"Read" More from Business Cards: Toward a Smart Social Contact Management System

Bin Guo, Daqing Zhang, Dingqi Yang
Institut TELECOM - TELECOM SudParis, France
Email: guobin.keio@gmail.com, {daqing.zhang, dingqi.yang}@it-sudparis.eu

***Abstract*— The ability to leverage the power of a network of social contacts is important to get things done. However, as the number of contacts increases, people often find it difficult to maintain their contact network by using merely memory, and are frequently encompassed with questions like "who is that person, I met him in Tokyo last year". Existing contact tools make up for the shortage of unreliable human memory by storing contact information in the digital format, but laying much burden on users on manually inputting contact data. This paper, however, presents a social contact management system called SCM, which supports the auto-collection of rich contact data by exploring the aggregated power of pervasive sensing and Web intelligence techniques. Regarding that people often need to leverage several associated things (e.g., meeting location) to fetch other information about a contact (e.g., his name), we also develop an associative contact retrieval method. The effectiveness and runtime performance of our system is validated through a set of experiments.**

Keywords- Social Contact Management, Pervasive Computing, Web Intelligence, HCI, Information Integration

I. INTRODUCTION

In modern life, people participate in various social activities and meet numerous people day by day. All the acquaintances form a social contact network (SCN) of a person. The ability to manage the SCN and leverage it to get things done, however, becomes a significant, yet difficult task. The major trouble here is that people find it impossible to maintain the ever-increasing contact information merely using their memory. Various aiding tools are thus exploited. Before the era of computing, it often takes the physical ways like address book writing and note taking, but they suffer from problems like possible loss and inefficient search support. Currently, the focus has been changed to digital contact tools. Though enabling reliable storage and enhanced search support, it still faces several issues. In the following, we take a typical socially-active community − the academic community − for example, to illustrate the issues.

Bob is an active researcher. He participates in various academic activities, such as attending conferences, giving talks, and so on, where he acquires numerous new contacts. To better maintain his SCN, Bob uses a digital contact book for storing contact information. However, he finds two issues.

(1) *Manually inputting contact info is a big burden*. In addition to *basic info* (e.g., affiliation, position, e-mail, etc.), Bob is interest to some *academic info* of a contact, such as his research interest, education history (e.g., his Ph.D. university), relationship with others, etc. Such info is useful because it can facilitate a set of further tasks like expert finding and contact grouping. Though valuable, the cost of manually collecting and inputting such info in a contact tool is considerably high.

(2) *Finding contact info can be difficult.* People often need to recollect info about a contact. Using a traditional tool, it is easy to find the needed info if we know the contact name. However, the problem is that we often forget contact names. For example, Bob has ever seen someone he met before in an airport. He wants to talk with that person but cannot recall his info, even his name. The only thing that is clear in his mind is "*I met him in a conference held in Tokyo in 2009; he is humorous*". Traditional contact tools do not work in such cases, and thus more time is used for info finding.

To address the above issues, we develop SCM (Social Contact Manager), which aims at allowing people to better manage their social contacts. The paper chose the academic community to do a showcase study, but the technologies developed can be applied to other communities. The main contributions of our work are two folds.

- *Tech-aided contact data gathering*. To lessen user effort on contact data recording, we leverage a combination of pervasive sensing and Web intelligence techniques for extracting needed information. Our solution is inspired by the general contact acquaintance process: in social occasions, our connection with new contact usually starts from exchanging *business cards*; after getting basic info from *business cards*, people try to fetch more information about the contact from the Web. An interesting phenomenon revealed here is that "business cards" play a key role in it, which triggers and leads the contact data gathering process. SCM explores techniques to automate this process: We employ a wearable card-scanner to extract basic info from the collected business cards, which is then used to extract other contact information from the Web, using a hybrid of heuristic rules and CRF (Conditional Random Field) [1] based information extraction method.
- *Associative contact search*. It is difficult to search a contact if we forget the name, so we need to find a way to search without knowing contact name. As reported in memory-related studies [2, 3], people often recount associated events or cues that go with the target item. In the context of contact search, it can refer to meeting-

event contexts (e.g., location), user impression, and profile info (e.g., position) that associates with the contact. To this end, we have also developed an interface that supports search by association of contacts leveraging memory cues.

II. RELATED WORK

Many systems have been developed to facilitate *contact communication*, such as Skype and Outlook. There are also systems that *recommend contacts* to a user based on his profile and context. In WhozThat [4], users within a public place (e.g., in a bar) can exchange their profiles to find someone interest. Our system differs from them on the purpose – we want to develop a system that supports collecting and recalling the information of human contacts.

To the best of our knowledge, our system is a pioneering one that explores the fusion of *sensed* and *Web-extracted* data for building pervasive apps. The two data sources have distinct strength: (1) Web is a major source to extract static or slowly changing information, such as user profile; (2) Pervasive sensing enables the detection of human activities and social interactions in the physical world [5]. Due to the diverse features, aggregation of data from the two distinct sources provides unique opportunities to pervasive applications. By gathering contextual and basic contact info using portable sensors from the real world, and extracting bio info from the Web, SCM makes an attempt to illustrate the aggregated effects of Web/Sensing data sources by solving a compelling human problem – social contact management.

III. THE SCM SYSTEM DESIGN

SCM is the first tool that fully supports *tech-aided contact data collection* and *associative contact recall*. The biggest challenge here is how to gather the needed contact data. It can be divided into two sub-issues: *what data to gather* and *how*. The data to be gathered is determined by two factors: (1) the needs of users in the application domain; (2) the needs of the contact search task. Based on the two factors and the use case presented in the introduction, we identify four sets of data to be gathered, as shown in Fig. 1.

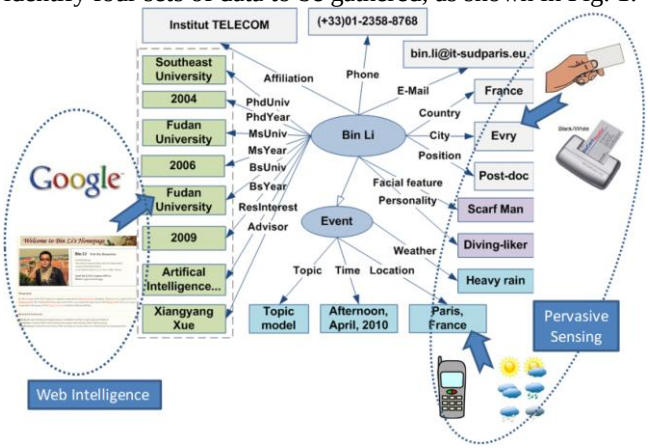

Figure 1. Contact data gathering.

- *Basic info*. It involves person name, affiliation, country, city, E-Mail address, and so on. They form common information to all types of business communications.
- *Selected bio-info*. Education history, research interest, and social relationship, is important information academic community. Eight types of biographical data are carefully selected, as illustrated in Fig. 1.
- *Contextual cues*. To support associative search, we need to gather contexts that refer to the physical events we obtain contact data. Four types of contexts are defined, which refer to when, where, in what weather the event happens and the relevant topic discussed.
- *Impression*. User impression to a contact, such as his facial features (e.g., beard) and personality (e.g., a diving-liker), is also crucial cues for contact recall.

IV. CONTACT DATA GATHERING

Tech-aided contact data gathering is the fundamental function supported by SCM. In the following, we first describe how *bio-info* is extracted, and then present the techniques used for *the other three data sets*.

A) Business Card Enhanced Homepage Finder

As presented in the introduction, we propose a business card triggered bio-info gathering method: the basic info from a business card is used to find the homepage of the contact; afterwards, a CRF-based method is used to extract needed info from the homepage (presented in next subsection).

We recognize personal homepages as a public repository to extract a researcher's bio-info, but the problem along with it is: *how to determine the homepage of a person*? Search engines like Google can provide us a set of candidates relevant to a person if we use his name as the keyword. However, it is still difficult to find the right homepage from it because (1) the namesake problem (e.g., there could be many people called "Bin Li" in China), and (2) the result set always includes several types of web pages (e.g., news, digital libraries of papers, social websites, etc.) that are noisy. In terms of the characteristic of our application field, we propose a business-card enhanced homepage finding method, which is implemented in two steps, as mentioned below.

1) Collection of a High-Relevance Candidate Set

Namesake is a well-known problem in Web people search field. Previous research mainly focuses on clustering the result set into *k* groups when a person name is provided [6]. As a Web people search system, the clustering result can be returned to the searcher for final decision.

Our system faces a somewhat different problem because we are not building a Web people search system; to reduce user intervention, there also lacks the user decision process. Then, is there a way to deal with this new situation? Let us first review what people will do when facing the namesake problem. Actually, if a person finds that there are more people with the same name on the Web, he may use additional keywords that can characterize the target user (e.g., his affiliation) to filter irrelevant results. Our solution is inspired by this, which attempts to employ a contact's basic information that can be obtained from his business card, to deal with the namesake problem. As shown in Fig. 2, four types of contact data obtained from a contact's name card,

including *E-Mail*, *country*, *city*, and *affiliation*, in combination with *person name*, are used for enhanced people search. For each keyword-pair, we retain the top eight search results returned from Google search API (http://code.google.com/apis/ajaxsearch/). Pre-processing to the four obtained result sets is then performed, as mentioned below:

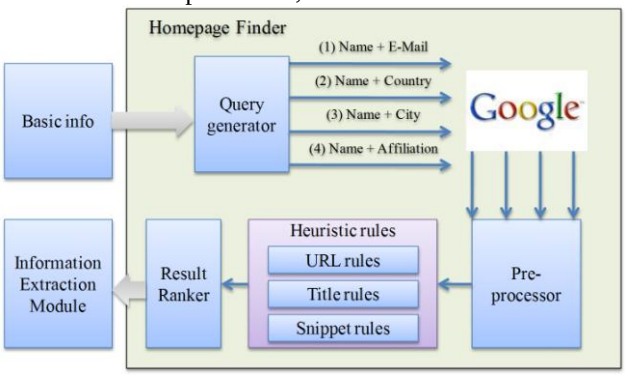

Figure 2. Homepage finding process.

- *Result filtering*. It removes some results that come from well-known websites, such as Facebook, DBLP.
- *Result scoring*. After filtering, the remained results from the four result sets will be combined into a unique set. For the web pages that occur in more than one result set, only one copy will be kept. However, as the occurrence time reflects the degree of correlation of a web page to a person, we assign *occurrence score* to the web pages, calculated by formula (1).

$$OccurrenceScore = 2 \times OccurrenceTime \quad (1)$$

2) Identification of the homepage

Next step we need to identify the homepage from the candidate set. A Google search result consists of three website metadata: *its title*, *URL*, and *snippet*. We find that homepages share many commons over these metadata. For example, the title of a homepage often involves the full name (e.g., Li Bin) or its variant of a person (e.g., bli); the URL and snippet may contain some positive characters or words (e.g., '~', website). We thus define a set of heuristic rules (with scores, e.g., *if URL contains full user name, +2*) for homepage identification. Finally, the web page with the highest score (the sum of *occurrence score* and *heuristic-rule score*) is identified as the contact's homepage.

B) Biographical Information Extraction Using CRF

Distilling structured info like '*PhdUniv*' from web pages is a typical information extraction problem. The Conditional Random Field (CRF) method is used here to extract bio-info. CRF is presented by Lafferty et al. in [1], and is often used for labeling sequential data, such as texts.

1) Information Extraction Workflow

The information extraction workflow is shown in Fig. 3, which consists of three major modules: *raw file processing module*, *model training module*, and *labeling module*.

M-1: Raw file processing module. This module accepts raw homepage files and transforms them into standard CRF input files. A standard CRF input file is composed of tokens, token features, and labels (when it is a training input file). Three units are included in this module.

- *Preprocessing unit*. We first segment the HTML page into 'logical sentences' or paragraphs delimited by structural HTML tags such as
, <p>, <hr>. After paragraph segmentation, we remove all HTML tags using rules represented by regular expressions
- *Feature computation unit*. The input file from the preprocessing unit is first tokenized into words, punctuation marks, and paragraph delimiters. After tokenization, we compute features of the tokens for subsequent model learning by CRF.
- *CRF input file construction unit*. This step parses the XML format file from the feature computation unit and transforms it into a standard CRF input file.

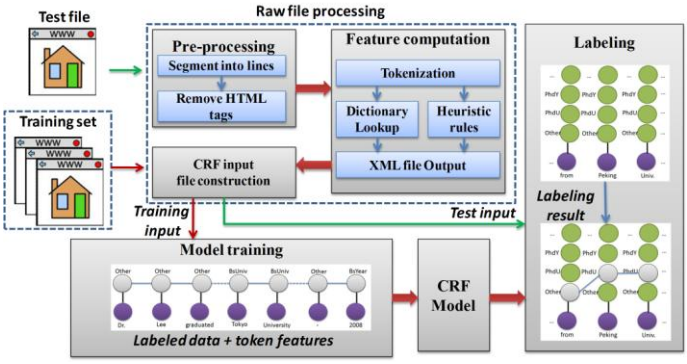

Figure 3. Biographical information extraction process.

M-2: Model training module. After getting the labeled training input file, we use the CRF++ toolkit (http://crfpp.sourceforge.net) to train the CRF model.

M-3: Labeling module. This module accepts CRF test files and labels them by using the trained CRF model.

2) Feature Selection

Four types of features are particularly calculated.

- *Token features*. The token itself, its morphology (e.g., capitalized), type (is it a word, number, or punctuation), and part of speech (POS) are used as token features.
- *Paragraph features*. It reports whether the paragraph is *long* or *short*. Paragraph feature is helpful for extracting some bio-info from homepages. For example, people often write research interests in informal *short* lines.
- *Dictionary features*. A number of dictionaries are used, including advisor, title, country, city, degree, etc.
- *Context features*. We group them into three types:

- *Local context*. It represents the dependency between the label of a token and its neighbor tokens. For example, in the sentence "*he obtained his Ph.D. degree from Tokyo Univ.…*" the word '*from'* is selected as a feature which indicates that the following words might be a university name.
- *Paragraph feature*. It represents a relative long-distance dependency within a paragraph. For example, to determine that "Tokyo Univ." should be labeled as '*PhdUniv*' while not '*BsUniv*' in the above example, we should explore the context determined by "Ph.D. degree".

- *Inter-paragraph context*. The informal nature of homepage texts often causes that one item of bio-info is written in several short paragraphs (e.g., research interests). Then, the context should be captured in the adjacent paragraphs.

C) Pervasive Sensing and User Editing

Physical- or software-sensor is another important source for gathering contact info. Here, we describe how the other three types of contact data identified in Section 3 are collected.

- *Basic info*. We use a portable mini card scanner, IRISCard (http://www.irislink.com), to extract it.
- *Context*. (1) *Time*. In addition to *year*, we record two types of "semantic" context: *season*, *morning*/*afternoon*/*evening*. The contexts are recorded when data from the business card is scanned (i.e., the contact is met for the first time) or a new event is created by the user (another meeting event with the same contact). (2) *Location context* is obtained from GPS-enabled mobile phones, which is always carried by the user. (3) *Weather*. The Yahoo Weather web service is used to fetch city weather context. (4) *Topic context* is highly subjective and should be defined by the user.
- *Impression data*. Human can sense richer information that is beyond the capacity of sensors. In our system, facial feature and personality information of a contact is derived from human observation and cognition.

V. Associative Contact Search

A keyword-based method is used for associative contact search. There are twenty contact metadata (or data fields in the contact database), and a *full-index* query method is used. In this way, when a keyword like "Tokyo" is inputted, the contacts graduated from "Tokyo Univ." and met in Tokyo can all be listed. The search interface is shown in Fig. 4.

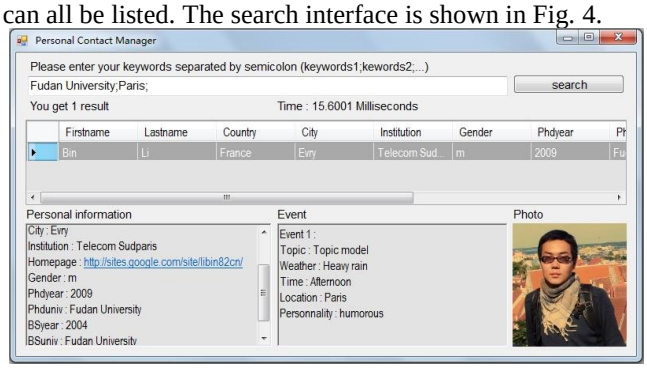

Figure 4. SCM search interface.

The search bar at the top allows users to input keywords. The search result is shown below the search bar.

VI. Evaluation

We have conducted a series of experiments to evaluate the usability and performance of SCM.

A) Homepage Finder

To test if the homepage finder can correctly find the right homepages of researchers, we collected 50 researcher names from the Arnetminer data set (http://arnetminer.org/lab-datasets/profiling/). We find that when homepage finder was not used, i.e., using person name as the keyword and recognizing the first recommended web page from Google API as the homepage, only 10 results (20%) were correct. However, when the homepage finder we proposed was used, the accuracy increased dramatically to 94%.

B) Biographical Data Extraction

To evaluate the performance of CRF-based bio-info extraction method, we made a three-fold cross validation with a total of 150 randomly selected homepages from the Arnetminer data set. We conducted evaluations in terms of precision, recall and F1-measure. The experimental results are shown in Table 1. The results show that the mean F1-score of our method attains to 80%, which indicates that using CRF to extract bio-info from homepages is feasible.

TABLE I. PERFORMANCE OF BIOGRAPHICAL DATA EXTRACTION (%)

Metadata	Full features (Atomic + Combination)		
	Precision	Recall	F1
PhdYear	78	91	84
PhdUniv	100	78	88
MsYear	91	61	73
MsUniv	79	61	69
BsYear	100	89	94
BsUniv	72	87	79
ResInterest	61	75	67
Advisor	100	55	71
Overall	85	75	80

VII. Conclusion

This paper reports our early effort on social contact management. To lessen user effort on manually recording contact information, we have explored a combination of pervasive sensing and Web intelligence techniques to auto-gather rich contact information from both real-world interactions and the Web. In terms that people often need to leverage several associated things to fetch other information of a contact (e.g., contact name), we have developed a user interface that supports search by association of social contact information. We plan to mine more semantic relationship (e.g., colleague, friend, weak/strong ties) info among the contacts in the future to augment social contact management.

References

[1] J. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML'01, pp. 282-289, 2001.

[2] D.H. Chau, B. Myers, and A. Faulring, What to do when search fails: finding information by association. In Proc. of CHI'08, 2008.

[3] D. Elsweiler, M. Baillie, and I. Ruthven. Exploring memory in email refinding. ACM Trans. Inf. Syst., Vol. 26, No. 4, pp.1-36, 2008.

[4] A. Beach et al., Whozthat? evolving an ecosystem for context-aware mobile social networks, IEEE Network, Vol. 22, No.4, 2008.

[5] D. Zhang, B. Guo, and Z. Yu, Social and community intelligence, IEEE Computer, 2011 (to appear).

[6] J. Tang, D. Zhang, L.M. Yao, Social Network Extraction of Academic Researchers, In Proc. of ICDM'07, pp. 292-301, 2007.