mpii max planck institut informatik

Common Sense Knowledge in Domain-Specific Knowledge Bases

Aparna S. Varde*

Joint work with: Niket Tandon, Sreyasi Nag Chowdhury and Gerhard Weikum

Max Planck Institute for Informatics (MPII), Saarbruecken, Germany

Technical Report on Research Visit in August 2015

*Aparna Varde is an Associate Professor of Computer Science at Montclair State University, NJ, USA. This report is based on the early work with MPII colleagues during her research visit in August 2015. Further work on this project is ongoing.

1. Introduction

Common sense knowledge (CSK) is so intuitive that it is often the hardest to capture and use for various real world applications. Prior work at MPII includes extracting and harnessing common sense knowledge in various contexts. This entails deriving a fact database for common sense on a Web scale, getting comparative common sense knowledge, discovering CSK from movie scripts and mining on specific activities that are CSK-based [7, 8, 9]. Notable among these is the WebChild project that involves creating a huge commonsense knowledge base from Web contents [7]. It includes a commonsense browser through which users can search information about various real-world concepts, their common properties and related terms along with illustrations.

This project propels future research in several avenues such as expanding the scope of the "activities" captured therein, exploiting functional dependencies in relations, hierarchical aspects such as ranking relations with confidence levels and the development of encyclopedic as well as SPARQL queries to extract information from the commonsense browser. It also sparks further work in areas such as machine translation, mobile devices, domain-specific knowledge bases, smart cities, domestic robotics and automated cars. We briefly describe some of these research avenues herewith. We then focus on the issue of creating and applying domain-specific knowledge bases along with potential usefulness in Smart Cities [4, 11, 12].

A domain-specific knowledge base (KB) serves the purpose of storing fundamental knowledge of a specific domain pertaining to basic concepts as well as the instantiation of those concepts into various entities, their properties and interactions with each other. Such a KB is useful in getting at-a-glance information about the domain; query processing, data mining and knowledge discovery within the domain; and the development of expert systems, intelligent tutors and other AI-based software. Common sense knowledge can be useful in the creation of domain-specific KBs and furthermore in the application of those KBs for social media mining and potentially some aspects of Smart Cities such as Smart Governance [11]. We thus describe some directions in domain-specific KB development and use. In particular, we consider the development of a KB in Urban Planning and Simulation using commonsense knowledge as a backbone. This would be useful to mine relevant data in Urban Planning and would also help to enhance existing work in the general area of data mining in domain specific applications [1, 3, 5, 10]. It would make contributions to knowledge discovery in Environmental Management as a whole [2, 6]. Such KB development would be useful in other domains as well for predictive analysis and decision support. We thus present certain issues relevant to domain-specific KB development and use based on CSK.

2. Potential Research Avenues in Common Sense Knowledge

A potential list of topics for further work in the area of Common Sense Knowledge is provided below with brief descriptions.

1. EXPANDING ACTIVITIES: The current work on WebChild considers ACTIVITY as a verb-noun pair. Some activities could potentially be conveyed through concepts other than verbs along with nouns. It would be useful to expand the notion of ACTIVITY herewith to include other terms as well, e.g., only nouns.
2. FUNCTIONAL DEPENDENCIES: It would be helpful to apply functional dependencies to common sense RELATIONS as defined in WebChild. This could reduce redundancy, provide more effective storage and make relations richer.

3. RANKING RELATIONS: Considering hierarchical aspects would be interesting, e.g., include ranking and confidence levels for RELATIONS, wherever applicable (air, oxygen etc.)
4. NUMERIC VALUES: It might be debatable to address the issue of RELATIONS potentially having numeric values, e.g., weight, size etc. Currently all values are categorical. This might need a domain-specific parser, also it might fall beyond the scope of common sense knowledge, since getting that specific may not be the best way to describe concepts in common parlance, e.g., a truck is big sounds more like common sense than giving its precise length, width and height estimates.
5. QUERY PROCESSING: Since WebChild provides such a huge common sense knowledge base, it would be very good to develop encyclopedic queries and SPARQL queries to extract information from the WebChild browser.
6. MACHINE TRANSLATION: We could adequately use CSK with WebChild in Machine Translation applications, e.g., the classic HAMLET quote: "The spirit is willing but the flesh is weak" translates to Russian as "The Vodka is good but the meat is rotten". CSK could certainly help here.
7. MOBILE DEVICES: Smartphone technologies could potentially benefit from common sense knowledge, e.g., SIRI searches in iPhones could yield better results if SIRI had more common sense.
8. DOMAIN-SPECIFIC KB: It would be extremely advantageous to create knowledge bases for specific domains using Common Sense Knowledge and WebChild. These could then be used for mining to derive further knowledge in applications such as decision support

and predictive analysis. Examples of domains that could benefit from this include Urban Planning and Simulation, Sports and Recreation, Music and Entertainment etc.

9. SMART CITIES: There is recent research in the area of smart cities that fits into the general theme of Data Mining for Social Good. This includes achieving the ultimate goal of urban sustainability where knowledge discovered by mining urban data can help users such as urban planners, environmental scientists etc. Common sense knowledge along with WebChild can be helpful here, especially in a multi-city context, since many terms used here may be in common parlance. Also, if mining from social media data, this could be useful in interpreting user posts and discovering knowledge from them. This fits into the ICT aspect of smart cities, namely, Information and Communication Technology.
10. DOMESTIC ROBOTICS: In the long run, CSK can be useful for training domestic robots to perform activities such as cooking, cleaning etc. This might fit into the highly optimistic, seemingly unreal 2020 vision of "One Robot per Household".
11. AUTOMATED CARS: It is estimated that automated cars are going to be available to the general public by 2017, so they are almost in the market. CSK can play a role in training them for recognizing events with an activity KB as pertinent to WebChild.
12. SOCIAL MEDIA: In addition to aspects such as veracity, it is also important to incorporate common sense knowledge in mining from social media sites, e.g., microblogs such as Twitter. This could have applications in product marketing, recommender systems etc. Furthermore, social media could be a rich source of information in various domain-

specific applications, thus the use of common sense knowledge in developing domain-specific KBs could be used in conjunction with knowledge discovery from social media.

3. Role of CSK in Domain-Specific Knowledge Bases

Common Sense Knowledge (CSK) defines concept classes which helps generate encyclopedic entities in a domain-specific context that could in turn be useful to capture relevant social media opinions in order to conduct domain-specific mining. We thus have the following.

a. Concept Classes: These are fundamental aspects consisting of attributes, relations and interactions as stated below.

   I. Attributes: These are properties of concepts, e.g., in Urban Planning - type of car, weight of car, amount of carbon emission etc.
   II. Relations: These relate one concept to another, e.g., one type of car v/s another
   III. Interactions: These denote the activities performed, e.g., driving a car, renting a bike etc.

b. Encyclopedic Entities: These are instances of concept classes and thus the attributes, relations and interactions can be instantiated as described below.

   i. Instantiate Attributes: These would denote the values of the properties, e.g., is a BMW, has carbon emission = 127g/km (0.45 lbs/mile) etc.

   ii. Instantiate Relations: These would involve specifics in the relations, e.g., compare a BMW with a Toyota for a car

        iii. Instantiate Interactions: These would lead to specific events, e.g., excessive road traffic, overcrowding of buildings etc.

c. Media Opinions: These capture the public reactions to specific entities and their attributes, relations and interactions and shown below.

    i. Opinions on Attributes: People could express opinions on properties and their values, e.g., “There is too much carbon emission in this city, many people wear masks outside” (Beijing)

    ii. Opinions on Relations: Opinions could entail comparisons, e.g., “I find a Toyota better than a BMW since spare parts are cheaper and easily available here” (Mumbai)

    iii. Opinions on Interactions: Social Media could reveal opinions such as “One of the best things about this area is that most buildings have 24/7 doorman service” (New York City)

4. CSK for Urban Planning and Simulation

We consider the area of Urban Planning and Simulation. Examples of concept classes here include Car, Pollutant and Apartment Complex as follows.

Example 1: Car

Attributes	Type, Production, Assembly, Body-Style, Engine, Length, Width, Height, Weight, Carbon-Emission
Relations	Bigger, Heavier, More-Emissions
Interactions	Drive, Buy, Sell, Wash, Oil-Change, Service

Example 2: Pollutant

Attributes	Diameter, AQI-Mapping, Health-Impact
Relations	Finer, Safer, More-Abundant
Interactions	Damage-Lungs, Cause-Skin-Disease, Lead-to-Cancer, Increase-Cough

Ref - AQI: Air Quality Index

Example 3: Apartment Complex

Attributes	Location, Number-of-Floors, Doorman-Hours, Laundry-Facility, Proximity-to-Public-Transport, Average-Cost-Per-Apt, Property-Tax
Relations	Taller, Safer, Cheaper, Easier-Transportation
Interactions	Buy-Apt, Sell-Apt, Rent-Apt, Get-Tenant, Pay-Tax, Regulate-Tax

These concept classes could be used to instantiate entities corresponding to encyclopedic knowledge which could then be used in mining as follows.

a. Map to specific entities, extract relevant data for mining: Consider for example that a Twitter file has millions of Tweets. Topical classification of tweets can be conducted using background knowledge. Thus, we have:

   Commonsense concept classes ➔Wiki categories ➔ Hashtags

b. Canonicalized semantic concepts: For instance, the term "Cloud" can be disambiguated. Does it refer to cloud computing for saving energy or it is the natural cloud relevant to weather? Both pertain to Urban Planning but would be different concepts in the KB.

5. CSK for Smart Cities

Recently there is much interest in the paradigm of developing and maintaining Smart Cities [4, 11, 12]. In general, a Smart City is typically defined by the following characteristics [11].

- Smart Economy: This refers mainly to competitiveness, including
  - Innovative Spirit
  - Productivity
- Smart People: This focuses on Social & Human Capital with aspects such as
  - Qualification
  - Creativity
- Smart Governance: Thus entails user participation consisting of
  - Decision-making
  - Transparent Governance
- Smart Mobility: This deals mainly with transport entailing
  - Local Accessibility
  - Sustainable & Safe Systems

- Smart Environment: This pertains to natural resources, important indicators being
    - Pollution Control
    - Sustainable Resources
- Smart Living: This determines the overall quality of life with factors such as
    - Health Conditions
    - Housing Quality

A well-developed Urban Planning KB could be useful in big context of Smart Cities. To make cities smarter, there is a need for easy access to knowledge. This fits into the general requirement of ICT (Information and Communication Technology) that applies to Smart Cities, for example, in the context of Smart Governance [4, 11]. An Urban Planning knowledge base can provide the data and metadata which is useful in aspects of planning such as decision-making. It thus seems useful in the broad scope of the Smart Governance characteristic.

Examples of Smart Cities include Amsterdam and Barcelona [12]. In Amsterdam, street lamps are such that they allow municipal councils to dim lights based on pedestrian usage. Barcelona has a bus network with smart traffic lights and buses run on routes that are designed so as to optimize the number of green lights. These pertain to the Smart Environment characteristic by conserving energy and hence providing more sustainable resource consumption as a whole. Clearly, there is a potential for enhancing such facilities to make cities even smarter. This is where a well-developed Urban Planning KB could play a role. It could be harnessed to provide relevant real-time data to monitor resources and make them more sustainable. Thus, it could potentially contribute to the Smart Environment characteristic of Smart Cities.

6. Preliminary Implementation

As an initial step of implementation, the domain-specific KB is being developed with ground truth constituting commonsense concepts pertaining to the Urban Planning domain. Figure 1 is a screenshot of the GUI from the Commonsense Browser with the Domain Specific KB.

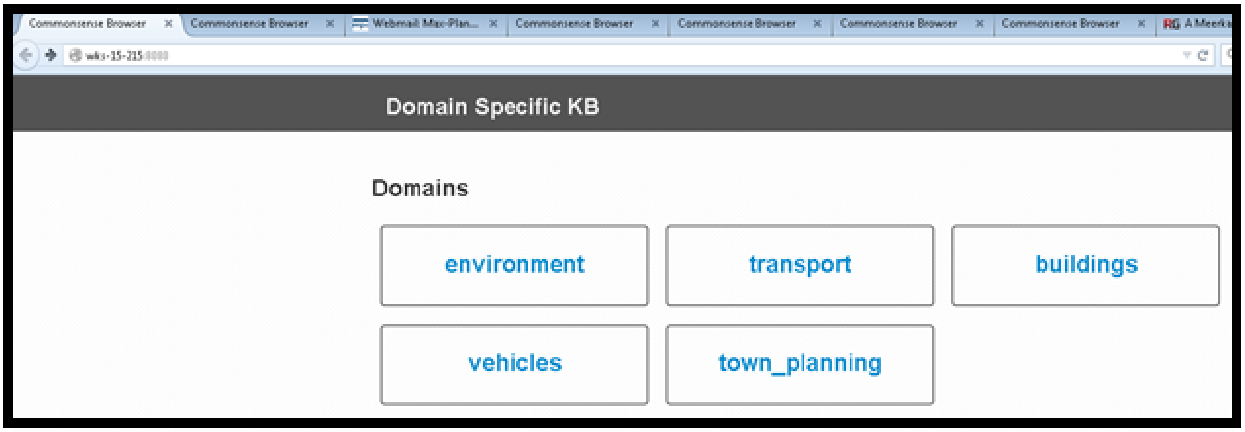
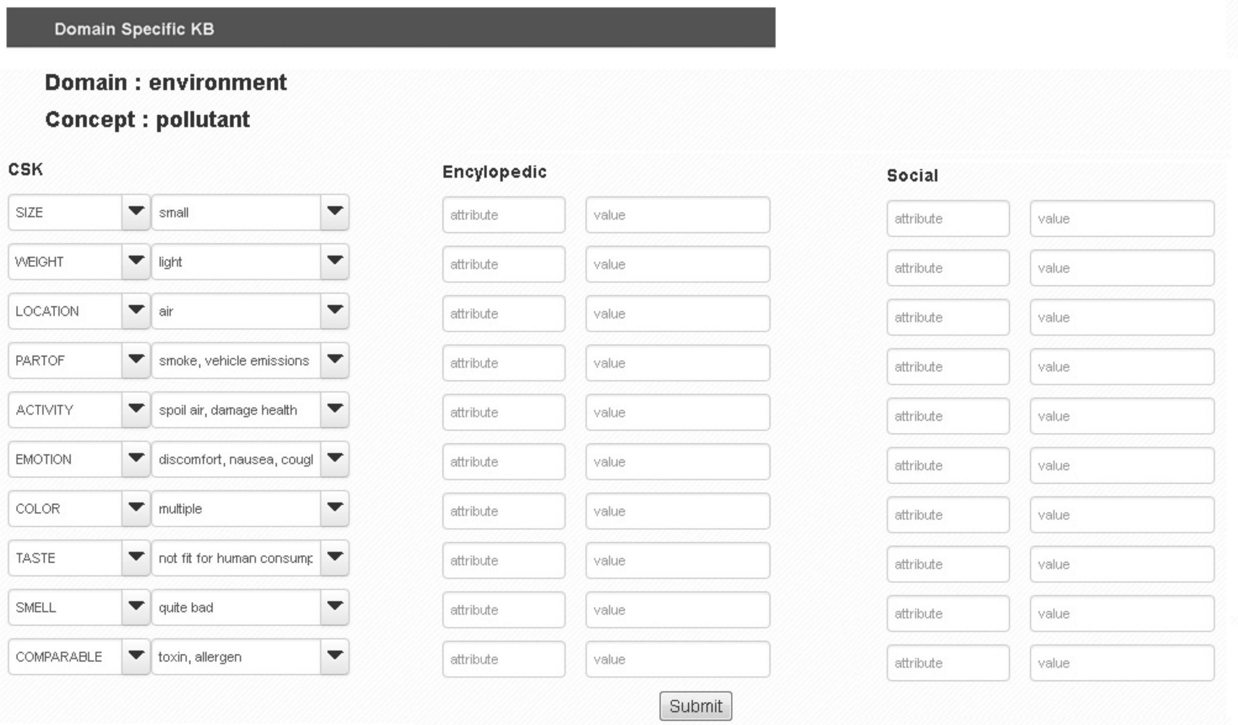


Figure 1: GUI for Domain Specific KB

Figure 2 is an example of commonsense knowledge pertaining to one concept in “environment”, namely, “pollution”. This can in turn be used to instantiate encyclopedic entities, e.g., a pollutant such as PM2.5 can be entered along with attribute-value details on the respective CSK terms.



Domain Specific KB

Domain : environment

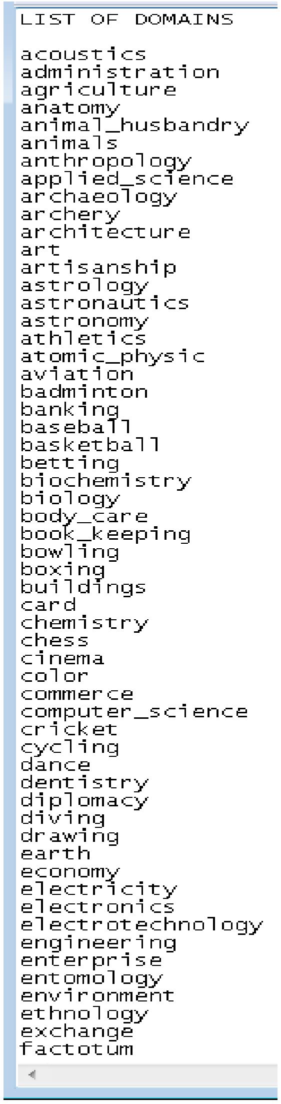
Concept : pollutant

CSK		Encylopedic		Social	
SIZE	small	attribute	value	attribute	value
WEIGHT	light	attribute	value	attribute	value
LOCATION	air	attribute	value	attribute	value
PARTOF	smoke, vehicle emissions	attribute	value	attribute	value
ACTIVITY	spoil air, damage health	attribute	value	attribute	value
EMOTION	discomfort, nausea, cougl	attribute	value	attribute	value
COLOR	multiple	attribute	value	attribute	value
TASTE	not fit for human consump	attribute	value	attribute	value
SMELL	quite bad	attribute	value	attribute	value
COMPARABLE	toxin, allergen	attribute	value	attribute	value

Figure 2: Example of commonsense concept in Domain Specific KB

Further this encyclopedic knowledge can be used to map to relevant social media terms that would be useful in mining.

The process of creating this KB is as follows. We have an initial file with terms from the source of a Probabilistic Domain Classifier. A partial snapshot of this is shown in Figure 3.



```
LIST OF DOMAINS

acoustics
administration
agriculture
anatomy
animal_husbandry
animals
anthropology
applied_science
archaeology
archery
architecture
art
artisanship
astrology
astronautics
astronomy
athletics
atomic_physic
aviation
badminton
banking
baseball
basketball
betting
biochemistry
biology
body_care
book_keeping
bowling
boxing
buildings
card
chemistry
chess
cinema
color
commerce
computer_science
cricket
cycling
dance
dentistry
diplomacy
diving
drawing
earth
economy
electricity
electronics
electrotechnology
engineering
enterprise
entomology
environment
ethnology
exchange
factotum
```

Figure 3: Partial Snapshot Domains from Probabilistic Domain Classifier

From this file, relevant domains are selected that relate to Urban Planning in general based on a fundamental knowledge of the broad realm of Environmental Management. These are used to develop the GUI for the KB as shown in Figure 1, which includes the domains, “environment”, “transport”, “buildings”, “vehicles” and “town-planning.

As a prototype, this knowledge has been mapped to hashtags and used for Social Media mining for opinions on Twitter data. As a starting point, a Twitter file consisting of billions of tweets has been reduced by orders of magnitude and filtered with relevant terms based on topical classification of tweets using background knowledge. Here are some details.

- *Contents of original Twitter file: 750 million tweets*
- *Reduced file with relevant data: 2.5 million tweets on relevant concepts in KB*
- *High confidence classification: ~175k tweets that matched multiple entries from dictionary*

Furthermore, the emotion classifier developed by Tandon et al. as part of the movie scene search project has been used on these tweets. This has revealed that majority of the tweets are polarized. The distribution is as follows.

- 1.17 Million *Positive*
- 0.45 Million *Neutral*
- 0.88 Million *Negative*

This provides an idea of how commonsense knowledge is useful in building domain-specific KBs and using them for data mining, especially from Social Media.

As immediate next steps, we propose the following.

- Continue to populate the domain-specific KB with relevant data in Urban Planning
- Further the use of this KB for topical classification from Social Media, especially Twitter
- Apply the discovered knowledge in the context of Urban Planning and Simulation for the purpose of predictive analysis and decision support
- Consider other sources of Social Media data for text
- Extend this procedure to other domains of interest with interesting applications
- Consider the potential use of images in Social Media in addition to text, for mining, as further discussed next.

7. Discussion

We can potentially outline an overall approach that is expressed in a nutshell in Figure 4.

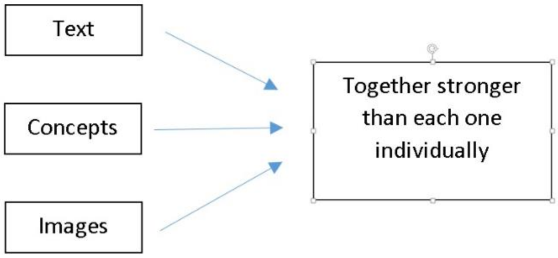


Figure 4: Potential Outline for Mining from Social Media

Since social media often has pictures in addition to text, it would be useful to harness that information as well, for enhancing the mining process. Thus, we try to claim that the knowledge discovered from the text itself along with concepts from CSK and the domain-specific KB and the images would be richer and more useful than that from each aspect alone. We need to further this idea by considering image sources, drilling down to more details and implementing this in multiple domains.

Useful sources of images include Instagram and Flickr. Developing this could involve potential interaction with colleagues in vision and graphics. Testing the implementation would require domain experts and other users who would benefit from the targeted applications.

Some further ideas discussed so far are presented below.

In general, we can consider Opinion Mining and Automated Learning as two broad areas of work after domain-specific KB development. We put together a few ideas in these categories.

Opinion Mining: This can be done using CSK, encyclopedic knowledge and hashtags. We list a few important points here.

- Consider posts in online communities, e.g., "Intercity buses in Melbourne are free"
- Potentially consider images to convey opinions, e.g., picture: polar bear in the middle of ice, message: global warming kills the planet (when ice melts, bear will die)
- Photo-blogs could be a useful source: Text + Images (with opinions)
- Incorporate Topical Classification as well as Polarity of Tweets
- Temporal issue: Cluster with Tweets from the same timestamp
- Spatial issue: Learning Tweet location
- Other sources of mining besides Twitter: Tumblr, Snopes
- Domain-specific media sources: Do we have specialized communities to discuss urban knowledge and planning

Automated Learning: This seems like a far-fetched goal. Some relevant aspects here are as follows.

- Start with domain, e.g., Urban Planning
- Go to entities in Wiki, Yago --- maybe assisted by human labor
- Open information extraction: useful in terms of classification
- Get a list of all Wiki Categories for a domain
- Consider Open IE (Information Extraction) and OLLIE (Open Language Learning for Information Extraction)
- Sources of data acquisition, e.g., Urban Maps
- Useful for applications, e.g., Green Energy, Electric Cars

General Roadmap: We put forth a list of tasks that would be useful in Opinion Mining and Automated Learning and can be pursued next.

- Use Wikipedia
- Derive KB by Open IE
- Construct Domain-Specific KB
- Capture Social Media Content
- Use KB for feature enhancement
- Consider Topical Classification and Polarity
- If content has images, consider a multimodal data space
- Go to comments on image and get polarity
- Organize the information such that it is a reachable interface
- Build the topical tree in a linear manner

8. Ongoing and Future Work

We outline a list of relevant tasks that can be accomplished in the near and distant future for further work in the use of Common Sense Knowledge for Domain-Specific Knowledge Bases.

1. Develop a detailed Urban Planning KB with concept classes and encyclopedic entities which could be mapped to terms in Social Media.
2. Use this KB in mining from Social Media to discover knowledge useful in the overall context Urban Planning and Simulation.
3. Extract pertinent knowledge discovered from Social Media mining to further enhance the Urban Planning KB, thus serving a dual purpose.
4. Fit the Urban Planning KB in broader context of Smart Cities, especially with reference to the Smart Governance and Smart Environment characteristics.

5. Develop domain-specific KBs for other suitable domains with targeted applications. Examples of domains include Sports and Recreation, Music and Entertainment.
6. Use these Domain-Specific KBs for Machine Learning and in the long run try to automate the learning process.
7. In general, expand on the areas of Opinion Mining and Automated Learning with CSK and domain-specific knowledge to drill down to more details and execute the ideas such as learning from photo-blogs and building topical trees.
8. Consider the use of CSK-derived Domain-Specific KBs in the development of Intelligent Tutors and Expert Systems.

9. Conclusions and Ongoing Work

In this project, we have addressed the importance of common sense knowledge (CSK) in the context of domain-specific knowledge bases. We have considered the use of CSK in developing KBs for specific domains and mining with that knowledge, especially from Social Media. We have outlined an approach that utilizes existing work from the WebChild project and furthers this to include domain-specific concepts that can then be instantiated into entities and mapped to relevant terms for Social Media mining. We have started building a domain-specific KB which is being further developed. As a simple prototype implementation, we have used concepts from the KB to map to hashtags in Twitter and produced an output that entails the topical classification of tweets.

The work in this project is ongoing. As immediate deliverables, we expect to further the KB development, use the KB for Social Media mining with text, and apply the discovered knowledge in a real world context. This is being currently implemented in the context of Urban Planning and Simulation. In the near future, we will also implement this in other suitable domains with targeted applications. Considering images in addition to text to further enrich the approach will also be addressed in the near future. Potential long term goals include proposing automated learning with the approach and consider more real world applications for example the use of an Urban Planning KB in the area of Smart Cities and the use of other domain-specific KBs in applications such as Intelligent Tutors and Expert Systems.

10. References

1. Hamidi, S., and Ewing, R., A Longitudinal Study of Changes in Urban Sprawl between 2000 and 2010 in the United States, *Journal of Landscape And Urban Planning*, 2014, Vol. 128, pp. 72-82.
2. Miller, H., and Han, J., *Geographic Data Mining and Knowledge Discovery*, London, UK, Taylor & Francis, 2001.
3. Nagy, R., and Lockaby, B., Urbanization in the Southeastern United States: Socioeconomic Forces and Ecological Responses along an Urban-Rural Gradient. *Journal* of *Urban Ecosystems*, 2014, Vol. 14, No. 1, pp. 71-86.
4. IEEE Smart Cities, http://smartcities.ieee.org/
5. Pampoore-Thampi, A., Varde, A. and Yu, D., Mining GIS Data to Predict Urban Sprawl, *ACM conference on Knowledge Discovery and Data Mining (KDD Bloomberg Track)* New York City, NY, 2014, pp. 118-125.
6. Pawlish, M., Varde, A., Robila, S. and Ranganathan, A., A Call for Energy Efficiency in Data Centers*, Journal of ACM's Special Interest Group on Management of Data Record (SIGMOD Record),* 2014, Vol. 43, No. 1, pp. 45-51.

7. Tandon, N., de Melo, G., Suchanek, F. and Weikum, G., WebChild: Harvesting and Organizing Commonsense Knowledge from the Web, *ACM international conference on Web Search and Data Mining (WSDM),* New York, NY, February 2014, pp. 523-532.

8. Tandon, N., de Melo, G. and Weikum, G., Acquiring Comparative Commonsense Knowledge from the Web, *International Conference of Association for Advancement of Artificial Intelligence, (AAAI),* Quebec City, Canada, July 2014, pp. 166-172.

9. Tandon, N., Weikum, G., de Melo, G. and De, A., Lights, Camera, Action: Knowledge Extraction from Movie Scripts, *International Conference on World Wide Web (WWW Companion Volume)*, Florence, Italy, May 2015, pp. 127-128.

10. Varde A., and Tatti, N., A Panorama of Imminent Doctoral Research in Data Mining, ACM SIGMOD Record, Vol. 43, No. 3, pp. 71 – 74.

11. Vienna University of Technology et al., European Smart Cities, www.smart-cities.eu

12. Wikipedia on Smart Cities, https://en.wikipedia.org/wiki/Smart_city