GRAPE: A System for Disambiguating and Tagging People Names in Web Search

Lili Jiang[†], Wei Shen[§], Jianyong Wang[§], Ning An[†]
[†]School of Information Science and Engineering, Lanzhou University, Lanzhou, China
jianglili06@lzu.cn; nan@lzu.edu.cn
[§]Department of Computer Science and Technology, Tsinghua University, Beijing, China
chen-wei09@mails.tsinghua.edu.cn; jianyong@tsinghua.edu.cn

ABSTRACT

Name ambiguity is a big challenge in people information retrieval and has received considerable attention, especially with the increasing volume of Web data in recent years. In this demo, we present a system, GRAPE, which is capable of finding people related information over the Web. The salient features of our system are people name disambiguation and people tag presentation, which effectively distinguish different people entities sharing the same name and uniquely represent each namesake with a cluster of tags, such as occupation, birthdate, and organization.

Categories and Subject Descriptors

H.3.1 [**Information Storing and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storing and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Performance, Experimentation

Keywords

Name Disambiguation, Clustering, Tag Presentation

1. INTRODUCTION

Finding people information through Web search engine is one of the most common activities. However, issued by a people name as query, the traditional search engines return a long list of pages, which often denote more than one persons in the real world. For example, given a query "John Smith", the top 100 returned Web pages from Google may refer to at least 10 namesakes, which challenge the users to quickly locate who they are looking for. Therefore, users are supposed to use a more specialized search engine, which effectively handles the issues of people name ambiguity and people information disperstiveness. Some approaches have been proposed to address these issues [5][6]. Motivated by the observation that a combination of people tag information, such as birthdate, organization, and email address appears to be informative and can almost identify a unique target people, we implement a Web people search system, GRAPE, which aims to disambiguate people name through a graph-based clustering framework, and provides tag information to uniquely represent each people entity (i.e., namesake). The underlying algorithm and experimental evaluations are described in detail in our previous work [3].

There have been some systems on people search, such as the commercial search engines, Spock (www.spock.com) and Wink (www.wink.com). Spock collects people information from Web and the social networks (e.g., LinkedIn and Facebook), while Wink allows users to find others who have similar interests in different social networks. However, both of them simply list all the people owning the queried people name instead of taking measures to distinguish namesakes. WEST is a new people search system [4]. Queried by a people name and some advanced people features (i.e. location and organization) or other keywords, WEST disambiguates the namesakes for the given people name, and returns the support URLs for each identified namesake. Additionally, some meta search engines, such as Vivisom's clusty (clusty.com) and Carrot2 (www.carrot2.org) could group the documents relevant with a people name from multiple search engines into different clusters, each of which focuses on a similar topic instead of a people entity.

The rest of this paper is organized as follows. Section 2 presents the overall architecture of GRAPE, and introduces the employed techniques including Webpage preprocessing, information extraction, clustering algorithm, and tag selection. Section 3 describes our demonstration scenarios.

2. AN OVERVIEW OF THE SYSTEM ARCHITECTURE

GRAPE is a Web people search system. As illustrated in Figure 1 and Figure 2, it enables the user to input a people name as query and disambiguates the queried people name based on the collected Web pages from Google. Finally it provides the essential tag information as well as the support documents about each people entity to users. In this section, we will give an overview of GRAPE including its architecture, working process, and employed techniques. As depicted in Figure 3, given a user query through user interface, the system is organized on three levels generalized in three bounding boxes respectively, Web data collection, people name disambiguation, and people tag presentation.

2.1 Web Data Collection

Instead of directly sending the user request to general search engines for search results. GRAPE formulates the

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

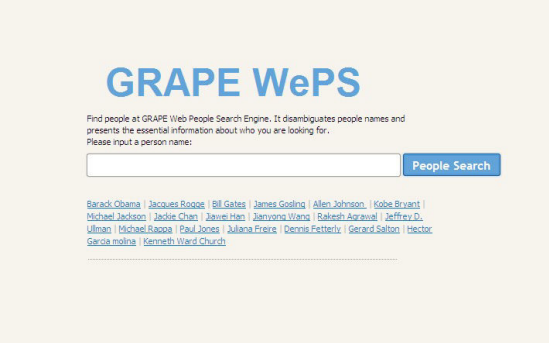

Figure 1: Input Interface of GRAPE

Figure 2: Result Interface of GRAPE

user query firstly, and then use Google Web search[1] as the source for data collection.

- **Query Formulator:** We have a function to validate the given query. It is observed that a people name containing more than three words hardly cause the problem of name ambiguity. For this reason, GRAPE simply supports bi-gram and tri-gram queries. Particularly, the system employs a fault-tolerance method, which firstly removes the repetitive whitespace within the query and the non-letter characters around it. And then only the queries containing letters and character "." are admitted (e.g., John Smith, John Kennedy Smith, John K. Smith, and J. Smith). Finally, the valid query is rewritten through capitalizing the leading letter of each word involved, while the invalid query causes a warning message without any results.

- **Data Collector**: GRAPE uses a data collector for fetching the top 100 search results from Google, which exclude the results with some file types, such as pdf, doc, and image for two reasons. a) Improve the effectiveness of name disambiguation by getting more available information about the queried people name in the top hundred results; b) Improve the system performance due to the time-consuming connection with these non-script pages.

2.2 People Name Disambiguation

We introduce in this section the components to address the primary problem of people name ambiguity.

- **Page Preprocessor**: Webpage preprocessing is an essential step for further development. As shown in Figure 3, the 100 fetched Web pages are processed by the component of Page Preprocessor through cleaning out all the unrelated pages that are non-English documents or do not contain the queried people name. Then each of the remained Web pages is parsed into a plain text. Most importantly, we employ a notion of chunk-window to tail a window of 2500 characters around the queried people name, which avoids extracting some noisy tags and balances the length of documents to some extent. After preprocessing, a corpus $D = \{d_1, d_2, \ldots, d_n\}$ is generated, where n is the total number of cleaned documents.

- **Tag Extractor and Formulator**: According to our previous work [3], eight common types of people tags are extracted from D, among which people name, organization, and location are detected using both the character language model and hidden Markov model in a natural language processing toolkit Lingpipe[2], other types of tags including email address, phone number, birthdate, occupation, and URL domain are extracted using a set of rules. In this paper, we make some improvements on the tag extraction rules introduced in [3]. It is found that some types of tags, such as occupation, birthdate, and phone number usually occur nearer from the queried people name compared with other types of tags, and therefore we restrict the rule-based extraction within certain bound around the queried people name for these types of tags. Besides, based on the extraction rule "username@domain-name", we propose two notions, *match_ratio* and *leap_count*, to accurately identify the tag of email. Take the query name "John Smith" and an extracted email "jsmith@lawsonlundell.com" for example, *match_ratio* is computed as S/N, where S is the number of characters contained in both email username and query name, N is the length of email username. In addition, supported by the intuition that most of the letters in username should keep the same order with those in the query name, we scan the email username, and increase the value of *leap_count* by one when the sequence of any adjacent letters in email username is different from that in the query name. Finally, we choose the email addresses, which meet the conditions of *match_ratio* above 0.75 and *leap_count* smaller than 2.

  Moreover, we observe that the tags of birthdate and telephone number have many variants, which decrease the accuracy in disambiguating different people entities sharing the queried people name, for example, the tag of "12, Nov. 1984" is equivalent with "November 12 1984", and "(355) 345 2334" is the same with "355-345-2334". To address this issue, we propose some rules to reformulate tags of each type in the same format. Furthermore, we found that Lingpipe often mistakenly extracts the organization or location starting with a preposition, such as "at University of Maryland", and therefore, we remove the leading preposition of the tag to generate an accurate tag "University of Maryland" for further processing. After employing these extraction techniques, we get a non-repetitive tag corpus $A = \{a_1, a_2, \ldots, a_m\}$.

- **Tag Filterer**: Due to the imperfect preprocessing and tag extraction, some tags which make little sense should be filtered out from the tag corpus A. a) Tags equal to the

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

[2]http://alias-i.com/lingpipe/

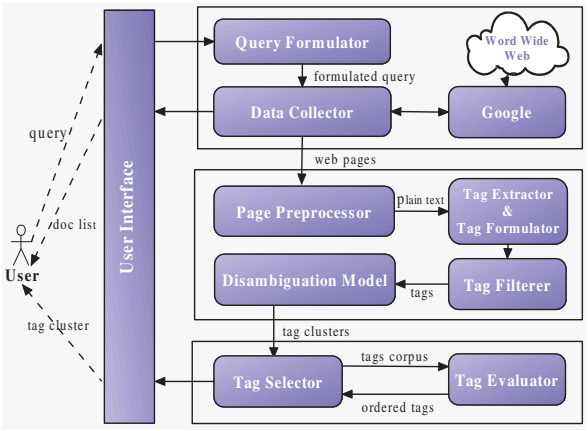

Figure 3: The Architecture of System GRAPE

commonly used English words, such as "the", "cache", and "audience"; b) Tags containing illegal characters; c) Tags equivalent with the queried people name or its variants, for example, "John smith", "smith John", and "J. smith" are often extracted as related people tags with respect to the query "John Smith"; d) Tags of organization or location that are also detected as related people; e) Special emails, which frequently occur as the contact of a Web site, such as webmaster@domain-name, support@domain-name, and feedback@domain-name. After removing these noisy tags, the corpus A is then presented as $A' = \{a_1, a_2, \ldots, a_{m'}\}$, where $A' \subseteq A$.

- **Disambiguation Model**: The core component of our system is a graph-based disambiguation model, which is developed based on the observation that the co-occurrence of people tags can help achieve high disambiguation quality. By exploring the relationships between tags, the proposed graph model is constructed based on tag corpus A' and finally partitions the graph into several subgraphs, each of which can uniquely characterize a specific people entity owning the queried people name. We give a general description about the graph-based disambiguation model as follows.

  Firstly, A' is modeled as a graph G with m' nodes, where a node is created for each unique tag $a \in A'$, and an edge is added between two tags when they co-occur in the same document. Secondly, we compute the edge weight and node weight respectively. Since the edge indicates the relevance of the connected nodes, we define edge weight as the number of documents where any two tags co-occur. Due to different characteristics of the eight types of tags surrounding the queried people name. We assign different type weights to nodes with different tag types under the following heuristic: the higher the number of unique tags of a certain type, the smaller the weight of each node with this tag type. Thirdly, the connectivity strength between any two tags in G are measured based on their edge weight and node weights. Finally, a clustering algorithm is performed on the graph to group these tags into clusters, while the tags in each cluster are used to represent a certain people entity. We have evaluated our clustering algorithm in comparison with the top five best methods in [1] [2] respectively, and the experimental results show that our method outperforms the state-of-the-art Web people name disambiguation approaches.

2.3 People Tag Presentation

After clustering, we can identify each people entity with a bag of tags with different types. Due to the great amount of tags in each cluster, we determine to select a quantity of tags for presentation, which can describe each people entity uniquely and help the users quickly locate the person they are actually looking for. In this section, we describe the process of tag evaluation and selection for each people entity.

- **Tag Evaluator**: Lingpipe is potentially useful for tag extraction with a high recall. However, the big drawback is its ambiguous extraction between organization and location, that is, organizations are often detected as locations. Thus, we do not display location information to users although they have beneficial impacts on the disambiguation model. To choose tags with high ability of representing each people entity uniquely, we propose $tag\_confidence$ to measure the significance of each tag, which takes both tag frequency and document rank into account. Intuitively, frequent tags are regarded more relevant with the people entity than the infrequent ones. Additionally, motivated by the fact that the returned documents from Google are ranked according to the importance and relevance with the queried people name, we give more priorities to the tags that are extracted from the documents of high ranking. The $tag\_confidence$ is defined as follows.

$$tag\_confidence(t) = \sum_{i=1}^{k} \alpha \times \frac{Freq(t_i)}{Rank(d_i)} \quad (1)$$

  Assume tag t is located in cluster c, which is composed of k support documents, $c = \{d_1, d_2, \ldots, d_k\}$, $Freq(t_i)$ denotes the frequency of tag t in d_i and $Rank(d_i)$ is the ranking value of document d_i. This formula means that the frequent tags in highly-rank documents are more reliable. Our experimental studies prove that this formula is extremely effective for choosing the representative tags.

  Additionally, the experiment results found that some accurately extracted organization names contain certain distinctive key words, such as "university", "center", "company", "base", "foundation", and "association", and thus we assign a confidence parameter α to the tags containing one of these key words. Finally, the tags of each type for each namesake are sorted in a descending order according to their $tag\_confidence$ values. Parameter α is acquired and verified through a large amount of experiments.

- **Tag Selector**: After evaluation and sorting, we present the representative tags with high $tag\_confidence$ to users. For each people entity, five related people names are presented, while the types of birthdate, phone number, and email are presented with one tag respectively if available. Usually a person is relevant with more than one organizations or occupations. Motivated by the intuition that if a people entity is related with multiple tags with the same type, the differences of $tag\_confidence$ values between these tags are not very big. We simply focus on the top two tags of these types in this demo. Assume a and b is the two top tags ranked by $tag\_confidence$ typed as organization or occupation, if the confidence value of a is once bigger than b, only a is displayed, otherwise both of them are displayed. In next section, some example queries and their corresponding tag clusters generated by the tag selector are demonstrated.

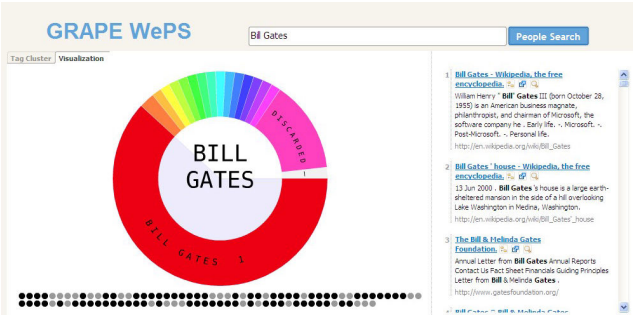

Figure 4: Visualization Interface of GRAPE

3. DEMONSTRATION DESCRIPTION

As shown in Figure 2, the result interface is composed of two panels, tag cluster presentation panel on the left and document presentation panel on the right. Also, a visualization panel is demonstrated in Figure 4.

- **Documents Presentation**: Similar with the general Web search systems, GRAPE presents the top 100 results from Google on the right panel when it is issued by a people name query, and each result includes title, snippet, and URL.
- **Tag Cluster Presentation**: Tag clusters for the namesakes are displayed individually and sorted descendingly according the number of support documents in each cluster. As shown in Figure 2, each people entity are labeled with the query name plus a serial number. Since there is a mapping between documents and the tag cluster, if users click any people entity label on tag cluster presentation panel, and the document list supporting this cluster will show up on the right panel accordingly. Additionally, the elapse time of data collecting and clustering is also showed below the panel. Figure 5 demonstrates three example queries and the corresponding tag clusters generated by our system online. Due to the limited space, we ignore the people entities without any representative tags and present the top-three people entities for each query.
- **Visualization Presentation**: For the purpose of both attractive visualization and statistics information presentation, we employ the flash-based visualization component of Carrot2, which is an open-source framework developed by Dawid Weiss [7]. As the example query "Bill Gates" presented in Figure 4, a circle graph is divided into several sectors, illustrating relative magnitudes of namesakes. The area of each sector is proportional to the quantity of support documents in each tag cluster. Moreover, there are a hundred solid points below this circle graph to denote the top hundred documents from Google. When the user click any sector, the corresponding documents for the represented people entity will be shown on the right panel, and meanwhile some of the solid points will be highlighted to show the corresponding documents rank distribution in Google search results. In this example, most of the documents are relevant with a "Bill Gates", the former executive of Microsoft. If we search for a common people, the circle graph will be divided into many sectors, where the number of documents in most sectors are probably equal.

Query	Tag Cluster Presenation
Bill Gates	Bill Gates 1: **Occupaiton**: executive, software architect **Birthdate**: October 28, 1955 **Email**: billig@microsoft.com **Related People**: Paul Allen, Steve Ballmer, Eristalis gatesi, David Boies, Christos Papadimitriou **Organizaiton**: Lakeside School, Harvard University Bill Gates 2: **Organizaiton**: malaria and education, TED Blog Bill Gates 3: **Phone**: 512-892-033
Barack Obama	Barack Obama 1: **Occupaiton**: president **Birthdate**: Aug 4,1961 **Related People**: John McCain, Hillary Clinton, Mitch Stewart, Ann Dunham, Michael Tomasky **Organizaiton**: Columbia University, Harvard Law School Barack Obama 2: **Related People**: J. Clifford Barack Obama 3: **Occupaiton**: actor
John K. Smith	John K. Smith 1: **Organizaiton**: EOL Biodiversity Synthesis Group John K. Smith 2: **Occupaiton**: technician **Birthdate**: June 24, 1961 **Related People**: Natalie Smith, Irene Linden, Francis Smith, Tracey Smith, Meg Smith **Organizaiton:** Trinity Lutheran Church John K. Smith 3: **Related People**: Joseph Nathan, Mahlon Kline, Thomas Beecham, Silas Burroughs **Organizaiton**: SmithKline Beecham, Glaxo Wellcome

Figure 5: Tag Cluster Presentation in GRAPE

4. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grant No. 60833003 and Grant No. 90924025, an HP Labs Innovation Research Program award, the Okawa Foundation Research Grant, and the Guangdong Provincial Government, State Education Ministry of China under Grant No. 0712226-100097.

5. REFERENCES

[1] A. Javier, J. Gonzalo, and S. Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for web people search task. In *Proceedings of Semeval 2007, Association for Computational Linguistics*, pages 9–16, 2007.

[2] A. Javier, J. Gonzalo, and S. Sekine. Weps 2 evaluation campaign: overview of the web people search clustering task. In *Proceedings of In 2nd Web People Search Evaluation Workshop (WePS), 18th WWW Conference*, 2009.

[3] L. Jiang, J. Wang, N. An, S. Wang, J. Zhan, and L. Li. Grape: A graph-based framework for disambiguating people appearances in web search. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2009.

[4] D. V. Kalashnikov, Z. Chen, R. Nuray-Turan, S. Mehrotra, and Z. Zhang. West: Modern technologies for web people search. In *Proceedings of the 25th International Conference on Data Engineering (ICDE)*, pages 1487–1490, 2009.

[5] D. V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra. Towards breaking the quality curse: a web-querying approach to web people search. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 27–34, 2008.

[6] G. S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the 7th Conference on Computational Natural Language Learning*, 2003.

[7] S. Osinski and D. Weiss. Carrot^2: Design of a flexible and efficient web information retrieval framework. In *AWIC*, pages 439–444. http://project.carrot2.org, 2005.