Front. Comput. Sci., 2017, 11(6): 1085–1097
DOI 10.1007/s11704-016-5088-3

RESEARCH ARTICLE

Personal summarization from profile networks

Zhongqing WANG, Shoushan LI, Guodong ZHOU (✉)

Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, Suzhou 215006, China

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 consistent organization confronted with the large amount of available information. Therefore, it is always a challenge for people to quickly find desired information from them. In this paper, we address the task of personal profile summarization by leveraging both textual information and social connection information in social networks from both unsupervised and supervised learning paradigms. Here, using social connection information is motivated by the intuition that people with similar academic, business or social background (e.g., co-major, co-university, and co-corporation) tend to have similar experiences and should have similar summaries. For unsupervised learning, we propose a collective ranking approach, called SocialRank, to combine textual information in an individual profile and social context information from relevant profiles in generating a personal profile summary. For supervised learning, we propose a collective factor graph model, called CoFG, to summarize personal profiles with local textual attribute functions and social connection factors. Extensive evaluation on a large dataset from LinkedIn.com demonstrates the usefulness of social connection information in personal profile summarization and the effectiveness of our proposed unsupervised and supervised learning approaches.

Keywords natural language processing, machine learning, social networks, personal profile summarization

Received March 7, 2015; accepted December 24, 2015

E-mail: gdzhou@suda.edu.cn

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 [1,2]. 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 with similar backgrounds [3,4]. However, as there always contains large-scale information in various kinds of profile fields, such as experience and education, it is always a challenge for people to quickly find desired information from a profile. Therefore, it is important to develop reliable methods to generate a summary of a person through his profile automatically.

A straightforward approach to handle the personal profile summarization problem is to consider it as a traditional document summarization problem, which treats each personal profile independently and generates 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 [5,6].

However, such a straightforward approach fails to benefit from the carrier of personal profiles. As the centroid of social networking, people are usually connected to others with similar background in profile network (e.g., co-major, and co-corporation). Therefore, it is reasonable to leverage various kinds of social connections to improve the performance of personal profile summarization. For example, if there is co-major, co-university, co-corporation or other academic and business relationships between two persons, they tend to share similar experiences and should have similar summaries. Hence, the challenge is how to integrate both the textual profile information and the social connection information in the social networks.

In this paper, we address the task of personal profile summarization by leveraging both textual information and social connection information in social networks from both unsupervised and supervised learning paradigms. Our intuition is to combine textual information in an individual profile and collective social context information among relevant profiles to generate a summary for a personal profile.

For unsupervised learning, we propose a graph-based collective profile summarization approach, called SocialRank, which builds a uniform graph to connect people according to their personal profiles. Specifically, we first model the connections of the sentences in the same profile according to textual information and the connections of the sentences in different profiles according to social context information, such as relevant work experience (title, company) and education (major, university) fields. Then, we rank all the sentences according to above uniform graph and extract the most important sentences to form personal profile summaries collectively.

For supervised learning, we propose a collective factor graph model called CoFG, to summarize personal profiles 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 a personal profile on one person. Specifically, we model each sentence in a profile as a vector. In the training phase, we use the vectors with social connections to build the CoFG model; while in the test 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 dataset from LinkedIn.com indicates the usefulness of social connection information and the effectiveness of our proposed unsupervised and supervised learning models in personal 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 data collection and corpus construction. In Section 4, we give an overview of our framework for profile summarization. We show the unsupervised profile summarization approach in Section 5, and supervised profile summarization approach in Section 6. In Section 7, we present our experimental results. We sum up our work and discuss future directions in Section 8.

2 Related work

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

2.1 Topic-based summarization

Generally, traditional topic-based summarization can be clustered into two categories: extractive [7] and abstractive [8] summarization.

Most of the work in text summarization has focused on extractive summarization, which forms summary by selection of important sentences from the documents [5,9]. Statistical methods are often used to find key words and phrases [10]. Discourse structure [11] also assists in specifying the most important sentences in the document. Various machine learning techniques, such as topic-model, centroid-based method have been applied for extracting features for salient sentences using training corpus [7,12,13].

A few research works have addressed single and multi-document abstractive summarization in academia. Abstractive summarization techniques are broadly classified into two categories: structured based approach and semantic based approach. Structured based approach encodes the most important information from the documents through cognitive schemas [14] such as templates, extraction rules and other structures [15,16]. In semantic based method, semantic representation of documents is used to feed into natural language generation system. This method focuses on identifying noun phrases and verb phrases by processing linguistic data [17,18].

Compared to above unsupervised learning studies, there are few supervised learning studies in summarization. For example, Shen et al. presented a conditional random fields based framework to treat the summarization task as a sequence labeling problem [19]. Wong et al. utilized various kinds of features, such as content, relevance and event features to train the learning model [20]. Meng et al. proposed a supervised entity-centric summarization framework to produce opinion summaries in accordance with topics and remarkably emphasizing the insight behind the opinions [21].

Different from all existing studies, our study attempts to consider both textual information and social connection information in both unsupervised and supervised summarization.

2.2 Social-based summarization

Due to the increasing popularity of Web 2.0 in empowering people to actively interact with each other, studies on social media have been drawing more and more attention recently [3,22]. Social-based summarization is exactly a special case of summarization where social information is employed to help generating the summary. Compared to topic-based summarization, there are only a few studies on social-based summarization.

For example, Hu et al. proposed an unsupervised PageRank-based social summarization approach by incorporating both document context and user context in the sentence evaluation process [23]. Meng et al. 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) [21].

Unlike above studies, this study addresses a novel task of personal profile summarization with focus on employing various kinds of social information in personal profiles, such as co-major, and co-corporation between two people.

2.3 Factor graph model

Among the increasing studies in social networks [2,24–26], factor graph model (FGM) has become a popular approach to describe the relationship in social networks [27,28]. In principle, factor graph model builds a graph to represent the relationship of nodes in social networks via various kinds of factor functions.

For example, Tang et al. and Zhuang et al. formalized the problem of social relationship learning into a semi-supervised framework, and proposed partially-labeled pair-wise factor graph model (PLP-FGM) to infer the types of social ties [27,28]. Dong et al. gave a formal definition of link recommendation across heterogeneous networks, and proposed a ranking factor graph model (RFG) to predict links in social networks [29]. Yang et al. 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 [30].

Different from above studies, this study proposes a pairwise factor graph model to collectively utilize both textual information and social connection information to generate the summary of a personal profile.

3 Data collection and corpus construction

As a novel task, there is lack data in personal profile summarization. In this study, we collect a data set containing personal summaries with corresponding information, such as self-introduction and personal profiles from social media.

3.1 Data collection

We collect the data set from LinkedIn.com[1)], which 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.

In this study, the data set is crawled as follows. To begin with, ten public profiles are randomly selected as seed profiles, and then those profiles from their "People Also Viewed" field are collected. In this way, we get a total of 3 182 public profiles[2)]. For privacy issue, we ignore personal names in public profiles.

Figure 1 shows an example of a personal profile from LinkedIn.com, which includes following fields:

• overview, which gives a structure description of a person's general information, such as current/previous position and workplace, brief education background and general technical background;

• summary, which summarizes a person's work, experience and education;

• experience, which details a person's work experience;

• education, which details a person's education background.

Among these fields, only the overview field is required, while the others, such as project, course and interest groups, are optional. Due to the importance of various fields to personal profile summarization, this study only considers the overview, summary, experience and education fields and ignores the others.

Table 1 shows the statistics of major fields in our data col-

1) http://www.linkedin.com
2) We collect all the data from LinkedIn.com on December 17, 2012

lection of 3 182 personal profiles. It shows that, most profiles are incomplete and that the experience field is much longer than other fields. It also shows that most people provide their experience and education information. However, only about 30% of people fill the summary field. This is mainly because writing summary is normally more difficult than writing other fields. Therefore, it is highly desirable to develop reliable automatic methods to generate a summary of a person through his profile.

John Smith (a pseudo example)	
Overview	
Current	Applied researcher at Apple Inc.
Previous	Senior research scientist at IBM …
Education	Massachusetts Institute of Technology (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–Nowadays Query recognition and relevance; …	
Education	
MIT *Ph D, electrical engineering*, 2002–2008 …	

Fig. 1 An example of a personal profile from LinkedIn.com

Table 1 Statistics of major fields in our data set

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

3.2 Corpus construction

Among 921 profiles which contain the summary, we manually select 497 profiles with high quality summary to construct the corpus for our evaluation. 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 could give an overview of a person, e.g., representing the education and experience information of a person. Since the data set is hard to be collected and analyzed, we only select limited data for this study. A large-scale data set will be collected as the future work.

Since the experience field contains most of information for a person, we treat the text in the experience field as the source of personal summarization for each profile. Besides, we collect social connection information from the education and experience fields, explicitly included by LinkedIn. Table 2 shows the average length of the summary and experience fields in our corpus, with the compression ratio of 1:10.

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

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

Figure 2 shows an example of a social network with the dotted line for the social connection of people from the profiles of LinkedIn. Obviously, people can be connected by various kinds of social connections. For example, John and Lucy are connected by co_unvi relationship, and Lily and Linda are connected by co_corp relationship. From LinkedIn, four kinds of social relationship of people are extracted from the education and experience fields:

• co_major, denoting that two persons have the same major at school

• co_univ, denoting that two persons are graduated from the same university

• co_title, denoting that two persons have the same title on corporation.

• co_corp, denoting that two persons work on the same corporation.

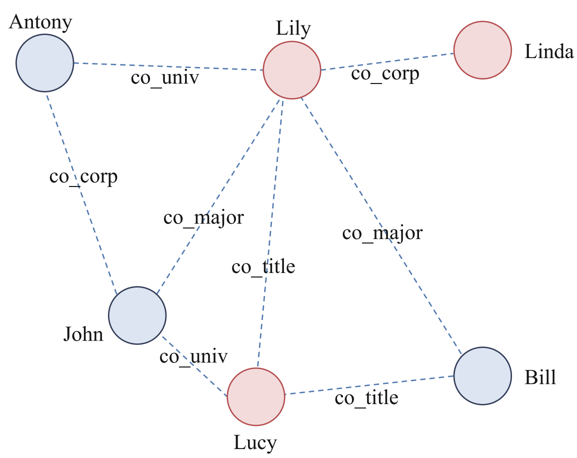

Fig. 2 An example of a personal profile network (red for female, blue for male, and the dotted line for social connection between two persons)

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

Table 3 gives the statistics over various kinds of social connections in our corpus. From Table 3, we can see that given 497 users, there exist 14 307 social connections. That is, a user has 29 social connections on average. We can also see that the number of social connections from the education field is comparable to the number of social connections from the experience filed. Besides, among all the four kinds of social connections, co_unvi is the most common.

Table 3 Statistics over various kinds of social connections in our corpus

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

4 Methodology

To generate summaries for personal profiles, a straightforward approach is to treat each personal profile independently and generate a summary for each personal profile individually. As mentioned in Section 3, we use the sentences of the experience field as a text document and consider them as the source of summarization for each profile.

Instead, we formalize the problem of personal profile summarization in the pair-wise graph model and propose graph-based approach to learn the model for generating the summary of the profile. Our basic idea is to combine textual information in an individual profile and collective social context information among different profiles to generate a summary for a personal profile.

The overview of the proposed method is shown in Figure 3. First, we treat each sentence of profiles as vectors with textual information (local textual attribute functions); Second, all the vectors are connected by social connection relationships and we model these vectors and their relationships into the graph; third, we propose a graph-based learning algorithm (SocialRank or CoFG) 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 how to define the graph model (SocialRank or CoFG) to connect profiles with social connection. We introduce both supervised and unsupervised learning approaches to generate summaries on the following sections.

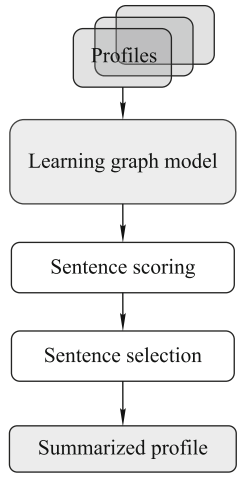

Fig. 3 The overview of our proposed framework

5 SocialRank: unsupervised ranking for personal summarization

In this section, we will introduce our unsupervised collective profile summarization framework.

5.1 Basic idea

We propose a collective profile summarization approach, called SocialRank, to combine the textual information in an individual profile and the collective social context information among different profiles to generate a summary for a personal profile. A uniform graph is proposed to connect profiles with social contextual information. Figure 4 presents an example. In the graph, each gray circle indicates a sentence of a profile. The two-headed arrow represents the text-based similarities (correlation) relationship between two sentences. The black square and gray square separately denote the experience and education connection between two profiles.

Specifically, sentences are connected by two situations: 1) the sentences in the same profile are connected according to textual information; 2) the sentences in different profiles are connected according to social context information from relevant profile fields (e.g., education, and experience). By learning such a graphical model, we propose a graph-based ranking model to rank which sentences of the profile are important (or informative) with social context to generate the summary for each profile.

Figure 5 illustrates the overall procedure of our proposed collective personal profile summarization approach. In the rest of this section, we explain in detailed how we incorporate social context to the graph-based sentence model, and rank the sentence through the graph.

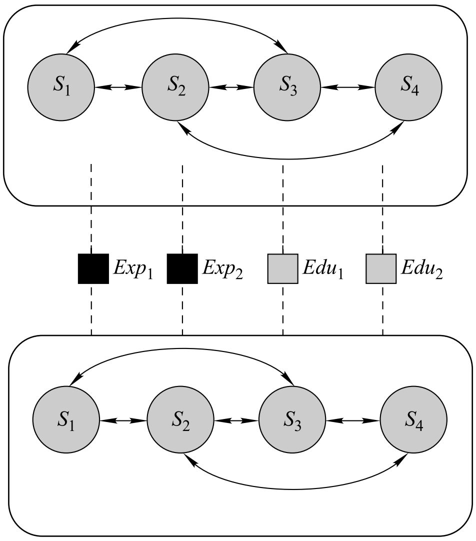

Fig. 4 An example of collective profile summarization with social networking

Input:
The sentences of all the profiles
Output:
The summary of each profile
Process:

1) Build a uniform graph G to combine textual information in an individual profile and collective social context information among different profiles.
2) Rank the sentences through the graph G.
3) Select the sentences with high ranking scores to generate the summary for each profile.

Fig. 5 The algorithm of the approach to collective personal profile summarization

5.2 Graph construction and ranking

We model the collective personal profile summarization problem in the graph-based ranking model. Each sentence of the profile is associated to a uniform graph with textual and social contextual information, and ranks the sentences through the graph to indicate the importance of the sentence to generate the summary.

• **Graph construction**

Formally, we construct a uniform graph $G = (V, E)$ to reflect the textual and social relationships between sentences of the profiles. In the graph G,

1) V is the set of vertices of sentences. For each vertex v_i,

• v_{TXTi} represents the v_i with textual features, when v_i connects with the sentence on the same profile;

• v_{SOCi} represents the v_i with social contextual features, when v_i connects with the sentence on different profiles.

2) E is the set of edges. Each edge e_{ij} in$E$ is associated with an affinity weight $f(i \rightarrow j)$ between sentences i and j ($i \neq j$). The affinity weight is computed with following two conditions:

• if sentences i and j are from the same profile, then the affinity weight $f(i \rightarrow j)$ is computed by v_{TXTi} and v_{TXTj} with the textual features;

• if sentences i and j are from different profiles, then the affinity weight $f(i \rightarrow j)$ is computed by v_{SOCi} and v_{SOCj} with the social contextual features.

Thus, the affinity weights of sentences i and j are computed by the standard cosine measure (Baeza-Yates and Ribeiro-Neto, 1999), i.e.,

$$f(i \rightarrow j) = \begin{cases} \dfrac{v_{TXTi} \cdot v_{TXTj}}{|v_{TXTi}||v_{TXTj}|}, & \text{if } i \text{ and } j \text{ are from the same profile;} \\ \dfrac{v_{SOCi} \cdot v_{SOCj}}{|v_{SOCi}||v_{SOCj}|}, & \text{if } i \text{ and } j \text{ are from different profiles.} \end{cases} \tag{1}$$

• **Feature representation**

The textual features of v_{TXTi} represent the textual information of sentence i, we employ the bag-of-words model to represent the textual information as a vector for sentence i.

The social contextual features of v_{SOCj} represent the social context information of sentence i. We employ four kinds of social contextual features to construct the vector. Table 4 shows the social contextual feature we use.

Table 4 illustrates the social contextual features we use. If two persons have the same education background or the same work experience, these two persons may be connected, and their summaries may be similar. As university and major could represent the education background of a person, and company and title could represent the work experience of a person, we extract these four features to construct the social feature vector of each sentence. All these social features are extracted from education and experience field of profiles (i.e., bold and italic style strings of the corresponding fields of Fig. 1).

• **Sentences ranking**

After computing the affinity weights, the transition probability from sentence i to sentence j is then defined by normalizing the corresponding affinity weight as follows.

Table 4 The social contextual features extracted from the education and experience fields

Field	Features	Description
Education	University	The graduated universities of the person
	Major	The major of the person studied
Experience	Company	The companies of the person currently worked or previously worked
	Title	The titles of the person when he is at the corresponding companies

$$p(i \to j) = \frac{f(i \to j)}{\sum\limits_{k \neq i} f(i \to k)}. \tag{2}$$

Given the graph G, the saliency score score(i) for sentence i can be deduced from those of all other sentences linked with it and it can be formulated in a recursive form.

$$\text{score}(i) = \mu \sum_{j \neq i} \text{score}(j) p(j \to i) + (1 - \mu). \tag{3}$$

In the implementation, μ is the damping factor and usually set to be 0.85 [31]. The initial scores of all sentences are set to one, and the iteration algorithm is adopted until convergence [5].

As long as the saliency scores of sentences are obtained, the sentences are ranked with the scores. The sentences of each profile with high ranking scores form the summary individually.

6 CoFG: supervised ranking for personal summarization

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.

6.1 Overview of our framework

We formalize the problem of personal profile summarization in a pair-wise factor graph model and propose an approach referred to 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 maximize the joint probability of the input graph.

6.2 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

$$P(Y|X, G) = \frac{P(X, G|Y) P(Y)}{P(X, G)}. \tag{4}$$

Here, G denotes all forms of network information. This probabilistic formulation indicates that labels of sentences depend on not only local attributes X, but also the structure of the network G. According to Bayes' rule [29,32], we have

$$\begin{aligned} P(Y|X, G) &= \frac{P(X, G|Y) P(Y)}{P(X, G)} \\ &\propto P(X|Y) P(Y|G), \end{aligned} \tag{5}$$

where $P(Y|G)$ represents the probability of labels given the structure of the network and $P(X|Y)$ 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, and thus we have

$$P(Y|X, G) \propto P(Y|G) \prod_i P(x_i|y_i), \tag{6}$$

where $P(x_i|y_i)$ is the probability of generating attributes x_i given the label y_i. Now, the problem becomes how to instantiate the probability $P(Y|G)$ and $P(x_i|y_i)$. We model them in a Markov random field, and thus according to the Hammersley-Clifford theorem [33], the two probabilities can be instantiated as follows:

$$P(x_i|y_i) = \frac{1}{Z_1} \exp\left\{\sum_{j=1}^{d} \alpha_j f_j\left(x_{ij}, y_i\right)\right\}, \tag{7}$$

$$P(Y|G) = \frac{1}{Z_2} \exp\left\{\sum_i \sum_{j \in NB(i)} g(i, j)\right\}, \tag{8}$$

where Z_1 and Z_2 are normalization factors. Equation (7) 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 jth attribute. Equation (8) 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.

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

1) 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 [34]. We can accumulate all the attribute functions and obtain local entropy for a person:

$$\frac{1}{Z_1}\exp\left(\sum_i\sum_k \alpha_k f_k\left(x_{ik}, y_i\right)\right). \tag{9}$$

The textual attributes include following features [19,30]:

• BOW: the bag-of-words of each sentence, we use unigram features as the basic textual features for each sentence;

• Length: the number of terms of each sentence;

• Topic_words: these are the most frequent words in the sentence after the stop words are removed;

• PageRank_scores: as shown in the related work section, a document can be treated as a graph and applying a graph-based ranking algorithm [5]. We thus use the PageRank score to reflect the importance of each sentence.

2) 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 [27,32], i.e.,

$$g\left(y_i, y_j\right) = \exp\left\{\beta_{ij}\left(y_i - y_j\right)^2\right\}. \tag{10}$$

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 Fig. 6. The left part of Fig. 6 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 part of Fig. 6 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. 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 pair-wise 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.

6.3 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 $L(\theta) = \log P_\theta(Y|X, G)$, i.e.,

$$\theta* = \arg\max L(\theta). \tag{11}$$

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

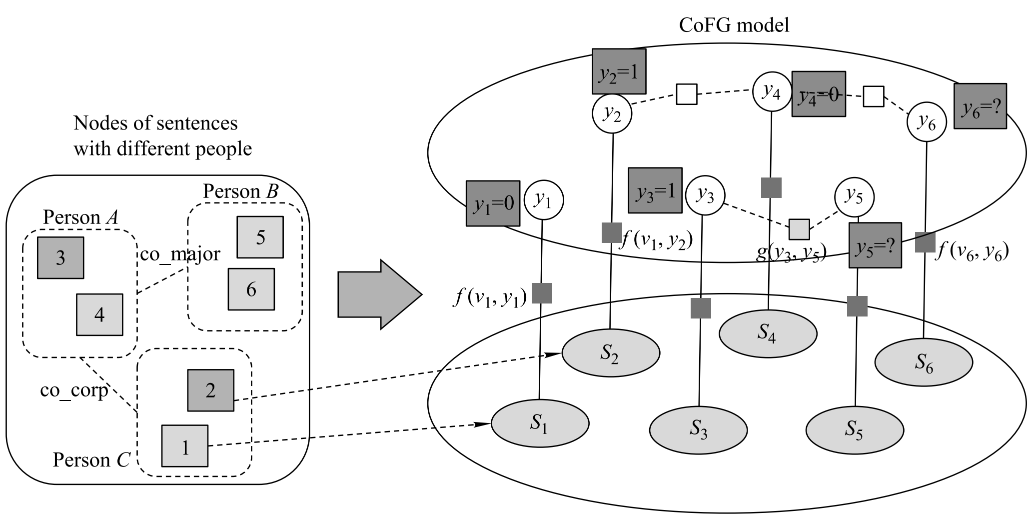

Fig. 6 Graph representation of CoFG

(Eq. (11)):

$$\frac{L(\theta)}{\beta_k} = E\left[g(i, j)\right] + E_{P_{\beta k}(Y|X,G)}\left[g(i, j)\right], \tag{12}$$

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 $P_{\beta k}(Y|X, G)$ given by the estimated model. A similar gradient can be derived for parameter a_j.

We approximate the marginal distribution $E_{P_{\beta k}(Y|X,G)}[g(i, j)]$ using LBP [28,32]. 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 rateη. The learning algorithm is summarized in Fig. 7.

Input: network G, learning rate η
Output: estimated parameters θ
Initialize $\theta \leftarrow 0$
Repreat to
1) perform LBP to calculate the marginal distribution of unknown variables, i.e., $P(y_i|x_i,G)$;
2) perform LBP to calculate the marginal distribution of each variables, i.e., $P(y_i,y_i|X_{(i,j)},G)$;
3) calculate the gradient of β_k according to Eq. (10) (for a 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

Fig. 7 The learning algorithm for CoFG model

6.4 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

$$Y^* = \arg\max L(Y|X, G, \theta). \tag{13}$$

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^* [29].

7 Experimentation

In this section, we describe the settings of our experiment and present the experimental results of the proposed both supervised and unsupervised learning models.

7.1 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 Section 3.3, the average length of text on summary field 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 for supervised learning (CoFG).

We use the ROUGE-1.5.5 [35] toolkit for evaluation, a popular tool that has been widely adopted by several evaluations such as DUC and TAC [5,6]. 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). We conduct significant tests using t-test to see whether an improvement is statistically significant.

7.2 Overall experimental results

Firstly, we compare the proposed unsupervised ranking approach SocialRank and supervised rank approach CoFG with four 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 [5]. In detail, we first consider the words as hubs and 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 [5]. 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 mallet toolkits[3)].

Table 5 shows the comparison results of our approaches (SocialRank, and CoFG) and the baseline approaches. From Table 5, 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) although SocialRank is an unsupervised learning approach, it outperforms the MaxEnt by considering social connection; 4) CoFG model performs best and it significantly outperforms both the unsupervised and supervised learning approaches in terms of the ROUGE-2/W/SU4 F-measure score (p-value<0.05). This result verifies the effectiveness of considering the social connection between the sentences in different profiles. Note that the ROUGE-L performance of CoFG is lower than SocialRank, which may be due to that the SocialRank tends to find the similar long common subsequence between sentences, while the CoFG tends to find more representative sentences.

Table 5 Performances of different approaches to profile summarization in terms of different measurements

Approach	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
Random	0.021 9	0.136 3	0.083 1	0.028 8
HITS	0.029 5	0.149 9	0.090 5	0.035 5
PageRank	0.033 8	0.176 6	0.088 0	0.039 6
MaxEnt	0.034 9	0.165 9	0.099 5	0.037 7
SocialRank	0.036 0	0.168 0	0.084 9	0.035 6
CoFG	0.038 3	0.169 6	0.101 5	0.041 5

Figure 8 illustrates the examples of the results with different approaches. From the figure, we could find both the results of SocialRank and CoFG could represent the summary of the personal experience, while the result of SocialRank focuses on the details, and the result of CoFG is more general.

Gold Summary
My goal is to continually improve my technical leadership skills, and hardware ownership experience. I have certified principal engineer (CPE) experience on the Orion program with the thrust vector control system. I also have extensive design experience from concept to production, and the extensive experience in creating and reviewing drawings and requirement specifications.
Result of SocialRank
I have experience on noise modeling of advanced CMOS devices, and device and process simulation for advanced CMOS technology. My projects are about stress simulation of advanced MOSFETs. I'm the RF modeling team lead for advanced bulk CMOS technology. My major is statistical modeling (MonteCarlo and fixed corners) of advanced MOSFETs
Result of CoFG
I have certified principal engineer. My responsible of the project is for electromechanically devices in the human space flight line of business. I have experience on conceptualize, prototype, procure, design, and support fabrication of hardware required for special tests, ground tests, and tooling.

Fig. 8 Example of results with different approaches

7.3 Experimental results of SocialRank

We then analyze the results of unsupervised learning approach SocialRank in detail.

We first analyze the effectiveness of social context features with ROUGE-2 F-measure scores on Table 6. Remember that our approach, i.e., SocialRank without social contextual features is degenerated to a PageRank approach. Thus, our approach is denoted PageRank in Table 6 when none social contextual features are employed.

Table 6 Performances of our approach in terms of ROUGE-2 F-measure when using different kinds of social context features

Approach	ROUGE-2
PageRank	0.033 8
+Exp	0.035 3
+Edu	0.032 6
+Exp+Edu	0.036 0

From Table 6, we can see that 1) both experience information (+Exp) and education information (+Edu) are very useful for profile summarization; 2) the experience information is shown to be more effective than the education information.

3) http://mallet.cs.umass.edu/

This is mainly because the experience information is capable of better representing the background of a person; 3) integrating these two kinds of information (+Exp+Edu) is superior to using any single one of them. It is shown that incorporating both experience and education information is very effective for summarizing the personal profiles.

We then analyze the sensitiveness of the weights of experience and education features. From Fig. 9, we can see that our approach of SocialRank is not very sensitive to the weights of social context features. It robustly outperforms the PageRank approach with different weights of social context features.

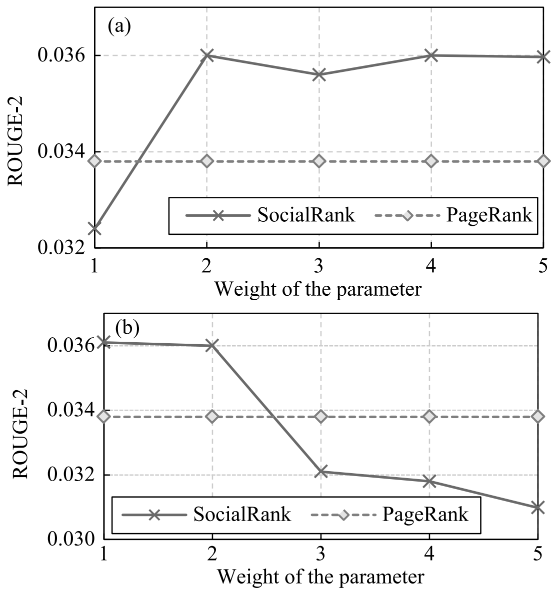

Fig. 9 Weight sensitive of SocialRank in terms of ROUGE-2 F-measure: Performances of SocialRank and PageRank when varying the weights of (a) experience features and (b) education features

7.4 Experimental results of CoFG

In this subsection, we analyze the results of supervised learning approach CoFG in detail.

Figure 10 shows the performance of our proposed CoFG model with different sizes of training data. From Fig. 9, 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, since the CoFG approach could get global optimization and the score of each sentence would be influenced by related ones, the performance of CoFG using only 10% training data achieves better performance than MaxEnt using 100% training data.

Table 7 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.

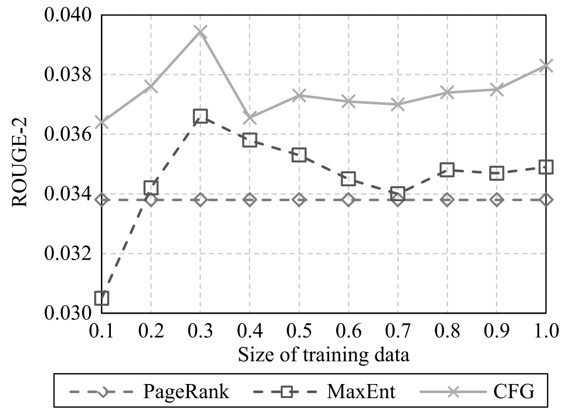

Fig. 10 The performance of CoFG with different training data sizes

Table 7 ROUGE-2 F-measure score of the contribution of social edges

	ROUGE-2
MaxEnt	0.034 9
CoFG	0.038 3
CoFG-edu	0.038 2
CoFG-exp	0.038 1

From Table 7, 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).

8 Conclusion and future work

In this paper, we present a novel task named profile summarization and propose both unsupervised and supervised learning approaches 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 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.

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

The work was supported by the National Natural Science Foundation of China (Grant Nos. 61273320, 61375073, and 61402314) and the Key Project of the National Natural Science Foundation of China (61331011).

References

1. Lappas T, Punera K, Sarlos T. Mining tags using social endorsement networks. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011, 195–204
2. Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P. User-level sentiment analysis incorporating social networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011, 1397–1405
3. Yang S H, Long B, Smola A, Sadagopan N, Zheng Z H, Zha H Y. Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the International Conference on World Wide Web. 2011, 537–546
4. Guy I, Zwerdling N, Ronen I, Carmel D, Uziel E. Social media recommendation based on people and tags. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010, 194–201
5. Wan X J, and Yang J W. Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008, 299–306
6. Wan X J. Using bilingual information for cross-language document summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011, 1546–1555
7. Radev D R, Jing H Y, Stys M, Tam D. Centroid-based summarization of multiple documents. Information Processing and Management, 2004, 40(6): 919–938
8. Radev D R, McKeown K R. Generating natural language summaries from multiple on-line sources. Computational Linguistics, 1998, 24(3): 469–500
9. Kupiec J, Pedersen J, Chen F. A trainable document summarizer. In: Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1995, 68–73
10. Luhn H P. The automatic creation of literature abstracts. IBM Journal of Research and Development, 1958, 2(2): 159–165
11. Knight K, Marcu D. Statistics-based summarization—step one: sentence compression. In: Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence. 2000, 703–710
12. Celikyilmaz A, Hakkani-Tur D. Discovery of topically coherent sentences for extractive summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011, 491–499
13. Wang H L, Zhou G D. Toward a unified framework for standard and update multi-document summarization. ACM Transactions on Asian Language Information Processing, 2012, 11(2)
14. Genest P E, Lapalme G. Framework for abstractive summarization using text-to-text generation. In: Proceedings of the Workshop on Monolingual Text-To-Text Generation. 2011, 64–73
15. Barzilay R, McKeown K R. Sentence fusion for multi-document news summarization. Computational Linguistics, 2005, 31(3): 297–328
16. Khan A, Salim N. A review on abstractive summarization methods. Journal of Theoretical and Applied Information Technology, 2014, 59(1): 64–72
17. Saggion H, Lapalme G. Generating indicative-informative summaries with sumUM. Computational Linguistics, 2002, 28(4): 497–526
18. Ganesan K, Zhai C X, Han J W. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd International Conference on Computational Linguistics. 2010, 340–348
19. Shen D, Sun J T, Li H, Yang Q, Chen Z. Document summarization using conditional random fields. In: Proceedings of the International Joint Conference on Artifical Intelligence. 2007, 2862–2867
20. Wong K F, Wu M L, Li W J. Extractive summarization using supervised and semi-supervised learning. In: Proceedings of the 22nd International Conference on Computational Linguistics. 2008, 985–992
21. Meng X F, Wei F R, Liu X H, Zhou M, Li S J, Wang H F. Entity-centric topic-oriented opinion summarization in Twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012, 379–387
22. Rosenthal S, McKeown K. Age prediction in blogs: a study of style, content, and online-behavior in pre- and post-social media generations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011, 763–772
23. Hu P, Sun C, Wu L F, Ji D H, Teng C. Social summarization via automatically discovered social context. In: Proceedings of the International Joint Conference on Natural Language Processing. 2011, 483–490
24. Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web. 2010, 641–650
25. Lu Y, Tsaparas P, Ntoulas A, Polanyi L. Exploiting social context for review quality prediction. In: Proceedings of the 19th International Conference on World Wide Web. 2010, 691–700
26. Guy I, Zwerdling N, Ronen I, Carmel D, Uziel E. Social media recommendation based on people and tags. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010, 194–201
27. Tang W B, Zhuang H L, Tang J. Learning to infer social ties in large networks. In: Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery. 2011, 381–397
28. Zhuang H L, Tang J, Tang W B, Lou T C, Chin A, Wang X. Actively learning to infer social ties. In: Proceedings of Data Mining and Knowledge Discovery. 2012, 270–297
29. Dong Y X, Tang J, Wu S, Tian J L, Chawla N V, Rao J H, Cao H H. Link prediction and recommendation across heterogeneous social net-

works. In: Proceedings of the 12th IEEE International Conference on Data Mining. 2012, 181–190

30. Yang Z, Cai K K, Tang J, Zhang L, Su Z, Li J Z. Social context summarization. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011, 255–264
31. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: bringing order to the Web. Technical Report, 1998
32. Tang J, Zhang Y, Sun J M, Rao J H, Yu W J, Chen Y R, Fong A C M. Quantitative study of individual emotional states in social networks. IEEE Transactions on Affective Computing, 2011, 3(2): 132–144
33. Hammersley J, Clifford J. Markov field on finite graphs and lattices. 1971, CiteULike: 8970271
34. Lafferty J, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning. 2001, 282–289
35. Lin C Y. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of ACL-04 Workshop on Text Summarization Branches Out. 2004

Zhongqing Wang received the Master's degree in July 2012 from the School of Computer Science and Technology, Soochow University, China. Since 2012, he has been a PhD candidate at the School of Computer Science and Technology, Soochow University. His current research interests include natural language processing, sentiment analysis, and social computing.

Shoushan Li received his PhD degree in 2008 from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, China. He is a full professor in the School of Computer Science and Technology, Soochow University, China. His current research interests include natural language processing, social computing, and sentient analysis.

Guodong Zhou received his PhD degree in 1999 from the National University of Singapore, Singapore. He is a full professor in the School of Computer Science and Technology, and the director of the Natural Language Processing Laboratory from Soochow University, China. His research interests include information retrieval and natural language processing.