RECOGNIZING REFERENCES TO PHYSICAL PLACES INSIDE SHORT TEXTS BY USING PATTERNS AS A SEQUENCE OF GRAMMATICAL CATEGORIES

A. Romero, D. Sol

Tecnológico de Monterrey en Puebla, Mexico - aromerog@itesm.mx, dsol@itesm.mx

KEY WORDS: Short messages, patterns, GSP, grammatical categories, location identification, crowdsourcing, reports

ABSTRACT:

Collecting data by crowdsourcing is an explored trend to support database population and update. This kind of data is unstructured and comes from text, in particular text in social networks. Geographic database is a particular case of database that can be populated by crowdsourcing which can be done when people report some urban event in a social network by writing a short message. An event can describe an accident or a non-functioning device in the urban area. The authorities then need to read and to interpret the message to provide some help for injured people or to fix a problem in a device installed in the urban area like a light or a problem on road. Our main interest is located on working with short messages organized in a collection. Most of the messages do not have geographical coordinates. The messages can then be classified by text patterns describing a location. In fact, people use a text pattern to describe an urban location. Our work tries to identify patterns inside a short text and to indicate when it describes a location. When a pattern is identified our approach look to describe the place where the event is located. The source messages used are tweets reporting events from several Mexican cities.

1. INTRODUCTION

Technology evolution allows message exchange in real time. Social networks provide tools to write short messages where a large community of users can access these messages. People use more and more social networks to report holes in the street, lights that do not work, traffic, construction sites on the street. In all these cases is important to indicate the event location. People that answer these reports need to organize the messages to support urban making process decision. We can observe another situation in humanitarian support to recover from a crisis. People write short messages to ask for help and rescuers try to identify where the people are located and organize medicaments, doctors, food and materials. To achieve this goal, a big amount of messages need to be processed. This methodology will contribute to make cities and data smart.

In both cases, the problem is that a big number of messages do not have GPS coordinates and rescue organizations need to read the messages to look for a place where the people sending the message is located. Many times the person location does not correspond with the event location. Our aim is to identify in the short message the words referencing a real world location. There is no method to do this work (Yu Z. 2007), (Finkel et al. 2009), (Romero et al. 2015), (Maya D. 2014), (Parikh M. et al 2013), (Srikant R et al. 1996), (Sun B. 2010).

2. RELATED WORK

Several works have been developed to get information from a text. Rodriguez and Simon (Rodriguez et al. 2013) presented a method to extract data from texts written in Spanish "Método para la extracción de información estructurada desde textos". The data is represented by a graph, more precisely by a conceptual map. They use a method which, combines syntactic analysis with a set of linguistics patterns. They use knowledge related with the domain to support information extraction. This is a general method and can be applied to several domains.

The work developed by S. Vieweg (Vieweg et al. 2011) is related to emergencies. The title of this work is "Natural Language Processing to the Rescue: Extracting 'Situational Awareness' Tweets During Mass Emergency". The approach consists of a classifier to detect different kind of events. The classifier helps to detect the kind of event described in a short message. This work looks for the identification of three dimensions when the short message is written: a) subjectivity, b) if the message is personal or impersonal and c) the style of the message if it is formal or informal. When someone writes a short message in case of crisis the message is impersonal, formal and objective.

A third approach is found in the work of D. Smith, "Detecting Events with Date and Place Information in Unstructured Text" (Smith D. 2002). The work looks to identify dates and names of physical places with only one difference: the texts are formal and structured and they were not written as a short message. In this case the structure of the document can be predicted and the analysis can be organized based on the formality of the document.

In Mexico, about 53 million people have access to internet and about 93% of them use social networks. The most important social network in Mexico are Instagram, Facebook, Twitter and Snapchat. From the total people using social networks about 69% have a Twitter account and most of them use it every day. Several projects in Mexico have been proposed to take advantage of social network and in particular, from Twitter. The work developed by D. Maya (Maya D. 2014) titled *"Sistemas de recuperación de información de eventos viales mediante el análisis de mensajes publicados en las redes sociales",* looks for traffic reports to show in a map the volume of vehicles by avenue or street. In this work, D. Maya affirms that most of the messages do not have coordinates to be located in a map so the location work is made manually.

3. STEPS TO LOCATE AN EVENT IN A MAP FROM A SHORT MESSAGE

CIC (Centro de Integración Ciudadana) is an example, in Mexico, of a platform that uses a social network to support citizens. This association has its Twitter account and several hashtags to support the communication. Several CIC Twitter accounts have been created to support citizens in several towns in Mexico. CIC support citizens in security, traffic and emergencies in Puebla, Monterrey and Saltillo (Internet M. 2015), (Integración C. 2016). CIC has created applications with maps to help people send citizen reports, however most of the people prefers to write a short message to report an event. It will be important to locate an event in real time, however if the report is sent in a short message the map application need to wait for a volunteer to read the message, to interpret the text and to identify the location of the event. The interval between the time when the message is written and its location in a map can be long and can have an important impact in the relevance of the report.

We can follow several steps to define if a text describes a physical place. The first step is to identify the words referencing a physical place. The reports referencing a physical place are published on the map and some data can be added to the report to improve the description of the event. Every message need to be analyzed by a volunteer. All the process need to be evaluated by a volunteer.

In humanitarian help some tools have been developed to support volunteers. For instance, the Application MicroMappers (MicroMappers 2016) which is a collection of web sites called "Clickers". Clickers are designed to help volunteers to improve a report. One of the clickers consists of the presentation of one tweet and a map to geo-locate the event on the map based on the text written in the tweet.

Based on this description, our aim is to develop a method to automate the process to identify a physical place based on the text presented in one tweet. The process will allow the pattern identification to describe a physical place in a short text. Our method will use morphological analysis to look for grammatical categories and sequence identification. The physical place description can follow patterns not exactly defined by the grammatical rules. This is also a discovery that we made in the development of this work. Our algorithm is based on the GSP (Generalized Sequence Pattern) (Balke W. 2015), (Srikant R 1996). This goal will be accomplished by three actions: a) analyzing the sequences of grammatical categories to identify patterns by using the GSP algorithm b) defining rules from the patterns to indicate a possible sequence of words referencing a physical place and c) applying the rules to several texts to evaluate the patterns.

4. SEQUENCES AND PATTERNS

Digital documents contain a big volume of text. It is difficult for a user to read a big number of documents to obtain information. Information Retrieval is charged to find texts according to querys and rules proposed by the user. When a text is found, information extraction is focused on the analysis of the text to describe its structure and to get information that the user is looking for (Romero A. et al. 2015) Texts are not structured data and information extraction look for the identification of patterns, rules and sequences of words.

The data analysis allows us to identify patterns and their frequency in a collection of short texts. Methods for text analysis look for a set of words included in a short text. The short text can be a paragraph or a sentence. In our case, the short text is a tweet. A set of words is important in our approach, but more important is an ordered set of words. An ordered set of words inside a short text is a sequence. We look for patterns described as sequences in a collection of tweets. We propose then to work with a collection of tweets and our method will analyze the collection to identify the frequency of sequences. The number of times that a sequence appears in a collection is known as the support. We can say that a sequence is frequent when the support is greater than the minimum required. A frequent sequence is known as a pattern. Our approach is based on the Generalized Pattern Sequence Algorithm, its acronym is GSP Algorithm. The GSP is based on the a-priori approach. Other approaches to discover sequences as patterns are FreeSpan and PrefixSpan based on grown pattern methods and Sequential Pattern Algorithm Mining in Vertical Data Format SPADE (Parikh M. 2013).

At the beginning, GSP algorithm considers all the candidates with length one and it analyzes the support to determine which candidates are patterns length one. The algorithm takes the patterns to build the candidates with length two and it analyzes the support to determine which candidates are patterns length two. This process continues iteratively until no more patterns length-n are found.

With the sequences identified as patterns inside a collection we can start the morphological analysis to find texts which describe a physical place.

An example of a one-length pattern is the label "C", which represents a coordinated conjunction. The support of this one length sequence is 42.89. A very common coordinated conjunction is "entre", which means "between" in English. An example of a two-length pattern is "BN" which represents a preposition followed by a proper name. The support of this sequence is 44.06.

Three pattern examples are shown in Table 1

Sequence in Spanish	Possible translation to English of the sequence
"NCN" "Name Conjunction Name", example: Hidalgo y Revolución	Example: Hidalgo and Revolución
"BN" "Preposition Name", example: en Avenida las Torres	Example: in Avenue las Torres
"C" "Coordinated conjunction", example: entre	Example: between

Table 1.Sequence example and its codification

5. METHODOLOGY

In the context of smart cities, previous work was published by the authors regarding traffic problem and spatial reasoning (Sol D. et al. 2009). The method and the results presented now is the thesis master presented by Alba Romero in 2016 (Romero A. 2016). This section describes the steps to retrieve fragments of text describing physical places and the algorithm to find patterns with sequence of words.

5.1 Steps

The methodology consists of the retrieval of the texts and the analysis of the information, and can be broken down into the following steps:

1. The retrieval of the texts. The texts are retrieved from messages posted on Twitter through a platform that

collects messages that report different events in the cities of Puebla, Guadalajara, Monterrey and Mexico City.

2. The pre-processing of the texts. This step includes removing some characters and words from the texts that are not useful for our analysis.

3. Morphological labeling. The words in the retrieved texts are labeled with its corresponding grammatical categories.

4. Obtaining patterns and statistics about the way that mentions to places are written based on the grammatical categories of the words in the texts.

5. Defining the algorithm to identify mentions to places in short texts based on the analysis results.

The number of retrieved texts used for the analysis was 1075. From that total, 203 correspond to events that happened in the city of Puebla, 285 to events in Guadalajara, 290 to events in Monterrey, and 297 to events in Mexico City. The reference to a physical place on each of the texts was manually identify to analyze the way in which they were written to discover patterns. Once the tweets are retrieved from Twitter, the text on each of them is pre-processed to remove special characters and URLs.

After that, the text is morphologically analyzed to obtain the labels that correspond to the grammatical categories of each word. Right after, the part of the text that corresponds to a physical place is manually determined to analyze the characteristics of the references to places in the texts and discover patterns that allow us to determine references to places in other short texts that are different from the analyzed.

Once the sequential patterns and sequential rules are found in the references to physical places of the analyzed texts, an algorithm is proposed to identify a reference to a place on a new text. Each label is defined with a letter and it represents a grammatical category. For example, letter B represents a preposition, letter C represents a coordinated conjunction, letter Q represents an adjective, letter Z represents a number, letter V represents a verb, letter F represents a punctuation sign and letter D a determinant.

Our algorithm is organized in 12 steps. The algorithm can work with any short message in Spanish. The following tweet will be used as an example:

“Ya vi choque doble en av Vallarta y robles gil”

The steps will show the evolution of this example to identify where the message has physical places references.

5.2 Algorithm to identify references to physical places in short texts

1. Obtaining the labels that may start a sequence (starting labels) according to the adjacency table. That is to say, those labels that have its father identifier as null. Two examples of labels that may start a sequence are C and R.
2. If the first label of the text is a starting label, a sequence is started. For the text that we are using as an example, the first label is R, so that means a sequence is started.
3. For each of the remaining labels on the text, in our example V S Q B S S C S S:

a. If there’s already a started sequence:
   i. Obtain the last label from the started sequence.
   ii. Obtain all the possible next labels (children) from the last one.
   iii. If the current label is a ‘B’ and there is already a ‘B’ or a ‘C’ on the started sequence and the last label is different from ‘B’, the started sequence is ended and a new sequence is started with the current label. Go to step i).
   iv. If the current label is ‘Z’ and the last label is ‘Q’, the started sequence is ended and a new sequence is started with the current label. Go to step i).
   v. If the current label corresponds to one of the possible next (children) labels, then the current label is added to the started sequence. On the contrary, the started sequence is ended. In our example, all of the labels were possible next children labels, so they were added to the started sequence.

b. If there is no started sequence:
   i. If the current label is a starting label, a new sequence is started with that label.

4. The sequences that end with two consecutive ‘V’ labels (verb) are modify by removing the last two labels. For example, “seguir funcionando” (“keep working”) are two consecutive verbs, which means their labels “V V” would be removed from the sequence.
5. The sequences that end with an ‘F’ o ‘D’ label are modify by removing the last label.
6. In the text that corresponds to the sequences, the word ‘in’ is searched for, and it is modified by removing the previous labels to the one corresponding to the word ‘in’. This only happens once, with the first sequence where that word is found. In our example this rule applies since there is the word “in” on our text. So the resulted sequence is: B S S C S S.
7. The value of each sequence is calculated according to the patterns and frequent words found on it. In our example, there is only one sequence found so there is no need to calculate a value for it.
8. The sequence with the highest value is chosen. In our example we only found one sequence so there is no need to choose.
9. The last label from the chosen sequence is chosen. In our example the last label is S.
10. For all the sequences starting from the one with the highest value:

b. If the last label is an ‘A’, ‘B’, ‘C’, ‘D’ o ‘E’, then the sequence is not completed and therefore that sequence is concatenated to the next one (if it exists). In our example the last label is not any of the mentioned on this step.

c. If not, and if the next sequence exists and it starts with a label ‘C’, or with a label ’B’ and the word that is represents is ‘between’, then the next sequence is concatenated to the current one. In our example, there are no more sequences found so there is no need for this step.

11. For the previous sequence (if it exists) to the one with the highest value, the first label is chosen and if it is a label ‘B’ that corresponds to the word ‘in’, then the highest valued sequence is concatenated to this

previous sequence. In our example, there is no previous sequence from the chosen one.

12. The words corresponding to the resulting highest sequence, are assigned as the mention to a place found on the text. In our example the words corresponding to the resulting sequence (B S S C S S) are "en av vallarta y robles gil" (at Vallarta avenue and robles gil), so those are the words that correspond to the place found.

The final codification of the example used in this section is shown in the table 2. The following sections will show several experimental examples with their results.

Short message: Tweet	Sequence codification
Ya vi choque doble en av vallarta y robles gil Reference to physical place found: en av Vallarta y robles gil	R V S Q B S S C S S Labels that corresponds to the phsycial place found: B S S C S S

Table 2.Short message example analyzed by our algorithm

6. TEST RUNS AND RESULTS

The result of the proposed algorithm was manually rated with a number from 0 to 6, depending on the result (Table 3).

Result	Description
0	A place was not found but it did exist on the text.
1	A place was found but it did not exist don the text.
2	The wrong words were found as a place on the text.
3	The place was found but it was incomplete.
4	The place was found but with extra words on it.
5	The place was found correctly.
6	A place was not found but it did exist.

Table 3. Text rating meaning based on the algorithm result.

To test the algorithm first it was applied to the same 1075 texts with which the analysis was made. Then the algorithm was applied to 102 new texts different from the ones that were analyzed. The identification of places in texts is not based in a predefined list of places, but instead it is based on the patterns found on sequences of grammatical categories in analyzed texts. The rating distribution between the analysed texts is presented in the table 4.

Result	Amount of texts	Percentage that it represents
0	0	0
1	0	0
2	67	6.23
3	54	5.02
4	161	14.98
5	793	73.77
6	0	0
Total	**1075**	**100**

Table 4. Amount of texts and percentage for each result of the algorithm on the same texts that were analysed.

On the next table (Table 5) several examples are stated from texts that obtain a result with value 5. Meaning that the places were correctly identified. The results are presented in Spanish because the analysed texts were written in this language. The examples on the table were translated to English to help the understanding of the results. The translation may not be word to word, but it tries to keep the original meaning.

Tweet	Place found	Result
@JavierLopezDiaz Cerrada 22 sur y sn Francisco, col. Sn. Manuel obras de agua de Puebla, rodar por rio conchos y reincorporarse a la 22 "@JavierLopezDiaz 22 south and st. Francisco closed, st. Manuel neighborhood due to repairments work, transit on rio conchos and incorporate on the 22"	cerrada 22 sur y sn francisco , col. sn . manuel "22 south and st. Francisco, st manuel neighborhood"	5
@SSPTM_Puebla mucho tráfico en cúmulo y atlixcayotl, el semáforo verde dura 5 segundos! @SSPTM_Puebla a lot of traffic in "Cúmulo" and Atlixcayotl, Green traffic light takes only 5 seconds!	en cúmulo y atlixcayotl in Cumulo and Atlixcayotl	5
@Trafico_ZMG semáforos en glorieta del Maíz sin funcionar (parres arias y Venustiano Carranza) ..un caos @Trafico_ZMG lights in roundabout del Maiz do not work (parres arias y Venustiano) a chaos	parres arias y venustiano_carranza parres arias and venustiano_carranza	5
@CICPue Bache profundo en la 32 norte entre 16 y 18 oriente Col. Humboldt. https://t.co/V9BolpYbji @CICPue big hole in 32 norte between 16 and 18 east neighborhood Humboldt	en la 32 norte entre 16 y 18 oriente col._humboldt hole in 32 norte between 16 and 18 east neighborhood Humboldt	5
Semáforo sin funcionar en crucero de Zavaleta y Forjadores @CICPue Light do not work in cross roads Zavaleta and Forjadores @CICPue	en crucero de zavaleta y forjadores in cross roads Zavaleta and Forjadores	5
@CICPue falta alumbrado en Av Las Torres entre vía Atlixcáyotl y Blvd Atlixco @CICPue lack of lighting in Avenue las Torres between way Atlixcayotl and Boulevard Atlixco	en av_las_torres entre vía atlixcáyotl y blvd_atlixco in Avenue las Torres between way Atlixcayotl and Boulevard Atlixco	5
@JavierLopezDiaz A VUELTA DE RUEDA	sobre la carretera federal puebla-atlixco	5

SOBRE LA CARRETERA FEDERAL PUEBLA-ATLIXCO SENTIDO PUEBLA TOMEN VIAS ALTERNAS https://t.co/yFXlr1R9S5 @JavierLopezDiaz the traffic is too slow on Puebla Atlixco federal road direction Puebla take alternatives	sentido puebla on Puebla Atlixco federal road direction Puebla	
@TV3Noticias trafico intenso en la pista mexico puebla altura finsa @TV3Noticias heavy traffic on the motorway Mexico Puebla facing finsa	en la pista mexico puebla altura finsa on the motorway Mexico Puebla facing finsa	5

Table 5. Examples of texts and the results.

The results from applying the algorithm to 102 new texts show that the algorithm was able to find mentions to places in the texts 87.25% of the times, that is to say, the result of the algorithm was successfully 4 on 5 examples. From that percentage, 85.39% of the times the mention to a place was found correctly and 14.61% of the times was found complete but with some extra words.
The next table (Table 6) shows the summary of the results for all the new texts.

Result	Amount of texts	Percentage that it represents
0	0	0
1	2	1.96
2	4	3.92
3	7	6.86
4	13	12.75
5	76	74.51
6	0	0
Total	**102**	**100**

Table 6. Amount of texts and percentage for each result of the algorithm on new texts.

The next table (Table 7) shows some examples translated to English of new analyzed texts and its results.

Tweet	Found place	Result
@retioMTY hamburger stand blocks half of the street “red sea” and “Tlatelolco” neighbourhood Loma Linda	of the street red sea and Tlatelolco, neighborhood Loma Linda	5
@Cicmty It was Heard strong shock in Vasconcelos corner Rosas river, neighborhood del Valle	Vasconcelos corner Rosas river neighborhood del Valle	5
@alertuxmxdf How its state is tragic in motorway Cuernavaca Mexico	in motorway Cuernavaca Mexico	5
@zona3noticias 8 people injured in overturning in the freed road Zapotlanejo by Green cross Tonalá and 6 by red cross Zapotlanejo https://t.co/ViKldiysyB	in the free road Zapotlanejo	4
@Cicmty loose wire risk at av.Benito juarez between 5 de mayo and azteca https://t.co/GYtlrGmUEk	risk at av.Benito juarez between 5 de mayo and azteca	4
@Cicmty Apologies I’m not sure if I hear bullets in neighborhood Rancho Viejo avenue Israel Cavazos	in neighborhood Rancho Viejo	3
Con el #NodoVialAtlixcayotl485 en Puebla sigue la transformación y el progreso. cc @TonyGali with the #node Via Atlixcayotl 1485 in Puebla the transformation and continuous progress	the transformation and continuous progress	2
@ManceraMiguelMX It is arbitrary and wrong the “does not circulate”, the solution is to fix trash truck and to allow new cars	to allow new cars	1

Table 7. Examples of new texts and the results.

The results show that our algorithm has a high level of performance to identify sequences inside a text describing a physical place. The sequences with a high percentage are used as patterns to identify new texts describing events and their location. It is interesting to precise how the people describe an event and its location by using not structured phrases and sentences outside the grammatical rules. We think to add constraints to the algorithm to consider outliers, that is to say when the people use conjunctions and prepositions not describing a physical place. We think that our algorithm can add patterns not yet discovered with the actual collection. We used short texts coming from different environments and different cities in Mexico. It should be more interesting to test our method with short messages coming from another Spanish speaking country.

7. CONCLUSIONS

The proposed methodology allows us to identify patterns on how the references to physical places are written in short texts, by using the grammatical categories of each word. The identification of references to physical places in texts is of great relevance for applications that use messages published on social

networks for specific interests, like the apps for getting traffic information, event reporting, or crisis management.

The patterns found allowed us to identify which words from a short text reference a physical place. The purpose of this identification of places is to be able later on to geo-locate this places to improve the efficiency of the applications that recollect and use messages published on social networks.

The results obtained on the tests prove that the algorithm proposed works on new texts just as much as it works on the texts that were used to do the analysis. This allows us to conclude that there are in fact patterns in the way in which the references to physical places are written in short texts where the intention is to report an event.

It is possible to apply this methodology on a different language or to Spanish coming from another country. To do that, first we have to identify the different grammatical categories in that language to define the labels. It is not necessary to know the formal structure of the language because the algorithm will identify patterns as people describe an event and its location. We think that our methodology could be applied to this sequences of labels from English (or other language) written texts. The results may vary as well as the accuracy.

8. FUTURE WORK

- Our methodology assumes that the texts presented contain a reference to a place. As a future work we can include a classification that allow us to separate texts that contain a place from those that do not.
- To improve the algorithm by using, besides the morphologic analysis, the syntactic analysis. The syntactic analysis will allow us to identify the sequence and hierarchical relations that the words have.
- Geo-locate the references to physical places identified on the texts.
- Integrate the analysis methodology and the algorithm into a web service so that it can be used by anyone interested as a tool for developing other web applications.
- Integrate the algorithm, along with the geo-location with existing applications that retrieve messages from social networks to support the requests for help during crisis.
- Add gazzeters to manage a list of place names depending the analyzed geographical territory.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the Instituto Tecnológico y de Estudios Superiores de Monterrey for supporting this research, which results were presented by Romero A. on her thesis for her Master's Degree on Science and Intelligent Systems (Romero A. 2016).

REFERENCES

Balke, W., Homoceanu, S., 2015. Data Warehousing & Data Mining. Institute for Information Systems at Technische Universität Braunschweig, Germany. http://www.ifis.cs.tu-bs.de/webfm_send/540 (2nd October 2015).

Finkel J., Manning C. 2009. Joint parsing and named entity recognition. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 326-334.

Integración, C. 2016. CIC (Centro de Integración Ciudadana). http://www.cic.mx/ (21 March 2016).

Internet, M. 2015. 11° Estudio sobre los hábitos de los usuarios de Internet en México 2015. Asociación Mexicana de Internet. https://www.amipci.org.mx/images/AMIPCI_HABITOS_DEL_INTERNAUTA_MEXICANO_2015.pdf (4th April 2015).

Maya D. 2014. Sistema de recuperación de información de eventos viales mediante el análisis de mensajes publicados en las redes sociales. Tesis Mtra. en Cien. Comp. Instituto Politécnico Nacional.

MicroMappers 2016. Digital Disaster Response. With a Single Click! http://micromappers.org/ (2nd April 2016).

Parikh M., Chaudhari B., Chand Ch. 2013. A Comparative Study of Sequential Pattern Mining Algorithms. International Journal of Application or Innovation in Engineering & Management, 2(2).

Rodríguez A., Simón A. 2013. Método para la extracción de información estructurada desde textos. Revista Cubana de Ciencias Informáticas, 7(1). ISSN: 2227-1899. pp. 55-67.

Romero A. Sol D. 2015. International Workshop: Intelligent Disaster Management Workshop. 27th February 2015, Querétaro, México.

Romero A. 2016. Análisis de textos cortos para la identificación de menciones a lugares mediante el reconocimiento de patrones en las secuencias de categorías gramaticales de las palabras. Tesis de Maestría, Tecnológico de Monterrey en Puebla, 2016.

Smith D. 2002. Detecting events with date and place information in unstructured text. In Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries (JCDL '02). ACM, New York, NY, USA, pp. 191-196.

Sol D., Rodriguez M., Zepeda C. 2009. The Car Traffic Problem to Evacuate People in the Popocatépetl Volcano. Chapter 33 in the Book. Numerical Modeling of Coupled Phenomena in Science and Engineering: Practical Use and Examples, Taylor and Francis, pp. 408-420. ISBN: 9780415476287.

Srikant R., Agrawal R. 1996. Mining sequential patterns: Generalizations and performance improvements. In Lecture Notes in Computer Science. Advances in Database Technology — EDBT '96. Springer Berlin Heidelberg. pp 1-17.

Sun B. 2010. Named Entity Recognition: Evaluation of existing systems. Master Thesis. Norwegian University of Science and Technology. Department of Computer and Information Science.

Vieweg S., Corvey W., Palen L., Martin J., Palmer M., Schram A., Anderson K. 2011. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During

Mass Emergency. International AAAI Conference on Web and Social Media.

Yu Z. 2007. High Accuracy Postal Address Extraction from Web Pages. Master Thesis, Dalhousie University, Halifax, Nova Scotia.