Collective Personal Profile Summarization with Social Networks

Zhongqing Wang, Shoushan Li*, Kong Fang, and Guodong Zhou
Natural Language Processing Lab, School of Computer Science and Technology
Soochow University, Suzhou, 215006, China
{wangzq.antony, shoushan.li}@gmail.com,
{kongfang, gdzhou}@suda.edu.cn

Abstract

Personal profile information on social media like LinkedIn.com and Facebook.com is at the core of many interesting applications, such as talent recommendation and contextual advertising. However, personal profiles usually lack organization confronted with the large amount of available information. Therefore, it is always a challenge for people to find desired information from them. In this paper, we address the task of personal profile summarization by leveraging both personal profile textual information and social networks. Here, using social networks is motivated by the intuition that, people with similar academic, business or social connections (e.g. *co-major*, *co-university*, and *co-corporation*) tend to have similar experience and summaries. To achieve the learning process, we propose a collective factor graph (CoFG) model to incorporate all these resources of knowledge to summarize personal profiles with local textual attribute functions and social connection factors. Extensive evaluation on a large-scale dataset from LinkedIn.com demonstrates the effectiveness of the proposed approach.

1 Introduction

Web 2.0 has empowered people to actively interact with each other, forming social networks around mutually interesting information and publishing a large amount of useful user-generated content (UGC) online (Lappas et al., 2011; Tan et al., 2011). One popular and important type of UGC is the personal profile, where people post detailed information on online portals about their education, experiences and other personal information. Social websites like Facebook.com and LinkedIn.com have created a viable business as profile portals, with the popularity and success partially attributed to their comprehensive personal profiles.

Generally, online personal profiles provide valuable resources for businesses, especially for human resource managers to find talents, and help people connect with others of similar backgrounds (Yang et al., 2011a; Guy et al., 2010). However, as there is always large-scale information of experience and education fields, it is hardly for us to find useful information from the profile. Therefore, it is always a challenge for people to find desired information from them. For this regard, it is highly desirable to develop reliable methods to generate a summary of a person through his profile automatically.

To the best of our knowledge, this is the first research that explores automatic summarization of personal profiles in social media. A straightforward approach is to consider personal profile summarization as a traditional document summarization problem, which treating each personal profile independently and generate a summary for each personal profile individually. For example, the well-known extraction and ranking approaches (e.g. PageRank, HITS) extract a certain amount of important sentences from a document according to some ranking measurements to form a summary (Wan and Yang, 2008; Wan, 2011).

However, such straightforward approaches are not sufficient to benefit from the carrier of personal profiles. As the centroid of social networking, people are usually connected to others with similar

* Corresponding author

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 715–725,
Seattle, Washington, USA, 18-21 October 2013.

background in social media (e.g. *co-major*, *co-corporation*). Therefore, it is reasonable to leverage social connection to improve the performance of profile summarizing. For example if there are *co-major*, *co-university*, *co-corporation* or other academic and business relationships between two persons, we consider them sharing similar experience and having similar summaries.

The remaining challenge is how to incorporate both the profile textual information and the connection knowledge in the social networks. In this study, we propose a collective factor graph model (CoFG) to summarize the text of personal profile in social networks with local textual information and social connection information. The CoFG framework utilizes both the local textual attribute functions of an individual person and the social connection factor between different persons to collectively summarize personal profile on one person.

In this study, we treat the profile summarization as a supervised learning task. Specifically, we model each sentence of the profile as a vector. In the training phase, we use the vectors with the social connection between each person to build the CoFG model; while in the testing phase, we perform collective inference for the importance of each sentence and select a subset of sentences as the summary according to the trained model. Evaluation on a large-scale data from LinkedIn.com indicates that our proposed joint model and social connection information improve the performance of profile summarization.

The remainder of our paper is structured as follows. We go over the related work in Section 2. In Section 3, we introduce the data we collected from LinkedIn.com and the annotated corpus we constructed. In Section 4, we present some motivational analysis. In Section 5, we explain our proposed model and describe algorithms for parameter estimation and prediction. In Section 6, we present our experimental results. We sum up our work and discuss future directions in Section 7.

2 Related Work

In this section, we will introduce the related work on the traditional topic-based summarization, social-based summarization and factor graph model respectively.

2.1 Topic-based Summarization

Generally, traditional topic-based summarization can be categorized into two categories: extractive (Radev et al., 2004) and abstractive (Radev and McKeown, 1998) summarization. The former selects a subset of sentences from original document(s) to form a summary; the latter reorganizes some sentences to form a summary where several complex technologies, such as information fusion, sentence compression and reformulation are necessarily employed (Wan and Yang, 2008; Celikyilmaz and Hakkani-Tur, 2011; Wang and Zhou, 2012). This study focuses on extractive summarization.

Radev et al. (2004) proposed a centroid-based method to rank the sentences in a document set, using various kinds of features, such as the cluster centroid, position and TF-IDF features. Ryang and Abekawa (2012) proposed a reinforcement learning approach on text summarization, which models the summarization within a reinforcement learning-based framework.

Compared to unsupervised approaches, supervised learning for summarization is relatively rare. A typical work is Shen et al., (2007) which present a Conditional Random Fields (CRF) based framework to treat the summarization task as a sequence labeling problem. However, different from all existing studies, our work is the first attempt to consider both textual information and social relationship information for supervised summarization.

2.2 Social-based Summarization

As web 2.0 has empowered people to actively interact with each other, studies focusing on social media have attracted much attention recently (Meeder et al., 2011; Rosenthal and McKeown, 2011; Yang et al., 2011a). Social-based summarization is exactly a special case of summarization where the social connection is employed to help obtaining the summarization. Although topic-based summarization has been extensively studied, studies on social-based summarization are relative new and rare.

Hu et al., (2011) proposed an unsupervised PageRank-based social summarization approach by incorporating both document context and user context in the sentence evaluation process. Meng et al., (2012) proposed a unified optimization framework to produce opinion summaries of tweets through

integrating information from dimensions of topic, opinion and insight, as well as other factors (e.g. topic relevancy, redundancy and language styles).

Unlike all the above studies, this paper focuses on a novel task, profile summarization. Furthermore, we employ many other kinds of social information in profiles, such as *co-major*, and *co-corporation* between two people. They are shown to be very effective for profile summarization.

2.3 Factor Graph Model

As social network has been investigated for several years (Leskovec et al., 2010; Tan et al., 2011; Lu et al., 2010; Guy et al., 2010) and Factor Graph Model (FGM) is a popular approach to describe the relationship of social network (Tang et al., 2011a; Zhuang et al., 2012). Factor Graph Model builds a graph to represent the relationship of nodes on the social networks, and the factor functions are always considered to represent the relationship of the nodes.

Tang et al. (2011a) and Zhuang et al. (2012) formalized the problem of social relationship learning into a semi-supervised framework, and proposed Partially-labeled Pairwise Factor Graph Model (PLP-FGM) for learning to infer the type of social ties. Dong et al. (2012) gave a formal definition of link recommendation across heterogeneous networks, and proposed a ranking factor graph model (RFG) for predicting links in social networks, which effectively improves the predictive performance. Yang et al., (2011b) generated summaries by modeling tweets and social contexts into a dual wing factor graph (DWFG), which utilized the mutual reinforcement between Web documents and their associated social contexts.

Different from all above researches, this paper proposes a pair-wise factor graph model to collectively utilize both textual information and social connection factor to generate summary of profile.

3 Data Collection and Statistics

The personal profile summarization is a novel task and there exists no related data for accessing this issue. Therefore, in this study, we collect a data set containing personal summaries with the corresponding knowledge, such as the self-introduction and personal profiles. In this section, we will introduce this data set in detail.

3.1 Data Collection

We collect our data set from LinkedIn.com[1]. It contains a large number of personal profiles generated by users, containing various kinds of information, such as personal overview, summary, education, experience, projects and skills.

John Smith[2]	
Overview	
Current	Applied Researcher at Apple Inc.
Previous	Senior Research Scientist at IBM …
Education	MIT, Georgia Institute of Technology, …
Summary	
Machine learning researcher and engineer on many fields: Query understanding. Automatic Information extraction…	
Experience	
Applied Researcher *Apple Inc.,* September 2012 ~ Query recognition and relevance …	
Education	
MIT *Ph.D., Electrical Engineering*, 2002 – 2008 …	

Figure 1: An example of a profile webpage from LinkedIn.com

In this study, the data set is crawled in the following ways. To begin with, 10 random people's public profiles are selected as seed profiles, and then the profiles from their "People Also Viewed" field were collected. The data is composed of 3,182 public profiles[3] in total. We do not collect personal names in public profiles to protect people's privacy. Figure 1 shows an example of a person's profile from LinkedIn.com. The profile includes following fields:

- *Overview*: It gives a structure description of a person's general information, such as current/previous position and workplace, brief

[1] http://www.linkedin.com

[2] The information of the example is a pseudo one.

[3] We collect all the data from LinkedIn.com at Dec 17, 2012.

education background and general technical background.

- *Summary*: It summarizes a person's work, experience and education.
- *Experience*: It details a person's work experience.
- *Education*: It details a person's education background.

Among these fields, the *Overview* is required and the others are optional, such as *Project*, *Course* and *Interest groups*. However, compared with *Overview*, *Summary*, *Experience*, *Education* fields, they seem to be less important for summarization of personal profiles. Thus, we ignore them in our study.

3.2 Data Statistics of Major Fields

We collected 3,182 personal profiles from LinkedIn.com. Table 1 shows the statistics of major fields in our data collection.

Field	#Non-empty fields	Average field length
Overview	3,182	45.1
Summary	921	25.8
Experience	3,148	192.1
Education	2,932	33.6

Table 1: Statistics of major fields in our data set, i.e. the number of non-empty fields and the average length for each field

From Table 1, we can see that,

- The information of each profile is incomplete and inconsistent, That is, not all kinds of fields are available in each personal's profile.
- Most people provide their experience and education information. However, the *Summary* fields are popularly missing (Only about 30% of people provide it). This is mainly because writing summary is normally more difficult than other fields. Therefore, it is highly desirable to develop reliable automatic methods to generate a summary of a person through his/her profile.
- The length of the *Experience* field is the longest one, and work experience always could represent general information of people.

3.3 Corpus Construction and Annotation

Among the 921 profiles that contain the summary, we manually select 497 profiles with high quality summary to construct the corpus for our research. These high-quality summaries are all written by the authors themselves. Here, the quality is measured by manually checking that whether they are well capable of summarizing their profiles. That is, they are written carefully, and could give an overview of a person and represent the education and experience information of a person.

After carefully seeing the profiles, we observe that the *Experience* field contains the most abundant information of a person. Thus, we treat the text of *Experience* field as the source of summary for each profile. Besides, we collect social context information from *Education* and *Experience* field, and these social contexts are including by LinkedIn explicitly. Table 2 shows the average length of summary and experience fields we used for evaluating our summarization approach.

Field	Average length
Summary (*the summary of the profile*)	37.2
Experience (*the source text for the summarizing*)	372.0

Table 2: Average length of the high-quality summary and corresponding experience fields

From Table 2, we can see that,

- Compared with the average length of 25.8 in Table 1, summaries of high quality have longer length because they contain more information of the profiles.
- The compression ratio of our proposed corpus is 0.1 (37.2/372.0).

4 Motivation and Analysis

In this section, we propose the motivation of social connection to address the task of personal profile summarization. To preliminarily support the motivation, some statistics of the social connection are provided.

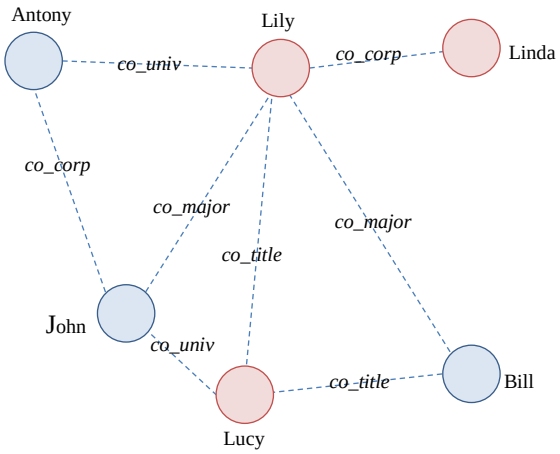

Figure 2: An example of personal profile network. *Red is for female, blue is for male, and the dotted line means the social connection between two persons.*

We first describe the social connections which we used. Figure 2 shows an example of social connection between people from the profiles of LinkedIn. We find that people are sometimes connected by several social connections. For example, John and Lucy are connected by *co_unvi* relationship, while Lily and Linda are connected by *co_corp* relationship. From LinkedIn, four kinds of social relationship between people are extracted from the *Education* field and *Experience* field. They are:

- *co_major* denotes that two persons have the same major at school
- *co_univ* denotes that two persons are graduated from the same university
- *co_title* denotes that two persons have the same title at corporation.
- *co_corp* denotes that two persons work at the same corporation.

Our basic motivation of using social connection lies in the fact that "connected" people will tend to hold related experience and similar summaries.

We then give the statistics of edges of social connection. Table 3 shows basic statistics across these edges. From Table 3, we can see that the number of users is 497 while the number of social connection edges is 14,307. The latter is much larger than the former. The number of the edges from *Education* field is similar with the number of the edges from *Experience* filed. Among all the relationships, *co_unvi* is the most common one.

	Numbers
# users	**497**
co_major	1,288
co_unvi	6,015
# education field	**7,303**
co_title	3,228
co_corp	3,776
# experience field	**7,004**
# total edges	**14,307**

Table 3: The statistic of edges for our main datasets

5 Collective Factor Graph Model

In this section, we propose a collective factor graph (CoFG) model for learning and summarizing the text of personal profile with local textual information and social connection.

5.1 Overview of Our Framework

To generate summaries for profiles, a straightforward approach is to treat each personal profile independently and generating a summary for each personal profile individually. As we mentioned on Section 3.3, we use the sentences of *Experience* field as a text document and consider it as the source of summary for each profile.

Instead, we formalize the problem of personal profile summarization in a pair-wise factor graph model and propose an approach referred to as Loopy Belief Propagation algorithm to learn the model for generating the summary of the profile. Our basic idea is to define the correlations using different types of factor functions. An objective function is defined based on the joint probability of the factor functions. Thus, the problem of collective personal profile summarization model learning is cast as learning model parameters that maximizes the joint probability of the input continuous dynamic network.

The overview of the proposed method is a supervised framework (as shown in Figure 3). First, we treat each sentence of the training data and testing data as vectors with textual information (local textual attribute functions); Second, all the vectors are connected by social connection relationships (social connection factors) and we model these vectors and their relationships into the collective factor graph; third, we propose Loopy Belief Prop-

agation algorithm to learn the model and predict the sentences of testing data; finally, we select a subset of sentences of each testing profile as the summary according to the models with top-*n* prediction score. Thus, the core issues of our framework are 1) how to define the collective factor graph model to connection profiles with social connection; 2) how to learn and predict the proposed CoFG model; 3) how to predict the sentences from the testing data with the proposed CoFG model, and generate the summary by the predict scores. We will discuss these issues on the following subsections.

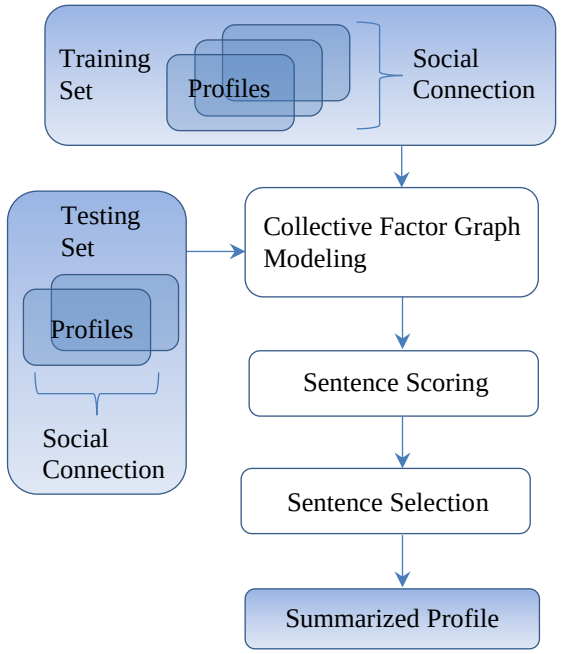

Figure 3: The overview of our proposed framework

5.2 Model Definition

Formally, given a network $G=(V,S^L,S^U,X)$, each sentence s_i is associated with an attribute vector x_i of the profile and a label y_i indicating whether the sentence is selected as a summary of the profile (The value of y_i is binary. 1 means that the sentence is selected as a summary sentence, whereas 0 stands for the opposite). V denotes the authors of the profiles, S^L denotes the labeled training data, and S^U denotes the unlabeled testing data. Let $X=\{x_i\}$ and $Y=\{y_i\}$. Then, we have the following formulation

$$P(Y \mid X,G)=\frac{P(X,G \mid Y)P(Y)}{P(X,G)} \quad (1)$$

Here, G denotes all forms of network information. This probabilistic formulation indicates that labels of skills depend on not only local attributes X, but also the structure of the network G. According to Bayes' rule, we have

$$P(Y \mid X,G)=\frac{P(X,G \mid Y)P(Y)}{P(X,G)} \propto P(X \mid Y)P(Y \mid G) \quad (2)$$

Where $P(Y \mid G)$ represents the probability of labels given the structure of the network and $P(X \mid Y)$ denotes the probability of generating attributes X associated to their labels Y. We assume that the generative probability of attributes given the label of each edge is conditionally independent, thus we have

$$P(Y \mid X,G) \propto P(Y \mid G)\prod_i P(x_i \mid y_i) \quad (3)$$

Where $P(x_i \mid y_i)$ is the probability of generating attributes x_i given the label y_i. Now, the problem becomes how to instantiate the probability $P(Y \mid G)$ and $P(x_i \mid y_i)$. We model them in a Markov random field, and thus according to the Hammersley-Clifford theorem (Hammersley and Clifford, 1971), the two probabilities can be instantiated as follows:

$$P(x_i \mid y_i)=\frac{1}{Z_1}\exp\left\{\sum_{j=1}^{d}\alpha_j f_j(x_{ij},y_i)\right\} \quad (4)$$

$$P(Y \mid G)=\frac{1}{Z_2}\exp\left\{\sum_i \sum_{j\in NB(i)} g(i,j)\right\} \quad (5)$$

Where Z_1 and Z_2 are normalization factors. Eq. 4 indicates that we define an attribute function $f(x_i,y_i)$ for each attribute x_{ij} associated with sentence s_i. α_j is the weight of the j^{th} attribute. Eq. 5 represents that we define a set of correlation factor functions $g(i,j)$ over each pair (i,j) in the network. $NB(i)$ denotes the set of social relationship neighbors nodes of i.

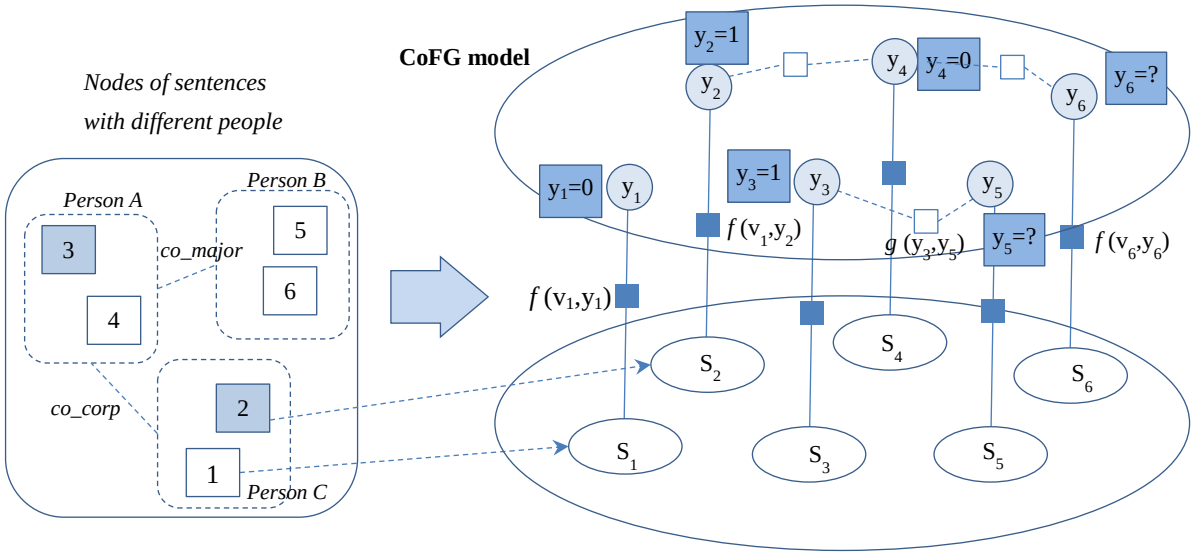

Figure 4: Graph representation of CoFG

The left figure shows the personal profile network. Each dotted line indicates a social connection. Each dotted square denotes a person, and the grey square denotes the sentence selected in the summary, and the white square denotes a sentence that is not selected as the summary..

The right figure shows the CoFG model derived from left figure. Each eclipse denotes a sentence vector of a person, and each circle indicates the hidden variable y_i. $f(v_i,y_i)$ indicates the attribute factor function. $g(y_i,y_j)$ indicates the social connection factor function.

We now briefly introduce possible ways to define the attribute functions $\{f(x_{ij}, y_i)\}_j$, and factor function $g(i, j)$.

Local textual attribute functions $\{f(x_{ij}, y_i)\}_j$: It denotes the attribute value associated with each sentence *i*. We define the local textual attribute as a feature (Lafferty et al., 2001). We can accumulate all the attribute functions and obtain local entropy for a person:

$$\frac{1}{Z_1}\exp\left(\sum_i\sum_k \alpha_k f_k(x_{ik}, y_i)\right) \qquad (6)$$

The textual attributes include following features (Shen et al., 2007; Yang et al., 2011b):

1) *BOW*: the bag-of-words of each sentence, we use unigram features as the basic textual features for each sentence.
2) *Length*: the number of terms of each sentence.
3) *Topic_words*: these are the most frequent words in the sentence after the stop words are removed.
4) *PageRank_scores*: as shown in the related work section, a document can be treated as a graph and applying a graph-based ranking algorithm (Wan and Yang., 2008). We thus use the PageRank score to reflect the importance of each sentence.

Social connection factor function $g(y_i, y_j)$: For the social correlation factor function, we define it through the pairwise network structure. That is, if the person of sentence *i* and the person of sentence *j* have a social relationship, a factor function for this social connection is defined (Tang et al., 2011a; Tang et al., 2011b), i.e.,

$$g(y_i, y_j) = \exp\left\{\beta_{ij}(y_i - y_j)^2\right\} \qquad (7)$$

The person-person social relationships are defined on Section 4, e.g. *co_major*, *co_univ*, *co_title*, and *co_corp*. We define that if two persons have at least one social connection edge, they have a social relationship. In addition, β_{ij} is the weight of the function, representing the influence degree of *i* on *j*.

To better understand our model, one example of factor decomposition is given in Figure 4. In this example, there are six sentences from three profiles. Among them, four sentences are labeled (two are labeled with the category of "1", i.e, $y=1$ and the other two are labeled with the category of "0', i.e., $y=0$) and two sentences are unlabeled (they are represented by y=?). We have six attribute functions. For example, $f(v_1, y_i)$ denotes the set

of local textual attribute functions of y_i. We also have five pairwise relationships (e.g., (y_2, y_4), (y_3, y_5)) based on the structure of the input personal profile social network. For example, $g(y_3, y_5)$ denotes social connection between y_3 and y_5, while they share the *co_major* relationship on the left figure.

5.3 Model Learning

We now address the problem of estimating the free parameters. The objective of learning the CoFG model is to estimate a parameter configuration $\theta = (\{\alpha\}, \{\beta\})$ to maximize the log-likelihood objective function $L(\theta) = \log P_\theta(Y \mid X, G)$, i.e.,

$$\theta^* = \arg\max L(\theta) \quad (9)$$

To solve the objective function, we adopt a gradient descent method. We use β (the weight of the social connection factor function $g(y_i, y_j)$) as the example to explain how we learn the parameters (the algorithm also applies to tune α by simply replacing β with α). Specifically, we first write the gradient of each β_k with regard to the objective function (Eq. 9) :

$$\frac{L(\theta)}{\beta_k} = E\left[g(i,j)\right] + E_{P_{\beta k}(Y|X,G)}\left[g(i,j)\right] \quad (10)$$

Where $E[g(i,j)]$ is the expectation of factor function $g(i,j)$ given the data distribution (essentially it can be considered as the average value of the factor function $g(i,j)$ over all pair in the training data); and $E_{P_{\beta k}(Y|X,G)}[g(i,j)]$ is the expectation of factor function $g(i,j)$ under the distribution $P_{\beta k}(Y \mid X, G)$ given by the estimated model. A similar gradient can be derived for parameter α_j.

We approximate the marginal distribution $E_{P_{\beta k}(Y|X,G)}[g(i,j)]$ using LBP (Tang et al., 2011; Zhuang et al., 2012). With the marginal probabilities, the gradient can be obtained by summing over all triads. It is worth noting that we need to perform the LBP process twice for each iteration: one is to estimate the marginal distribution of unknown variables $y_i = ?$ and the other is to estimate the marginal distribution over all pairs. In this way, the algorithm essentially performs a transfer learning over the complete network. Finally, with the obtained gradient, we update each parameter with a learning rate η. The learning algorithm is summarized in Figure 5.

Input: Network G, Learning rate η
Output: Estimated parameters θ
Initialize $\theta \leftarrow 0$
Repeat
1) Perform LBP to calculate the marginal distribution of unknown variables, i.e., $P(y_i \mid x_i, G)$
2) Perform LBP to calculate the marginal distribution of each variables, i.e., $P(y_i, y_j \mid X_{(i,j)}, G)$
3) Calculate the gradient of β_k according to Eq. 10 (for α with a similar formula)
4) Update parameter θ with the learning rate η

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \frac{L(\theta)}{\theta}$$

Until *Convergence*

Figure 5: The Learning Algorithm for CoFG model

5.4 Model Prediction and Summary Generated

We can see that in the learning process, the learning algorithm uses an additional loopy belief propagation to infer the label of unknown relationships. With the estimated parameter θ, the summarization process is to find the most likely configuration of Y for a given profile. This can be obtained by

$$Y^* = \arg\max L(Y \mid X, G, \theta) \quad (11)$$

Finally, we select a subset of sentences of each testing profile as the summary according to the trained models with top-*n* prediction scores by Y^* (Tang et al., 2011b; Dong et al, 2012).

6 Experimentation

In this section, we describe the settings of our experiment and present the experimental results of our proposed CoFG model.

	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
Random	0.0219	0.1363	0.0831	0.0288
HITS	0.0295	0.1499	0.0905	0.0355
PageRank	0.0307	0.1574	0.0944	0.0383
MaxEnt	0.0349	0.1659	0.0995	0.0377
CoFG	0.0383	0.1696	0.1015	0.0415

Table 4: Performances of different approaches to profile summarization in terms of different measurements

6.1 Experiment Settings

In the experiment, we use the corpus collected from LinkedIn.com that contains 497 profiles (see more details in Section 3). The existing summaries in these profiles are served as the reference summary (the standard answers). As discussed in subsection 3.3, the average length of summary is about 40 words. Thus, we extract 40 words to construct the summary for each profile. We use 200 personal profiles as the testing data, and the remaining ones as the training data.

We use the ROUGE-1.5.5 (Lin and Hovy, 2004) toolkit for evaluation, a popular tool that has been widely adopted by several evaluations such as DUC and TAC (Wan and Yang, 2008; Wan, 2011). We provide four of the ROUGE F-measure scores in the experimental results: ROUGE-2 (bigram-based), ROUGE-L (based on longest common subsequences), ROUGE-W (based on weighted longest common subsequence, weight=1.2), and ROUGE-SU4 (based on skip bigram with a maximum skip distance of 4).

6.2 Experimental Results

We compare the proposed CoFG approach with three baselines illustrated as follows:

- **Random**: we randomly select sentences of each profile to generate the summary for the profile.
- **HITS**: we employ the HITS algorithm to perform profile summarization (Wan and Yang, 2008). In detail, we first consider the words as hubs the sentences as authorities; Then, we rank the sentences with the authorities' scores for each profile individually; Finally, the highest ranked sentences are chosen to constitute the summary.
- **PageRank**: we employ the PageRank algorithm to perform profile summarization (Wan and Yang, 2008). In detail, we first connect the sentences of the profile with cosine text-based similar measure to construct a graph; Then, we apply PageRank algorithm to rank the sentence through the graph for each profile individually; Finally, the highest ranked sentences are chosen to constitute the summary.
- **MaxEnt**: as a supervised learning approach, maximum entropy uses textual attribute as features to train a classification model. Then, the classification model is employed to predict which sentences can be selected to generate the summary. For the implementation of MaxEnt, we employ the tool of *mallent toolkits*[4].

Table 4 shows the comparison results of our approach (CoFG) and the baseline approaches. From Table 4, we can see that 1) either HITS or PageRank outperforms the approach of random selection; 2) The supervised approach i.e. MaxEnt, outperforms both the HITS algorithm and the PageRank approach; 3) CoFG model performs best and it greatly outperforms both the unsupervised and supervised learning baseline approaches in terms of the ROUGE-2 F-measure score. This result verifies the effectiveness of considering the social connection between the sentences in different profiles,

Figure 6 shows the performance of our proposed CoFG model with different sizes of training data. From Figure 6, we can see that CoFG model with social connection always performs better than MaxEnt, and the performance of our approach descends slowly when the training dataset becomes small. Specifically, the performance of CoFG using only 10% training data achieves better performance than MaxEnt using 100% training data.

[4] http://mallet.cs.umass.edu/

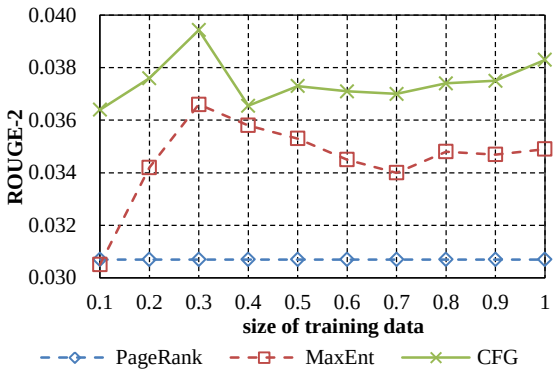

Figure 6: The performance of CoFG with different training data size

Table 5 shows the contribution of the social edges with CoFG. Specifically, CoFG is our proposed approach with both education and experience information, CoFG-*edu* means that the CoFG model considers the social edges of education field (*co_major*, *co_univ*) only, and CoFG-*exp* means that the CoFG model considers the social edges of work experience field (*co_title*, *co_corp*) only. MaxEnt can be considered as using textual information only.

	ROUGE-2
MaxEnt	0.0349
CoFG	0.0383
CoFG-*edu*	0.0382
CoFG-*exp*	0.0381

Table 5: ROUGE-2 F-Measure score of the contribution of social edges

From Table 5, we can see that all of our proposed approaches, i.e., CoFG-*edu*, CoFG-*exp*, and CoFG, outperform the baseline approach, i.e., MaxEnt. However, the performance of CoFG-*edu*, CoFG-*exp* and CoFG are similar. This result is mainly due to the fact that the information of social connection is redundant. For example, two persons who are connected by *co_major* (education field) might also be connected by *co_corp* (experience field).

7 Conclusion and Future Work

In this paper, we present a novel task named profile summarization and propose a novel approach called collective factor graph model to address this task. One distinguishing feature of the proposed approach lies in its incorporating the social connection. Empirical studies demonstrate that the social connection is effective for profile summarization, which enables our approach outperform some competitive supervised and unsupervised baselines.

The main contribution of this paper is to explore social context information to help generate the summary of the profiles, which represents an interesting research direction in social network mining. In the future work, we will explore more kinds of social context information and investigate better ways of incorporating them into profile summarization and a wider range of social network mining.

Acknowledgments

This research work is supported by the National Natural Science Foundation of China (No.61273320, No.61272257, No.61331011 and No.61375073), and National High-tech Research and Development Program of China (No.2012AA011102).

We thank Dr. Jie Tang and Honglei Zhuang for providing their software and useful suggestions about PGM. We acknowledge Dr. Xinfang Liu, Yunxia Xue and Yulai Shen for corpus construction and insightful comments. We also thank anonymous reviewers for their valuable suggestions and comments.

References

Baeza-Yates R. and B. Ribeiro-Neto. 1999. Modern Information Retrieval. *ACM Press and Addison Wesley*, 1999

Celikyilmaz A. and D. Hakkani-Tur. 2011. Discovery of Topically Coherent Sentences for Extractive Summarization. In *Proceeding of ACL-11*.

Dong Y., J. Tang, S. Wu, J. Tian, N. Chawla, J. Rao, and H. Cao. 2012. Link Prediction and Recommendation across Heterogeneous Social Networks. In *Proceedings of ICDM-12*.

Elson D., N. Dames and K. McKeown. 2010. Extracting Social Networks from Literary Fiction. In *Proceeding of ACL-10*.

Erkan G. and D. Radev. 2004. LexPageRank: Prestige in Multi-document Text Summarization. In *Proceedings of EMNLP-04*.

Guy I., N. Zwerdling, I. Ronen, D. Carmel, E. Uziel. 2010. Social Media Recommendation based on People and Tags. In *Proceeding of SIGIR-10*.

Hammersley J. and P. Clifford. 1971. Markov Field on Finite Graphs and Lattices, *Unpublished manuscript*. 1971.

Hu P., C. Sun, L. Wu, D. Ji and C. Teng. 1011. Social Summarization via Automatically Discovered Social Context. In *Proceeding of IJCNLP-11*.

Lafferty J, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML-01*.

Lappas T., K. Punera and T. Sarlos. 2011. Mining Tags Using Social Endorsement Networks. In *Proceeding of SIGIR-11*.

Leskovec J., D. Huttenlocher and J. Kleinberg. 2010. Predicting Positive and Negative Links in Online Social Networks. In *Proceedings of WWW-10*.

Lin, C. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of ACL-04 Workshop on Text Summarization Branches Out.*

Lu Y., P. Tsaparas, A. Ntoulas and L. Polanyi. 2010. Exploiting Social Context for Review Quality Prediction. In *Proceeding of WWW-10*.

Meng X，F. Wei， X. Liu， M. Zhou， S. Li and H. Wang. 2012. Entity-Centric Topic-Oriented Opinion Summarization in Twitter. In *Proceeding of KDD-12*.

Murphy K., Y. Weiss, and M. Jordan. 1999. Loopy Belief Propagation for Approximate Inference: An Empirical Study. In *Proceedings of UAI-99*.

Radev D. and K. McKeown. 1998. Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics*, 24(3):469–500.

Radev D., H. Jing, M. Stys, and D. Tam. 2004. Centroid-based Summarization of Multiple Documents. *Information Processing and Management*. 40 (2004), 919-938.

Rosenthal S. and K. McKeown. 2011. Age Prediction in Blogs: A Study of Style, Content, and OnlineBehavior in Pre- and Post-Social Media Generations. In *Proceeding of ACL-11*.

Ryang S. and T. Abekawa. 2012. Framework of Automatic Text Summarization Using Reinforcement Learning. In *Proceeding of EMNLP-2012*.

Shen D., J. Sun, H. Li, Q. Yang and Zheng Chen. 2007. Document Summarization using Conditional Random Fields. In *Proceeding of IJCAI-07.*

Tan C., L. Lee, J. Tang, L. Jiang, M. Zhou and P. Li. 2011. User-Level Sentiment Analysis Incorporating Social Networks. In *Proceedings of KDD-11*.

Tang W., H. Zhuang, and J. Tang. 2011a. Learning to Infer Social Ties in Large Networks. In *Proceedings of ECML/PKDD-11*.

Tang J., Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. Fong. 2011b. Quantitative Study of Individual Emotional States in Social Networks. I*EEE Transactions on Affective Computing*. vol.3(2), Pages 132-144.

Wan X. and J. Yang. 2008. Multi-document Summarization using Cluster-based Link Analysis. In *Proceedings of SIGIR-08.*

Wan X. 2011. Using Bilingual Information for Cross-Language Document Summarization. In P*roceedings of ACL-11*.

Wang H. and G. Zhou. 2012. Toward a Unified Framework for Standard and Update Multi-Document Summarization. *ACM Transactions on Asian Language Information Processing*. vol.11(2).

Xing E, M. Jordan, and S. Russell. 2003. A Generalized Mean Field Algorithm for Variational Inference in Exponential Families. In *Proceedings of UAI-03*.

Yang S., B. Long, A. Smola, N. Sadagopan, Z. Zheng and H. Zha. 2011a. Like like alike — Joint Friendship and Interest Propagation in Social Networks. In *Proceeding of WWW-11*.

Yang Z., K. Cai, J. Tang, L. Zhang, Z. Su and J. Li. 2011b. Social Context Summarization. In *Proceeding of SIGIR-11*.

Zhuang H, J. Tang, W. Tang, T. Lou, A. Chin, and X. Wang. 2012. Actively Learning to Infer Social Ties. In *Proceedings of Data Mining and Knowledge Discovery (DMKD-12)*, vol.25 (2), pages 270-297.