Appeared in IEEE ICDE 2009 Conference.

WEST: Modern Technologies for Web People Search

Dmitri V. Kalashnikov Zhaoqi Chen Rabia Nuray-Turan Sharad Mehrotra Zheng Zhang

Computer Science Department
University of California, Irvine

I. Introduction

In this paper we describe WEST (Web Entity Search Technologies) system that we have developed to improve people search over the Internet. Recently the problem of Web People Search (WePS) has attracted significant attention from both the industry and academia. In the classic formulation of WePS problem the user issues a query to a web search engine that consists of a name of a person of interest. For such a query, a traditional search engine such as Yahoo or Google would return webpages that are related to *any* people who happened to have the queried name. The goal of WePS, instead, is to output a set of clusters of webpages, one cluster per each distinct person, containing all of the webpages related to that person. The user then can locate the desired cluster and explore the webpages it contains.

The WePS approach offers significant advantages. For example, consider searching for a person who is a namesake of the former President Bill Clinton. The webpages of the less famous person will be overshadowed in today's search engines and will appear far in the search. WePS systems address this problem by first presenting to the user the set of clusters, among which the user then can select the cluster containing the webpages of the namesake of interest.

The key technology of any WePS system, including WEST, is that of Entity Resolution. In a setting of Entity Resolution problem, a dataset contains information about objects and their interactions. The objects are referred to via (textual) descriptions/references, which might not be unique identifiers of the objects, leading to ambiguity. The task of Entity Resolution algorithms is to identify all of the references that co-refer, i.e., refer to the same real-world entity. In WePS the webpages returned by a search engine can be viewed as references. The overall task can be viewed as that of finding the webpages that refer to the same namesake.

We have developed three different Entity Resolution algorithms that can be employed by WEST:

1) *GraphER* approach extracts the Social Network (people, organizations, locations) off the webpages along with hyperlink and email information. It represents the resulting Entity-Relationship network as a graph. The approach then analyzes this graph and the webpage textual similarity to determine which webpages co-refer [4], [5]. GraphER will be covered in Section III-A.
2) *EnsembleER* approach combines results of multiple "base" ER systems to produce the overall clustering. During the training phase, EnsembleER approach employs supervised learning to study how well the base ER systems perform in terms of their quality under variety of conditions/contexts by training a meta-level classifier. It then uses this classifier during the actual query processing to compute its final clustering [3]. EnsembleER will be covered in Section III-B.
3) *WebER* approach, unlike the above two (and many other) approaches, does not limit its processing to analyzing the relevant webpages only. Instead, it leverages a powerful external data source to gain its advantage. Specifically, like GraphER it first extracts social network off the webpages. But then it queries the Web to collect additional information on the various components of this network [6]. WebER will be covered in Section III-C.

Each of these three algorithms has been demonstrated to outperform the current state of the art techniques on a variety of datasets [3]–[6]. The comparison includes 18 approaches that have been part of WePS Task competition on a large dataset which is now considered to be a de facto standard for testing WePS solutions [1].

WEST provides multiple interfaces to search. The input and output interfaces of WEST are illustrated in Figures 1 and 2 respectively. Naturally, WEST supports the standard WePS interface where the user provides a person name as the query. It also supports additional functionality, where the user can specify *context queries* to help locate the namesake of interest quicker. The context can be specified in the form of location, people, and/or organizations associated with the namesake of interest. Notice that the context here is not used as additional keywords to query the Web, but is used to identify the right namesake the user is looking for. This means that the webpages in the cluster does not have to each contain the context keywords, and some of them might even contain none of these additional context keywords.

Besides the UI for searching for a single individual, WEST offers a Group Search interface to support the Group Identification query capabilities. In a Group Identification task, the input is multiple names of people that are known to be related in some way. For instance, a query might be "Michael Jordan"

This research was supported by NSF Awards 0331707 and 0331690, and DHS Award EMW-2007-FP-02535.

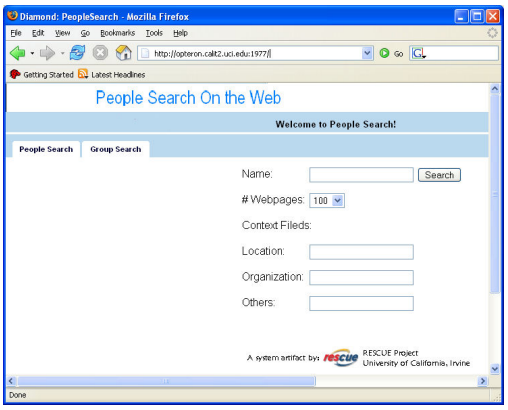

Fig. 1. Input Interface of WEST.

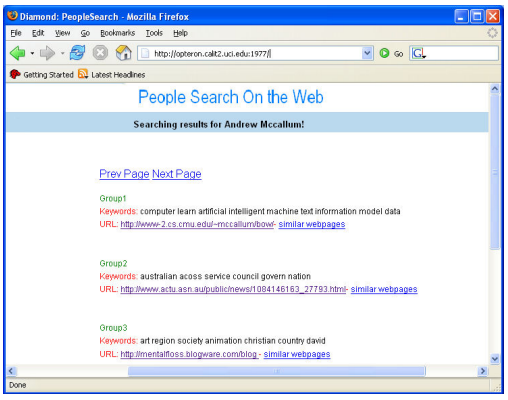

Fig. 2. Output Interface of WEST.

and "Magic Johnson", implying that the meant namesakes are basketball players. The objective is to retrieve the webpages of the meant namesakes only.

While the demonstration will illustrate both the single person search and group search capabilities, the subsequent discussion will focus on a single person search. The algorithmic details of the Group Search can be found in [4]. The rest of this paper is organized as follows. Section II presents the steps of the overall WEST approach. Then Section III covers the three Entity Resolution algorithms. Finally, Section IV describes the functionality of WEST that will be displayed during the demo.

II. Overall Algorithm

The steps of the overall WEST approach, in the context of a *middleware* architecture, are illustrated in Figure 3. They include:

1) *User Input.* The user issues a query via the WEST input interface.
2) *Top-K Retrieval.* The system (middleware) sends a query consisting of a person name to a search engine, such as Google, and retrieves the top-K returned web pages. This is a standard step performed by most of the current WePS systems.

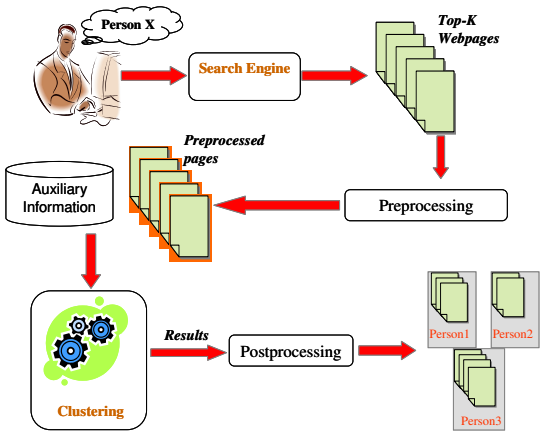

Fig. 3. Overview of the WEST Processing Steps.

3) *Pre-processing.* These top-K webpages are then preprocessed. The main two pre-processing steps are:
   a) *TF/IDF.* Pre-processing steps for computing TF/IDF are carried out. They include: stemming, stop word removal, noun phrase identification, inverted index computations, etc.
   b) *Extraction.* Named Entities, including people, locations, organizations are extracted using a third party named entity extraction software. Hyperlinks and emails addressed are extracted as well. Some auxiliary data structures are built on this data.
4) *Clustering.* One of the three Entity Resolution algorithms is applied to the data to cluster the web pages. The algorithms will be explained in Section III.
5) *Post-processing.* The post-processing steps include:
   a) *Cluster Sketches* are computed.
   b) *Cluster Rank* is computed based on (a) the context keywords, if present and (b) the original search engine's ordering of the webpages.
   c) *Webpage Rank* is computed to determine the relative ordering of webpages inside each cluster.
6) *Visualization.* The resulting clusters are presented to the user, which can be interactively explored.

We next discuss the key component of any WePS system: the Entity Resolution algorithms.

III. Entity Resolution Algorithms

This section presents an overview of the three entity resolution algorithms used by the WEST system for clustering the webpages.

A. GraphER

To determine whether two references u and v co-refer traditional approaches at the core analyze similarity of features of u and v according to some feature-based similarity function $f(u, v)$. The GraphER approach has been developed based on the observation that many datasets are relational in nature. They contain not only objects and their features but also information about relationships in which they participate.

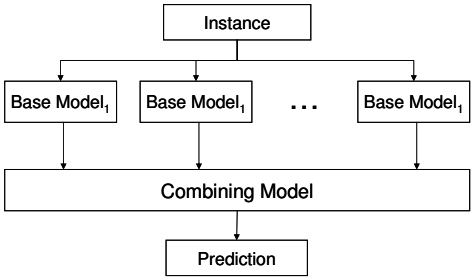

Fig. 4. A General Framework for Combining Multiple Systems.

GraphER utilizes the information stored in these relationships to improve the disambiguation quality.

The approach views the dataset being analyzed as an Entity-Relationship Graph of nodes (entities) interconnected via relationships (edges). For the WePS domain, the nodes are the named entities, hyperlinks, and emails extracted off the webpages during the pre-processing as well as the webpages themselves. The relationships are co-occurrence relationships, and those that are derived from hyperlink and decompositions. The graph creation procedure is discussed in detail in [4].

The entity relationships graph in this case is a combination of the Social Network extracted from the webpages as well as the hyperlink graph. To decide whether two references u and v co-refer, GraphER analyzes how strongly u and v are connected in this graph according to a *connection strength* measure $c(u, v)$. To compute $c(u, v)$, the algorithm discovers the set P^L_{uv} of all L-short simple u-v paths.[1] The value of $c(u, v)$ is computed as the sum of the connections strength contributed from each path p in P^L_{uv}: $c(u, v) = \sum_{p \in P^L_{uv}} c(p)$. A supervised learning procedure, formulated as a linear programming optimization task, is used to learn $c(p)$ function from data [4], [5]. The similarity function $s(u, v)$ is then defined as a combination of $c(u, v)$ and $f(u, v)$. The output of this function is used by a correlation clustering algorithm to generate the final clusters of webpages.

B. EnsembleER

EnsembleER approach is motivated by the observation that often there is no single entity resolution (ER) technique always perform the best. Rather, different ER solutions perform better in different *contexts*. EnsembleER is a stacking-like framework that combines the clustering results of multiple base-level ER systems so that the final clustering quality is superior to that of any single base ER system.

The key idea is to transform the output of base-level ER systems, together with context, into a meta-level feature set. A supervised learning approach is utilized to train a classifier on the meta-level data. The algorithm then applies the meta-level classifier to the dataset being processed to create the final clustering results. Figure 4 shows a general framework of combining multiple systems.

Similar to GraphER approach, EnsembleER also utilizes a graph representation of the dataset. The graph however is different. The nodes are the top-K webpages. Edge (u, v) between two webpages u and v is created only if a certain number of the base-level ER systems decide that u and v should be in the same cluster. Edge (u, v) represents a possibility that u and v might co-refer. With respect to the graph that task of EnsembleER can be viewed as deciding for each edge whether u and v should be put in one cluster.

Let $S_1, S_2, \ldots, S_n$ be the n base-level ER systems. For each edge $e_i = (u, v)$, each S_j output its decision $d_{ij} \in \{0, 1\}$. Here, if u and v are placed in the same cluster by S_j then $d_{ij} = 1$ otherwise $d_{ij} = 0$. Then, for each edge e_i we can define a *decision feature* vector as $\mathbf{d}_i = \{d_{i1}, d_{i2}, \ldots, d_{in}\}$.

For edge e_i its local context is also encoded as a multi-dimensional *context feature* vector $\mathbf{f}_i = \{f_{i1}, f_{i2}, \ldots, f_{im}\}$. One of the interesting aspects of EnsembleER solution is that it creates context features in a predictive way, based on first *estimating* some unknown parameters of the data being processed. For instance, let $K_1, K_2, \ldots, K_n$ be the number of clusters that systems $S_1, S_2, \ldots, S_n$ output. One of the features used by EnsembleER is computed by applying a regression to this data to estimate the number of namesakes $K^\star$, where the true number of namesakes K^+ is unknown beforehand to the algorithm. EnsembleER then converts the difference between $K^\star$ and K_j into a feature, based on the intuition that the closer the K_j to $K^\star$, the more confidence can be placed in the answer of system S_j.

The goal of EnsembleER reduces to finding a mapping $\mathbf{d}_i \times \mathbf{f}_i \rightarrow a^\star_i$. Here, $a^\star_i = \{0, 1\}$ is the prediction of the combined algorithm for edge $e_i = (u, v)$, where $a^\star_i = 1$ if the overall algorithm believes u and v belong to the same cluster, and $a^\star_i = 0$ otherwise. The details of the Ensemble algorithm can be found in [3].

C. WebER

WebER approach is considerably different from most of the other WePS solutions. Unlike many other WePS systems, WebER does not limit its processing to analyzing only the information stored in the top-K returned webpages. Rather it employs the Web as an external data source to get additional information, which ultimately leads to higher quality results.

WebER is primarily intended to be a server-side solution. That is, its code is executed at a search engine (server) side. Because of that, most of the pre-processing can be accomplished in bulk before query processing starts, including extraction and TF/IDF computations. The queries to the search engine are carried out internally without going via the Internet thus making their processing much faster.

Let $D = \{d_1, d_2, \ldots, d_K\}$ be the set of the top-K returned webpages. WebER first merges some of the webpages into *initial clusters* using Named Entity (NE) clustering with a conservative thresholds. The document- document similarity is computed using TF/IDF approach with cosine similarity. Only a few webpages that have overwhelming evidence that they represent the same people are merged during this process. Let P_i and O_i be the set of people and organizations extracted from webpage d_i. For each pair webpages d_i and d_j that

[1] A path is L-short if its length does not exceed L. A path is simple if it does not contain duplicate nodes.

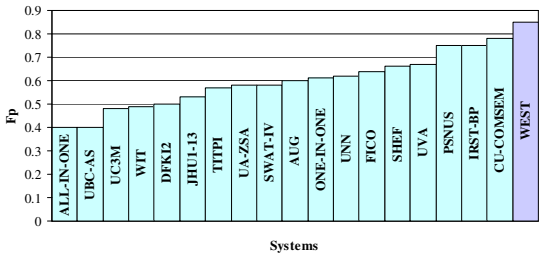

Fig. 5. The Experiment results on WePS dataset.

are not yet put in the same cluster the approach forms and issues queries to the Web to collect the co-occurrence statistics, which in this case is the number of the pages returned for a given query. WebER uses two main types of queries:

- $\mathcal{N}$ AND $\mathcal{C}_i$ AND $\mathcal{C}_j$
- $\mathcal{C}_i$ AND $\mathcal{C}_j$

Here $\mathcal{N}$ is the name of the person being queried by the user, and $\mathcal{C}_i$ and $\mathcal{C}_j$ are the context of pages d_i and d_j. Context $\mathcal{C}_i$ can be either (a) an OR combination of people from P_i, or (b) an OR combination of organizations from O_i. The same holds for $\mathcal{C}_i$ resulting in eight queries for d_i and d_j pair. These co-occurrence counts are indicative of how often the elements of the two social networks co-occur on the web and thus how strongly they are related. These counts are then transformed into features, which are then used to compute the similarity between webpages d_i and d_j.

One of the key contributions of this work is a new Skyline-based classifier for deciding which d_i and d_j webpages should be merged based on the corresponding feature vector. It is a specialized classifier that we have designed specifically for the clustering problem at hand. Skyline-based classifier gains its advantage due to a variety of functionalities built into it, including:

- It takes into account dominance that is present in the features space.
- It also fine tunes itself to the quality measure being used.
- It takes into account transitivity of merges: that is, accounts for the fact that two large clusters can be merged by a single merge decision, and, thus, one direct merge decision can lead to multiple indirect ones.

These properties allow it to easily outperform other classification methods (which are generic), such as DTC or SVM. The approach is discussed in detail in [6].

IV. Demonstration

The ER algorithms used by WEST are known to produce highly competitive results. Figure 5 presents the comparison results of the WEST with 18 other WePS solutions that have been part of the WePS Task challenge [1]. The quality of clustering is evaluated in terms of F_p measure (harmonic mean of Purity and Inverse Purity [1]). For the group identification we have compared WEST with the state of the art approach published in [2]. The average F-measure on this dataset achieved by WEST is 92% which is nearly 12% improvement over the result reported in [2].

The WEST system will be demonstrated through two applications built over the base system.

- **Single Person Search** (illustrated in Figure 1): wherein a user can enter a person name and context in the form of people, locations, and/or organizations associated with the person being queried. The results will be a set of clusters. Each cluster will have a set of keywords attached to indicate the main aspect of the corresponding namesake. The clusters will be presented in a ranked order based on the original ranks of the web pages in the clusters and the context keywords. Figure 2 shows sample resulting clusters for the query "Andrew McCallum". The first returned group corresponds to Andrew McCallum the UMass CS professor, the second to the president of the Australian Council of Social Services, the third to a Canadian musician, etc. The user will be able to click on the clusters and explore their clusters interactively. The webpages in a cluster will be presented in a ranked order as well.
- **Group Search**: Another interface will be used to demonstrate the Group Identification search capabilities of WEST. In group query interface, the user can input several person names. The result will be the web pages that are related to the meant namesakes.

These applications will be demonstrated both in the online and offline modes. In the online mode, the query input by the user will be translated into a corresponding (set of) queries over Internet search engines (specifically over Google). WEST allows the user to specify the number of web pages to retrieve from the search engine, which will be disambiguated into corresponding clusters. In the online mode, WEST uses only GraphER and EnsembleER approaches since WebER is a server-side approach and is not amenable for realization as a middleware. The demonstration will allow observers to do diverse searches (perhaps, of their own names) and perceive both the quality as well as efficiency of WEST.

In the offline mode, WEST will use preconstructed "canned" examples where we have already crawled the web to retrieve the search results and constructed the corresponding clusters. In the offline mode, in addition to illustrating the GraphER and EnsembleER approaches, we will also demonstrate the disambiguation power of the WebER approach.

References

[1] J. Artiles, J. Gonzalo, and S. Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *SemEval*, 2007.

[2] R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In *WWW*, 2005.

[3] Z. Chen, D. V. Kalashnikov, and S. Mehrotra. Combining entity resolution techniques with application to web people search. In *Under submission*.

[4] D. V. Kalashnikov, Z. Chen, S. Mehrotra, and R. Nuray. Web people search via connection analysis. *IEEE TKDE*, 2008. to appear.

[5] D. V. Kalashnikov, S. Mehrotra, S. Chen, R. Nuray, and N. Ashish. Disambiguation algorithm for people search on the web. In *ICDE*, 2007.

[6] D. V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra. Towards breaking the quality curse. A web-querying approach to Web People Search. In *Proc. of Annual International ACM SIGIR Conference*, Singapore, July 20–24 2008.