Building Knowledge Graph from Public Data for Predictive Analysis - A Case Study on Predicting Technology Future in Space and Time

Weiwei Duan
Department of Computer Science
University of Southern California
weiweidu@usc.edu

Yao-Yi Chiang
Spatial Sciences Institute
University of Southern California
yaoyic@usc.edu

ABSTRACT

A domain expert can process heterogeneous data to make meaningful interpretations or predictions from the data. For example, by looking at research papers and patent records, an expert can determine the maturity of an emerging technology and predict the geographic location(s) and time (e.g., in a certain year) where and when the technology will be a success. However, this is an expert- and manual-intensive task. This paper presents an end-to-end system that integrates heterogeneous data sources into a knowledge graph in the RDF (Resource Description Framework) format using an ontology. Then the user can easily query the knowledge graph to prepare the required data for different types of predictive analysis tools. We show a case study of predicting the (geographic) center(s) of fuel cell technologies using data collected from public sources to demonstrate the feasibility of our system. The system extracts, cleanses, and augments data from public sources including research papers and patent records. Next, the system uses an ontology-based data integration method to generate knowledge graphs in the RDF format to enable users to switch quickly between machine learning models for predictive analytic tasks. We tested the system using the Support Vector Machine and Multiple Hidden Markov Models and achieved 66.7% and 83.3% accuracy on the city and year levels of spatial and temporal resolutions, respectively.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Spatial Databases and GIS*

Keywords

Data integration; Geo-temporal ontology; Machine learning; Predictive analytics

1. INTRODUCTION

Integrating heterogeneous data sources is typically the first step in any data analysis task. After the data is mapped into a unified representation, the next step is to generate the required metadata or feature vectors for different types of analytic tools. In this paper, we present a semi-automatic approach that streamlines the data analysis workflow from heterogeneous data sources to analytic tasks. The approach integrates data of various types from a number of sources into knowledge graphs so that the user can use the SPARQL language to easily generate feature vectors for different prediction methods. The main contribution of our approach is that the user can quickly add a new data source and switch data analysis methods to test different hypotheses. By allowing the user to switch different prediction methods efficiently, the system enables low-cost data exploration and stimulates new ideas for data analysis.

We demonstrate our approach using a case study on fuel cell related technologies in which we predict the future center(s) of fuel cell technologies in both the geospatial and temporal dimensions using two types of predictive models. The case study includes an end-to-end system as an implementation of our approach, which takes in publicly available data, models the data using an ontology to generate knowledge graphs in the RDF (Resource Description Framework) format, and uses the knowledge graphs for predictive analytic tasks. The system generates the prediction results using the spatiotemporal resolution of city and year level granularity. The definition of a "technology center" can vary depending on the available data sources and the user requirements. For example, a technology center can be a manufacturer cluster, a distributing center, or a research center. The case study described in this paper does not distinguish between these centers.

Predicting the evolvement of a technology is an important but typically manual and expert-intensive task [2]. This type of prediction task can be summarized as by the solicitation of the IARPA (Intelligence Advanced Research Projects Activity) FUSE program (Foresight and Understanding from Scientific Exposition), which seeks to develop automated methods "that aid in the systematic, continuous, and comprehensive assessment of technical emergence using information found in published scientific, technical, and patent literature" [2]. In addition to the intelligence community, for investors, accurate prediction of technology evolvement in time and space can help them choose profitable geographic locations or filter out irrelevant locations for maximizing their investment returns. For professionals, the results can help them to choose the places (e.g., a city) that have more job opportunities matching their background.

In the remainder of this paper, Section 2 uses a case study on the fuel cell technologies to present the overall system for the predictions of future technology centers, Section 3 presents the related work, and Section 4 discusses the future work.

BigSpatial_16, October 31-November 03 2016, Burlingame, CA, USA © 2016 ACM.
ISBN978-1-4503-4581-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/3006386.3006388

2. CASE STUDY: PREDICT THE NEXT FUEL CELL TECHNOLOGY CENTERS

This section presents our system architecture and uses a case study to show how our system works in practice with real data. In the case study, we used our system to predict the future centers of fuel cell technologies in 2014 using the 2008 - 2013 data collected from IEEE (Institute of Electrical and Electronics Engineers) and USPTO (United States Patent and Trademark Office) as well as geospatial data from OpenStreetMap and GeoNames. We verified the prediction results using industry reports.[1]

Our overall system includes four major components (Figure 1). The first component collects data from online sources and cleans the data, the second component models the data with an ontology to generate a knowledge graph in RDF, the third component queries the RDF graph to obtain the features for training machine learning models for the predictive analytic tasks, and the last component performs the predictive analytic tasks.

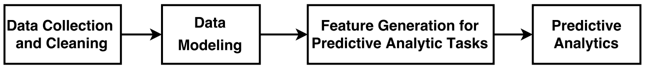
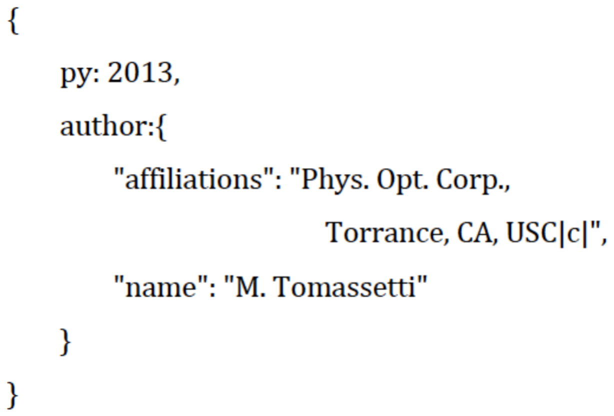

Data Collection and Cleaning
Data Modeling
Feature Generation for Predictive Analytic Tasks
Predictive Analytics

Figure 1. Overall system architecture

2.1 Data Collection, Cleaning, and Metadata Augmentation

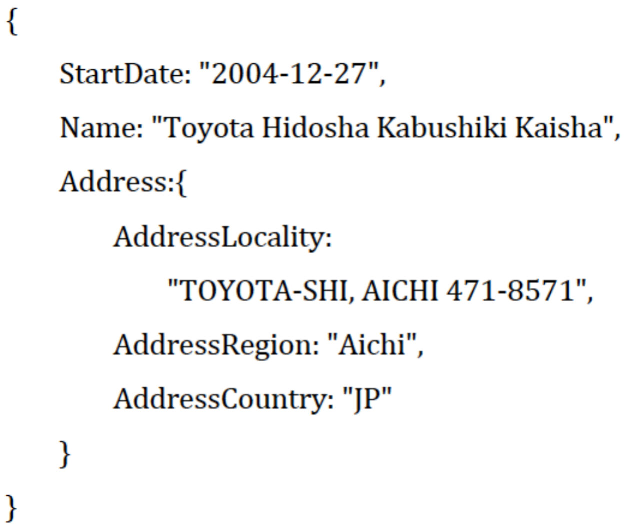
The data sources we used in the case study were research papers from the IEEE API (from 2005 to 2015), patent records from the USPTO website (from 2008 to 2014), administrative boundaries from OpenStreetMap (OSM), city populations, names, and their name variations (in different language) from GeoNames. After collecting the dataset from each of the sources, the system cleaned the data (e.g., remove ill-formatted address and affiliation information) and extracted metadata from the cleaned data. Figures 2 and 3 shows an instance of the metadata of the raw datasets of the IEEE research papers and USPTO patents, respectively. Tables 1 and 2 show the descriptions of their metadata. A portion of the collected data did not contain temporal and geospatial information, such as the location of the author affiliation, and these data were discarded. After data cleaning, there were 15,062 paper instances (IEEE) and 16,738 patent instances (USPTO).

The system also used the OSM and GeoNames data to infer the geospatial relationships and properties the IEEE and USPTO records to add new metadata. The system used the OSM administrative boundary to generate the adjacent city list for the cities where the paper authors and patent assignees located automatically. Figure 4 shows an instance in the adjacent city list. Table 3 is the description of the keys in Figure 4. GeoNames provides a gazetteer that lists the cities whose population is over 15,000. In the data modeling step, the system searched the city name in the gazetteer from GeoNames, and if the population of a city was over 15,000, the system labeled the city as a “big city”. This “big city” is an example of a user-defined property that can be useful in the final predictive analytic step. Figures 5 and 6 shows the metadata of IEEE research papers and USPTO patents after data cleaning and metadata augmentation, respectively. Tables 1 and 2 shows the description of the metadata. All the raw, cleaned, and augmented data were stored in ElasticSearch for efficient access with scalability.



```
{
    py: 2013,
    author:{
        "affiliations": "Phys. Opt. Corp.,
                        Torrance, CA, USC|c|",
        "name": "M. Tomassetti"
    }
}
```

Figure 2. An instance of the metadata in the raw IEEE data



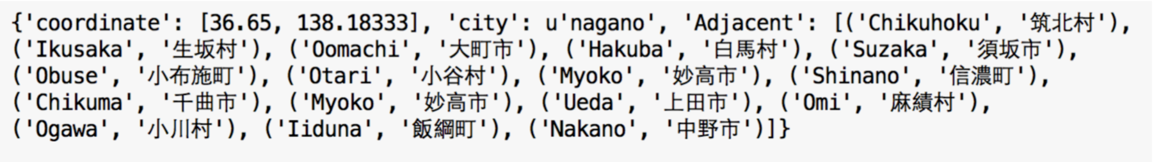
```
{
    StartDate: "2004-12-27",
    Name: "Toyota Hidosha Kabushiki Kaisha",
    Address:{
        AddressLocality:
            "TOYOTA-SHI, AICHI 471-8571",
        AddressRegion: "Aichi",
        AddressCountry: "JP"
    }
}
```

Figure 3. An instance of the metadata in the raw USTPO data

Table 1. Key description about IEEE data

Raw Data	
Key	**Description**
py	The publication year of the paper
author-affiliations	The affiliation of the author
author - name	The name of the author
Cleaned and Augmented Data	
py	The publication year of the paper
Author’s Address- City	The city where the author’s affiliation locates
Author’s Address-Country	The city where the author’s affiliation locates
AdjacentCity	The name of the cities adjacent to the location of the author affiliation

[1] A demonstration video of this case study can be accessed on: http://spatial-computing.github.io/video/TechPredictSys.mov



```
{'coordinate': [36.65, 138.18333], 'city': u'nagano', 'Adjacent': [('Chikuhoku', '筑北村'),
('Ikusaka', '生坂村'), ('Oomachi', '大町市'), ('Hakuba', '白馬村'), ('Suzaka', '須坂市'),
('Obuse', '小布施町'), ('Otari', '小谷村'), ('Myoko', '妙高市'), ('Shinano', '信濃町'),
('Chikuma', '千曲市'), ('Myoko', '妙高市'), ('Ueda', '上田市'), ('Omi', '麻績村'),
('Ogawa', '小川村'), ('Iiduna', '飯綱町'), ('Nakano', '中野市')]}
```

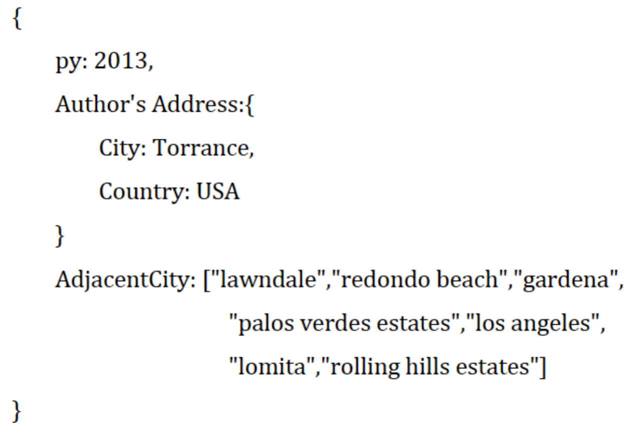
Figure 4. An adjacent city instance

Table 2. Key description about USPTO data

Raw Data	
Key	**Description**
StartDate	The date when the patent was assigned to the assignee
Name	The name of the assignee
Address	The address of the assignee
AddressRegion	The region (state level) where the assignee locates
AddressCountry	The country that where assignee locates
Cleaned and Augmented Data	
StartDate	The date when the patent was assigned to the assignee
EndDate	The date when the patent was assigned to another assignee
Author's Address- City	The city where the assignee locates
Author's Address-Country	The country where the assignee locates
AdjacentCity	The name of the cities adjacent to the location of the assignee

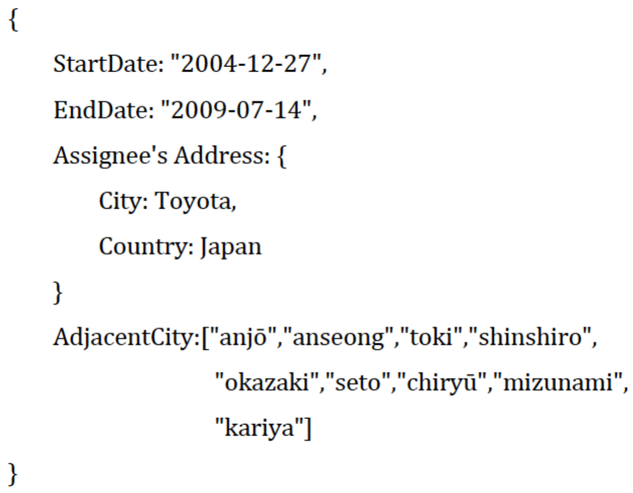
Table 3. The description of the dictionary in Figure 4

Key	**Description**
coordinate	The longitude and latitude of the city
city	The name of the city
Adjacent	All cities that are adjacent to the city stored in a list of tuples. The first element in the tuple is the city name in English, and the second element is the city name in the local language.



```
{
    py: 2013,
    Author's Address:{
        City: Torrance,
        Country: USA
    }
    AdjacentCity: ["lawndale","redondo beach","gardena",
                   "palos verdes estates","los angeles",
                   "lomita","rolling hills estates"]
}
```

Figure 5. An instance of the metadata in the IEEE data after data cleaning and metadata augmentation



```
{
    StartDate: "2004-12-27",
    EndDate: "2009-07-14",
    Assignee's Address: {
        City: Toyota,
        Country: Japan
    }
    AdjacentCity:["anjō","anseong","toki","shinshiro",
                  "okazaki","seto","chiryū","mizunami",
                  "kariya"]
}
```

Figure 6. An instance of the metadata in the USTPO data after data cleaning and metadata augmentation

2.2 Data Modeling

The goal of this component is to generate a RDF graph representing the data collected from the Internet. We combined and extended two existing ontologies to create a simple ontology for representing spatiotemporal datasets. For the geospatial properties, we used the ontology in Karma [7] and added the adjacent relationship. The adjacency relationship existed between two entities if they shared the same border. For the temporal properties, we used the ontology created by Hobbs and Pan [8].

Figure 7 shows the basic ontology in our system, called geo-temporal ontology. The geo-temporal ontology contains one class, seven data properties, and one object property. An entity represents a record in the collected datasets. If the entity lasts for a period of time, we use the "interval" property to model the temporal dimension. If an entity happens at a specific time, we use the "instance" property. The "hasCity" and "hasCountry" properties indicate the administrative jurisdictions of the entity location. The "isAdjacentTo" property connects the entity and its adjacent cities. This ontology is only a preliminary design to capture the spatiotemporal relationships between data records for the case study. One important component of this ontology is that it captures the adjacency relationship between cities, which helps to model the First Law of Geography, according to Tobler, "everything is related to everything else, but near things are more related than distant things."

For the case study, we needed to distinguish between big and small cities based on the population. Hence, we added a "isBigCity" property into the basic ontology to accommodate the metadata for city sizes derived from the data sources. This also demonstrates the flexibility of using a knowledge graph for integrating new data sources that might have useful, unique metadata. Figure 8 shows the ontology we used for the case study (the updated geo-temporal ontology).

With the updated geo-temporal ontology, in the case study, we used a Python script to map the structured data collected and cleaned in the previous step to a RDF graph. The user can access the RDF graph by issuing SPARQL queries.

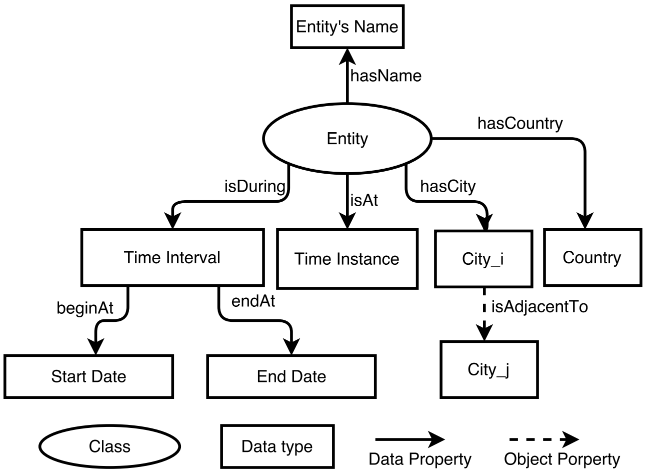


Figure 7. Geo-temporal ontology

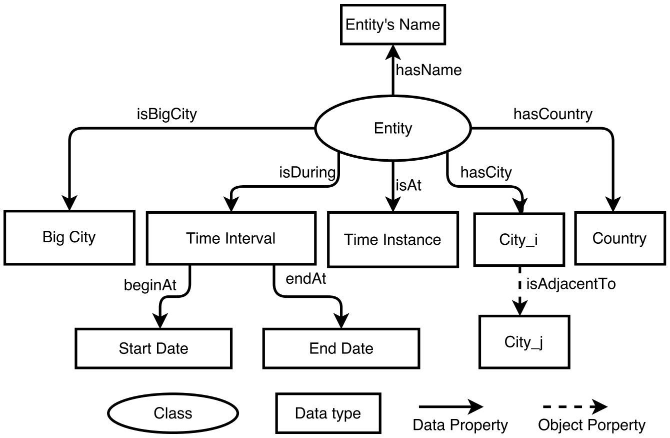
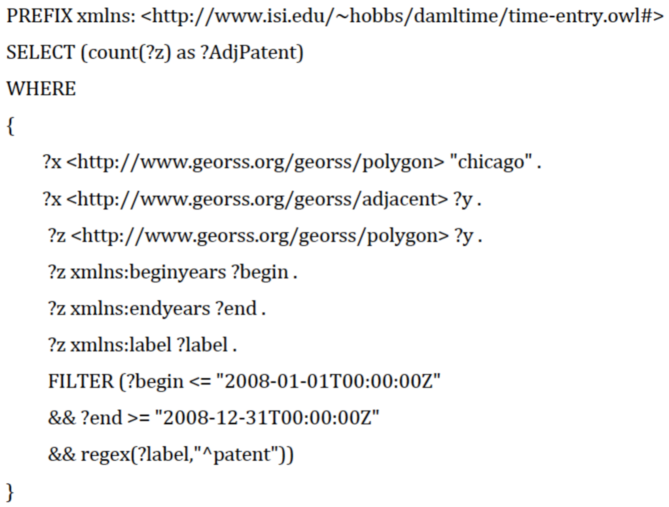


Figure 8. Updated Geo-temporal ontology for the case study

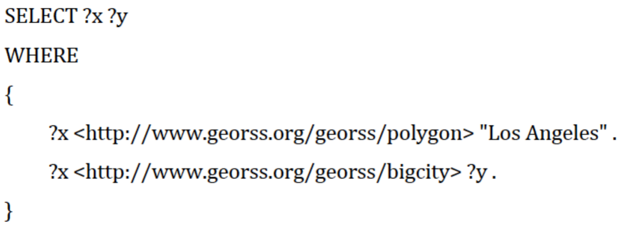
2.3 Feature Generation for Predictive Analytic Tasks

In this component, the user writes SPARQL queries against the RDF graph created in the previous step (Section 2.2) to generate features for training a predictive model. Since at this step, all data sources are already integrated into the RDF graph by the ontology, the user only has to be familiar with the ontology for writing queries to prepare data for different predictive models and does not need the knowledge of the structures of the original sources. Figure 9 shows an example of the SPARQL query used in the case study to generate the feature for training the predictive models (Section 2.3). In Figure 9, the SPARQL returns the total number of patents in Toyota's adjacent cities in 2008. (Toyota is a city in Japan.) Figure 10 is another example of querying features for prediction analysis. The query checks the knowledge graph to see whether or not Los Angeles is a big city. The query returns 1 as a big city, 0 as not a big city.



```
PREFIX xmlns: <http://www.isi.edu/~hobbs/damltime/time-entry.owl#>
SELECT (count(?z) as ?AdjPatent)
WHERE
{
     ?x <http://www.georss.org/georss/polygon> "chicago" .
     ?x <http://www.georss.org/georss/adjacent> ?y .
      ?z <http://www.georss.org/georss/polygon> ?y .
      ?z xmlns:beginyears ?begin .
      ?z xmlns:endyears ?end .
      ?z xmlns:label ?label .
      FILTER (?begin <= "2008-01-01T00:00:00Z"
      && ?end >= "2008-12-31T00:00:00Z"
      && regex(?label,"^patent"))
}
```

Figure 9. Querying the knowledge graph for the total number of patents in the adjacent cities of the city of Toyota



```
SELECT ?x ?y
WHERE
{
      ?x <http://www.georss.org/georss/polygon> "Los Angeles" .
      ?x <http://www.georss.org/georss/bigcity> ?y .
}
```

Figure 10. Querying the knowledge graph to check if Los Angeles is a big city

2.4 Predictive Analytics

In the case study, we used two machine learning models to predict the future centers of fuel cell technologies and compared their performance. We used a non-time-series model – Support Vector Machine (SVM) [13] and a time-series model – Multiple Hidden Markov Models (Multi-HMMs) [14].

For both models, we queried the RDF graph to obtain data from 2008 to 2013 as the training data. Since we did not have the ground truth of the training data (i.e., the true fuel cell centers), we used New York Times articles between 2008 and 2013 to help us label the training data. If a New York Times article mentioned "fuel cell" and a city name, we gave the city a positive count. If a city accumulated more than three positive counts, we said this city was the technology center of fuel cell technologies (a positive example). If a city was mentioned once but had less than three positive counts, we used it as a negative example for training both predictive models.

For testing the predictive models, we manually used a series of the industry reports [9 - 12] to determine the ground truth in 2014. Our steps were as follows. In [9], the report provides the market share of companies in the fuel cell industry on the country level in 2014. We used the countries with the market share over 10% as (country level) centers for fuel cell technologies. These were three countries, Japan, Germany, and United States.

For Japan, the industry report [10] contains the number of installations of fuel cell stations in major Japanese cities in 2014. We used the Japanese cities with more than five stations as the cities of fuel cell centers in 2014. For Germany, the industry report [11] provides the geospatial distributions of the fuel cell infrastructure in Germany in 2014. We manually identified the city in the clusters of fuel cell infrastructure as the cities of fuel cell centers in 2014. For the United States, the industry report [12] shows the locations of the fuel cell companies that the U.S. Department of Energy (DOE) has funded. We selected the cities with companies having more than 4.5 million funding from the US DOE as the cities of fuel cell centers.

In sum, according to the industry reports [9 - 12], we selected a total of six instances (cities) of fuel cell centers in 2014, which were Tokyo and Nagoya in Japan, Livermore and Danbury in the US, and Stuttgart and Frankfurt in Germany. Our goal was to successfully identify these six cities using the model trained with data from 2008 to 2013.

2.4.1 Support Vector Machine Performance

Table 4 shows the features used in the SVM model. Figure 11 shows the format of the feature vector for SVM. These features were obtained by using a SPARQL query against the RDF graph. The SVM model successfully predicted four fuel cell centers with an accuracy of 66.67%. The SVM model missed Nagoya, Japan, and Frankfurt, Germany. The reason that the model did not successfully identify these two cities could be that the data sources could not directly infer the center definition in Japan and Germany. The center definition for Japanese and German cities was based on the installations of fuel cell infrastructure, and the RDF graph was about the fuel cell manufacturing and research centers (i.e., patents and research papers).

Table 4. Feature components for SVM

Feature	Description
Big City	If the city candidate has a population over 15,000, the value of this feature component is 1; otherwise, the value of is 0.
# Papers	The number of IEEE research papers produced in the city candidate each year
# Patents	The number of patents produced in the city candidate each year
#Paper of adjacent cities	The sum of the numbers of IEEE research papers produced in the adjacent cities of the city candidate each year
# Patents of adjacent cities	The sum of numbers of patents produced in the adjacent cities of the city candidate each year

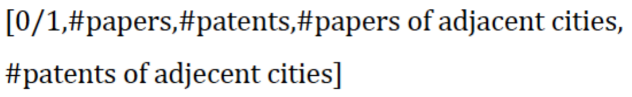

[0/1,#papers,#patents,#papers of adjacent cities, #patents of adjecent cities]

Figure 11. The format of feature vectors

2.4.2 Multiple Hidden Markov Models Performance

We also tested the Multiple Hidden Markov Models [13] to predict the future fuel cell centers. We trained four HMMs, namely, HMM$_1$, HMM$_2$, HMM$_3$, and HMM$_4$. The time component *t* was at the yearly level. The observations for HMM$_1$ were the number of IEEE research papers in a city, for HMM$_2$ the number of patents in a city, for HMM$_3$ the sum of numbers of IEEE research papers in adjacent cities, and for HMM$_4$ the sum of numbers of patents in adjacent cities. Figure 12 shows the format of the observation sequence of HMM$_2$. Figure 13 shows Toyota's (a city in Japan) observation sequence of HMM$_2$. The formats of other HMMs are similar to the format of HMM$_2$. The format of the observation sequences for HMMs are very different from the feature vectors used in SVM. By using SPARQL to query the knowledge graph, we could generate various types of feature vectors for efficiently testing different predictive models. The 0 state for HMMs meant that the city was not the fuel cell center at time *t*. The 1 state for HMMs meant that the city was the fuel cell center at time *t*. The hidden state chain and observation chain from 2008 to 2013 were put into the model to learn the transition and emission probability matrixes. To test the model, we put the observation chain from 2008 to 2014 into the model, and each model gave us the probability of being the fuel cell center in 2014. We also assumed that the if the fuel cell technology was promising in adjacent cities of a city, the city was more likely to be a center. Therefore we assigned higher weights to the HMM$_3$ and HMM$_4$. The weights of HMM$_1$ and HMM$_2$ were both 0.2. The weights of HMM$_3$ and HMM$_4$ were both 0.3. The accuracy of the Multi-HMMs was 83.3%. The model successfully predicted every center correctly but not the city of Frankfurt, Germany. Again, just like the SVM result, this error could be a result from the fact that the data sources did not fully support the center definition in our ground truth.

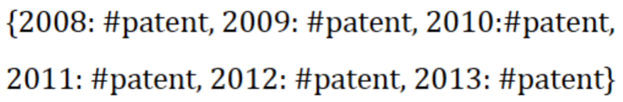
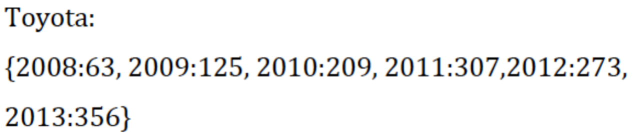

{2008: #patent, 2009: #patent, 2010:#patent, 2011: #patent, 2012: #patent, 2013: #patent}

Figure 12. The format of the observation sequence for HMM2


Toyota:

{2008:63, 2009:125, 2010:209, 2011:307,2012:273, 2013:356}

Figure 13. Toyota's observation sequence for HMM$_2$

2.4.3 Comparison of Two Models

Figure 14, 15, 16 show a map visualization of the prediction results in Japan, Germany, and the United States, respectively compared with ground truth. The spots in the maps represent the locations of the fuel cell technology centers. The performance of Multi-HMMs was better than SVM by one city. The reason could be that the temporal evolvement of the feature vector was important in the analysis and using the differences between feature vectors (i.e., the numbers of IEEE research papers and patents) at different years were more robust than the absolute value. Intuitively, if for a given year, the values of the feature component increased from the previous year, the city was more likely to be the center of that year. While if for a given year, the values of the feature components stayed the same or had small differences from the previous year, the city might not be the center of that year. From this preliminary experiment, the time-series model could be more suitable for the problem of identifying the center of fuel cell technologies.

To sum up, the case study showed that a user could easily test different predictive models and compare the results without tackling with the raw data from the sources.

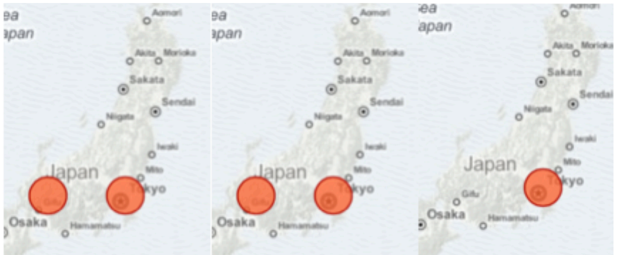


Figure 14. The fuel cell centers in Japan: from left to right, they are the ground truth, prediction result of multi-HMMs, and prediction result of SVM

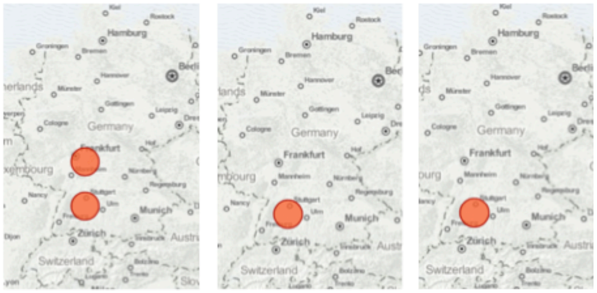


Figure 15. The fuel cell centers in Germany: from left to right, they are the ground truth, prediction result of multi-HMMs, and prediction result of SVM

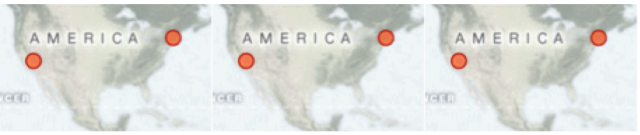


Figure 16. The fuel cell centers in the United States: from left to right, they are the ground truth, prediction result of multi-HMMs, and prediction result of SVM

3. RELATED WORK

There exist many studies demonstrating the benefit of integrating heterogeneous data sources for predictive analysis, especially in the biology domain. Myers and Troyansakaya [15] integrated diverse genomic data using a Bayesian network to capture the context-dependent reliability variation. After integrating the genomic data, the performance of the network predictions can be improved significantly. Allen et al. [16] showed that utilizing three complementary types of data would afford predictive models that outperform traditional models built using fewer data types. They demonstrated that using integrative technique on predictive toxicological studies can improve predictive power. Kim et al. [17] proposed a graph-based framework for integrating multi-omics data and genomic knowledge to improve the prediction performance of clinical outcomes based on experiments on an ovarian cancer dataset to predict the stage, grade, and survival outcomes.

Our system provides an efficient workflow for integrating heterogeneous data sources for data analysis. The user can easily switch different types of analysis methods by using SPARQL to generate feature vectors for a specific analysis method.

In the domain of predicting the future centers of a certain technology, there are a few existing studies about technology evolvement concerning both the geospatial and temporal dimensions. Leydesdorff and Rafols [6] collected data about a specific technology and detected patterns underlying the technology development. In [6], the proposed method plots data on Google Maps by using the Science Citation Index as the data source for two types of technologies. They detected the small world and preferential attachment characteristics, but do not go further to support predictive analytics.

Other studies focus on the prediction of the technology evolvement in the temporal dimension but not the geospatial dimension (e.g., [1, 4, 5]). Kim et al. [1] used Elsevier research papers and European Patent Office (EPO) patent records as their data sources to predict whether or not a technology will emerge in the future. Their work did not take the location dimension into consideration.

The Korean Institute of Science and Technology Information developed a system called *InSciTe Advanced* [4], which used Semantic Web technologies to integrate a variety of data sources to discover the technology life cycle and forecast technology maturity. The system *InSciTe* provided five major services: (1) trends and predictions, (2) technology levels, (3) relationship paths, (4) roadmaps, and (5) competitors and collaborators. Similar to the work in [1], the system *InSciTe* does not predict possible geographic locations of the technologies.

In contrast, we presented a system that integrates data from public sources and enables users to efficiently perform predictive analytic tasks on the geospatial and temporal dimensions.

4. DISCUSSION AND FUTURE WORK

This paper presented a semi-automatic approach that streamlines the data analysis workflow from heterogeneous data sources to analytic tasks. We demonstrated this system in a case study that integrates heterogeneous data sources to a knowledge graph in RDF to efficiently support predictive analytic tasks. The advantages of our system are (1) users can easily add new data sources to the knowledge graph by mapping the source to the ontology or extending the ontology if needed; (2) different prediction analysis methods can be efficiently tested on the integrated data by using SPARQL. To verify the feasibility of our system, we used a case study for predicting the future centers of the fuel cell technologies in both the geospatial and temporal dimensions. In the case study, two different types of predictive models were used. We showed that our system provided an end-to-end approach that extracted data from public sources, integrated the data using a domain independent ontology, published the integrated data as an RDF graph, and queried the graph to enable predictive analytics with two types of machine learning models.

We plan to improve the work presented in this paper in several ways. First, the current data extraction method only supports keyword search for finding relevant data given a data source (e.g., querying the patent records using “fuel cell” to find patent entries of fuel cell technologies). We plan to use the data returned by keyword search to further build a knowledge base of a certain technology and use the knowledge base to find more relevant records in the data sources. Second, we plan to add more data sources, such as news articles from Google News or New York Times or technology specific industry reports (e.g., [9,10,11,12]) to build a more comprehensive knowledge graph. Third, we plan to test the overall system with other technology domains, such as the solar power technologies.

5. ACKNOWLEDGMENTS

This work is supported in part by BAE Systems. We thank Ishan Alok, Vijayan Balasubramanian, Saarthak Khanna, Sumedha Kucherlapati, Amandeep Singh, Sanjay Singh, and Jakapun Tachaiya for their work in helping building the system.

6. REFERENCES

[1] Seonho, K., Woondong, Y., Woondong, Y., Byong-Youl, C., Waqas, R., & Jaewoo, K. (2011). A Semi-Automatic Emerging Technology Trend Classifier Using SCOPUS and PATSTAT. In *Symposium on Information and Knowledge Management*.

[2] Foresight and Understanding from Scientific Exposition (FUSE): 2013. http://www.iarpa.gov/images/press/28_iarpa_fuse_research.pdf. Accessed: 2016- 03- 10.

[3] Kim, J., Hwang, M., Jeong, D. H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. Expert Systems with Applications, 39(16), 12618-12625.

[4] Lee, M., Lee, S., Kim, J., Seo, D., Kim, P., Jung, H., ... & Sung, W. K. (2011). InSciTe Advanced: Service for Technology Opportunity Discovery. In International Semantic Web Conference–Semantic Web Challenge, Germany.

[5] Kim, J., Hwang, M., Jeong, D. H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. Expert Systems with Applications, 39(16), 12618-12625.

[6] Leydesdorff, L., & Rafols, I. (2011). Local emergence and global diffusion of research technologies: An exploration of patterns of network formation. Journal of the American Society for Information Science and Technology, 62(5), 846-860.

[7] Knoblock, C. A., Szekely, P., Ambite, J. L., Goel, A., Gupta, S., Lerman, K., ... & Mallick, P. (2012). Semi-automatically mapping structured sources into the semantic web. In Extended Semantic Web Conference (pp. 375-390). Springer Berlin Heidelberg.

[8] Hobbs, J. R., & Pan, F. (2004). An ontology of time for the semantic web. ACM Transactions on Asian Language Information Processing, 3(1), 66-85.

[9] The Fuel Cell and Hydrogen Annual Review, 2015: 2015. http://www.4thenergywave.co.uk/wp-content/plugins/datavisualisation/data/FuelCell-and-Hydrogen-Annual-Review-2015.pdf. Accessed: 2016- 04- 05.

[10] Country update Japan December 2nd, 2014 22nd IPHE SC Meeting Rome, Italy: 2014. http://www.iphe.net/docs/Meetings/SC22/ MeetingJapan_SC22.pdf. Accessed: 2016- 04- 07.

[11] The Energy Storage Market in Germany: 2014. https://www.gtai.de/GTAI/Content/EN/Invest/_SharedDocs/Downloads/GTAI/Fact-sheets/Energy-environmental/fact-sheet-energy-storage-market-germany-en.pdf?v=6. Accessed: 2016- 04- 07.

[12] Fuel Cell Technologies Market Report 2014: 2014. http://energy.gov/sites/prod/files/2015/10/f27/ fcto_2014_market_report.pdf. Accessed: 2016- 04- 07.

[13] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[14] Park, H. S., & Lee, S. W. (1996). Off-line recognition of large-set handwritten characters with multiple hidden Markov models. Pattern Recognition, 29(2), 231-244.

[15] Myers, C. L., & Troyanskaya, O. G. (2007). Context-sensitive data integration and prediction of biological networks. Bioinformatics, 23(17), 2322-2330.

[16] Allen, C. H., Koutsoukas, A., Cortés-Ciriano, I., Murrell, D. S., Malliavin, T. E., Glen, R. C., & Bender, A. (2016). Improving the prediction of organism-level toxicity through integration of chemical, protein target and cytotoxicity qHTS data. Toxicology Research, 5(3), 883-894.

[17] Kim, D., Joung, J. G., Sohn, K. A., Shin, H., Park, Y. R., Ritchie, M. D., & Kim, J. H. (2015). Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. Journal of the American Medical Informatics Association, 22(1), 109-120.