Vitae and Map Display System for People on the Web

Harumi Murakami[1], Chunliang Tang[1], Suang Wang[1], and Hiroshi Ueda[2]

[1] Graduate School for Creative Cities, Osaka City University
3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585 Japan
harumi@media.osaka-cu.ac.jp
http://murakami.media.osaka-cu.ac.jp/
[2] ATR Creative Inc.
2-2-2, Hikaridai, Seika, Soraku, Kyoto 619-0288 Japan

Abstract. We present a system that displays a curriculum vitae with a map to understand people. Our method is based on the following processes: (1) creating curriculum vitae using related work [1], (2) extracting the names of places where the person studied and worked from the vitae, (3) getting such location information as latitudes, longitudes, and addresses from the place names using Google Maps API, and (4) displaying a vitae along with a map using Google Maps JavaScript API. We developed a prototype and evaluated our algorithms that extract place names and convert them into location information from web search results for 56 person names.

Keywords: web people search, curriculum vitae, resume, map display, location information.

1 Introduction

Due to an increase in the number of people about whom the web can provide information, the popularity of web searches that identify people continues to rise. Such detailed and organized information might provide valuable insight into search targets.

In this paper, we develop an interface that extracts such detailed information as personal histories and organizes it into a vitae with a map to understand people. We obtain important, relatively long-term location information for the person and focus on such life stages as schools and workplaces.

Extracting schools and workplaces from web search results is difficult. We must solve the following main problems: (a) finding an "event-time-place" tuple for the person, (b) removing redundant information, and (c) obtaining location information that matches the extracted place names.

For (a) and (b), we use related work [1], which creates a curriculum vitae when a person's name is input and gives web search results. We create a vitae and extract place names (schools and workplaces) from it. For (c), we use Google Maps API to obtain location information from the extracted place names.

A. Moonis et al. (Eds.): IEA/AIE 2014, Part II, LNAI 8482, pp. 348–359, 2014.

Below, we explain our approach and give examples of our implemented prototype in Section 2 and describe our experiments in Section 3. We discuss the significance of our research in Section 4. The examples in this paper were translated from Japanese into English for publication.

2 Approach

2.1 Overview

Our approach is based on the following processes: (1) creating curriculum vitae using related work [1], (2) extracting the names of places where the person studied and worked from the vitae, (3) getting location information from the place names using Google Maps API, and (4) displaying a vitae that includes location information along with a map using Google Maps JavaScript API.

Figure 1 shows an overview of our approach.

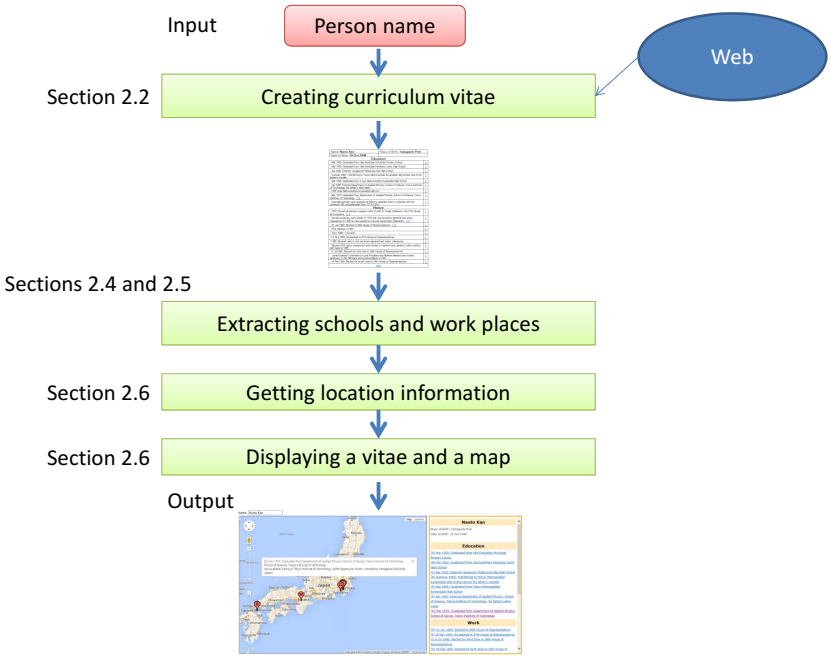

Fig. 1. Overview

2.2 Creating Vitae

The related work [1] is based on the following: (1) extracting event sentences (a character string that includes at least one time and one event) using heuristics and filtering them, (2) judging whether the event sentences are related to the

person by SVM mainly using the patterns of HTML tags, (3) classifying these sentences into four categories (background, education, history, and awards) by SVM, and (4) clustering the event sentences including both identical times and events. The history categories contain event sentences except for the background, education, and award categories. Finally, (5) such background information as date and place of birth is generated.

Extracting Event Sentences. First, sentences are extracted from web pages by cutting-sentence heuristics and event sentences are selected from those sentences using sentence-extraction heuristics that include the time (or judging year) listed below.

- The following cutting-sentence heuristics segment character strings:
  - those ending with period
  - those included in *tr*, *li*, *h1-h5*, *title*, *p*, or *div* tags without periods
  - those ending with *br* tags without periods
- The following sentence-extraction heuristics extract sentences that include time (judging year):
  - those including four sequential numbers (e.g., 2000)
  - those including an apostrophe with two sequential numbers (e.g. '00)
  - those including names of Japanese calendars with two numbers with year

Next, 21 stop words (e.g., copyright, post) and 20 stop patterns (e.g., expressing such detailed times as 2008 12-31 12:00:00) were used to filter unnecessary event sentences.

Judging Relation to a Designated Person. Event sentences are judged on the degree to which they are related to a designated person using SVM.

Five person names including politicians and researchers were used as queries and 200 search results (web pages) for each person name (i.e., 5 * 200 = 1,000) were obtained from Google Web APIs. 7,211 event sentences were obtained through the event-sentence-extraction process. They were judged manually and 2,266 correct and 4,474 incorrect training data were obtained.

We learned 31 patterns for SVM as the training data. Example patterns include: "The event sentence includes both the family and given names," "the nearest *h1* to *h5* tags include the family and given names," "the number of nouns inside the *title* tag," and "the first *tr* tag in the *table* tag includes the family name."

Classifying Event Sentences into Four Categories. Using the one-versus-rest method of SVM, event sentences, which are judged to be related to a designated person, are classified into four categories: background, education, history, and awards.

We obtained 500 web pages that were expected to include profile information for the training data: 200 for the query "personal history site:ja.wikipedia.org", 200 for the query "personal history inurl:profile", and 100 for the query "alma

mater site:read.jst.go.jp". After the event-sentence-extraction process, 14,974 event sentences were obtained and classified into the four categories. We morphologically analyzed the event sentences and calculated the tf-idf values. The idf values were the number of event sentences. The tf-idf values and the number of terms were used to form SVM patterns. The non-linear SVM was used for each category.

Clustering Event Sentences Including Both Identical Times and Events. Next we group event sentences with the same meanings. For example, we must combine the following sentences: "A entered X university in 1982" and "In April 1982, A enrolled at Univ X." We must also distinguish the combined sentences from such examples as, "A graduated from X university in 1986." Since these sentences share many words, standard clustering algorithms may not be able to distinguish them.

We focus on time and cluster event sentences. The tf-idf, cosine, and single-path methods were used for the clustering. The following is the algorithm.

Step 1 Event sentences are obtained that include "year, month, and day." Each year-month-day cluster is generated, and event sentences are clustered in year-month-day clusters.

Step 2 Event sentences, which include "year and month" and years/months that are identical with the existing clusters, are clustered by the single-path method.

Step 3 Event sentences, which include "year" and years identical with the existing clusters, are clustered by the single-path method.

Step 4 Each event sentence, which includes a "year and month" and does not belong to any existing clusters, becomes a year-month cluster. The clusters with identical years/months are clustered by the single-path method. Event sentences that include "year" and years identical with the generated clusters are clustered by the single-path method.

Step 5 Each event sentence, which includes a "year" and does not belong to any existing clusters, becomes a year cluster. The clusters with identical years are clustered by the single-path method.

Generating Background Information. We extracted the dates of birth and death from the background category by simple heuristics.

- date of birth: select a cluster that includes the most frequent "date of birth," "birthday," or "being born" and extract a date as "date of birth" from the cluster.
- date of death: select a cluster that includes the most frequent "date of death," or "being dead" and extract a date as "date of death" from the cluster.
- place of birth: extract the most frequent prefecture name in the cluster from which the date of birth was extracted.

Example of Created Vitae. Figure 2 shows the process for creating a vitae for *Naoto Kan*, a former Japanese prime minster. On the left is shown a created vitae from the implemented related work [1]. The background, education, and history categories are displayed. The education and history categories include event sentence clusters.

When a user selects [+], the clustered event sentences are displayed. When a user selects "L", the original web page containing the event sentence is displayed.

Since the amount of event sentence clusters in a history category is large, more than 11 clusters are hidden. When a user selects "more," the hidden clusters will be displayed.

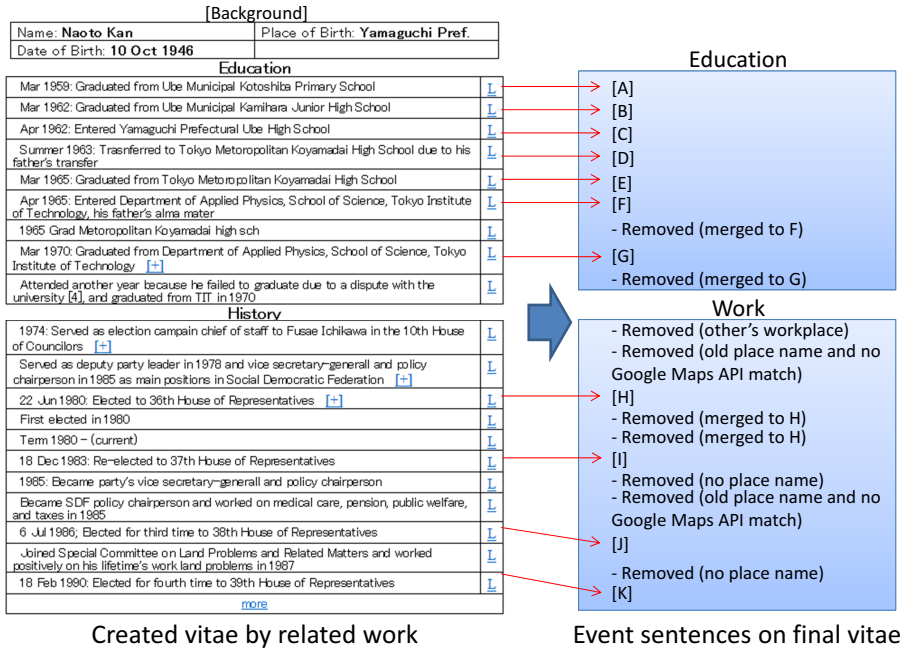

Fig. 2. Creating vitae for Naoto Kan

2.3 Issues of Using Vitae Created by Related Work

There are three main issues for extracting place names from the created vitae: extracting (1) schools from the education category, (2) schools from the history category, and (3) workplaces from the history category.

(1) Extracting schools from the education category is the easiest among the three since the schools themselves are relatively easy to extract because character strings for them are typical and such events are also typical (e.g., enter, graduate, transfer). The main problem here is filtering the same event sentences as those that weren't filtered by the previous step (creating vitae). Identifying locations from ambiguous place names (e.g., universities with multiple campuses) is another important problem; however, this paper ignores it.

(2) Extracting schools from the history category (as workplaces) is slightly more difficult, since such events include various positions (e.g., RA, researchers, professors), degrees, and institutes.

(3) Extracting workplaces from the history category is the most difficult step in our research. This is why the related work used the "history" category instead of the "work" category. Working history is ambiguous, and workplaces are ambiguous as well. For example, when a politician joins a party, is this treated as working history? If he is elected, is this working history? If so, what are appropriate working places, party offices, private offices, or parliament, etc.?

Extracting schools and workplaces from event sentences are described in Sections 2.4 and 2.5.

2.4 Extracting Schools

The process for extracting schools is listed below.

Step 1 Event sentences are extracted that contain the following: "(a piece of) school," or "(a piece of) graduation."

Step 2 Morphological analysis was performed on the event sentences.

Step 3 School candidates were extracted from the event sentences.

- **3-a** When nouns are judged to be organizations or locations in the event sentences:
  The following nouns are connected: those starting from being judged to be organizations or locations and those ending at these noun phrases: (a) "graduate school" when it contains "graduate school;" (b) "university" or "college" when it contains "university" or "college" but not "school;" (c) otherwise, "school."
  e.g.) "Ube Municipal Kotohira Primary School"
- **3-b** When no noun exists that is judged to be an organization or a location in the event sentences:
  The following nouns are connected (with backtracking) starting at these noun phrases: "graduate school," or "school."
  e.g.) "Azabu High School"

Step 4 Identical school candidates were filtered.

- **4-a** The school candidates are grouped into three categories if they contain the following terms: (1) "enter," (2) "transfer," or (3) "graduation," or "graduate."
- **4-b** The character strings of school candidates in each category are compared, and when all the characters of a school candidate are included in other school candidates, the shorter candidate is removed.
  e.g.) "Tokyo Metoropolitan Koyamadai High School" and "Metoropolitan Koyamadai High School" are compared, and the latter is removed.

2.5 Extracting Workplaces

The process for extracting workplaces is listed below.

Step 1 Event sentences are extracted that contain the following keywords: "elected," "position," "resignation" (for politicians), "university," "school," "research" (for researchers), "join," "traded," "retirement" (for athletes), "role," "being in charge," (for entertainers), "employed," "joining a company," "leaving a company" (for company workers).

Step 2 Morphological analysis was performed on the event sentences.

Step 3 Event sentences are removed that do not belong to the designated person. When a noun is judged to be a "family name" that is different from the designated person with Japanese particles that link a subject with nouns, the event sentence is removed.

Step 4 When one or more nouns are judged to be organizations in the event sentences:

4-a When there is only one noun, it is extracted.
e.g.) "House of Representatives"

4-b With more than one noun, the nearest one to the keyword is extracted.

Step 5 When no noun is judged to be an organization in the event sentences:

5-a When "company limited" or "Co. Ltd."are found, the nouns before and after the character strings are connected.
e.g.) "Mikasakosan Co. Ltd." and "Kurehaboseki Company Limited"

5-b When "minister" is found, a minister's name is extracted from a minister dictionary.
e.g.) "Transportation Minister"

5-c When "mayor" or "governor" and nouns judged to be locations are found, the nearest noun and "city" (for mayor) or "prefecture" (for governor) are connected.
e.g.) "Akune City"

Step 6 When "university," "school," or "research" (for researchers) are found in the event sentences:

6-a The same process as in Section 2.4 was performed. For 3-b, the noun phrases of the starting point of the backtracking were "graduate school" and "research center."
e.g.) "Graduate School of Information Science and Technology, Nara Institute of Science and Technology" and "National Institute of Informatics"

2.6 Getting Location Information and Displaying Vitae and Maps

Each place name obtained in Sections 2.4 and 2.5 was converted to location information (latitudes, longitudes and addresses) using the Google Maps API v3.

Next, the vitae created in Section 2.2 was modified. We extracted the event sentence clusters that include the above place names and location information

(Fig. 2 (right)). Seven out of the nine event sentence clusters in the education category are extracted as A to G. The seventh event sentence clusters are filtered since they are combined with the previous one. For the history category, many event sentence clusters were removed. For example, the first-event sentence cluster was removed since there was no workplace for the designated person. The second-event sentence cluster was filtered since there is no workplace, or exactly speaking, there was an old work place name without a hit in the Google Maps API.

Although related work [1] failed to create work categories, this research successfully identified important work history by extracting workplaces.

Finally, a vitae that includes location information along with a map is displayed using Google Maps JavaScript API v3.

2.7 Prototype

Based on the person name input, we developed a prototype that displays (1) a vitae with background information, education, and work information that includes location information in the event sentences, and (2) a map on which icons express places included in the event sentences.

When the user selects an event sentence in the vitae or an icon on the map, the event sentence, the place name, and the address hit in the Google Maps API are displayed.

Figure 3 shows an example for *Naoto Kan*. Almost all (7/9, 78%) the event sentence clusters in the vitae's education category are displayed in the education category, and the selected (19/41, 46%) event sentences in the vitae's history category are displayed in the work category.

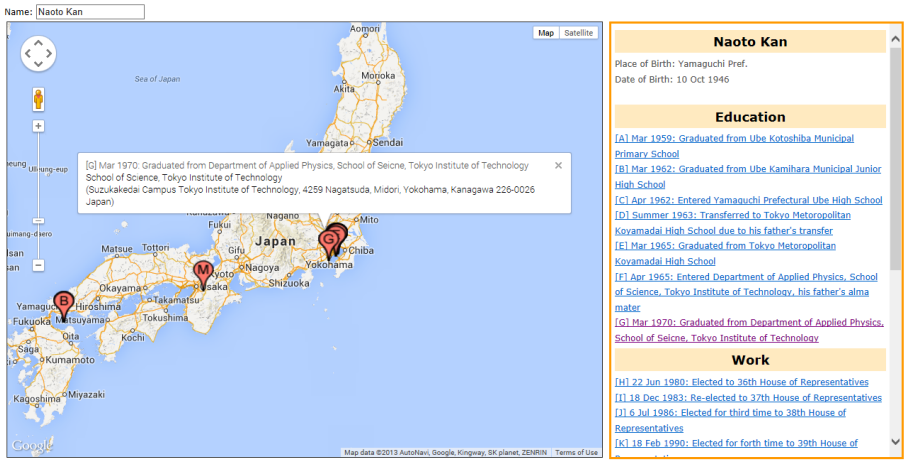

Fig. 3. Prototype

3 Experiment

3.1 Dataset

We used the names (mostly famous) of 56 Japanese people: 17 politicians, 14 athletes, 12 entertainers, 10 researchers, 1 entrepreneur, 1 manga artist, and 1 historical figure. We obtained 50 web search results for each person name and removed unrelated web pages. Our implemented [1] method generated 56 vitae.

We identified the correct data (place names) included in the vitae.

3.2 Experiment 1: Extracting Place Names

The following are the evaluation measures for extracting place names (schools and workplaces):

$$Precision = \frac{\text{correct extracted place names}}{\text{extracted place names}} \tag{1}$$

$$Recall = \frac{\text{correct extracted place names}}{\text{correct place names to be extracted}} \tag{2}$$

Note that correct place names are duplicated. For example, when person A entered and graduated from X university, it counts as two in the education category. If person B also entered X university, it counts as three in the education category. If person A worked at X university, it counts as three in the education category and one in the work category.

As stated before, since workplaces are ambiguous, we defined some guidelines to judge them in this research.

In addition, we classified their levels of correctness. 1 is exact: extracted place name is identical as the correct answer. 2 is partially correct: e.g., part of the correct answer, abbreviation for the correct answer. An example of level 2 includes "Marinos (soccer team name)" for "Yokohama F Marinos (complete soccer team name)."

Table 1 shows the result. The results of extracting schools are fine (Precision: 91-97% and Recall: 84-90%), however, those of the workplaces are insufficient (Precision: 65-73% and Recall: 70-90%). Since our guidelines for workplaces are not strict enough, these figures show maximum values.

3.3 Experiment 2: Getting Location Information

We used the correct answers to get location information by the Google Maps API v3.

We judged the top ranked results. For schools, 0.90 (53/59), and for workplaces, 0.53 (81/154) were correct.

The most common error for workplaces was the name of political parties, e.g., LDP. Party offices were sorted from north to south in the Google Maps API

Table 1. Experiment 1 results

	Precision		Recall	
	p1	p2	r1	r2
Schools	0.91 (59/65)	0.97 (63/65)	0.84 (59/70)	0.90 (63/70)
Work places	0.65 (154/241)	0.73 (166/241)	0.70 (154/233)	0.90 (166/233)

and the northernmost offices were top ranked. However, most correct offices are located in Tokyo, in central Japan.

If school names were correctly extracted, Google Maps API returned good results. For workplace names, the result was poor. We need to try different geocoding services or consider better algorithms and create dictionaries using Google Maps API.

4 Related Work and Discussion

WePS-3 conducted a competitive evaluation on person attribute extraction on web pages [2]. Schools were included in the extraction candidates, but not workplaces since they are ambiguous.

Murakami et al. [3] only extracted "one representative" piece of location information, which is typically the most current places from the web search results. Our research extracts schools and workplaces about the designated person.

Other research assigns information to learn about people. Wan et al. assigned titles (similar to vocations) [4]. Ueda et al. assigned vocation-related information including vocations, organizations, and works [5], Mori et al. assigned keywords [6], and Murakami et al. assigned library classification numbers [7] to person clusters. This research is part of a project to develop interfaces to select and understand people on the web [8].

Some research extracted event sentences from web pages. Kimura et al. [9] extracted event sentences to create personal histories and listed them. Our research creates vitae and displays schools and workplaces on a map.

Although related work [1] failed to create work categories, this research successfully identified important work history by extracting workplaces. Therefore, we not only used the related work but also improved it.

Some geocoding services (e.g. [10]) convert place names to location information. In our survey, Google Maps API is currently the best service among them.

We believe that our work's main contribution is that it displays study and work places on a map with curriculum vitae for people on the web. To the best of our knowledge, this is the first research to do so.

Although our research is limited to Japanese, the idea is easily applicable to other languages.

Future work must improve the extraction of workplaces. We also need to consider algorithms to find exact workplaces for geocoding services. The algorithms should be evaluated using different datasets (e.g., ordinary people).

5 Conclusions

We presented a system that displays curriculum vitae on a map to understand people. Our method is based on the following processes: (1) creating curriculum vitae using related work [1], (2) extracting the names of places where the person studied and worked from the vitae, (3) getting location information (latitudes, longitudes and addresses) from the place names using Google Maps API, and (4) displaying a vitae along with a map using Google Maps JavaScript API. We evaluated our algorithms that extract place names and converted them into location information from 56 person name search results.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number 22500219, 25330385.

References

1. Ueda, H., Murakami, H., Tatsumi, S.: Creating Curriculum Vitae for Understanding People on the Web. Transactions of the Japanese Society for Artificial Intelligence 25(1), 144–1565 (2010)
2. Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., Amigo, E.: WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In: CLEF 2010 (2010)
3. Murakami, H., Takamori, Y., Ueda, H., Tatsumi, S.: Assigning Location Information to Display Individuals on a Map for Web People Search Results. In: Lee, G.G., Song, D., Lin, C.-Y., Aizawa, A., Kuriyama, K., Yoshioka, M., Sakai, T. (eds.) AIRS 2009. LNCS, vol. 5839, pp. 26–37. Springer, Heidelberg (2009)
4. Wan, X., Gao, J., Li, M., Ding, B.: Person Resolution in Person Search Results: WebHawk. In: Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005), pp. 163–170. ACM Press, New York (2005)
5. Ueda, H., Murakami, H., Tatsumi, S.: Assigning Vocation-Related Information to Person Clusters for Web People Search Results. In: Proceedings of the 2009 Global Congress on Intelligent Systems (GCIS 2009), vol. 4, pp. 248–253. IEEE Press, New York (2009)
6. Mori, J., Matsuo, Y., Ishizuka, M.: Personal Keyword Extraction from the Web. Journal of Japanese Society for Artificial Intelligence 20, 337–345 (2005)
7. Murakami, H., Ura, Y., Kataoka, Y.: Assigning Library Classification Numbers to People on the Web. In: Banchs, R.E., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., Lang, J. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 464–475. Springer, Heidelberg (2013)
8. Murakami, H., Ueda, H., Kataoka, S., Takamori, Y., Tatsumi, S.: Summarizing and Visualizing Web People Search Results. In: Proceedings of the Second International Conference on Agents and Artificial Intelligence (ICAART 2010), vol. 1, pp. 640–643. INSTICC Press (2010)

9. Kimura, R., Oyama, S., Toda, H., Tanaka, K.: Creating Persnal Histories from the Web using Namesake Disambiguation and Event Extraction. In: Baresi, L., Fraternali, P., Houben, G.-J. (eds.) ICWE 2007. LNCS, vol. 4607, pp. 400–414. Springer, Heidelberg (2007)
10. Geocoding Tools & Utilities, http://newspat.csis.u-tokyo.ac.jp/geocode/