SUMMARIZING AND VISUALIZING WEB PEOPLE SEARCH RESULTS

Harumi Murakami[1], Hiroshi Ueda[2], Shin'ichi Kataoka[1], Yuya Takamori[1] and Shoji Tatsumi[2]
[1]*Graduate School for Creative Cities, Osaka City University, Japan*
[2]*Graduate School of Engineering, Osaka City University, Japan*

Keywords: Web people search, Summarization, Visualization, Interface.

Abstract: People search is one major search activity on the Web. If the list of people search results is merely "person 1, person 2, . . . and so on," users have difficulty determining which person clusters they should select. In this paper, we present a project that summarizes and visualizes Web people search results to help users select person clusters more easily. We explore three ways of summarizing people: (a) selecting terms from the extracted information, (b) combining the extracted information, and (c) obtaining information from external databases referring to the extracted information. To visualize people, we present three types of interfaces: (a) tables, (b) two-dimensional space, and (c) map interfaces. We report the two results of the project. (1) We investigated algorithms for distinguishing individuals with identical names and three ways of summarizing people: extracting keywords, prefectures and vocations; combining vocation-related information; and obtaining locations. (2) We developed prototypes to display separated individuals by three types of interfaces.

1 INTRODUCTION

People search is one major search activity on the Web. According to (Guha and Garg, 2004), 30% of queries in Web searches include person names. Person name disambiguation, or distinguishing people with identical names, is becoming more and more important in Web searches. If the list of search results is merely "person 1, person 2, …, and so on," users have difficulty determining which person clusters they should select.

This research helps users select person clusters that are separated into different people from the results of person searches on the Web.

Below, in Section 2 we explain the overview of our project and describe the developed algorithms and prototypes in Sections 3-6. Related work is discussed in Section 7.

2 OVERVIEW OF THE PROJECT

The aim of this project is to develop interfaces to select and understand people on the Web.

First, we create person clusters by distinguishing individuals from Web people search results. Second, we present interfaces to help users select and understand people.

The main feature of this research is summarizing people. We extract various kinds of attribute information related to people. Instead of displaying all extracted information like traditional information extraction research, we present one representative piece of information by referring to the extracted information.

For example for the first author, from the following list, which is the one representative piece of information?: *faculty staff*, *researcher*, *professor*, *Osaka City University professor*, or *computer scientist*? We explore three ways to summarize people: (a) selecting terms from the extracted information, (b) combining the extracted information, and (c) obtaining information from external databases that refer to the extracted information.

We present three types of interfaces to visualize people: (a) tables, (b) two-dimensional space in which similar individuals are displayed close together, and (c) map interfaces. See Figure 1.

Murakami H., Ueda H., Kataoka S., Takamori Y. and Tatsumi S. (2010).
SUMMARIZING AND VISUALIZING WEB PEOPLE SEARCH RESULTS.
In *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence - Artificial Intelligence*, pages 640-643
DOI: 10.5220/0002760706400643

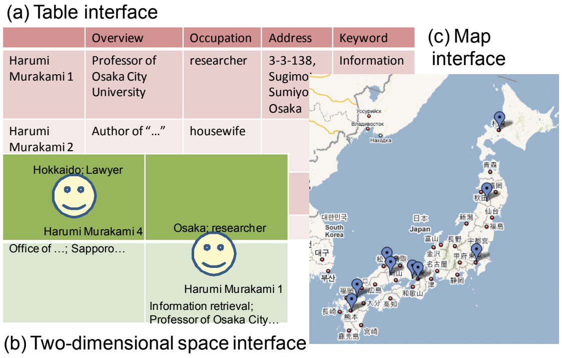


Figure 1: Project overview.

3 INITIAL PROTOTYPES

We developed initial prototypes of a table and a two-dimensional-space interface (Ueda et al. 2007) that separate individuals from Google people search results and automatically extract attribute information (keywords, prefectures, and vocations).

The table interface displays a list of people with attribute information (Figure 2), and the two-dimensional-space interface displays person icons with part of the extracted attribute information (Figure 3).

3.1 Algorithms

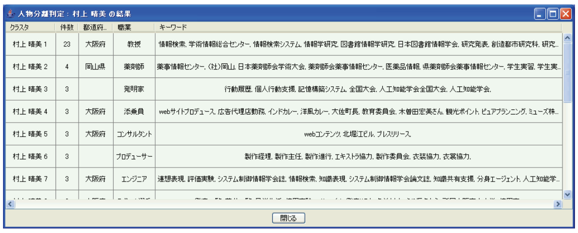
The algorithms that extract keywords, prefectures, and vocations are described here. They are based on the type of summarization: (a) selecting terms from the extracted information.

3.1.1 Extracting Keywords

We extracted nouns near person names using morphological analysis, collocated, and ranked them with TermExtract. For example, for the first author, *information retrieval* was the top ranked keyword.

3.1.2 Extracting Prefectures

We extracted prefecture names by pattern-matching using prefecture dictionaries from texts near the person names. One of the most frequent prefectures is displayed. For example, for the first author, *Osaka Prefecture* was the answer.



人物分離判定：村上 晴美 の結果

クラスタ	件数	都道府…	職業	キーワード
村上 晴美 1	23	大阪府	教授	情報検索, 学術情報総合センター, 情報検索システム, 情報学研究, 図書館情報学研究, 日本図書館情報学会, 研究発表, 創造都市研究科, 研究…
村上 晴美 2	4	岡山県	薬剤師	薬事情報センター, (社)岡山, 日本薬剤師会学術大会, 薬剤師会薬事情報センター, 医薬品情報, 県薬剤師会薬事情報センター, 学生実習, 学生実…
村上 晴美 3	3		発明家	行動履歴, 個人行動支援, 記憶構築システム, 全国大会, 人工知能学会全国大会, 人工知能学会,
村上 晴美 4	3	大阪府	添乗員	webサイトプロデュース, 広告代理店勤務, インドカレー, 洋風カレー, 大佐町長, 教育委員会, 木曽田宏美さん, 観光ポイント, ピュアプランニング, ミューズ株…
村上 晴美 5	3	大阪府	コンサルタント	webコンテンツ, 北堀江ビル, プレスリリース,
村上 晴美 6	3		プロデューサー	製作経理, 製作主任, 製作進行, エキストラ協力, 製作委員会, 衣装協力, 衣裳協力,
村上 晴美 7	3	大阪府	エンジニア	連想表現, 評価実験, システム制御情報学会誌, 情報検索, 知識表現, システム制御情報学会論文誌, 知識共有支援, 分身エージェント, 人工知能学…

Figure 2: Table interface.

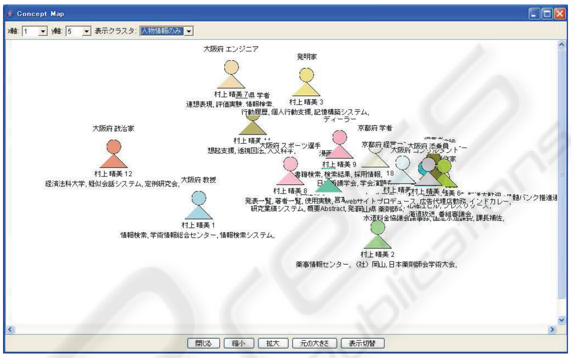

Concept Map
x軸: 1
y軸: 5
表示クラスタ: 人物情報のみ
大阪府 エンジニア
発明家
村上 晴美 7
連想表現, 評価実験, 情報検索,
村上 晴美 3
行動履歴, 個人行動支援, 記憶構築システム,
ディーラー
大阪府 政治家
京都府 学者
村上 晴美 11
想起支援, 透視図法, 人文科学,
大阪府 スポーツ選手
京都府 経営
大阪府 添乗員
大阪府 コンサルタント
村上 晴美 9
書籍検索, 検索結果, 採用情報, 18
村上 晴美 12
経済法科大学, 疑似会話システム, 定例研究会,
大阪府 教授
村上 晴美 8
発表一覧, 著者一覧, 使用実験,
研究業績システム, 概要Abstract,
webサイトプロデュース, 広告代理店勤務, インドカレー,
岡山県 薬剤師
海道放送, 番組審議会,
水道料金協議会事務, 課長補佐,
村上 晴美 1
情報検索, 学術情報総合センター, 情報検索システム,
村上 晴美 2
薬事情報センター, (社) 岡山, 日本薬剤師会学術大会,
閉じる
縮小
拡大
元の大きさ
表示切替

Figure 3: Two-dimensional space interface.

3.1.3 Extracting Vocations

We extracted vocations by pattern-matching using dictionaries from texts near the person names. The vocation dictionaries were created from Wikipedia. One of the most frequent vocations is displayed. For example, for the first author, *professor* was the answer.

Preliminary experiments suggested that vocations and keywords were useful to select people.

4 DISTINGUISHING DIFFERENT PEOPLE

We investigated two approaches to separate different people: non-hierarchical clustering and hierarchical clustering.

First in the initial prototype, to simulate how people separate Web pages, non-hierarchical clustering (single path method) was examined. We checked Web pages from the top of the search results and added similar pages to the existing clusters. Since this approach produced too many clusters (Ueda et al. 2007), we need to group them in the future.

As an alternative, we examined a two-step clustering method using person names to distinguish individuals (Kataoka et al. 2008). The key idea was to use person names to create initial clusters by combining mechanisms. Our method was comprised of two processes: (1) first step clustering based on the co-occurrence of person names and (2) second step clustering based on hierarchical clustering using keywords. Overall, our method outperformed two baselines and was ranked 2nd in the WePS (Artiles et al. 2007) systems.

In seeking state-of-the art person name disambiguation, since we have found it difficult to automatically distinguish different people, we decided to abandon person name disambiguation and concentrate on summarizing and visualizing manually separated person clusters. In the following sections, we investigate two important bits of attribute information: vocations (Section 5) and locations (Section 6).

5 ASSIGNING VOCATION-RELATED INFORMATION

Although vocations are the most informative ways to select people, defining vocations and obtaining good vocation dictionaries was difficult (Ueda et al. 2007). We therefore proposed a method to extract vocation-related information (VRI) from Web pages in person clusters. VRI is information related to vocation, which is more useful for identifying people (Ueda et al. 2009). We selected three types of VRI: (1) vocation, (2) organization and position, and (3) publication title and role.

The method is comprised of two processes: (1) extraction of VRI candidates using HTML structure and heuristics, and (2) VRI generation using term frequency, clustering synonyms, and calculation using a Web search engine. The method is based on the type of summarization: (b) combining the extracted information.

For example, two pieces of extracted information, *Tozai newspaper* and *Cultural Affairs Department journalist*, are combined, and *Cultural Affairs Department journalist, Tozai newspaper* is generated as a VRI.

The main advantage of our method is to label VRIs to person clusters based on the context, not vocations based on term frequencies.

VRI can be used as one bit of attribute information (e.g., vocations, titles, occupations, or companies) and as one good source of an overview of people.

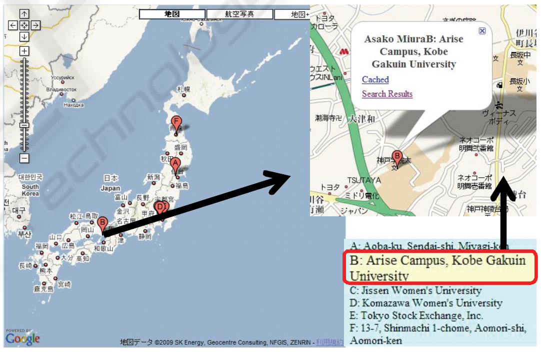
6 ASSIGNING LOCATION INFORMATION ON A MAP

This research (Murakami et al. 2009) aims to display person icons on a map to help users select person clusters that are separated into different people from the results of person searches on the Web. We proposed a method to assign person clusters with one piece of location information.

Our method is comprised of two processes: (1) extracting location candidates from Web pages and (2) assigning location information using a local search engine. The method is based on the type of summarization: (c) obtaining information from external database referring to the extracted information.

For example, the location information of *Arise Campus, Kobe Gakuin University*, is obtained from Yahoo! Local search using one of the extracted terms: *Kobe Gakuin University*.

Our main idea exploits search engine rankings and character distance to obtain good location information among location candidates.



Asako MiuraB: Arise Campus, Kobe Gakuin University
Cached
Search Results

A: Aoba-ku, Sendai-shi, Miyagi-ken
B: Arise Campus, Kobe Gakuin University
C: Jissen Women's University
D: Komazawa Women's University
E: Tokyo Stock Exchange, Inc.
F: 13-7, Shinmachi 1-chome, Aomori-shi, Aomori-ken

地図
航空写真
地図+
Japan
TSUTAYA
神戸学院大
Google
地図データ ©2009 SK Energy, Geocentre Consulting, NFGIS, ZENRIN - 利用規約

Figure 4: Map interface prototype system.

Figure 4 shows the prototype interface. When a person name is input as a query, Web pages are classified into person clusters, and individual icons that express appropriate locations are displayed on a map. Users can select icons to display their location information to access searched Web pages.

Experimental results revealed the usefulness of our proposed method.

7 RELATED WORK

Most people search services (e.g., MyLife) provide manually constructed database input by users or operators. When a user types a person name as a query, a list or table is displayed to select a person.

Person name disambiguation is manually solved. Ages, locations (usually states in US), titles, and companies are often displayed as attribute information.

Some services and research automatically execute Web people search and display a list or a table. Zoominfo extracts and displays such attribute information as titles and companies. Wan et al. (2005) also separated Web people search results and assigned titles to person clusters to provide a list interface to select people.

We present three kinds of interfaces and explore three types of summarization methods.

Even though much work (Artiles et al., 2007) separates Web pages into person clusters, it seldom assigns labels to person clusters. The WWW2009 WePS-2 workshop evaluated a technique to extract attribute information. Although it aims to extract all attribute information related to people, we assign representative attribute information to person clusters and call this summarization.

Concerning the visualization of people search, Matsuo et al. (2006) visualizes a human network to select people. Mori et al. (2008) presents an interface design for people search that includes a human network and two-dimensional space. Our work is different; we concentrate on summarizing people.

We showed a scenario for people search by names; however, our proposal can be applied to other types of people search such as by keywords. For example, *Bill Clinton*, *George W. Bush*, and *Barack Obama* are summarized and visualized by an input query: president of US.

8 CONCLUSIONS

We presented a project that summarizes and visualizes Web people search results to help users select person clusters more easily. We explored three ways to summarize people: (a) selecting terms from the extracted information, (b) combining the extracted information, and (c) obtaining information from external databases referring to the extracted information. We present three types of interfaces to visualize people: (a) tables, (b) two-dimensional space, and (c) map interfaces. We reported the project results.

REFERENCES

Guha, V., and Garg, A., 2004. Disambiguating People in Search. Stanford University.

Ueda, H., Murakami, H., and Tatsumi, S., 2007. A system that distinguishes different people with identical names on the Web and displays a list by attribute information. In *The 21st Annual Conference of the Japanese Society for Artificial Intelligence*. (CD-ROM).

Kataoka, S., Ueda, H., Murakami, H., and Tatsumi, S., 2008. Two-step Clustering based on Person Names to Identify Different People with Identical Names on the Web. In *The 22nd Annual Conference of the Japanese Society for Artificial Intelligence*. (CD-ROM).

Artiles, J., Gonzalo, J., and Sekine, S., 2007. The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, 64-69.

Ueda, H., Murakami, H., and Tatsumi, S., 2009. A ssigning Vocation-Related Information to Person Clusters for Web People Search Results. In *Proceedings of the 2009 Global Congress on Intelligent Systems (GCIS 2009),* 4, 248-253.

Murakami, H., Takamori, Y., Ueda, H., and Tatsumi, S., 2009. Assigning Location Information to Display Individuals on a Map for Web People Search Results. In *Proceedings of The Fifth Asia Information Retrieval Symposium (AIRS 2009),* 26-37.

Wan, X., Gao, J., Li, M., Ding, B., 2005. Person resolution in person search results: WebHawk. In *CIKM2005, Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management*, 163-170.

Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hashida, K., Ishiduka, M., 2006. Polyphonet: An advanced social network extraction system, In *WWW2006*, 397-406.

Mori, J., Basselin, A. Kroner, and A. Jameson, 2008. Find me if you can: Disigning Interfaces for People Search, In *Proceedings of the 13th international conference on Intelligent user interfaces*, 377-380.