STA: a Spatio-temporal Thematic Analytics Framework for Urban Ground Sensing

Guizi Chen[1], Liang Yu[2], Wee Siong Ng[1], Huayu Wu[1] and Usha Nanthani Kunasegaran[3]

[1] Institute for Infocomm Research, A*STAR, Singapore
[2] Alibaba Cloud, Hangzhou, China
[3] Urban Redevelopment Authority, Singapore
chengz, wsng, huwu@i2r.a-star.edu.sg; liangyu.yl@alibaba-inc.com
Usha_NANTHANI@ura.gov.sg

Abstract. Urban planning has always involved getting feedback from various stakeholders and members of public, to inform plans and evaluation of proposals. A lot of rich information comes in textual forms, which traditionally have to be read manually. With advancements in machine learning capabilities, there is potential to tap on it to aid planners in synthesizing insights from large amount of textual feedback data more efficiently. In this paper, we developed a more general urban-centric feedback analysis framework, which encompasses the spatio-temporal thematic of ground sensing. Three essential methods: geotagging, topic modeling, and trend analysis are proposed and a prototype has been implemented. The results of experiments indicate that the proposed framework could not only accurately extract precise geospatial information, but also efficiently analyze the semantic themes based on a probabilistic topic modeling with Latent Dirichlet Allocation. Importantly, the spatial and temporal trends of detected topics indicate the effectiveness of our proposed algorithm and then benefit domain experts in their routine work and reveal many interesting insights on ground sensing matters.

Keywords: Geotagging, Topic Modeling, Urban Planning, Spatio-temporal Analysis.

1 Introduction

Singapore, a 710 km^2 island with more than 5 million people, is one of the most densely populated countries in Asia, even in the world. In order to build a high quality living environment, establishing a smart and integrated urban planning framework is one of the many important efforts by the government agencies and research communities in Singapore [1]. The development in artificial intelligence and big data analytics provides a pretty good opportunity for smart urban planning [2-4]. In recent years, the Urban Redevelopment Authority of Singapore (URA) initiated several projects with the Institute for Infocomm Research (I2R) under Singapore's Agency for Science, Technology and Research on leveraging data analytics for better urban planning.

Urban planning has always involved getting feedback from various stakeholders and members of public, to inform plans and evaluation of proposals. A lot of rich information comes in textual forms, which traditionally have to be read manually. With advancements in machine learning capabilities, there is potential to tap on it to aid planners in synthesizing insights from large amount of textual feedback data more efficiently. Public feedback in the form of text are typically informative and helpful. However, a great deal of tedious manual work has been done to perform useful thematic analysis, such as detecting events or topics over a designated planning area or a period of time in practice [3]. Moreover, if planners need to investigate the ground sensing problem in different domains, similar efforts have to be made repeatedly. Thus, it is necessary and important to automate this process.

Driven by the business needs, the objective of this study is to develop a geo-tagging and theme modeling system, which not only infers locations of messages but also detects underlying events or topics and monitors their trends from user feedback data. The system can automate the process of learning about the ground concerns and thus improve the productivity of urban planners. For the past years, many researchers have been working on smart urban planning using artificial intelligence and data analytics techniques [2-4, 6-8]. However, limited research has been done on mining unstructured textual user feedback data. In this paper, we design and develop a spatio-temporal thematic analytics (STA) system. Specifically, we have made the following contributions:

- Practical ground sensing solutions and the results are validated by domain experts.
- Extracting precise geospatial information by developing a geotagging tool which integrates machine learning techniques.
- Analyzing themes and providing a quick and comprehensive understanding of some of the ground concerns using a probabilistic topic modelling.
- Exploring the trend of the detected topics both spatially and temporally to analyze the changes of event or topic popularity over times at different locations.
- Using real-world urban planning data for experiments and effectiveness validation.

The rest of this paper is organized as follows. In Section 2, we describe the location identification and theme detection problems which we would address in this paper. The methodology for spatio-temporal thematic analytics (STA) engine is presented in details in Section 3. In Section 4, we present the implementation of STA system. The practical application and evaluation of our system on the real-life user feedback data are described in Section 5. After that, we review the related work in Section 6 and conclude this paper in Section 7.

2 Problem Statement

2.1 Location Identification

In fact, the analysis of the GS datasets shows that most of the textual data contain location information embedded within the feedback, *i.e.*, the issues are related to some certain geo-locations. The analytical results can be visualized clearly using a map. Fig. 1

shows the planning area map in five regions in Singapore. The central region with a high population density is one of the major concerns in urban planning, and hence it is further divided into 3 sub-regions, i.e., central area, central west and central east. Thus, the urban planning subdivisions of Singapore include 7 regions. Each region can be divided into several planning areas. For example, Table 1 lists the corresponding planning areas for Central Area, Central East and Central West. By identifying locations, each textual feedback would be annotated with a location of concern. Syntactically, we need to identify all location names from the textual feedback data. It is similar to a geocoding task, but the input is a full textual paragraph rather than pure address text. Semantically, we need to identify the most appropriate location especially when multiple locations co-occur in the data.

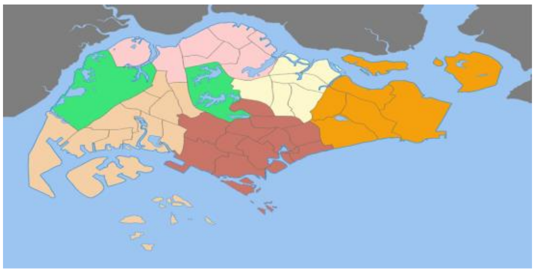


Fig. 1. Urban planning subdivisions of Singapore [9]

Table 1. Subdivisions and Corresponding Planning Areas

Regions	Planning Areas
Central Area	Orchard, Rochor, Museum, Singapore River, Downtown Core, Outram, Straits View, Marina South, Marina East
Central East	Marine Parade, Newton, Geylang, Kallang, Toa Payoh, Novena
Central West	Bishan, Bukit Timah, Queenstown, Bukit Merah, Tanglin, River Valley

2.2 Theme Detection

Considering the great amount of textual information, it would be difficult for urban planners to have a quick overview and insight into some of the rising concerns of the public. The thematic analytics module *via* probabilistic topic modeling provides an automatic way to understand and respond to public concerns timely. Combining with the detected location information above, the urban planning officers can quickly understand what the concerns are in a planned area/region, and offer the corresponding feedback in a timely manner. It will help to reduce the manual efforts needed to read through all the feedback data for gathering an overview of the key ground concerns. To achieve the objectives, a well-established topic modeling approach via Latent Dirichlet Allocation (LDA) [4] is introduced to infer a set of latent events/topics from given textual feedback data.

3 Methodology

3.1 STA Framework

In this section, an overview of the proposed spatio-temporal thematic analytics (STA) framework, including five major components in the framework, *i.e.,* data preprocessing, geo-tagging, thematic analysis, spatial analysis, and temporal analysis. Since the input data to the system are informal feedback emails, in the first step we clean the data and then extract the representative words with thematic indicator. Next, we analyze the specific geo-locations by using a Gazetteer based method followed by an information extraction method, which can discover and annotate the large archives of public feedback to explore what aspects of the key ground concerns are mentioned. As noted above, urban planners may have interests in the evolving trends of particular topics over certain time periods or locations, rather than merely focusing on detected list of topics and the feedbacks that are closely related to the topics. Therefore, we further analyze the trend of the detected particular topics over given timelines and geographic locations. The essential components will be described in details in the following section.

3.2 Geo-tagging

Location extraction is one of the fundamental tasks of Named Entity Recognition (NER) [11] in Natural Language Processing (NLP). However, general NER method may suffer from two real issues related to user feedback data. One is that feedback data typically consists of short text or even grammatically incorrect content. The other is that many local location names are implicitly mentioned in the feedback text. In contrast, we import a number of maps shown in Table 2 as a gazetteer to look up the location names. All address names are indexed in a tree structure to speed up the searching. Since they are extracted from maps, each address name is coupled with coordinates. For those maps containing polygon and polyline shapes, such as planning area and road map, we use their internal center points as the locations.

Table 2. Imported map layers

Map Name	Number	Granularity	Map Name	Number	Granularity
	of Features			of Features	
Postal Code	197214	1	Planning Area	55	5
Cadastral Parcel	143210	2	Map Name	Number	Granularity
				of Features	
Road Network	8413	3	Map Name	Number	Granularity
				of Features	
Subzone	311	4			

If there is only one location extracted, we just treat it as the location for the feedback. However, in fact, there are often more than one addresses mentioned in the feedback. For example, one might complain about the noise around his office, but he may attach his residential address to the end of the feedback. To locate the appropriate location name from multiple candidates, a machine learning based method is developed in order to estimate the probability of location candidates.

3.2.1 Feature Engineering. Whether a location is predicted correctly depends on the context where it is mentioned. We model the context by the following attributes:

- Surrounding words (uni-gram [7]). We use the words in the same sentence where the address is mentioned.
- Position of the address in the whole text. The position is denoted by the percentage, *i.e.,* beginning - 0% and ending - 100%.
- Layer granularity level. We use integers to denote the granularity values for each map layer. For example, postal map has higher granularity than planning area, using the central point of a planning area often introduces higher error range.
- Roles of the address words. Since all location names should be noun, we use Part-of-Speech (POS) [12] tag as an attribute and only the cases in which the address words are all noun are retained.

3.2.2 Training and Prediction. In reality, there are feedback already annotated with geo-locations by users who addressed the feedbacks, we use the feedback data as the training set and treat the known locations as the ground truth. If there are multiple locations identified from a feedback, the one near the true location is more appropriate, which are annotated 1 and others are annotated by 0.

As well known, random forest is a bagging algorithm based on decision tree while it corrects for overfitting habit of decision trees. Additionally, the introduction of randomized node optimization and bootstrapping procedure leads to robust model performance in many practical applications [13]. In this paper, random forest is used as a classifier to predict the probability for location candidates.

3.3 Thematic Analysis Method

We employ a probabilistic thematic analysis method via latent Dirichlet allocation model (LDA) to infer the location-based topics/themes. LDA model is a generative probabilistic algorithm that discovers latent semantic structure of a given collection of text documents. And it assumes that each document is associated with a mixture of various underlying topics, and each word in a document is attributable to one of the document's topics [14]. Fig. 2 show the graphical representation of LDA model, and Table 3 lists the meanings of the notations used in the model.

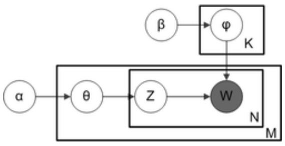


Fig. 2. Graphical representation of LDA model. The boxes refer to plates that indicate replicates. The outer plate refers to document, while the inner plate refers to the repeated selections for latent topics and words within each document.

Table 3. Notations in LDA

Notations	Meanings	Notations	Meanings
M	Number of documents	θ_m	Topic distribution for document *m*
N_m	Number of words in a document *m*	φ_k	Word distribution for topic *k*
K	Number of hidden topics	z_{mn}	Topic for the n^{th} word in document *m*
α	Parameter of Dirichlet prior for per-document topic distribution	w_{mn}	A specific word
β	Parameter of Dirichlet prior for per-topic word distribution		

Given a collection of text documents as training data, by using Gibbs sampling method [15], we can infer the model parameters of LDA based on the drawn samples. In particular, we can compute the per-document topic distribution θ, an $M×K$ matrix, where each entry means the probability of the document *m* belonging to the topic *k*:

$$\vartheta_{m,k} = \frac{n_m^{(k)}+\alpha_k}{\sum_{k-1}^{K} n_m^{(k)}+\alpha_k} \tag{1}$$

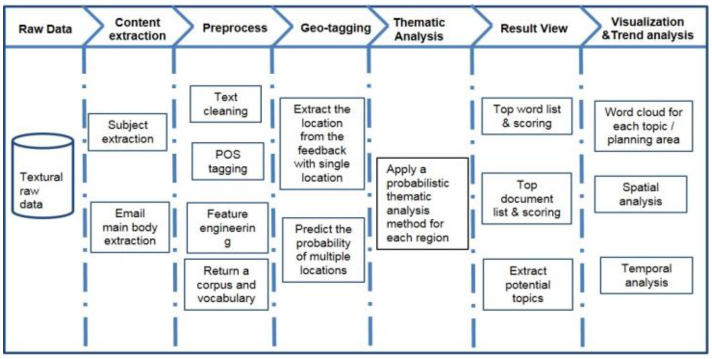
where $n_m^{(k)}$ means the number of times that the words in a document m are allocated to a topic *k* [2-3]. Then, we compute the per-topic word distribution φ, a $K×V$ matrix, where each entry means the probability of clustering a word t into the topic *k*:

$$\varphi_{k,\mathrm{t}} = \frac{n_k^{(t)}+\beta_t}{\sum_{t-1}^{V} n_k^{(t)}+\beta_t} \tag{2}$$

where *V* means the size of the vocabulary generated from the training data, and $n_k^{(t)}$ means the number of times that a term (word) in the vocabulary is allocated to a topic *k* across all the documents [2-3].

4 Implementation of System

In this section, we present the implementation information of our proposed spatio-temporal thematic analytics (STA) system, as illustrated in Fig. 3.



Raw Data
Content extraction
Preprocess
Geo-tagging
Thematic Analysis
Result View
Visualization &Trend analysis
Textural raw data
Subject extraction
Email main body extraction
Text cleaning
POS tagging
Feature engineerin g
Return a corpus and vocabulary
Extract the location from the feedback with single location
Predict the probability of multiple locations
Apply a probabilistic thematic analysis method for each region
Top word list & scoring
Top document list & scoring
Extract potential topics
Word cloud for each topic / planning area
Spatial analysis
Temporal analysis

Fig. 3. Flow chart of STA system with main modules.

From Fig. 3, it can be seen that the major modules of the STA systems, i.e., Input data, Content extraction, Preprocess, Geo-tagging, Thematic analysis, Result view, Visualization & Trend analysis. Each of the modules would be briefly described below.

4.1 Email Content Extraction & Preprocessing

In the customer feedback collection system, the input data are 5596 emails from 55 planning areas of Singapore. Email is one of convenient and common means for communication *via* text, but it has many noisy words and needs to be in-depth cleaned before high quality email mining. Hence, a specially designed cascade information extraction method (CIEM) is proposed to extract the email subject and main body. The specific rule for subject extraction is that if a line begins with a pattern of “Subject:” and ends with line break, then extract the middle part of characters. For email main body extraction, the rule is that if a line begins with a pattern of “Dear sir”, “Hi Officer”, *et.al.*, remove these pattern characters, as well as the words ahead of them. After that, if a line begins with a pattern of “Regards”, “Cheers”, *et.al.*, remove these pattern characters and the words behind them. In this way, the signature, email disclaimer and confidentiality declaration can be removed and only the email main body is retained. As a consequence, the meaningful information could be extracted by combining the email subject and main body parts. Generally, only the content words that possess actual meanings, such as noun, will pose an important influence on semantics or meaning. Therefore, in the pre-process, the part-of-speech (POS) tagging technique [12] is also applied to label the cleaned email text. The meaningless words are filtered and the words with actual meaning are retained for the following process. Finally, normal text cleaning process is implemented to remove common and stop words, punctuation, numbers and non-English words.

4.2 Geo-tagging & Thematic Analytics

We propose a meaningful list of features, and then develop a random forest classification method to identify location names from preprocessed feedback data. Sections 3.2 presents the method in details. Also, we employ a probabilistic thematic analysis method via latent Dirichlet allocation to detect semantic events or themes from processed feedback data. Section 3.3 describes the detailed information. Moreover, the trends of the detected topics from the public’s feedback/complain emails could be tracked over issue time and geographic locations. The demonstration form in STA system is described in details in Section 4.3.

4.3 User Interface

Our system provides efficient user interface for urban planners to upload data, perform analysis, view insights and save results. Fig. A in appendix shows how resident submit issues to the system. Geotagging of individual issue is shown in Fig. B, where top 3 possible locations are recommended. In Fig. C, urban planners can upload the issues and perform topic by specifying the number of topics. When the training of topic modeling is done, the top words list are shown automatically and the top documents list can be saved locally. The spatial and temporal trend analysis is illustrated in Fig. D, where the urban planners are free to choose any combination of locations and period of interests to see how the topic is changing over location or time.

5 Experimental Results

In this section, we use real-world feedback data to evaluate the major tasks in our spatio-temporal thematic analytics system.

5.1 Geo-tagging

In the 5596 emails, 75% feedbacks have locations mentioned in the text. We extracted those which have the true location annotation to generate a training dataset, and ran a 10-fold cross validation to estimate the performance according the criteria below:

- Correct Recommendation. We use the Planning Area map (Table 2) to validate, *i.e.*, if the recommended location fall into the same planning area of the true location, it is considered correct.
- Wrong Recommendation. If the recommended location is not in the same planning area of the true location, but at least one candidate location is, it is considered wrong. We may try to improve the machine learning algorithms to reduce this percentage.
- Wrong Candidates. All candidates are not matched to the true locations. This part can be considered as "base error" which is very hard to improve.

Finally, we have achieved 78.6% for the correct recommendation, 10.7% for the wrong recommendation, and 12.4% for the wrong candidates.

5.2 Topic Detection

5.2.1 Probabilistic topic modeling based method. In this section, the topic modeling with Latent Dirichlet allocation (LDA) is applied in each region to detect the potential topics. As discussed with domain experts, the number of topics is set as 10, 8 common topics and 1 or 2 new topics, based on their domain knowledge and experience. As mentioned in Section 3, in the topic modeling, each document can be viewed as a mixture of various topics and each topic has a probability distribution over the vocabulary. Therefore, topic distribution per document, distribution over the vocabulary per topic and the per-document per-word topic assignment can be derived from hidden variables. Also, based on the distribution over the vocabulary per topic, we can get the top words those are assigned to every topic. Table 4 illustrates some partial results of top words, in decreasing order of the corresponding scores (as computed from Eq. (1)). As we can see, given this list of words, the potential topic is quite distinct from each other and it could be easily derived. The first entry in each column is a title suggested by domain experts from URA, *e.g*., enquiries on sale sites, feedback on noise, sub-letting of apartment, enquiries on outdoor structure, which are issues of public's most concerned.

Table 4. Partial results of top words for the clustered topics (central west region).

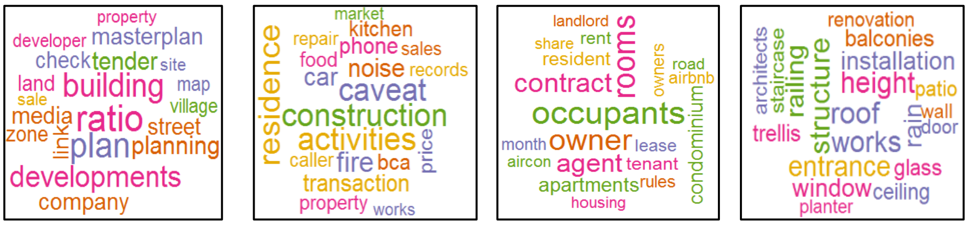
Topic #1 Enquiries on Sale Sites	Topic #2 Feedback on Noise	Topic #3 Sub Letting of Apartment	Topic #4 Outdoor Structures
ratio	construction	rooms	roof
plan	fire	owner	structure
tender	noise	agent	height
planning	car	condominium	entrance
masterplan	activities	apartments	trellis
check	caller	tenant	glass
land	repair	owners	ceiling
zone	street	lease	staircase
map	residence	airbnb	door
sale	caveat	occupants	renovation

Also, from the topic modeling with LDA, we can get the top documents closely related to the topics. Partial results of top articles (index of documents for space saving) for the clustered topics are listed in Table 5, in decreasing order of the corresponding scores (as computed from Eq. (2)). Based on these documents, we could verify the above inferred topics and have a further understanding of these topics.

Table 5. Partial results of top documents for the clustered topics (central west region).

Topic #1 Enquiries on Sale Sites	Topic #2 Feedback on Noise	Topic #3 Sub Letting of Apartment	Topic #4 Outdoor Structures
Doc #1816	Doc #848	Doc #883	Doc #2303
Doc #2566	Doc #850	Doc #4970	Doc #1233
Doc #1056	Doc #945	Doc #4732	Doc #1920
Doc #3770	Doc #707	Doc #3838	Doc #1026

5.2.2 Topic visualization. In order to have a direct and clear visualization of these inferred topics for each region, a very handy way is through a word cloud. Fig. 4 provides a quick visualization for top words of 4 reprehensive topics in central west region. The size of the words in the word cloud represents the possibility of the word given the topic. We also named each topic with a title manually which help us gain an overview of the inferred topics.



a). Topic #1: enquiries on sale sites

b). Topic #2: feedback on noise

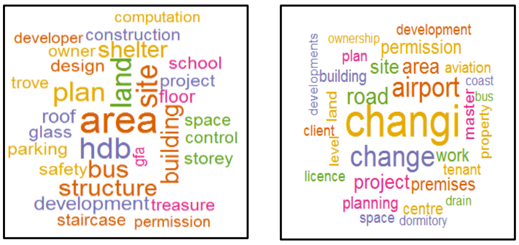
c). Topic #3: sub-letting of apartment

d). Topic #4: outdoor structures

Fig. 4. Thematic words with word cloud (central west)

From Fig. 4, we can see that, while visualizing the clustered words with a word cloud, it is clear for the major topic in the input text. For example, in Fig. 4(a), the words of “planning”, “ratio”, “masterplan”, “building”, “sales” and “development” automatically clustered together and the potential topic can be distinctly derived, which is enquiries on sale sites related. And we could summarize what the residents are talking about is sub-letting of apartment when the words of “occupants”, “owner”, “agent”, “tenant” and so on appeared together in one cluster.

To facilitate the urban planning and management, the regions are further divided into 55 planning areas and thus the word clouds with highly frequent words are plotted based on the planning area for refining the urban event/topic. For instance, Fig. 5 shows the highly frequent words in Punggol and Changi, respectively. As we know, Punggol is one of residential eco-towns and provides a high-quality living environment. Hence, more words like HDB, outer door structure, and parking appear in Fig. 5(a). Mention of Changi, however, the first one that probably pops into your head is the airport. Thus, in Fig. 5(b), there are a lot words about airport, aviation and dormitory/tenant. From Fig. 6, it can be seen the clear characteristics of a particular urbanized planning area.

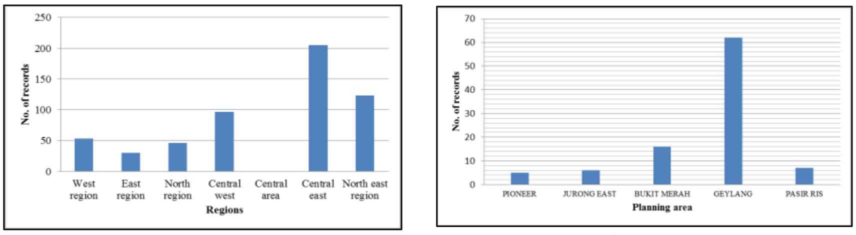


a). Punggol b). Changi

Fig. 5. Word cloud with highly frequent words in planning area.

5.3 Trend Analysis

5.3.1 Spatial analysis. A spatial analysis is done for displaying location-based topic popularity and thereby providing more efficient guidelines for urban policy or planning. For example, Fig. 6(a) shows the evolution of an example topic (outdoor structures) in these seven urbanized regions, indicating the central east region has the highest number of issues while the central area has very few issues. This information could efficiently help urban planner make the decision to investigate the central east for ways to avoid issues, and to look into break down of locational issue numbers. The break down analysis is also supported by our system as shown in Fig. 6(b), where a combination of sub-planning areas is chosen from west to east. The results of spatial analysis offer us wider perspective for the detected topics to understand public concerns and help officers to take corresponding actions and make localized plan or policies.



a). The variation of an example topic popularity over regions.

b). The variation of the example topic popularity over partial planning areas.

Fig. 6. Locational topic trend (Example topic: outdoor structures)

5.3.2 Temporal Analysis. The detected topics are also analyzed with the issued date and the variation plot of the topic popularity is supported in our system. Fig. 7 shows the temporal trend in Geylang from Jan 2015 to Dec 2015. As illustrated in Fig. 7, the topic of outdoor structures in Geylang has local peaks in March June, September and December 2015. The urban planner could further investigate these four months. These clear trends of detected topics could not only indicate the effectiveness of our proposed algorithm, but also benefit domain experts in their routine work and reveal many interesting insights on the key ground concerns.

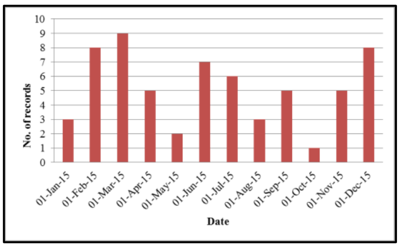


Fig. 7. Time series of detected topics. (topic: outdoor structures; planning area: Geylang)

6 Related Work

In recent years, the challenge of text-based geotagging has been receiving significant interest from research area due to the emergence of large amount of social media data such as tweets, posts and emails. Most existing studies of text-based geotagging aimed to identify the location where the text is talking about. For some social media with user profile and networks such as tweets, multi-indicator approach was proposed in Schulz *et al.* [16] to consolidate different user-profiled spatial indicators to identify the location of the tweet and the home location. Moreover, additional features from connections such as friends' location [8] can be incorporated to approximate the location of the text. Other than tweets, Flickr is another example of social media which is used to link photos with popular places where the photos were taken for the purpose of geotagging [17]. However, such techniques do not apply in our case because there is no user information. On the other hand, one advantage of the emails we have collected is that most of the texts contain words or phrases indicating the region. Therefore, the texts talking about the same region can be gathered to extract insights about this region. Along this direction, Quercia *et al.* [18] applied sentiment analysis to tweets in London to identify “well being” areas.

7 Conclusions

In addressing the requirements from urban planners on geotagging and analyzing extensive textural data on ground concerns, we propose a spatial-temporal thematic analytics (STA) framework to detect the specific locations of citizen’s feedbacks and analyze the semantic themes. Importantly, a prototype implementation was tested with real data. The result shows that the geospatial information could be effectively extracted and urban planning subdivisions could be accurately geolocated. Furthermore, the topics extracted from the feedbacks have clear locational representation, and time trend with tremendous benefits to the procedure of locational and timely effective urban planning-making. With the output from our system, the urban planners can take corresponding actions more timely and make localized plans or policies based on the characteristics of the regions.

Acknowledgments

This work was partially supported by the A*STAR Science and Engineering Research Council (SERC) Grant No. 1524100032.

References

1. Peter Q., Exploit Technology for Smart Urban Planning, Urban Redevelopment Authority, Singapore (2014).
2. Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022 (2003).
3. Chang, J. "Collapsed Gibbs sampling methods for topic models." (2012).
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
5. Lefevre, P., Kolsteren, P, De Wael, M.P., Byekwaso, F., Beghin, I.: Comprehensive Participatory Planning and Evaluation. Antwerp, Belgium: IFAD. Retrieved 2008-10-21(2000).
6. Bolelli, L., Ertekin, S¸., Giles, C.L.: Topic and trend detection in text collections using latent dirichlet allocation. In: European Conference on Information Retrieval. pp. 776–780. Springer (2009).
7. Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G.: Syntactic clustering of the web. Computer Networks and ISDN Systems 29(8), 1157–1166 (1997).
8. Compton, R., Jurgens, D., Allen, D.: Geotagging one hundred million twitter accounts with total variation minimization. In: Big Data (Big Data), 2014 IEEE International Conference on. pp. 393–401. IEEE (2014).
9. https://en.wikipedia.org/wiki/Regions_of_Singapore

10. Ma, B., Zhang, N., Liu, G., Li, L., Yuan, H.: Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. Information Processing & Management 52(3), 430–445 (2016).
11. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007).
12. Voutilainen, A.: Part-of-speech tagging. The Oxford handbook of computational linguistics, 219-232 (2003).
13. Liaw, A., and Matthew W.: Classification and regression by randomForest. R news 2.3, 18-22(2002).
14. Blei, David M., Andrew Y. Ng, and Michael I. Jordan: "Latent dirichlet allocation." Journal of machine Learning research 3.Jan, 993-1022 (2003).
15. Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., & Moreau, Y.: A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. Journal of Computational Biology, 9(2), 447-464 (2002).
16. Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., M¨uhlh¨auser, M.: A multi-indicator approach for geolocalization of tweets. In: ICWSM (2013)
17. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th international conference onWorld wide web. pp. 761–770. ACM(2009).
18. Quercia, D., Ellis, J., Capra, L., Crowcroft, J.: Tracking gross community happiness from tweets. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. pp. 965–968. ACM (2012).

APPENDIX:

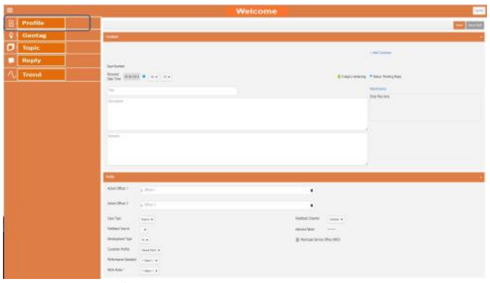
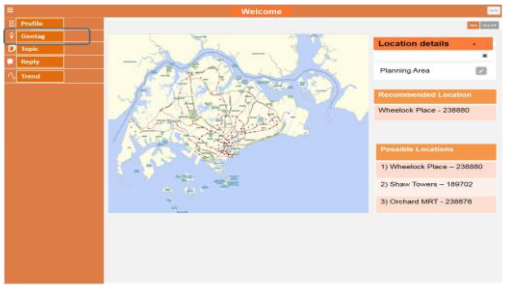


Figure A. Profile

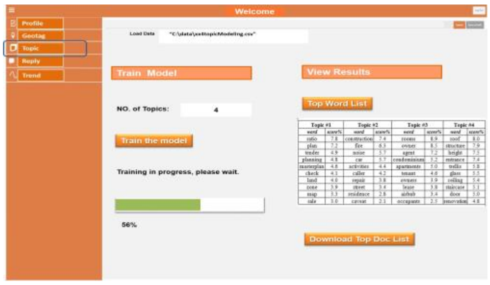


Planning Area

Wheelock Place - 238880

1) Wheelock Place – 238880
2) Shaw Towers – 189702
3) Orchard MRT - 238878

Figure B. Geotag



Load Data "C:\data\xxtopicModeling.csv"

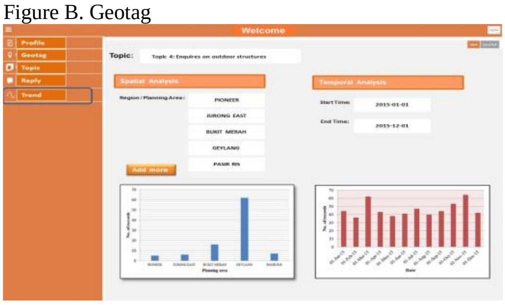
NO. of Topics: 4

Training in progress, please wait.

56%

Topic #1		Topic #2		Topic #3		Topic #4	
word	score%	word	score%	word	score%	word	score%
ratio	7.8	construction	7.4	rooms	8.9	roof	8.0
plan	7.2	fire	6.3	owner	8.5	structure	7.9
tender	4.9	noise	5.7	agent	7.2	height	7.5
planning	4.8	car	5.7	condominium	5.2	entrance	7.4
masterplan	4.8	activities	4.4	apartments	5.0	trellis	5.8
check	4.1	caller	4.2	tenant	4.6	glass	5.5
land	4.0	repair	3.8	owners	3.9	ceiling	5.4
zone	3.9	street	3.4	lease	3.8	staircase	5.1
map	3.3	residence	2.8	airbnb	3.4	door	5.0
sale	3.0	caveat	2.1	occupants	2.5	renovation	4.8

Figure C. Topic



Topic: Topic 4: Enquires on outdoor structures

Region / Planning Area: PIONEER, JURONG EAST, BUKIT MERAH, GEYLANG, PASIR RIS

Start Time: 2015-01-01

End Time: 2015-12-01

Figure D. Trend