Scalable Spatio-temporal Knowledge Harvesting

Yafang Wang Bin Yang Spyros Zoupanos Marc Spaniol Gerhard Weikum
Max-Planck-Institut für Informatik, Saarbrücken, Germany
{ywang, byang, zoupanos, mspaniol, weikum}@mpi-inf.mpg.de

ABSTRACT

Knowledge harvesting enables the automated construction of large knowledge bases. In this work, we made a first attempt to harvest spatio-temporal knowledge from news archives to construct trajectories of individual entities for spatio-temporal entity tracking. Our approach consists of an entity extraction and disambiguation module and a fact generation module which produce pertinent trajectory records from textual sources. The evaluation on the 20 years' New York Times news article corpus showed that our methods are effective and scalable.

Categories and Subject Descriptors

H.1.0 [**Information Systems**]: Models and Principles—*General*

General Terms

Algorithms, Experimentation, Management, Performance

Keywords

Knowledge Harvesting, Spatio-Temporal Facts, News Archive, Entity Disambiguation, MapReduce

1. INTRODUCTION

Who attended the 2008 G-20 Washington Summit? Which countries did George W. Bush visit during 2006? How many times did President Bush meet with Prime Minister Brown in 2007?

Such questions can be easily answered if *persons*' trajectories are known. We propose and develop a scalable and effective framework for harvesting the spatio-temporal knowledge from news archives. A spatio-temporal fact holds the "*person* `wasIn` *location* `at` *time*" relationship, where the *person* entity indicates a particular person, the *location* entity indicates a spatial location, and the *time* entity indicates a temporal period. For example, given the sentence "When President George H. W. Bush visited Mainz in 1989, he made a landmark appeal for a special relationship between the United States and Germany.", a spatio-temporal fact "George H. W. Bush `wasIn` Mainz `at` 1989" can be extracted. The pertinent spatio-temporal facts of individual *persons* constitute their trajectories. Based on these trajectories, more complicated analytic jobs can be conducted. For example, it is interesting to explore the relationship of *who* meets *whom*, *where*, and *when*. Furthermore, these trajectories can be used for identifying important events by capturing several politicians' trajectories intersected in the same place at a particular time.

This work distinguishes itself from existing works by considering both spatial and temporal annotations associated with the harvested facts. YAGO [4], is a simple knowledge base and T-YAGO [6] is its temporal enhancement. In the NewsStand [5] project, the spatial focus is identified and associated with a cluster of news articles. Our work associates both spatial and temporal annotations to specific entities.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. METHODOLOGY

Our spatio-temporal knowledge harvesting framework is composed of an entity extraction and disambiguation module as well as a fact generation module. In this section, we describe the methodologies that are applied in each module and how the needed trajectories are produced.

Entity extraction and disambiguation: We aim to identify *person*, *location* and *time* entities from text. *Person* and *location* entities are identified by a named entity recognizer, and *time* is recognized by regular expressions. An identified *person* or *location* may be ambiguous and may correspond to more than one entity. For example, "President Bush" may refer to "George H. W. Bush" or "George W. Bush"; and "Paris" may refer to the capital of France or to Paris in Texas, U.S., etc. We propose a multi-stage disambiguation process.

Person disambiguation: (1) Entities having only first or last name are disambiguated with the use of the full name entities. Suppose that a full name, of the form "*firstname lastname*", appears in an article or in the metadata of an article. For the same article, an entity annotated either with "*firstname*" or "*lastname*" should be disambiguated as the full name entity. If there are more than one proper full names, we take the closest one. (2) If the aforementioned procedure does not provide us with the needed entities for disambiguation, we simply use the most popular entity at the article's publication time. For example, for an article published in 1991, "President Bush" is disambiguated as "George H. W. Bush", since 1991 is in George H. W. Bush's president term, as it can be found in T-YAGO. Therefore, it is more probable that "President Bush" refers to "George H. W. Bush" than it refers to "George W. Bush".

Location disambiguation: (1) Group the identified *location* entities by text locality. (a) If these *location* entities satisfy a containment relationship, this relationship can be used for

location disambiguation. For example, if "Paris in Texas, U.S. " appears in an article, "Paris" can be disambiguated as the one in Texas, U.S.. (b) If these *location* entities do not imply possible containment relationships with each other, we use the most frequent candidate containing country of these *location*s as their containing county. For example, suppose that "Paris" and "Rouen" appear in the same paragraph. A city with the name "Paris" exists in France and in the U.S. Moreover, "Rouen" is a city of France and Barbados. Since the most frequent candidate containing country of the two cities is France, "Paris" and "Rouen" can be disambiguated as "Paris, France" and "Rouen, France" respectively. (2) The location name may change because of changes in country boundaries and regimes. For example, the city Saint Petersburg in Russia was called Leningrad in USSR. The obsolete location names and their corresponding current location names can be obtained by a specific dictionary from GeoNames [1]. Based on the publication time of an article, such location names can be disambiguated to the same *location* entity.

Fact generation: Spatio-temporal facts, in the form of (*person*, *location*, *time*), can be generated using all possible combinations of the disambiguated entities. To increase precision, we only consider the entities which appear in a pre-defined window, e.g. within a sentence. Additionally, for further fact cleaning, we have defined two pruning rules. Sometimes *location* may appear in noun phrases, meaning that it does not indicate a real spatial location. (1) If *location* appears before a *person*, we expect that some prepositions, e.g. "in" or "at", exist before the *location*. (2) If *location* appears after a *person*, we expect to find at least a verb between them. For instance, consider the sentence "Two New York advisors sent a signal to Senator Barack Obama, Democrat of Illinois in 2004." Since there is no "in" or "at" before New York, and no verbs between Barack Obama and Illinois, all generated facts will be pruned.

Execution outline: In our implementation, we used MapReduce, a framework for large-scale data-intensive computations, to execute our algorithm and to generate the trajectories from the extracted facts. The news archives are distributed and stored in different nodes. During the map phase, each article is processed by the entity extraction and disambiguation module and the fact generation module. The output of the map phase is a set of spatio-temporal facts. Articles located at different nodes, can be processed in parallel. In the reduce phase, the facts are grouped by *person* first, and sorted according to the *time*. *This way, the trajectory records of each individual persons are generated.*

3. EXPERIMENTAL RESULTS

Experimental setup: We evaluate our methods on the New York Times Annotated Corpus [3], which contains more than 1.8 million articles published between 1987 and 2007. The raw data size of the textual content is about 8 GB. All methods in this paper were implemented in Java using Sun JDK 1.6. LingPipe [2] is the named entity recognizer that we chose and Hadoop 0.21.0 is the distributed computing infrastructure that we used. All the experiments are run on a local cluster. Each node of the cluster has two Intel Xeon 2.40GHz, 4-core CPUs.

Visualization: Since two strict pruning rules are employed in fact generation module, only 79321 facts in total are extracted from the whole corpus, which indicates that the overall recall is not high. However, at the same time, the pruning rules make the survived facts have relatively high precision. We visualize part of the trajectory of George W. Bush on a map, as shown in Figure 1. The number of each marker represents the location visit order.

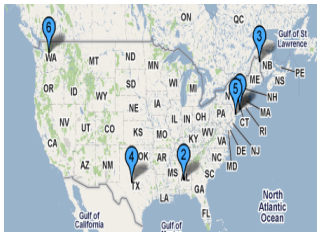

Figure 1: Visualization

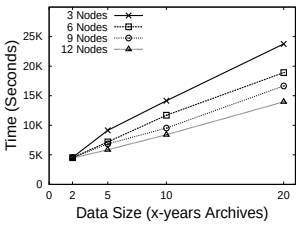

Figure 2: Scalability

Scalability: We configure different data and cluster size to test the scalability of our algorithm. We run our algorithm on each configuration twice, and report the average execution time. Note that we focus on harvesting the spatio-temporal knowledge from the whole corpus rather than querying the individual trajectories. The results, as shown in Figure 2, indicate that the larger the data size is, the more performance benefit the bigger cluster can gain, which indicate that our method is scalable.

4. CONCLUSION AND FUTURE WORK

This work harvests spatio-temporal facts from textual archives. The initial results show that our method is effective and scalable. As a future work, we plan to employ our (T-)YAGO ontology database to identify pertinent entities, including the entities that implicitly indicate *location* and *time* annotations, such as *event* and *organization* entities. More entity disambiguation and fact pruning rules will be defined to increase precision. Furthermore, more experiments with different window sizes (e.g. *n*-sentence level, paragraph level, etc) in the fact generation module will be conducted. We are also interested to apply our algorithm to other sources, like large volumes of web archived data where the MapReduce framework is really useful. Versioned Wikipedia can be used for a more sophisticated entity disambiguation.

Acknowledgements

This work is supported by the 7^{th} Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105.

5. REFERENCES

[1] GeoNames, http://www.geonames.org/.
[2] LingPipe, http://alias-i.com/lingpipe/index.html.
[3] New York Times Annotated Corpus, http://corpus.nytimes.com/.
[4] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. *J. Web Sem.*, 6(3):203–217, 2008.
[5] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In *GIS*, page 18, 2008.
[6] Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *EDBT*, pages 697–700, 2010.