Minerva II: A Novel Entity Discovery Tool

Andrew Jie Zhou

Department of Computer
Science
Georgetown University
Washington, DC
jz398@georgetown

Hui Yang

Department of Computer
Science
Georgetown University
Washington, DC
huiyang@cs.georgetown.edu

Hongkai Wu

Department of Computer
Science
Georgetown University
Washington, DC
hw271@georgetown.edu

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).

CHI'16 Extended Abstracts, May 7–12, 2016, San Jose, CA, USA.

ACM 978-1-4503-4082-3/16/05.

<http://dx.doi.org/10.1145/2851581.2892520>

Abstract

Entity discovery is a long-lasting interest in governments, enterprises, and the research community. It is a complex task that requires retrieving, extracting, linking, and displaying entities. Algorithms to support entity discovery have been proposed across disciplines including Information Retrieval (IR), Information Extraction (IE), Natural Language Processing (NLP), and Data Mining (DM). However, there is little study on User Interface (UI) for supporting effective entity discovery. This paper presents Minerva II, a novel entity discovery tool, to tackle this challenge. In the paper, we illustrate the UI design and how it effectively supports the typical work flow when a user performs entity discovery. We also describe a new visualization algorithm for entity networks. Our user study shows that Minerva II is able to greatly increase users' efficiency.

Author Keywords

Entity Information System; Interface; Visualization

ACM Classification Keywords

H.5.0 [Information interfaces and presentation (e.g., HCI)]: General

Introduction

Entity discovery is a long-lasting interest in governments, enterprises, and the research community. Entities include

persons, organizations, locations and other names that can be linked to a physical thing [1]. Entity discovery requires retrieving, extracting, linking, and displaying entities. Entity discovery is a complex task lying at the intersection of Information Retrieval (IR), Information Extraction (IE), Natural Language Processing (NLP), and Data Mining (DM).

Over the past decade, entity search attracted a great deal of attention from the research community. For instance, from 2005 to 2008, the TREC (Text REtrieval Conference) Enterprise Track was made for expert finding which is focused on a single kind of entity: people. Since 2007, the INEX XML Entity Ranking Track addressed retrieving a ranked list of entities when given a description such as "Impressionist art in the Netherlands". In 2009, TREC launched again a entity track to find related entities (REF) given a source entity. Research on entities includes entity extraction [5], entity linking [8], entity retrieval [6], entity resolution [7], and entity evolution [4].

Despite the abundant research on algorithms to support entity discovery, little has been done on User Interface (UI) to support entity discovery. In this paper, we present a novel entity discovery tool called Minerva II to tackle this challenge.

Figure 1 shows a snapshot of Minerva II. Our tool contains a range of features that exclusively designed entity discovery. The features include entity management, entity linking, similar entity searching and entity network visualization. A user study shows that our UI design can significantly increase users' work efficiency.

Our paper is arranged as the following: In section 2, we describe the data used in our entity search system (Minerva II). In section 3, we present the overview of Minerva II's interface. In section 4, we present the workflow of Minerva II.

In section 5, we discuss the algorithms that are powering Minerva II. In section 6, we evaluate Minerva II's usefulness. In section 7, we draw our conclusion.

Data

An entity search tool is designed for users to investigate entities. Minerva II is initially designed for law enforcement agencies to find suspicious sellers and buyers in the counterfeit electronic market. The raw data used by Minerva II was collected from U.S. law enforcement agencies. The data contain a list of the Defense Logistic Agency's microelectronics suppliers, a list of the buildings in China which are crowded with individual microelectronics suppliers, a list of elements that OSI/HSI are most interested in and are widely available on the counterfeit market, and a list of elements that are suspicious counterfeit. Besides, the raw data also includes primarily corporate websites, merchant sites for ordering parts, social media sites, onion sites and some defunct sites. The dataset contains a total of 1,727,202 documents.

Overview of User Interface

Minerva II has a user interface (Figure 1) specially designed for entity discovery. The interface contains three parts: search frame, entity management frame and network visualization frame. Their functions are described below:

1. Search Frame

The search frame is similar to a regular document search interface. Users can enter an entity into the text-entry field. The documents that contain the entered entity will be listed on the left panel. If the documents contain non-English contents, their English translations will also be displayed for users' convenience. The content of a specific document and all of the entities in the same document will be

Figure 1: The user interface of Minerva II.

displayed in the right panel. Worth to mention, users who conducting entity search usually encounter a specific user case: they want to look up a list of entities simultaneously. So, Minerva II enable users to upload a list of entities as batch queries, and the queried entities will be highlighted in various colors.

2. Entity Management Frame

The entity management frame is for users to manage the entities they are interested in. The functions of entity management frame include: a) The users can drag entities from the returned documents to the entity management frame for future reference. b) The users can tag the entities' types such as location, email address, name, and later on to find more entities of the same type. c) The users can also specify the links between two entities. This information will be

used to visualize entity network, which will be illustrated later.

3. Network Visualization Frame

The network visualization frame illustrates the network formed by the entities. For example, a counterfeit electronic seller may have 3 addresses and linked to 1 buyer. So that the network contains 5 entities (2 people and 3 addresses) and 4 links (3 "belong to" relations and 1 "buy-sell" relation). We construct the network in an interactive 3D interface which provides users with three unique features: a) The users can investigate the network structure from every perspective, as well as zoom in and zoom out from it. b) The users can show/hide the less important entities in a network. By "less important" we mean they only belong to another entity and have no connection to any others. In this

way, users can easily inspect the main structure of the entity network. c) The various types of links have their unique representations. Users can distinguish one kind of link from another instinctively. The representations include embracement between entities and attributes, point-to-point links between two entities, and tails for a list of attributes of the same type.

Typical Workflow

The users can search entities and visualize entity network through Minerva II. The workflow of Minerva II is as Figure 2 shows. The user starts by entering a name of an entity into the text-entry field or by uploading a file contains a list of entity names. After the user clicks the query button, the left panel shows a list of documents that are relevant to the queried entities. Each document has a vote button beside it, by which the user can vote up or down for like or dislike. By gathering the user's votes, our search algorithm can learn user's preference to optimize the search results. When the user browses the returned documents, one of the documents may get her attention so she clicks it. Then, the right panel will show the details of the documents, including all of the entities inside the document. As the user inspecting the details of the document, she may find some entities of interest and drag them into the entity management frame. Assuming now the entity management frame contains one or more entities, the user has three options to do with them. First, the user can tag the type for the entities. For example, if the entity is "1300 M Street NW, DC", the user may tag it as address. The tags will be stored in the database for future reference. After tagging, the user can right click the entity and choose "find more this type" to get the same type of entities someone tagged before. Second, the user can specify the links between two entities. For example, a phone number and a seller entity would be a "belong to" link. The link information will be used to visual-

Figure 2: The typical workflow of Minerva II.

ize the entity network. Third, the user can click the "visualization" button to view the 3D entity network. The structure that represents the network is created inside a 3D space and photoed by a visual camera. The user can view the structure through the camera and drag the view to change the camera's position. The user can also right click an entity to activate the hide/show menu. The menu provides the user options to hide/show the less important entities that only belong to a primary entity but has no connection to any other entities.

Algorithms Behind the Scene

1. Link Discovery Algorithm

The visualization tool utilizes the entities and the links between them to generate a 3D network. Our link discovery algorithm is based on bootstrapping [2]. It is designed mainly to help users find entities more easily and efficiently. There are six steps in the algorithm.

- A The user finds a good text fragment in a document as a seed. Then the user drags and drops the fragment in into the entity management frame.
- B Minerva II use Stanford NLP tool to parse the text fragment and get all the nouns using pos (part-of-speech) tagging.
- C The user is then provided with a list of suggested noun pairs and the user will tell the system in which pair there may be a connection.
- D Next, we extract the part between the noun pair and use the pos tagging to filter all other terms except the adjectives, numbers and nouns.
- E The filtered words are concatenated to form a lexico-syntactic text pattern.
- F The pattern is used as a query to re-search in the corpus and provide the user with a new list of documents. These documents contain sentences conforming to this pattern, which are for potential entities. It is expected that the user can find more good entity sentences in these documents and the whole process repeats.

Our algorithm works efficiently since a user only needs to do two things: drag-and-drop, and choose the meaningful noun pair. Then, the system automates to find a list of documents containing highly relevant potential entities, which saves the user's time comparing with searching using Google.

2. Arrangement Algorithm

The visualization tool of Minerva II utilizes 3D technology to visualize the entity network. As showed in Figure3, the

Figure 3: Visualize the entity network.

entities are represented by spheres. In our case, the counterfeit electronic network has three types of links, so we use three different designs to represent them: the embracement design indicates "belong to" relationships; a sequential of entities indicates that they are the same type; two entities linked by cylinder means that they have transaction with each other. With different designs, the user can instantly distinguish one kind of link from another, and thus get an intuitive impression of the network structure.

To allocate the spheres and links in a visually pleasant way, we use an arrangement algorithm to calculate the layout. The arrangement algorithm is derived from force-directed drawing algorithm [3] and adjusted for the 3D network's needs. The arrangement algorithm is described as follows:

We replace the nodes with electrically charged particles based on Coulomb's law, and replace the edges with springs based on Hooke's law. The particles are initially placed in some random place and let go so that the spring forces and electrically repulsive forces will move the system to a mechanical equilibrium state. The spring forces is calculate by $F = C_1 * \log(d)$ where C_1 is a constant. The algorithm use logarithmic force because Hookes Law springs are too

Figure 4: Average time cost of each task.

Figure 5: Average number of operations of each task.

Figure 6: Users' perception of traditional method and Minerva II

strong when the edges are long. The repulsive forces between two particles are calculated by $F = C_2 * (n_1 * n_2) / d^2$, where C_2 is a constant; n_1 and n_2 are the number of entities each particle has. We simulate the mechanical system as follows where C_3 and C_4 are both constants:

Divide the nodes into two sets: one is a set of primary nodes set, the other is a set of subordinate nodes. The nodes that have edges to other nodes and the nodes that have "has a" relationship with other nodes are primary nodes. The others are subordinate nodes.

Calculate n_i (number of nodes that each primary node has).

Place the primary nodes in random locations.

Repeat

Repeat C_3 times

Calculate the force on each nodes.

Move the node by $C_4 * (\text{force on node})$

Place the subordinate nodes evenly surrounding the primary nodes. By calculating the primary node's relative location with the center point of the 3D space, we can locate the longer sequential subordinate nodes at the farther side of the primary nodes from the center.

If we detect a collision between two nodes, we increase n of the corresponding primary nodes by C_5 to separate them more.

This step will be repeated until no collision is detected.

The values $C_1 = 2, C_2 = 0.2, C_3 = 50, C_4 = 0.2, C_5 = 3$ are appropriate for most of our experiments.

Evaluation

We evaluate the utility of Minerva II by comparing it with the counterfeit electronic agents' current search process. For

now, the agents' entity collection process is as following: enter a query in Google, read the document list, click an interesting document, look for entity, copy the entity, switch to Excel window, paste the entity, then switch back to Google to look for more entity. In contrast, Minerva II reduces the process to fewer steps: first enter the query, read the document list, looking for entity and drag them into the entity management frame.

We studied with 6 users whose ages range from 24 to 30, 3 male and 3 female, all of them possess basic computer skill. For each user, we asked them to accomplish tasks in both ways. We spent two minutes to introduce how to use Minerva II. Then we asked her to accomplish six tasks. The tasks include finding a certain entity and the similar entities, finding a seller's products, etc. In the end, he rates the usefulness, ease of use, and confidence of his accurate rate in a 1-10 scale. For the traditional method, we provide her with Google and excel, and ask her to accomplish the same tasks and rating. Three of the users used Minerva II first, and three of them uses the traditional method first.

We recorded both the time and the number of operations that were used to accomplish a task. The result shows that by using Minerva II, the time cost is reduced by 62 percents (Figure 4), and the number of operations is reduced by 47 percents (Figure 5). As Figure 6 shows, Minerva II also receives higher ratings than the traditional method.

Conclusion

This paper presents Minerva II, a novel entity discovery tool with an exclusively designed interface. Minerva II provides users with entity search, entity management and network visualization, which significantly simplifies investigation process and saves time for law enforcement agencies and other interested parties.

Acknowledgment

The research is supported by DARPA FA8750-14-2-0226, NSF IIS-1453721, and NSF CNS-1223825. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor.

REFERENCES

1. Krisztian Balog, Edgar Meij, and Maarten de Rijke. 2010. Entity search: building bridges between two worlds. In *Proceedings of the 3rd International Semantic Search Workshop*. ACM, 9.
2. Sheer El-Showk, Miguel F Paulos, David Poland, Slava Rychkov, David Simmons-Duffin, and Alessandro Vichi. 2012. Solving the 3D Ising model with the conformal bootstrap. *Physical Review D* 86, 2 (2012), 025022.
3. Stephen G Kobourov. 2004. Force-directed drawing algorithms. (2004).
4. Arturas Mazeika, Tomasz Tylenda, and Gerhard Weikum. 2011. Entity timelines: visual analytics and named entity evolution. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 2585–2588.
5. Yusuke Shinyama and Satoshi Sekine. 2004. Named entity discovery using comparable news articles. In *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, 848.
6. Avirup Sil and Alexander Yates. 2013. Re-ranking for joint named-entity recognition and linking. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*. ACM, 2369–2374.
7. Norases Vesdapunt, Kedar Bellare, and Nilesch Dalvi. 2014. Crowdsourcing algorithms for entity resolution. *Proceedings of the VLDB Endowment* 7, 12 (2014), 1071–1082.
8. Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. 2010. Mining advisor-advisee relationships from research publication networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 203–212.