EntityBases: Compiling, Organizing and Querying Massive Entity Repositories

Craig A. Knoblock
José Luis Ambite
Kavita Ganesan, Maria Muslea
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

Steven Minton
Greg Barish, Evan Gamble
Claude Nanjo, Kane See
Fetch Technologies
2041 Rosecrans Ave, Suite 245
El Segundo, CA 90245

Cyrus Shahabi
Ching-Chien Chen
Geosemble Technologies
2041 Rosecrans Ave, Suite 245
El Segundo, CA 90245

Abstract – *The current approaches for linking information across sources, often called record linkage, require finding common attributes between the sources and comparing the records using those attributes. This often leads to unsatisfactory results because the sources are often missing information or contain incorrect or outdated information. We are addressing this problem by developing the technology to build massive entity knowledgebases, which we call EntityBases. The key idea is to create a comprehensive knowledgebase for the entities of interest (e.g., companies). In order to build such a knowledge base, we must address the issues of linking entities with multi-valued attributes obtained from heterogeneous sources and providing a virtual repository that can be efficiently queried. This paper describes how we have addressed these issues and shows how an EntityBase™ can be used for understanding and linking text documents.*

Keywords: Record linkage, entity extraction, information integration, geospatial reasoning, EntityBases

1 Introduction

Recent advances in networking technology, especially the Internet, have made a huge amount of data available about entities, such as people, places and organizations. Even so, our ability to use the vast quantities of data on-line for identifying the references in text documents or linking information across sources remains primitive. Finding entities of interest in real-time is challenging, due to the difficulty of integrating and querying multiple databases, web sites, and document repositories.

We are developing technology that will make it possible to rapidly create large-scale, well-organized, entity knowledgebases -- which we refer to as EntityBases. EntityBases will enable information to be integrated on a scale that far exceeds current capabilities. The resulting EntityBases can then be used for a variety of applications (e.g., document understanding or data mining). The EntityBases design represents a novel approach to integrating information from numerous heterogeneous sources.

An EntityBase consolidates data, so that references to the same entity in multiple information sources can be resolved. The consolidation process can resolve references in different formats (e.g. "Joe Smith" vs. "Smith, J.E.", or "IBM" vs. "International Business Machines Corp."), represent uncertainty, accommodate aliases, and support continuous updates (so information is not lost when two consolidated references are later determined to refer to two distinct entities).

1.1 Representing EntityBases

There are several fundamental representation issues that must be addressed in building an EntityBase. First, since an EntityBase is constructed from multiple heterogeneous sources, we must be able to support multi-valued attributes for describing an entity. Second, different sources will describe these attributes at different levels of detail. Hence, an EntityBase must also be able to support the representation of the attributes at different levels of granularity. For example, one source may represent an address as a single unit, while another source may break it down into the street address, city, country, and postal code.

1.2 Organizing EntityBases

The capability to consolidate multiple references to the same individual entity collected from different information sources is a central aspect of the EntityBase architecture. Previous research has developed a foundation for statistically linking references across multiple databases, referred to variously as record linkage, consolidation or object identification (Fellegi and Sunter 1969; Winkler 1994; Tejada et al. 2001; Bilenko and Mooney 2003). The challenge here is to build this technology into a practical architecture for large-scale information repositories. We have developed an integrated EntityBase system that supports the statistical consolidation process “invisibly” as an EntityBase is populated, enables users to easily understand and analyze results, enables queries to be executed efficiently,

and is robust to updates so that references can be both consolidated as new information becomes available.

1.3 Querying EntityBases

In previous research, we and others (Duschka 1997; Ambite et al. 2001; Halevy 2001) have addressed many of the theoretical problems underlying virtual databases (i.e. mediator systems that integrate distributed, heterogeneous sources). Nevertheless, building large-scale virtual databases remains challenging in practice because it is difficult to model complex data relationships and potentially expensive to execute arbitrary queries against virtual databases. Our goal is to address these specific problems. By focusing only on entities, our architecture simplifies the modeling issues and improves the tractability of query processing.

1.4 Exploiting Geospatial Context

Most entities in the world are associated with some geospatial location. This location could be a point or even a region and we refer to the associated location as the geospatial extent of an entity. We automatically determine the geospatial extent of an entity and then use this extent as an additional source of information for linking new sources of information into the system. For example, if a new record is added to the system and the area code of the phone number indicates the record is in a particular region, then it would be less likely to match against an entity located in a different region.

In the remainder of this paper we present a motivating example, describe our approach to representing, organizing, and querying an EntityBase, and present our approach to exploiting the geospatial context.


U.S. Immigration and Customs Enforcement

ICE

ABOUT NEWS ROOM BORDER SECURITY FINANCIAL & TRADE INVESTIGATIONS PUBLIC SECURITY CAREERS

News Release

August 11, 2002

DUBAI TRADER AND IRANIAN COMPANY OFFICER INDICTED FOR SCHEME TO ILLEGALLY SHIP U.S. GOODS TO IRAN

Washington, D.C. - United States Attorney Kenneth L. Wainstein, Department of Homeland Security Assistant Secretary for Immigration and Customs Enforcement (ICE) Michael J. Garcia, U.S. Department of Commerce Undersecretary for Industry and Security Kenneth I. Juster, and jointly announced that a federal grand jury in the District of Columbia has returned an indictment charging Khalid Mahmood, also known as Khalid Mahmood Chaudhary, 52, of Dubai, United Arab Emirates, and Mohammad Ali Sherbaf, age unknown, of Iran, with violations of the International Economic Powers Act, the Iranian Transactions Regulations, and the Export Administration Regulations.

...

The Indictment alleges that Mahmood was doing business as Sharp Line Trading with offices in the United Arab Emirates, and Sherbaf was a principal officer of Sepahan Lifter Company, a forklift manufacturing firm located in Iran. According to the Indictment, in early June 2004, an employee of Sepahan Lifter Company contacted a United States company by email, and requested a price quotation for particular radiators for heavy-duty 5-ton capacity forklift trucks manufactured in Iran by

February 24, 2006

LEXINGTON, KENTUCKY MAN SENTENCED TO 39 MONTHS IN PRISON FOR VIOLATING TRADE EMBARGO AGAINST IRAN

Washington, D.C. - United States Attorney Kenneth L. Wainstein, Darryl W. Jackson, United States Department of Commerce Assistant Secretary for Export Enforcement, and Mark Gerrand, Acting Special Agent-in-Charge for U.S. Immigration and Customs Enforcement (ICE), Department of Homeland Security, announced yesterday that the Honorable John D. Bates sentenced Robert E. Quinn, 54, of Lexington, Kentucky, to 39 months of incarceration and a fine of $6,000. Quinn was found guilty by a federal jury on December 7, 2005, of one count of conspiring to violate the U.S. trade embargo against Iran and five counts of illegal exports to Iran.

...

In October 2005, a federal grand jury in the District of Columbia returned a six-count superseding indictment against Quinn, 54, and Michael H. Holland, also of Lexington, Kentucky, and Mohammed A. Sharbaf, of Iran, charging them with violating the United States embargo on trade with Iran. Quinn and Holland were sales executives employed by Clark Material Handling Corporation ("CMHC"), a Kentucky-based forklift truck manufacturer. Sharbaf is President and Managing Director of Sepahan Lifter, a forklift truck manufacturer in Esfahan, Iran.

Figure 1: Different documents with the same real-world entities appearing with different names

2 Motivating Example

Consider the following real-world example of the need for EntityBases. Figure 1 shows extracts of two news releases from the U.S. Immigration and Customs Enforcement, an agency of the U.S Department of Homeland Security. The documents describe a case involving several individuals and companies accused of illegal exports to Iran.

Entity extraction software, like Inxight's SmartDiscovery system, can automatically extract entities, such as company, person or location names, from these documents. For example, "Mohammad Ali Sherbaf", "Kenneth L. Wainstein", and "Khalid Mahmood Chaudhary", etc., would be recognized as person names. Similarly, "Sepahan Lifter Company", "Sharp Line Trading", and "Clark Material Handling Corporation" would be labeled as companies, and "Esfahan" as a city and "Iran" as a country.

However, simple entity extraction is not enough. The relationship between these two documents cannot be established without the kind of record linkage reasoning that our EntityBase provides. In particular, note that the 2002 document refers to one of the key persons involved in the case as "Mohammad Ali Sherbaf" while the 2006 document as "Mohammad A. Sharbaf" (even though the documents come from the same government agency). Different transliterations of foreign names, such as in this case, would foil simple match techniques. The multiplicity of names that refer to the same real-world entity is not limited to people --- other entities, such as companies and locations, exhibit the same phenomenon. For example, both "Isfahan" and "Esfahan" are common transliterations for the same Iranian city.

The EntityBase approach uses previously gathered knowledge to help differentiate the entities that appear in documents like these and provide additional information. Our EntityBase can recognize that "Mohammad Ali Sherbaf" and "Mohammad A. Sharbaf" are the same person and that "Sepahan Lifter Company", "Sepahan Lifter", "Sepahan Lifter Co." refer to the same company. Moreover, the EntityBase also shows that this company has its headquarters in "Nos. 27 and 29, Malekian Alley, North Iranshahr Ave., Tehran (15847)" and its factory in "Mahyaran Indus-

trial Town, Isfahan"; that its commercial manager is "Mohammad Kharazi" and its headquarters' phone and fax numbers are (+98-21) 8830360-1 and (+98-21) 8839643, respectively. At the same time, the EntityBase shows that "Sepahan Lifter Company", "Behsazan Granite Sepahan Co.", or "Rahgostar Nakhostin Sepahan Co." are different companies that are all located in Isfahan, Iran.

3 Representing EntityBases

In order to decide how to represent entities, two fundamental issues must be addressed. The first is the multi-valued nature of entity information. For example, a company may have multiple phone numbers or multiple addresses. This complicates the issue of "record" linkage, because the data is no longer a "record" (i.e., a tuple or row in a database), but an object with multi-valued attributes. The second issue is the level of schema granularity. Normalizing information into finer levels of granularity – while seemingly more precise – is not always possible and can potentially result in a loss of information. In the following subsections, we describe our solutions to these issues.

3.1 Multi-valued attributes

In practice, many entity attributes are multi-valued. For example, most businesses have multiple phone numbers. Many people are known by multiple names (e.g., maiden name and married name). Many publications have multiple authors. This feature of real world entities requires a more general representation than the traditional records. As a more detailed example, consider Figure 2 that shows some of the multi-valued attributes of our running example.

Company name	Key person
Sepahan Lifter Company	Mohammad A. Sharbaf
Sepahan Lifter Co.	Mohammad Ali Sherbaf
Sepahan	M.A. Sherbaf

Figure 2: Entity with multi-valued attributes

We represent an entity as a set of set-valued attributes. For example, we represent the above information as {[NAME: {"Sepahan Lifter Company", "Sepahan Lifter Co.", "Sepahan"}], [KEYPERSON: {"Mohammad Ali Sherbaf", "Mohammad A. Sharbaf", … }], … }. This representation reduces the amount of data to be stored to only the unique attribute values, but requires more sophisticated matching techniques. We need to move from "record" linkage to "entity" linkage.

3.2 Level of schema granularity

Another representational issue is the level of schema granularity, an issue that arises due to the heterogeneous origins of the data and the inability to precisely parse all types of real-world information. Data in the EntityBase comes from different sources which have different schemas. For example, one source may break down address into street, city, and state while another may just have all of this data in one attribute (e.g., address). Because of this, we must decide at what level of granularity the EntityBase will normalize the information. Generally speaking, there are two possible options: fine-grain or coarse-grain.

The fine-grain option, where one might capture attributes such as street, city, state, suite number, and zip provides more information about a match (i.e., identifies the specific attribute a match occurs on), but assumes that all information can be neatly deconstructed, or that it is possible to store ambiguous information when information cannot be reliably parsed. For example, consider the sequence of tokens "Mohammad Ali Sherbaf". The fine-grain option assumes that we can parse this name, in particular that we know the first name is "Mohammad" and that the last name is either "Ali Sherbaf" or that the middle name is "Ali" and the last name is "Sherbaf".

The coarse grain option, on the other hand, eliminates the need to unambiguously parse the data. If we simply treat the sequence of tokens as just that – a sequence of tokens (i.e., a document) – we have no need to resolve ambiguous parses when storing the data. However, this does mean that we must parse the data at run-time (i.e., query time), somewhat troubling from a performance standpoint.

In EntityBases, we chose to use a hybrid approach that exploits both the coarse and fine-grained representation. The coarse-grained representation is used during the initial phase of generating candidate matches since this initial matching is based on token overlap. And then both the coarse-grained and fine-grained representations are available for reasoning in the detailed matching process. Blocking is designed to be efficient, relying on simpler (e.g., token-based) metrics in order to identify a concise, quality candidate set. In contrast, linkage is designed to focus on accuracy, performing a more careful analysis of each candidate, including evaluation of the parsed data.

4 Organizing EntityBases

An EntityBase provides two main capabilities: (1) entity matching, that is, the ability to match the relevant entities given a query, and (2) entity creation/update, that is, the ability to decide whether newly acquired information belongs to a existing entity or constitutes a new entity. Note that entity creation or update requires matching as part of its process. We now present an overview of both the match and update processes.

4.1 Matching

There are two major phases of entity matching: blocking and linkage. The purpose of the blocking phase (Michelson and Knoblock 2006) is to very quickly identify the most promising candidates from a much larger set of possible candidates. In our system, blocking relies on simple, yet efficient techniques for reducing the space of possible candidates, for example by using token-based distance metrics (Jaccard coefficient, TF-IDF, etc). The purpose of the linkage phase is to do a more detailed evaluation of the incoming record/query to the candidate entity. This is accomplished through a variety of more sophisticated trans-

formations (e.g., alignment of parsed representations of the data), which can be more accurate but require more computational resources.

Blocking and linkage complement each other, with the former focusing on performance and the latter focusing on quality. A key challenge is to ensure that the efficiency of blocking does not result in false negatives, as this would handicap the linkage phase. At the same time, blocking cannot produce too many false positives, as that would swamp the linkage phase with computational demands.

Figure 3 shows an example of the matching process. An incoming news document mentions the company "Sepahan Lifter Corp" as well as "Mohammad Sherbaf". This information can be used to query the EntityBase (which consists of millions of company entities). The blocking process efficiently identifies candidates that appear consistent with the information we know. As the example shows, many of the candidates have tokens that also appear in the query, so even applying a Jaccard-style metric (i.e., token overlap) or TF-IDF would be sufficient to yield the candidates shown.

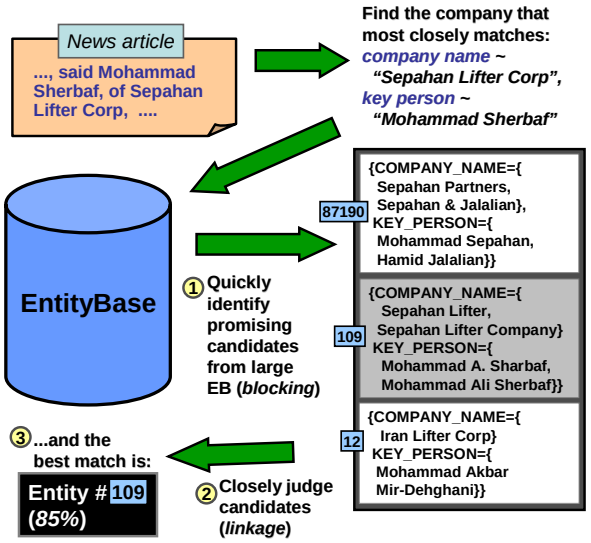

Figure 3: Entity blocking and linkage

The linkage process (Minton et al. 2005) then compares the data in greater detail, parsing the incoming query to realize that "Corp" is a previously unseen term associated with the company's name, and that "Ali" is missing from Mohammad Sherbaf's name. It also evaluates the other candidates and identifies similar differences. In evaluating the candidates, the linkage phase associates metric scores to quantify the similarity (or lack there of). A second part of the linkage process evaluates the similarities/dissimilarities and then judges the implications of such scores. For example, the linkage process could have identified that Corp is just a common company formation acronym (like "Inc." or "LLP") and that the missing "Ali" from the person's name is not critical (as opposed to a mismatch on last name, for example).

4.2 Updating

Recall that the EntityBase supports both queries and updates. Thus far, we have discussed the matching process, which plays a key role in both phases. However, during the update phase, we must also insert new data. This is not necessarily trivial, as new information may cause us to re-evaluate current entities. In particular, new data may cause two previously distinct entities to merge. Typically, the merging scenario arises when new information contains strong matches to two different entities. For example, in Figure 3 above, notice that entity #12 (Iran Lifter Corp) is different from entity #109 (Sepahan Lifter Company). However, suppose that we update the EntityBase with a new source of Iran company information and that one of those incoming records suggests that Mohammad Akbar Mir-Dehghani is a key person of Sepahan Lifter. The entity match phase would result in both entity #12 and entity #109 receiving high match confidences. At that point, the EntityBase could decide to merge those two entities together.

5 Querying EntityBases

Our EntityBase architecture provides access to the available information about entities from both local and remote sources. Even with the rapidly declining cost of storage, it is not possible to materialize all the entity information in a local warehouse due to policy, control, and security considerations. In addition data may be too volatile to store and must be accessed live when queried, such as the changing stock price of a company. Therefore, our EntityBase is organized as a virtual repository that integrates both locally materialized data and remote data.

Figure 4 shows the EntityBase query architecture. The Local Entity Repository (LER) materializes the identifying attributes of the entities in order to perform record linkage reasoning efficiently. The concept of an entity-identifying attribute is broader than the concept of a key in relational databases. For example, the name, address and phone number are useful entity-identifying attributes, but none of them are keys.

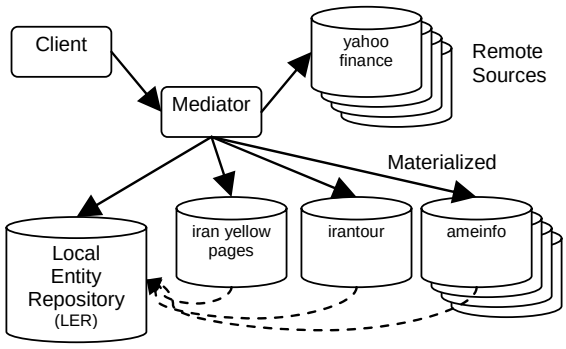

Figure 4: EntityBase data integration architecture

Additional information about entities can also be materialized, but it is not copied into LER for performance. For example, images and reports may be associated with enti-

ties, but they would not be useful for efficient record linkage. Finally, other information may reside in remote sources.

The EntityBase uses a mediator to orchestrate all these local and remote sources. The crucial idea in a mediator system is to use a mediated schema to assign common semantics to the data from the diverse sources. A human analyst, or a client program, queries the entity base using the mediated schema without worrying how the information is represented in the sources. We have built upon our previous work in the Prometheus mediator (Thakkar et al. 2005) to define our mediator for the EntityBase. In our mediator the contents of sources are declaratively expressed as Datalog rules.

The EntityBase supports two query scenarios:

- Free-Form Querying: A human analyst or a client program can pose arbitrary queries over the mediated schema to the EntityBase.
- Document Matching: Some partial entity information is extracted from a document and additional information about the entity is required.

The EntityBase mediator handles both cases uniformly. First, the mediator invokes the entity matching module with the constraints appearing in the query. In free-form querying the constraints are the selections on entity-identifying attributes appearing in the query. In document matching, the partial entity information triggers entity matching. Next, the mediator retrieves the requested information from local or remote sources corresponding to the set of candidate entities produced by the entity matching module.

Figure 5 shows the analyst query interface to the EntityBase. This graphical interface was developed using the Heracles system (Ambite et al. 2005). The Figure shows some detailed data about the Sepahan Lifter Company including geospatial locations. Note that the company has two addresses, one in Teheran and one in Isfahan, and thus the map shows these two locations.

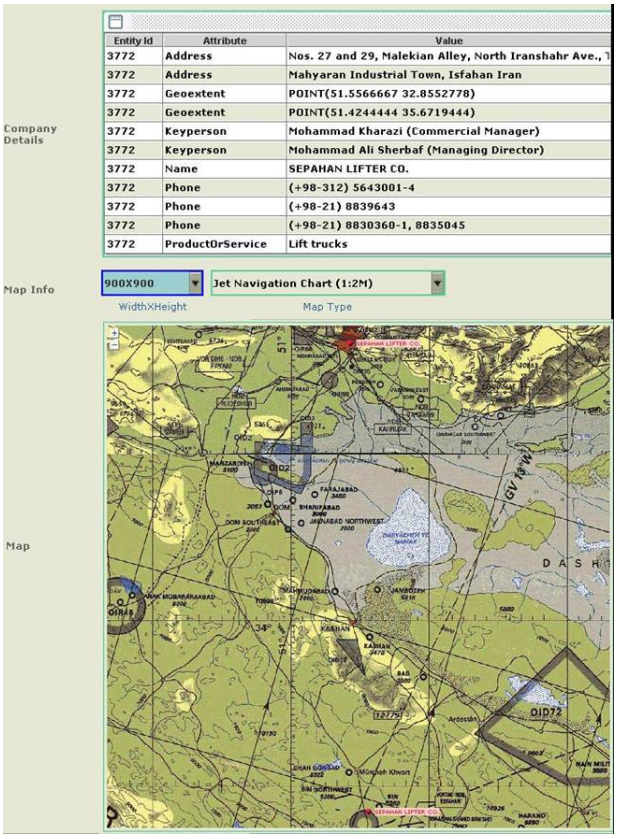

Figure 5: EntityBase analyst query interface

6 Exploiting Geospatial Context

When querying and matching entities, we can exploit the geospatial extents of the entities to help identify and assess possible matches. Consider our running example shown in Figure 3, where a document understanding system queries the EntityBase to match the company "Sepahan Lifter Corp" and the person "Mohammad Sherbaf," which are extracted from a document. The EntityBase may provide many candidates (e.g., entity #12, #109 and #87190) even after applying the blocking process. However, by exploiting the geospatial extents of these entities, the system can deduce additional information to narrow down the relevant entities. For example, the company mentioned in the document is located within area X shown in Figure 6. The system infers that entity #109 is located within area X as well (based on its "phone" attribute). It also infers that the companies of the entities #12 and #87190 are located within area Y (based on their associated "phone" attribute). This in turn implies lower similarity between the company mentioned in the document and the two entities #12 and #87190.

Unfortunately, identifying the exact geospatial extent of an entity is not straightforward. Essentially, we need to transform a record's textual geographic information (e.g., mailing addresses) to spatial extents (e.g., geocoordinates). A straightforward way to compute geocoordinates of a company is using a geocoder with the mailing address as input (Bakshi et al. 2004). Typically, a geocoder determines the geocoordinates of an address by utilizing a comprehensive spatial database (e.g., labeled road network data). However, such a comprehensive, well-formatted spatial database does not exist or is not accessible for many countries. Additionally, addresses are non-standard (e.g., "No. 1780, Opp.to The Main Gate of England Embassy Garden, Off the Dolat St., Shariati Ave., Tehran, Iran"), incomplete, and sometimes do not exist for a given record (e.g., only the phone number exists).

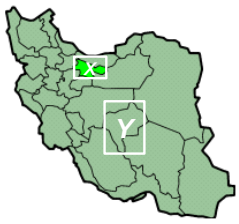

Figure 6: Geocoordinates of the companies in Iran

Toward this end, we utilize various techniques to build a geospatial knowledgebase of an area from available public data. The

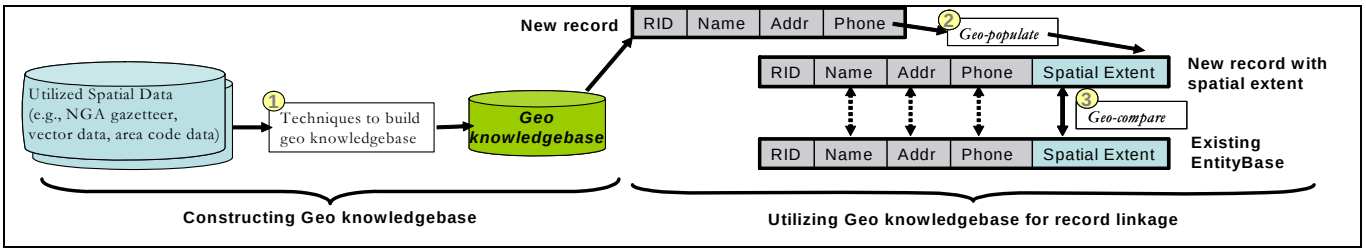

Figure 7: Exploiting geospatial data for entity linkage

geospatial knowledgebase contains abundant (inferred) spatial datasets, such as landmarks, road network data, zip code maps, and area code maps. To illustrate, consider the scenario that the area code data for Iran is not available. We can use the technique proposed in (Sharifzadeh et al. 2003) to approximate the area code regions and store them in the geospatial knowledgebase. This technique utilizes classification techniques (such as Support Vector Machines) based on a set of training data to build approximate thematic maps (e.g., area code maps). For example, the training data can be cities with spatial coordinates and telephone area code attributes. Spatial classification of the training data (geocoordinates labeled with phone area code) produces an approximate thematic map of the phone area code regions.

There are many online public data sources, such as the NGA gazetteer database[1] that can provide the labeled training points. Again, consider building area code maps for the country of Iran as in our case study. We apply the technique for building thematic maps to three datasets we collected for Iran: (1) Iran area codes and corresponding cities, which are available from IranAtom[2], (2) NGA gazetteer database that provides the coordinates of populated points (including cities) around the world, and (3) Iran province information[3] that provides the spatial bounding box for every province in Iran. Finally, we store the approximate area code vector maps into the geospatial knowledgebase in Oracle 10g, because Oracle fully supports spatial data types and queries, as well as R-tree index.

To utilize the geospatial knowledgebase for comparing two entities based on their geocoordinates, the system needs to assign the best spatial extent possible to a new incoming record or query. It also needs to support the efficient comparisons between two entities based on their assigned spatial extents. To achieve this, we developed two functions Geo-populate and Geo-compare. Consider populating the EntityBases. Geo-populate analyzes the phone number of a given new record to obtain its area code; it then queries the geospatial knowledgebase with the given area code to discover the spatial extents (a point or a region) for the record (the second step in Figure 7). Geo-compare then utilizes Oracle spatial APIs to compute how close is the new record to the records stored in the EntityBases based on their spatial extents (the third step in Figure 7). Figure 7 summarizes the framework for utilizing geospatial information for record linkage.

7 Related Work

Previous research on record linkage has developed a foundation for statistically linking references across multiple databases, referred to variously as record linkage, consolidation or object identification (Fellegi et al. 1969; Winkler 1994; Goldberg et al. 1995; Tejada et al. 2001). While similar to traditional record linkage in the goal of identifying consolidating objects, the EntityBase design contains some key differences. In particular, the matching (linkage) process of EntityBases that we describe here is distinctly different than it is for traditional record linkage. In traditional record linkage, the goal is to identify the unique objects between two sources, where the objects are generally encapsulated as rows. In contrast, an EntityBase is composed of multiple sources, where each entity can have multiple attribute values (e.g., multiple names). Unlike record linkage, EntityBases keeps track of all object aliases and views the entire data collection as part of the entity.

A critical focus of the EntityBases work is on scale. We are attempting to build EntityBases of millions of entities, which can be queried and updated tens to hundreds of times per second. There is some past work on parallel record linkage (Christen et al. 2004) and blocking techniques (Baxter et al. 2003). However, these systems assume that the sources to be consolidated are tables in a relational database and do not address the issue of multi-valued attributes, which can have a serious performance impact. Furthermore, they do not consider the issues of entity merging and splitting, as we need to do in EntityBases.

Our data integration approach builds upon our previous work on the Prometheus mediator (Thakkar et al. 2005). Prometheus supports both the global-as-view and local-as-view approaches to data integration (Halevy 2001). For local-as-view integration, Prometheus follows the Inverse Rules algorithm (Duschka 1997) with additional optimizations (Thakkar et al. 2005). We use the mediator to assign common semantics to the data coming from the sources and map such data into the mediated schema of the EntityBase. Our mediation approach introduces two novel techniques. First, the mediator integration program is declaratively expressed in datalog including predicates that call the re-

[1] http://earth-info.nga.mil/gns/html/
[2] http://irantom.ru
[3] http://www.iranembassy.hu/province.htm

cord linkage routines when a new source record is processed by the EntityBase. Second, the mediated schema is designed to accommodate the multivalued nature of attributes in the EntityBase.

There are many studies related to exploiting geospatial context to infer geospatial knowledge (Bruin 2000; Krieger et al. 2001; Sharifzadeh et al. 2003). For example, (Sharifzadeh et al. 2003) proposed a technique to construct thematic maps for non-natural features (e.g., zip code maps or area code maps). In general, we are not aware of previous work on exploiting these geospatial extents for matching entities.

8 Conclusion

In this paper we have described our approach to building massive entity repositories. In particular, we have described the representation of the entities, the approach to linking new data into an EntityBase, a framework for querying the large amount of data potentially available on an entity, and a technique for exploiting the geospatial context to improve the linking of entities.

One can build an EntityBase for just about any type of entity, including people, organizations, companies, terrorist groups, and so on. These knowledge bases of entities can then be used for a wide variety of applications. In this paper we described how an EntityBase can be used for associating and linking text documents with the actual entities that are mentioned in the documents. An EntityBase could also be used to process data in a database or to reason about the relationships between entities (such as finding all organizations that are located in the same region and are mentioned in the same document).

In future work, we plan to further develop the techniques and algorithms described in this paper. In particular, we plan to improve the accuracy of the linking process, extend the techniques for rapidly incorporating new sources of data, and exploit additional types of geospatial information. We also plan to refine the existing system to support the efficient construction and querying of EntityBases of millions of entities.

Acknowledgements

This research was sponsored by the Air Force Research Laboratory, Air Force Materiel Command, USAF, under Contract number FA8750-05-C-0116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

References

Ambite, J. L., C. A. Knoblock, I. Muslea and A. Philpot (2001). "Compiling Source Descriptions for Efficient and Flexible Information Integration." Journal of Intelligent Information Systems **16**(2): 149--187.

Ambite, J. L., C. A. Knoblock, M. Muslea and S. Minton (2005). "Conditional Constraint Networks for Interleaved Planning and Information Gathering." IEEE Intelligent Systems **20**(2).

Bakshi, R., C. A. Knoblock and S. Thakkar (2004). Exploiting online sources to accurately geocode addresses. The 12th ACM International Symposium on Advances in Geographic Information Systems (ACMGIS'04), Washington, D.C.

Baxter, R., P. Christen and T. Churches (2003). A Comparison of fast blocking methods for record linkage. ACM SIGKDD'03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, Washington, DC.

Bilenko, M. and R. J. Mooney (2003). Adaptive Duplicate Detection Using Learnable String Similarity Measures. The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003).

Bruin, S. D. (2000). "Predicting the areal extent of land-cover types using classified imagery and geostatistics." Remote Sensing of Environment **74**(3): 387–396.

Christen, P., T. Churches and M. Hegland (2004). A Parallel Open Source Data Linkage System. The 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04), Sydney, Australia.

Duschka, O. M. (1997). Query Planning and Optimization in Information Integration. Department of Computer Science, Stanford University.

Fellegi, I. P. and A. B. Sunter (1969). "A theory for record-linkage." Journal of the American Statistical Association **64**: 1183-1210.

Halevy, A. Y. (2001). "Answering queries using views: A survey." The VLDB Journal **10**(4): 270-294.

Krieger, N., P. Waterman, K. Lemieux, S. Zierler and J. Hogan (2001). "On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research." American journal of public health **91**(7): 1114-1116.

Michelson, M. and C. A. Knoblock (2006). Learning Blocking Schemes for Record Linkage. The 21st National Conference on Artificial Intelligence (AAAI-06), Boston, MA.

Minton, S. N., C. Nanjo, C. A. Knoblock, M. Michalowski and M. Michelson (2005). A Heterogeneous Field Matching Method for Record Linkage. The 5th IEEE International Conference on Data Mining (ICDM-05), Houston, TX.

Sharifzadeh, M., C. Shahabi and C. A. Knoblock (2003). Learning Approximate Thematic Maps from Labeled Geospatial Data. the International Workshop on Next Generation Geospatial Information, Cambridge (Boston), Massachusetts, USA.

Tejada, S., C. A. Knoblock and S. Minton (2001). "Learning Object Identification Rules for Information Integration." Information Systems **26**(8).

Thakkar, S., J. L. Ambite and C. A. Knoblock (2005). "Composing, Optimizing, and Executing Bio-informatics Web Services." VLDB Journal, Special Issue on Data Management, Analysis and Mining for Life Sciences **14**(3): 330-353.

Winkler, W. E. (1994). Advanced Methods for Record Linkage. Proceedings of the Section of Survey Research Methods of the American Statistical Association**:** 467--472.