* dataset with 1000, 5000 and 10,000 users, respectively, to test the performance and scalability of our approach. As part of the preprocessing step, we clean the interest set to remove symbols, numbers, foreign language. Interests with frequency less than 5 in the dataset are also removed. Strings of words in a multi-word interest are concatenated into a single "word," so that MALLET treats them as a single entity. For example, the interest 'artificial neural networks' is transformed into 'ArtificialNeuralNetworks' after preprocessing. Users whose in-degree and out-degree is zero, as well as users who do not have any interests declared are removed from the dataset. We are left with 801, 4026 and 8107 users in the three datasets, respectively, and approximately 14,000, 32,000 and 39,700 interests for each dataset after preprocessing. Furthermore, there are around 4,400, 40,000, 49,700 declared friendship links in the three datasets. We generate topic distributions for the users in the dataset using LDA; hyper-parameters (α, β) are set to the default values.

We make the assumption that the graph is complete, i.e. all declared friendship links are positive examples and all non declared friendships are negative examples [10], although this assumption does not hold in the real world. The user network graph is partitioned into two subsets with $2/3^{rd}$ of the users in the first set and $1/3^{rd}$ of the users in the second set (this process is repeated five times for cross-validation purposes). We used the subset with $2/3^{rd}$ of the users for training and the subset with $1/3^{rd}$ of the users for test. We ensure that the training and the test datasets are independent by removing the links that go across the two datasets. We also balance the data in the training set, as the original distribution is highly skewed towards the negative class.

5.2 Experiments

The following experiments have been performed in this work.

1. **Experiment 1**: In the first experiment, we test the performance of several predictive models trained on interest features constructed from topic distributions. The number of topics to be modeled is varied from 20 to 200. The 1000 user dataset described above is used in this experiment.
2. **Experiment 2**: In the second experiment, we test several predictive models that are trained on graph features, for the 1000 user dataset. To be able to construct the graph features for test data, we assume that a certain percentage of links is known [8] (note that this is a realistic assumption, as it is expected that some friends are already known for each user). Specifically, we explore scenarios where 10%, 25% and 50% links are known, respectively. Thus, we construct features for the unknown links using the known links.
3. **Experiment 3**: In the third experiment, graph based features are used in combination with interest-based features to see if they can improve the performance of the models trained with graph features only on the 1000 user dataset. For the test set graph features constructed by assuming 10%, 25% and 50% known links, respectively, are combined with interest features.

We repeat the above mentioned experiments for the 5000 user dataset. The corresponding experiments are referred to as **Experiment 4**, **Experiment 5** and **Experiment 6**, respectively. For the 10,000 user dataset, we build predictive models using just interest based features (construction of graph features for the 10,000 user dataset was computationally infeasible, given our resources). This experiment is referred to as **Experiment 7**. We use results from **Experiments 1**, **4** and **7** to study the performance and the scalability of the LDA approach to link prediction based on interests, as the number of users increases. For all the experiments, we used WEKA implementations of the Logistic Regression, Random Forest and Support Vector Machine (SVM) algorithms.

5.3 Results

Importance of the Interest Features for Predicting Friendship Links. As mentioned above, several experiments have been conducted to test the usefulness of the topic modeling approach on user interests for the link prediction problem in *LiveJournal*. As expected, interest features (i.e., topic distributions obtained by modeling user interests) combined with graph features produced the most accurate models for the prediction task. This can be seen from Tables 1 and 2. In both tables, we can see that interest+graph features with 50% known links outperform interest or graph features alone in terms of AUC values[1], for all three classifiers used. Interesting results can be seen in Table 2, where interest features alone are better than graph features alone when only 10% links are known, and sometimes better also than interest+graph features with 10% links known, thus, showing the importance of the user profile data, captured by LDA, for link prediction in social networks. Furthermore, a comparison between our results and the results presented in [21], which uses an ontology-based approach to construct interest features, shows that the LDA features are better than the ontology features on the 1,000 user dataset. As another drawback, the ontology based approach is not scalable (no more than 4,000 users could be used) [21].

Figure 3 depicts the AUC values obtained using interest, graph and interest+graph features with Logistic Regression and SVM classifiers across all numbers of topics modeled for the 1,000 and 5,000 user datasets, respectively. We can see that the AUC value obtained using interest+graph features is better than the corresponding value obtained using graph features alone across all numbers of topics, for all scenarios of known links, in the case of the 5000 user dataset. This shows that the contribution of interest features increases with the number of users. Also based on Figure 3, it is worth noting that the graphs do not show significant variation with the number of topics used.

Performance of the Proposed Approach with the Number of Users. In addition to studying the importance of the LDA interest features for the link prediction task, we also study the performance and scalability of the approaches considered in this work (i.e., graph-based versus LDA interest based, and combinations) as the number of users increases. We are interested in both a) the

[1] All AUC values reported are averaged over five different *train* and *test* datasets.

Table 1. AUC values for Logistic Regression (LR), Random Forests (RF) and Support Vector Machines (SVM) classifiers with interest, graph and interest+graph based features for the 1,000 user dataset. k% links are known in the test set, where k is 10, 25 and 50, respectively. The known links are used to construct graph features.

Exp#	Features	Logistic Regression	Random Forest	SVM
1	Interest	0.625 ± 0.03	0.5782 ± 0.04	0.6198 ± 0.04
2 (10%)	Graph 10%	$\mathbf{0.74 \pm 0.08}$	0.578 ± 0.04	$\mathbf{0.7738 \pm 0.05}$
3 (10%)	Interest+Graph 10%	0.6226 ± 0.05	$\mathbf{0.6664 \pm 0.04}$	0.6606 ± 0.02
2 (25%)	Graph 25%	$\mathbf{0.7684 \pm 0.07}$	0.7106 ± 0.05	$\mathbf{0.8104 \pm 0.05}$
3 (25%)	Interest+Graph 25%	0.7406 ± 0.04	$\mathbf{0.8188 \pm 0.03}$	0.7983 ± 0.03
2 (50%)	Graph 50%	0.8526 ± 0.03	0.8008 ± 0.03	0.8692 ± 0.03
3 (50%)	Interest+Graph 50%	$\mathbf{0.8648 \pm 0.03}$	$\mathbf{0.877 \pm 0.04}$	$\mathbf{0.8918 \pm 0.03}$

Table 2. AUC values similar to those in Table 1, for the 5,000 user dataset.

Exp#	Features	Logistic Regression	Random Forest	SVM
4	Interest	$\mathbf{0.6954 \pm 0.01}$	0.6276 ± 0.01	$\mathbf{0.7008 \pm 0.01}$
5 (10%)	Graph 10%	0.649 ± 0.03	0.5936 ± 0.02	0.692 ± 0.02
6 (10%)	Interest+Graph 10%	0.6718 ± 0.02	$\mathbf{0.6566 \pm 0.01}$	0.6998 ± 0.01
5 (25%)	Graph 25%	0.7022 ± 0.05	0.6716 ± 0.02	0.7896 ± 0.03
6 (25%)	Interest+Graph 25%	$\mathbf{0.7384 \pm 0.03}$	$\mathbf{0.7846 \pm 0.03}$	$\mathbf{0.7986 \pm 0.03}$
5 (50%)	Graph 50%	0.8456 ± 0.02	0.7086 ± 0.02	0.883 ± 0.02
6 (50%)	Interest+Graph 50%	$\mathbf{0.8696 \pm 0.02}$	$\mathbf{0.8908 \pm 0.02}$	$\mathbf{0.9046 \pm 0.01}$

quality of the predictions that we get for the *LiveJournal* data as the number of users increases; and b) the time and memory requirements for each approach.

From Figure 4, we can see that the prediction performance (expressed in terms of AUC values) is improved in the 5,000 user dataset as compared to the 1,000 user dataset, across all numbers of topics modeled. Similarly, the prediction performance for the 10,000 user dataset is better than the performance for the 5,000 user dataset, for all topics from 20 to 200. One reason for better predictions with more users in the dataset is that, when we add more users, we also add the interests specified by the newly added users to the interest set on which topics are modeled using LDA. Thus, we get better LDA probability estimates for the topics associated with each user in the dataset, as compared to the estimates that we had for a smaller set of data, and hence better prediction results. However, as expected, both the amount of time it takes to compute features for the larger dataset, as well as the memory required increase with the number of users in the data set. The amount of time it took to construct features for the 10,000 user dataset for all numbers of topics modeled in the experiments is around 14 hours on a system with Intel core 2 duo processor running at 3.16GHz and 20GB of RAM. This time requirement is due to our complete graph assumption (which results in feature construction for 10,000*10,000 user pairs in the case of a 10,000 user dataset) and can be relaxed if we relax the completeness assumption. Still the LDA feature construction is more efficient than the construction of graph features, which was not possible for the 10,000 user dataset used in our study.

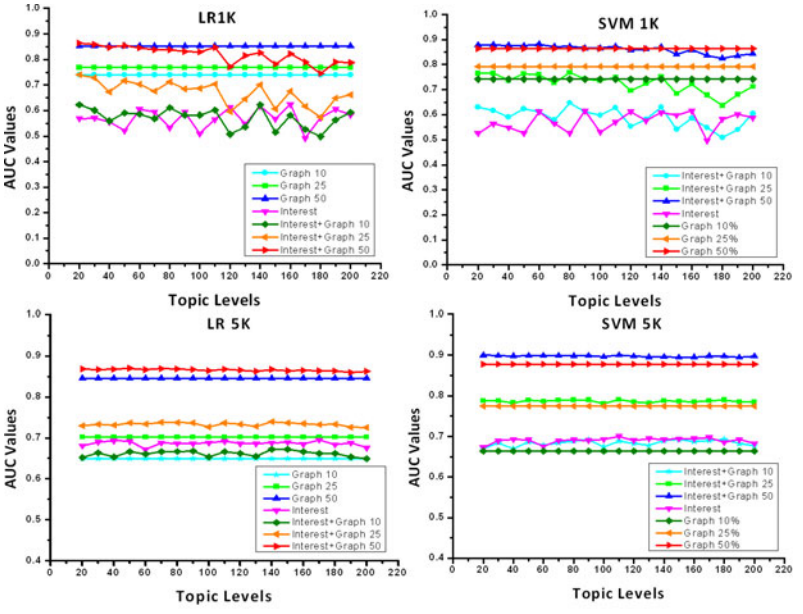

Fig. 3. Graph of reported AUC values versus number of topics used for modeling, using Logistic Regression and SVM classifiers, for the 1,000 user dataset (top-left and top-right, respectively) and 5,000 user dataset (bottom-left and bottom-right, respectively)

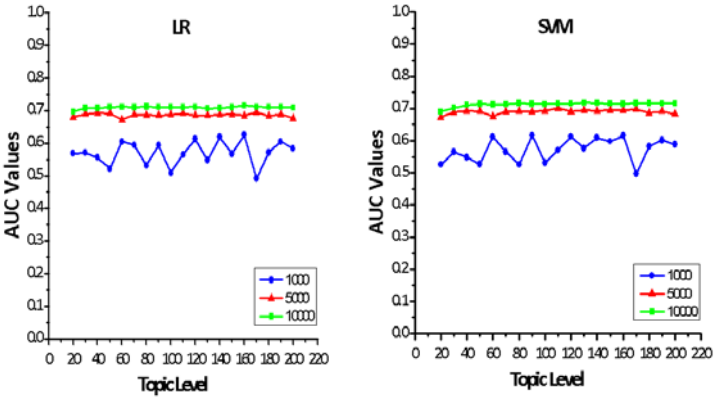

Fig. 4. AUC values versus number of topics for LR (left) and SVM (right) classifiers for the 1,000, 5,000 and 10,000 user datasets using interest-based features

6 Summary and Discussion

We have proposed an architecture, which takes advantage of both user profile data and network structure to predict friendship links in a social network. We have shown how one can model topics from user profiles in social networks using

LDA. Experimental results suggest that the usefulness of the interest features constructed using the LDA approach increases with an increase in the number of users. Furthermore, the results suggest that the LDA based interest features can help improve the prediction performance when used in combination with graph features, in the case of the *LiveJournal* dataset. Although in some cases the improvement in performance due to interest features is not very significant compared with the performance when graph features alone are used, the fact that computation of graph features becomes intractable for 10,000 users or beyond emphasizes the importance of the LDA based approach.

However, while the proposed approach is effective and shows improvement in performance as the number of users increases, it also suffers from some limitations. First, adding more users to the dataset increases the memory and time requirements. Thus, as part of the future work, we plan to take advantage of the MapReduce framework to support distributed computing for large datasets. Secondly, our approach takes into account, the static image of the *LiveJournal* social network. Obviously, this assumption does not hold in the real world. Based on user interactions in the social network, the graph might change rapidly due to the addition of more users as well as friendship links. Also, users may change their demographics and interests regularly. Our approach does not take into account such changes. Hence, the architecture of the proposed approach has to be changed to accommodate the dynamic nature of a social network. We also speculate that the approach of modeling user profile data using LDA will be effective for tasks such as citation recommendation in scientific document networks, identifying groups in online scientific communities based on their research/tasks and recommending partners in internet dating, ideas that are left as future work.

References

1. Boyd, M.D., Ellison, B.N.: Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication 13 (2007)
2. comScore Press Release, `http://www.comscore.com/Press_Events/Press_Releases/2007/07/Social_Networking_Goes_Globa`
3. TechCrunch Report, `http://eu.techcrunch.com/2010/06/08/report-social-networks-overtake-search-engines-in-uk-should-google-be-worried`
4. Fitzpatrick, B.: LiveJournal: Online Service, `http://www.livejournal.com`
5. Geetor, L., Lu, Q.: Link-based Classification. In: Twelth International Conference on Machine Learning (ICML 2003), Washington DC (2003)
6. Na, J.C., Thet, T.T.: Effectiveness of web search results for genre and sentiment classification. Journal of Information Science 35(6), 709–726 (2009)
7. Castillo, C., Donato, D., Gionis, A., Murdock, V., Silvestri, F.: Know your Neighbors: Web Spam Detection using the web Topology. In: Proceedings of SIGIR 2007, Amsterdam, Netherlands (2007)
8. Taskar, B., Wong, M., Abbeel, P., Koller, D.: Link Prediction in Relational Data. In: Proc. of 17th Neural Information Processing Systems, NIPS (2003)
9. Hsu, H.W., Weninger, T., Paradesi, R.S.M., Lancaster, J.: Structural link analysis from user profiles and friends networks: a feature construction approach. In: Proceedings of International Conference on Weblogs and Social Media (ICWSM), Boulder, CO, USA (2007)

10. Caragea, D., Bahirwani, V., Aljandal, W., Hsu, H.W.: Link Mining: Ontology-Based Link Prediction in the LiveJournal Social Network. In: Proceedings of Association of the Advancement of Artificial Intelligence, pp. 192–196 (2009)
11. Haridas, M., Caragea, D.: Link Mining: Exploring Wikipedia and DMoz as Knowledge Bases for Engineering a User Interests Hierarchy for Social Network Applications. In: Proceedings of the Confederated International Conferences on On the Move to Meaningful Internet Systems: Part II, Portugal, pp. 1238–1245 (2009)
12. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. In: Landauer, T., Mcnamara, D., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
13. Steyvers, M., Griffiths, T., Tenenbaum, J.B.: Topics in Semantic Representation. American Psychological Association 114(2), 211–244 (2007)
14. Steyvers, M., Griffiths, T.: Finding Scientific Topics. Proceedings of National Academy of Sciences, U.S.A, 5228–5235 (2004)
15. Blei, D., Ng, Y.A., Jordan, I.M.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
16. Blei, D., Boyd-Graber, J., Zhu, X.: A Topic Model for Word Sense Disambiguation. In: Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Comp. Natural Language Learning, pp. 1024–1033 (2007)
17. Guo, J., Xu, G., Cheng, X., Li, H.: Named Entity Recognition in Query. In: Proceedings of SIGIR 2009, Boston, USA (2009)
18. Krestel, R., Fankhauser, P., Nejdl, W.: Latent Dirichlet Allocation for Tag Recommendation. In: Proceedings of RecSys 2009, New York, USA (2009)
19. Chen, W., Chu, J., Luan, J., Bai, H., Wang, Y., Chang, Y.E.: Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior. In: Proceedings of International World Wide Web Conference (2009)
20. McCallam, K.A.: Mallet: A Machine Learning for Language Toolkit (2002), `http://mallet.cs.umass.edu`
21. Phanse, S.: Study on the Performance of Ontology Based Approaches to Link Prediction in Social Networks as the Number of Users Increases. M.S. Thesis (2010)