Analysis of Author-Selected Keywords in Urban Planning and Urban Management Papers

Masafumi Ono[1] and Ryosuke Shibasaki[2]

[1] Earth Observation Data Integration and Fusion Research Initiative, the University of Tokyo, Cw503 Institute of Industrial Science 4-6-1 Komaba, Meguro-ku, Tokyo, Japan
[2] Center for Spatial Information Science, the University of Tokyo, Cw503 Institute of Industrial Science 4-6-1 Komaba, Meguro-ku,Tokyo, Japan
maono@iis.u-tokyo.ac.jp; shiba@csis.u-tokyo.ac.jp

Abstract

In this paper, we use a topic model called Labeled Latent Dirichlet Allocation (L-LDA), which is an extension of the LDA model often used for text mining, to analyze the keywords selected by authors in urban planning and urban management papers. We use keywords, sessions and authors related to the Computers in Urban Planning and Urban Management (CUPUM) conference as inputs for the model. We then evaluate the performance by comparing the training and target sets. The results are displayed using Web-based technologies. Our method extracts the characteristics of selected keywords, and reveals other relevant topics undetected by the author. Thus, our results can support writers and readers of research papers in the field of urban planning and urban management.

1. Introduction

The buzz term ‘big data’ currently refer to exponential data growth that is widespread. The term often appears in the field of data engineering, as well as other academic fields, such as urban data management (UDMS,

2013). Various mining theories and techniques for online documents and databases are constantly being applied. One consequent effect is that it has become commonplace to quickly access Web content of interest through services such as Google.

The topic model (Wagner, 2010) is one example of a text mining technique. It is a statistical model used, after being primed with a training set, to calculate the probability of a document's topic. Latent Dirichlet Allocation (LDA) (Blei, 2003) is a popular and standard topic model with several extensions often used for analyzing Web content and social media (Pennacchiotti, 2011) for the purpose of clustering, categorizing or tagging.

The use of the topic model and its variations in research is growing. They are adopted not only for Web content, but also for academic content, such as e-learning or journal databases. Sekiya (Sekiya, 2010) proposes a support tool for updating the curriculum in university education. Wang (Wang, 2011) provides an effective recommender system for scientific articles.

Currently, many research papers have been published. Then, the total quantities of published papers are increasing. So, when researchers access to large archives of scientific papers, it becomes more difficult to find their interesting papers than past. This tendency will continue in future. Thus, topic model comes to be used for promoting that researchers can effectively find information and knowledge of their interests.

Generally, the topic model has been found to work well in many cases where topics are extracted from text data. However, we cannot determine if it will work with any specific domain until the model is applied because, in many cases, there are specific conditions, such as a unique vocabulary used only in the domain. Therefore, this paper tries to analyze research papers in the domain of urban planning and urban management using L-LDA (Labeled Latent Dirichlet Allocation) which is an extension of the LDA model.

A research paper usually contains explicit relevant keywords which are designed to alert readers to the paper's contents. These keywords are generally considered suitable for the paper because they are selected by the author, but readers cannot judge which keywords represent the paper well and are unable to determine the suitability of each keyword until they have finished reading the entire paper. However, the topic model can improve

this situation as it is able to determine the probabilistic distribution of keywords in the paper. This would be useful information for readers. In addition, the topic model can reveal undetected keywords which are not directly written in the paper, but are relevant, through statistical analysis from other papers, which would be helpful for authors. Thus, we conduct an analysis regarding the topic model in this paper.

This paper is organized as follows. In section 2, we describe our methodologies and models used. In section 3, the experimental results are presented, after which, we evaluate the performance of L-LDA in terms of accuracy. In section 4, we display the results using Web-based technologies. We conclude in Section 5 and propose recommendations.

2. Methodology

In this section, we explain our methodology and the models used in this paper. An outline of our approach is shown in 2.1. Models that are relevant to our research are described in sections 2.2 to 2.4. Section 2.5 shows the method used on the experiments presented in the next section.

2.1. Outline

The purpose of this research is to reveal the characteristics in keywords selected by authors in the field of urban planning and management. To achieve this, we will progress through the following steps.

P1: Count the underlying statistics, such as the total number of documents, the number of documents per session, the number of authors, and unique keywords in papers from the CUPUM 2009 and 2011 conferences (CUPUM, 2009, 2011).

P2: Input the parameters given by P1 into L-LDA and an alternative model, the Naïve Bayes estimator. Then compare the accuracy of the results given by each model.

P3: Visually display the probabilistic distribution of keywords by document and author.

P4: Discuss and interpret the results.

2.2. Latent Dirichlet Allocation (LDA)

LDA is a general probabilistic model widely used as a benchmark topic model. The basic premise of LDA is that a document is treated as bag of words. LDA then assumes that a document consists of words which are each generated from one topic. The pre-processes of LDA include counting the number of words and vocabularies for all documents under several conditions.

The LDA model is structured as a three-level hierarchical Bayesian model. The parameters in the model are estimated though a training process. Gibbs sampling is a typical algorithm used for training in LDA. In Gibbs sampling, the probability $P(z_i)$ on a topic assignment z for the i-th word w_i in a document d for a topic j, is sampled by

$$P(z_i = j|z_{-i}, w_i, d) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{v=1}^{V} C_{vj}^{WT} + V\beta} \times \frac{C_{dj}^{DT} + \alpha}{\sum_{t=1}^{T} C_{dt}^{DT} + T\alpha}. \quad (1)$$

where $C_{w_i j}^{WT}$ is the number of times the word token w_i was assigned to the topic j across all documents, but does not include the current instance z_i. $\sum_{v=1}^{V} C_{vj}^{WT}$ is the number of times all other words were related with topic j. V is the size of the vocabulary of all documents. C_{dj}^{DT} is the number of times the topic j was already assigned to some words in the document d, but it does not include the current instance z_i. Further, $\sum_{t=1}^{T} C_{dt}^{DT}$ is the number of times all other topics were related with the document d. T is the size of the topic of all documents. α and β are hyper-parameters. In addition, it should be noted that the value of the probability is not normalized. Accordingly, when we require the distribution across all topics, we must perform a normalization for each score.

In LDA, we can estimate $\varphi_{w_i}^{(j)}$, which is the probability of word w_i used in topic j, and $\theta_j^{(d)}$, which is the probability of topic j in document d as

$$\varphi_{w_i}^{(j)} = \frac{C_{w_i j}^{WT} + \beta}{\sum_{v=1}^{V} C_{vj}^{WT} + V\beta}, \quad (2)$$

and

$$\theta_j^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{t=1}^{T} C_{dt}^{DT} + T\alpha}, \quad (3)$$

respectively. The probabilities for all topics are calculated through the learning process of Gibbs sampling. The probability for a topic *j* is then updated by the probabilities for all other topics excluding topic *j*. The model will find the optimal values after this updating process is repeated several times.

One problem with LDA is that the model does not have an obvious way to receive supervised information since LDA is an unsupervised algorithm. For example, even if we attempt to learn from some tagged Web documents through LDA, we cannot directly set the tag information in LDA. This problem is addressed in L-LDA described next.

2.3. Labeled Latent Dirichlet Allocation (L-LDA)

L-LDA is an extension of LDA (Ramage, 2009). In contrast to standard LDA, another parameter is added to the model to associate labels and sets for a topic. The use of this extension improves the method to the point that we can set any number of labels as supervised information into the model. L-LDA suits our purpose of evaluating the characteristics of keywords in a paper.

The probability of an *i*-th word w_i in a document *d* for a topic *j* using Gibbs sampling as the parameter estimation for L-LDA is

$$P\left(z_i = j \middle| z_{-i}, w_i, d, \alpha_j, T_j\right) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{v=1}^{V} C_{vj}^{WT} + V\beta} \times \frac{C_{dj}^{DT} + \alpha_j}{\sum_{t=1}^{T} C_{dt}^{DT} + T_j \alpha_j}. \quad (4)$$

Although similar to LDA, L-LDA is constrained to the set of possible topics of the observed labels. In other words, topics in the document are restricted to their own labels. T_j, therefore, indicates the amount of observed labels in the documents while α_j represents the Dirichlet parameter α, when the topic *j* associates with the *j*-th label.

In relation to LDA, the variable $\varphi_{w_i}^{(j)}$, which is the probability of a word for a topic, is the same. In contrast, the variable $\theta_j^{(d)}$, which is the probability of a topic for a document, is given by

$$\theta_j^{(d)} = \frac{C_{dj}^{DT} + \alpha_j}{\sum_{t=1}^{T} C_{dt}^{DT} + T_j \alpha_j}. \quad (5)$$

2.4. Naïve Bayes

Naïve Bayes (NB) is a traditional and typical classifier, which is also a statistic bag-of-words model. In the learning process, the NB learner creates an occurrence table of words bound to every class (topic) t_j from tokenized documents. This is used as the training set, *T,* for classification. In the classification process, the classifier takes the word occurrence from *T* and counts each word $w_i \in W$ in a target document *d*. In the principle of Multinomial Naïve Bayes (MNB), which is a basic model of NB, we estimate a suitable class using the estimating formula (Rennie, 2003) as follows:

$$argmax_t \left[\log P(t_j) + \sum_i f_i \log \frac{N_{w_i} + \alpha}{N_W + \alpha |V|} \right]. \quad (6)$$

where $P(t_j)$ is a prior class probability used to estimate the documents belonging to topic class t_j for all documents; f_i indicates the frequency of word w_i in target document *d*; N_{w_i} is the number of word w_i assigned to class t_j in the training set; N_W is the total number of all word occurrences assigned to class t_j in *T;* $|V|$ is the total number of vocabularies in training set *T*, except overlapped words; and α is a smoothing parameter.

NB is easy to apply and quick to execute. For this reason, it is often used as initial training for students or in applications to real-time analysis.

2.5. Evaluating Accuracy

In order to evaluate the performance of L-LDA, we use two sets of data, the conference papers from CUPUM 2009 and 2011. We conduct the following three types of experiment on the datasets.

E1: Our first experiment is designed to infer the correct topics of one year's dataset after using itself as the training set. This is to check whether the model outputs calculated from the papers are correctly assigned to each session. In other words, this accuracy reflects the essential reliability for

the model.

E2: Our second experiment is to infer the topics of one year's dataset as a test set after learning another year's dataset as a training set. This is to check whether the papers categorized in the session "Public Participants" in CUPUM 2009 are correctly assigned to the common session "Public Participants" in 2011. The accuracy of this case purely reflects the inferring ability of the model. In this experiment, a significant amount of common topics between two data sets are required.

E3: In our final experiment, we infer the topics of one year's data, set as the test set, after learning from both datasets. This is to check whether the papers categorized in a topic in each year are mapped to a suitable topic in the whole corpus. If each semantic boundary between topics is clearly separated, such as "Astronomy" and "Medicine", it is meaningful to check the classification accuracy mapped to the original topic. However, if the boundary is ambiguous and cannot be clearly separated such as "Urban Planning and Design" and "Planning Support System", evaluating the accuracy of the original mapping is not so meaningful. In this case, it is more important to extract the characteristics of the topics over the entire corpus instead of the accuracy of the original mapping. The main purpose of this experiment is to reveal the differences in statistics between the original topics and inferred topics over the entire corpus.

In all cases (E1 to E3), we evaluate the accuracy of the results of L-LDA by comparing with the results of the NB model. We use two measures to perform this evaluation.

The first is the F measure, which is a popular measure for information retrieval or classification (Lewis, 1995). The F measure represents the accuracy of classification and categorization between inferred results and the true topics. If the value of the F measure is high, the inferred topics are similar to the topics in the training data. The F measure is calculated as follows:

$$F = \frac{2RP}{R+P}, \quad (7)$$

where R is the recall given by A/(A+C), and P is the precision given by A/(A+B). Here, A is the number of documents which the model can assign to the correct topic, B is the number of documents which the model assigns to an incorrect topic in the inferred result, and C is the number of documents that the model fails to assign to the correct topic in the true result.

The second measure is called MP and is the average of the Maximum Probabilities of most high ranked topics in a document. This measure shows whether the model can extract a strong feature from a document. If this score is high, the model outputs are well-characterized. If the score is low, the bias for all topics in the document is small.

$$MP = \frac{1}{N}\sum_{N} P(bestTopic|document) \ . \quad (8)$$

3. Experiments

In this section we present results for the three experiments described above, using the CUPUM 2009 and 2011 conference papers. Section 3.1 shows the basic statistics from the CUPUM proceedings. Sections 3.2 to 3.4 show the accuracy when setting the model inputs per session, and the author and keywords as supervised information for the topic model.

In all experiments, the Dirichlet parameters for L-LDA are set as α = 0.01 and β = 0.01. For NB, the smoothing parameter is set to α = 0.01.

3.1. Basic statistics of the CUPUM conferences

The basic statistics of the digital data of the CUPUM 2009 and 2011 proceedings are given in Table 1.

Table 1. Original statistics of the CUPUM 2009 and 2011 proceedings

	2009	**2011**	**Total**
All titles	146	152	298
Valid titles	144	145	289
Sessions	26	20	45{1}

In the above table, the number in the curly brackets {} indicates the number of common items across both years. We define titles not considered valid as follows:

- The title exists, but there is no digital file for the title.
- The digital file exits, but we cannot retrieve text information from the file because it is not a Word or PDF file, but an image file.

- The digital file exits, but there is no title for it in the CUPUM program.

In addition, both CUPUM programs contained a "poster session", which we exclude from our analysis. This is because the poster session is not categorized and the documents are relatively smaller than documents in oral sessions. Accordingly, we use the statistics in Table 2, which exclude the poster session information. The values in round brackets () indicate the standard deviation of the average value.

Table 2. Basic statistics of the CUPUM 2009 and 2011 proceedings data used in this paper

	2009	**2011**	**Total**
Valid documents	140	123	263
Sessions	25	19	43{1}
Average number of documents in a session	5.60 (3.00)	6.73 (3.41)	6.23 (3.33)
Maximum number of documents in a session	14	17	17
Minimum number of documents in a session	3	4	3

Table 3 shows the frequency of words and vocabularies (set of unique words) that appear in a document. We estimate that the pure number of words in a sentence in a document is smaller than the number in the table because these numbers include some noise consisting of the following factors:

- Numerical values which do not have a meaning themselves and are used in the formula or result of the experiments described in the paper.
- No relevant information to the document, generated by the tool used for extracting text from the original PDF or Word file, such as page number, header or footer information.
- Specific characters which cannot be handled by our script or failed to be converted through the tool.

We use a simple filter to address some of these problems.

Table 3. Statistics of words and vocabularies in a document

	2009	2011	Total
Average number of words in a document	5889.6 (1985.2)	6088.2 (2331.0)	5982.5 (2156.1)
Maximum number of words in a document	13556	15205	15205
Minimum number of words in a document	1854	1599	1599
Average number of vocabularies in a document	2023.3 (551.1)	1920.3 (620.2)	1975.1 (586.7)
Maximum number of vocabularies in a document	4395	4146	4395
Minimum number of vocabularies in a document	985	659	659

Table 4 shows the statistics regarding the authors of the papers in the dataset "Authors" counts all members who are first author and co-authors in a paper. "First authors" counts only the number of first authors in the data set and does not contain co-authors. If an author has written multiple distinct papers, they are only counted once for this purpose.

The measures "Duplicate authors" and "Maximum frequency of duplicate authors" allow duplications for co-authors. The difference between "Valid documents" in Table 2 and "First authors" in Table 4 indicates the number of first authors who submitted several papers in the same year. The curly brackets represent the common data over both years, as before.

Table 4. Statistics of authors in CUPUM 2009 and 2011

	2009	2011	Total
Authors	326	281	557 {50}
Duplicate authors	35	40	97* {8}
Maximum frequency of duplicate authors	4	10	10
First authors	135	120	236 {19}

*There were a number of authors (=30) who were not duplicate authors in a single year, but contributed in both years and are hence duplicate authors over both years. Thus, the total value for Duplicate Authors was calculated as Duplicate Authors in 2009(35) + Duplicate Authors in 2011 (40) + Duplicate Authors over both years (30) – Authors who were duplicated in both years (8) i.e., 35+40+30-8=97.

From Table 4, we can determine the number of common first authors over both years (19), indicating that about 14 % of participants in CUPUM 2009 contributed to CUPUM 2011.

As an example, Table 5 shows a list of top ranked authors who appeared more than four times in the merged CUPUM 2009 and 2011 data. The list

contains co-authors.

Table 5. The top ranked authors in CUPUM 2009 and 2011

Rank	Author	Counts
1	JD Hunt	10
2	David Simmonds	6
	Kazuaki Miyamoto	
	Ryosuke Shibasaki	
5	Akiko Kondo	4
	Carlo Ratti	
	Dick Ettema	
	John Abraham	
	Keiichi Kitazume	
	Margaret Horne	
	Nao Sugiki	
	Qingming Zhan	
	Shin Yoshikawa	
	Stan Geertman	
	Varameth Vichiensa	

The final set of statistics calculated in the pre-process is the selected keywords for each paper in the CUPUM proceedings data and is shown in Table 6. As can be seen, duplicate keywords comprise less than 12% (=106/888) of all vocabularies. In other words, most keywords are unique.

Table 6. Statistics of the selected keywords in CUPUM 2009 and 2011 proceedings papers

	2009	2011	Total
Valid documents	138	112	150
Unique keywords	509	440	888 {61}
Duplicate keywords	50	40	106* {14}
Max frequency of duplicate keywords	19	9	27

* There were a number of keywords (=30) who were not duplicate keywords in a single year, but contributed in both years and are hence duplicate keywords over both years. Thus, the total value for Duplicate Keywords was calculated as Duplicate Keywords in 2009(50) + Duplicate Keywords in 2011 (40) + Duplicate Keywords over both years (30) – Keywords which were duplicated in both years (14) i.e., 50+40+30-14=106.

We list the top ranked keywords from the data in Table 7 and Table 8. As shown, the keywords that were also used in the session name such as “Decision support systems” in 2009 or “Location choice” in 2011 often ranked highly.

Table 7. The list of top ranked popular keywords in CUPUM 2009

Rank	Keywords	Counts
1	GIS	19
2	Land use	8
3	Urban planning	7
4	Decision support system	6
	Planning support systems	
6	Remote sensing	5
7	Cellular automata	4
	Simulation	
	Sustainability	
	Visualization	

Table 8. The list of top ranked popular keywords in CUPUM 2011

Rank	Keywords	Counts
1	GIS	9
2	Location choice	6
	Visualization	
	Urban planning	
5	Cellular automata	5
6	Spatial analysis	4
	Land use	
8	Built environment	3
	Microsimulation	
	Modeling	
	Planning support systems	
	Simulation	
	Transportation planning	

3.2. Evaluating the accuracy of session information

In this section, we present the experimental results when inputting session information to the L-LDA and NB classifiers as supervised information.

A paper in the CUPUM program is assigned to a single session. However, to apply a single label to L-LDA is generally not recommended due to the performance of L-LDA. It is accepted that the outputs are similar to NB. Therefore, in this experiment we added the year “2009” or “2011” as a class for all papers.
In addition, there was only one common session as seen in Table 2. If we have too few samples, calculating the F measure has no meaning.

Table 9. The list of all sessions

	2009	2011	Common
Number	25	19	1
Names	Spatial Analysis: Methodology, Visualizing Sustainable Planning, Environmental Planning & Policy, Land Use and Transport, Urban Modelling, Spatial Analysis, CAD & Visualization, Remote Sensing, Planning Support System, Urban Planning and Design, Travel Demand Analysis, Agent Based Model: Urban Growth, Geocomputation, Planning and Urban Space, Agent Based Model: Pedestrian, Spatial Analysis: Public Facility, Urban Management Systems, Transport & Logistics, Agent Based Model: Economic Activity, Public Participation Disaster Mitigation and Evacuation Survey & Data Collection Decision Support System Safety & Health Care Transport Planning	Cellular Automata, Decision Support Systems, Land Use Forecasting, Land se Modelling, Location Choice, Measuring Built Form, Pedestrians and Local Services, Public Participation, Public Transit, Remote Imaging, Residential Location Choice, Simulation, Statistical Methods, Sustainable Transport, System Analysis, Transport Demand Modelling, Transportation Planning, Urban Planning, Visualization and Animation	Public participants

Table 9 lists all sessions in the CUPUM programs. In the table, there are several similar yet distinct sessions between years. For example, "Decision Support System" in 2009 and "Decision Support Systems" in 2011 are almost identical. As another example, we could regard "Transport Planning" in 2009 and "Transportation Planning" in 2011 as the same topic semantically.

Therefore, in this session-based experiment, we assume several similar sessions are the same in order to conduct our second experiment, E2. The similar sessions are shown in Table 10. For example, when inferring a paper classified as "CAD & Visualization" in 2009, if it is mapped to "Visualization and Animation" in the 2011 classification, we regard this as a correct answer when calculating the F measure for experiment E2.

Using the assumptions of Table 10, the result of each experiment for session classification is given in Table 11. We calculate the distribution for all topics based on the supervised information in the training set. The value

for each represents the average performance on the test documents. The value in the circle brackets represents standard deviation. Through the three experiments, we found that both models performed well when the training set was learned from the entire corpus including the test set. In contrast, the performance for E2 was not as good As the values of MP in L-LDA are higher than those in NB, it is clear that L-LDA characterizes the inferred results more accurately than NB in this research.

Table 10. Mapping of similar sessions between CUPUM 2009 and CUPUM 2011 programs

2009	**counts**		**2011**	**counts**
Cad & Visualization	12	≒	Visualization and Animation	8
Decision Support System	8	≒	Decision Support Systems	9
Remote Sensing	3	≒	Remote Imaging	4
Transport Planning	3	≒	Transportation Planning	4
Urban Planning and Design	3	≒	Urban Planning	4
Public Participation	3	=	Public Participation	9
Total	32		Total	38

Table 11. Result of accuracy measures using session classification as supervised information

					L-LDA		NB	
	Train	Infer	N	T	F	MP	F	MP
E1.	2009	2009	140	25	100.0(0.00)	97.04(5.51)	100.0(0.00)	23.05(3.99)
	2011	2011	123	19	100.0(0.00)	95.71(7.59)	100.0(0.00)	27.44(5.82)
E2.	2009	2011	32	6	16.17(13.80)	30.75(8.54)	13.62(15.87)	13.62(15.87)
	2011	2009	38	6	15.90(22.53)	26.73(22.53)	15.0(21.40)	9.40(1.72)
E3.	All	2009	140	25	100.0(0.00)	97.75(4.55)	100.0(0.00)	13.87(2.75)
	All	2011	123	19	100.0(0.00)	97.07(6.36)	100.0(0.00)	12.09(2.78)

Train: a training set for the topic model, Infer: a test set used for inferring, N: number of document files, T: number of topics, F: the result of the F measure (%), MP: the result of MP (%), All: merged CUPUM 2009 and 2011 data sets.

3.3. Evaluating the accuracy of author information

In this section, we evaluate the accuracy when we input author information as supervised information to the models. For E1, we use the 326 authors in 2009 as well as the 281 authors in 2011, given in Table 4. For E2, we use the 50 common authors over 2009 and 2011 referred by Authors-Total cell in Table 4. In this experiment, there is no specific limitation like that seen in section 3.1. For E3, we use all 557 authors as the training set. When there are three co-authors in a paper, if all of them are mapped to anywhere in the top three highest probabilities in the ranking of all authors, we regard that as true (In other words, we do not regard the order of the co-authors like first or second author in this experiment).

We can roughly state that the scores of E2 in Table 12 are higher than E2 in Table 11, although this case handles a greater number of topics. We suspect one reason is due to the session name being a generic word, yet the author name is a proper noun so it tends to be clearly characterized more easily.

Although NB is generally known to suffer from poorer performance than some more sophisticated models, it worked better than we expected. However, all probabilities per topic in NB remain at a similar level and it is then difficult to characterize them, since the MP score is low.

Table 12. Result of accuracy measures when the author is used as the supervised information

					L-LDA		NB	
	Train	Infer	N	T	F	MP	F	MP
E1.	2009	2009	140	326	96.20(17.14)	40.03(38.96)	99.83(2.26)	2.16(0.39)
	2011	2011	123	281	93.51(22.54)	40.42(34.65)	95.65(19.03)	1.57(0.36)
E2.	2009	2011	42	50	34.21(38.30)	21.84(11.57)	12.89(29.29)	1.27(0.16)
	2011	2009	42	50	37.21(41.05)	20.28(8.06)	17.40(33.22)	0.95(0.152)
E3.	All	2009	140	326	88.18(30.27)	42.74(40.01)	96.11(18.96)	0.72(0.12)
	All	2011	123	281	86.32(32.69)	42.96(37.60)	88.99(30.86)	0.72(0.13)

3.4. Evaluating the accuracy of keywords

As a final experiment to evaluate the accuracy, we input the keyword information as supervised information to the models. For the case of E1, we use the 509 keywords in 2009 as well as the 440 keywords in 2011, given in Table 6. For E2, we use the 61 common keywords between 2009 and 2011. Again, in this experiment, there is no limitation like that seen in 3.1. For E3, we use all 888 keywords.

Table 13. Result of accuracy measures when keywords are used as supervised information

					L-LDA		NB	
	Train	Infer	N	T	F	MP	F	MP
E1.	2009	2009	138	509	87.67(1.42)	25.18(29.28)	99.39 (7.66)	1.40 (0.22)
	2011	2011	112	440	87.96(31.01)	24.16(28.80)	100.0(0.0)	1.52(0.29)
E2.	2009	2011	66	61	11.66(22.87)	13.86(8.44)	3.29(11.66)	0.96(0.08)
	2011	2009	76	61	11.68(20.99)	13.91(6.60)	6.85(20.77)	0.91(0.11)
E3.	All	2009	138	509	74.79(41.95)	28.25(31.54)	76.41(42.18)	0.72(0.13
	All	2011	112	440	69.75(44.06)	28.82(31.97)	76.41(42.18)	0.72(0.13)

In general, the values in this experiment are lower than the other two experiments. This reason is obvious since the experiment handles a greater number of topics for the number of documents than the previous experiments. This is a notable result as we found the topic model can work better for academic papers than we expected.

An interesting result of this experiment is seen from the scores of E3 where about 70 % of F measures are marked. This means that 70 % of the keywords inferred from the entire corpus are statistically equal to keywords selected by the authors. In addition to this, 30 % of them would be selected from other vocabularies. Since these scores represent statistics for all documents, the details for an individual document are not presented in this result. In the next section we will display the probability distribution by using the results of E3.

4. Visualization

In this section, we visually display the probability distribution of keywords, mainly given by the result of section 3.4.We use a radar chart and tag-cloud to show the results. These implementations are realized by Web-based technology including HTML 5, CSS and Javascript. Through the use of Web-based technology, it is simple to provide our results for many communities including CUPUM.

4.1. Visualizing the probability distribution of keywords per a paper

In section 3.4, we calculated the probability distribution of all keywords per paper. This calculation was performed for all papers, after which we separated the output to the keywords selected in the paper and any others. The probability distribution is shown in Figure 1 through a radar chart. A probability distribution with the top six ranked keywords is shown in Figure 2. Each axis in the chart represents the probability for a keyword. In this implementation, we used a Javascript library (HTML5.jp, 2012).

In this visual example, we use the paper titled “Nowcast of Urban Population Distribution using Mobile Phone Call Detail Records and Person Trip Data” written by Horanont (Horanont, 2011). In his paper, he selected five keywords. We find the most characterized keyword is “Ubi GIS”. The probabilities for “Population dynamics” and “Mobile simulation” do not take zero values, yet are too small to be seen on the chart.

Figure 2 shows six undetected keywords which Horanont did not select. The keywords were inferred from the whole corpus. These undetected probabilities are much smaller than the selected keywords. In other words, this may indicate that the author selected suitable keywords for their paper.

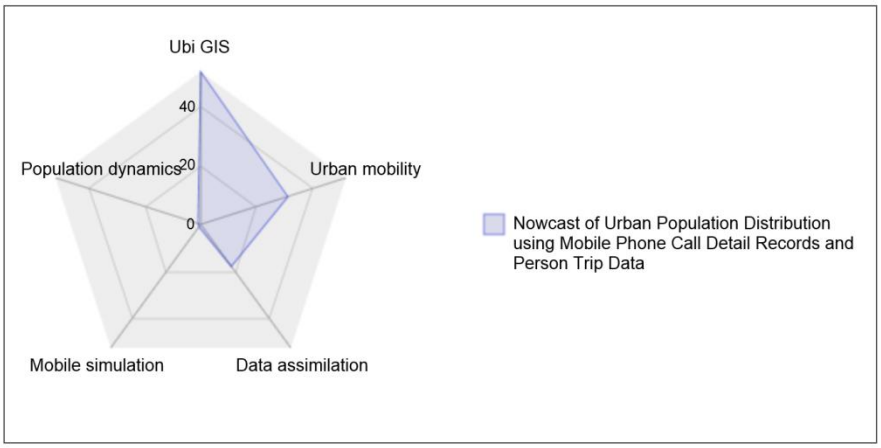

Fig. 1. An example of the probability distribution of selected keywords for a paper

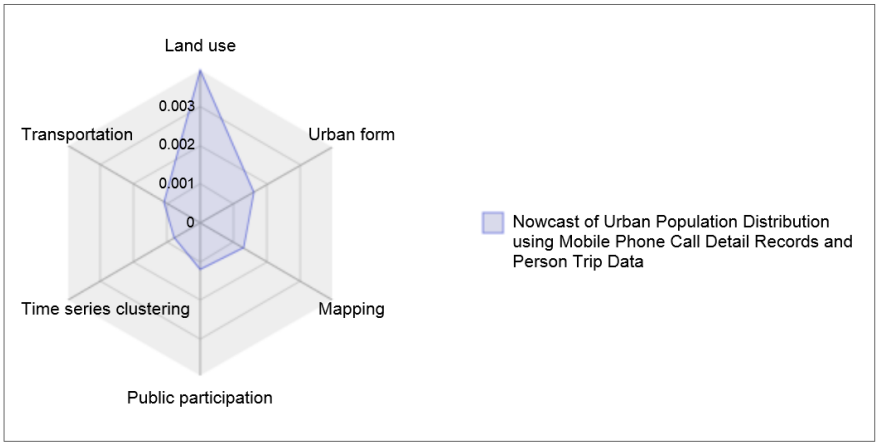

Fig. 2. An example of the probability distribution of undetected keywords for a paper

4.2. Visualizing the probability distribution of keywords for an author

In this section, we visualize the probability distribution of keywords for an author. From the distribution of a document, we can retrieve the distribution for an author with the following calculation:

$$\theta_a = \frac{1}{N}\sum_n P_n^{(a)}(keyword|doc)\ , \quad (9)$$

where $P_n^{(a)}(keyword|doc)$ represents the probability of the *n*-th paper related to an author a and *N* is the number of papers related to the author a. From θ_a, we can calculate the probability distribution of keywords for each author.

We show the distribution of the top six ranked keywords related to an author in Figure 3, using Eq. (9). Figure 4 shows the distribution of undetected keywords. "Ryosuke Shibasaki" is one of the second ranked authors in Table 5. His name appeared in six papers as co-author. The figure shows the overall tendency merged from all his papers.

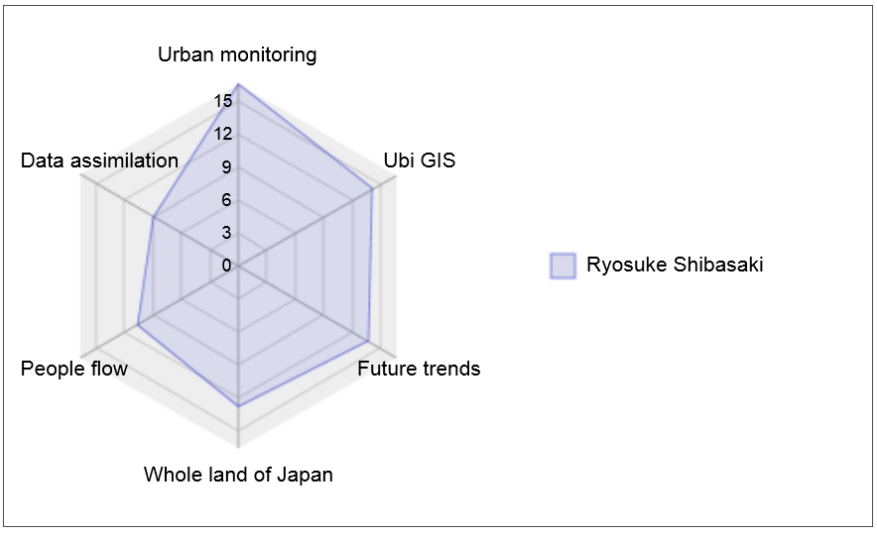

Fig. 3. An example of the probability distribution of keywords per author

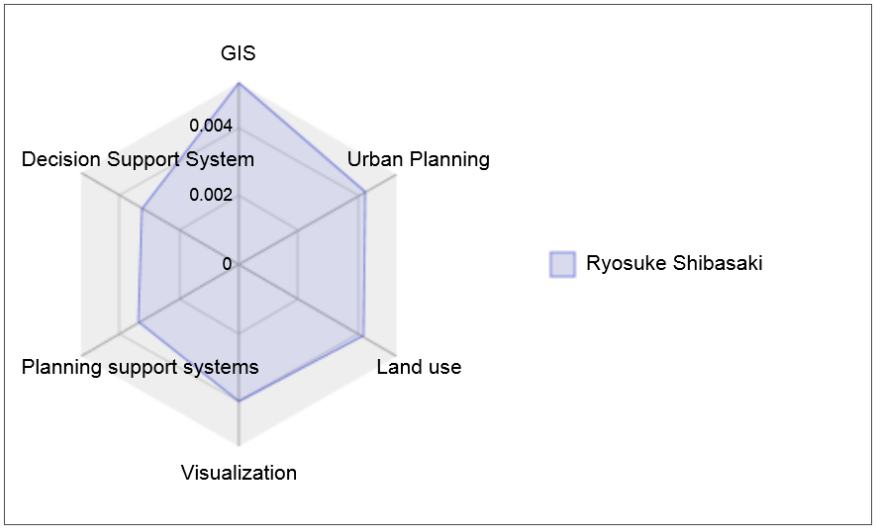

Fig. 4. An example of the probability distribution of undetected keywords per author

4.3. Visualizing the relevant authors

The distribution per author in 4.2 enables us to calculate the similarity between one author and another. The similarity is calculated using the Kullback–Leibler divergence, which is a popular index for similarity (Bigi 2003). The formula is given as

$$(\theta_{a_1} \| \theta_{a_2}) = \sum_i \theta_{a_1 i} log \frac{\theta_{a_1 i}}{\theta_{a_2 i}}, \quad (10)$$

where θ_{a_1} is the keyword probability distribution for author a_1 and θ_{a_2} is the keyword probability distribution for author a_2. This is not a symmetric formula, i.e., it only represents the similarity a_1 to a_2 and not the reverse.

If the calculation of an author against all authors is computed, all ranks are determined. Figure 5 shows an example of the top ranked relevant authors similar to "Ryosuke Shibasaki" through a tag-cloud interface. A JQuery plugin (Ongaro, 2013) was used in the implementation.

Ryosuke Shibasaki

Fig. 5. A sample of the relevant authors on an author through tag-cloud

5. Conclusion and recommendation

In this paper, we analyzed sessions, authors and particularly keywords selected by each author in the CUPUM community using L-LDA. Through visual representations of our analysis, we showed several potential applications. We believe that our results can contribute to the readers as well as writers of research papers published in CUPUM.

This research is strongly driven by the available dataset of the conferences. In the era of big data, this type of statistical approach will be more important in the future.

In future work, we intend to develop the extended vocabulary database or ontology in the field of urban planning and management from this result. Furthermore, by improving the accuracy of similarity between authors, an extended recommendation or reviewer matching system may be possible.

References

Bigi, B. (2003) Using Kullback-Leibler Distance for Text Categorization, In Proceedings of the ECIR-2003, vol. 2633 of Lecture Notes in Computer Science, pp. 305-319, Springer-Verlag.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003) Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol.3, pp.993-1022

CUPUM (2009), 11th International Conference on Computers in Urban Planning and Urban Management (CUPUM) Homepage http://www.dupad.hku.hk/cupumhk/, Last date accessed 02.2013.

CUPUM (2011), 12th International Conference on Computers in Urban Planning and Urban Management (CUPUM) Homepage http://cupum.hbaspecto.com/CUPUM_2011/Home.html, Last date accessed 02.2013.

HTML5.jp (2012), HTML5.jp Homepage http://www.html5.jp/library/graph_radar.html, Last date accessed 02.2013.

Lewis, D. (1995). Evaluating and optimizing autonomous text classification systems, In Proceedings of SlGIR-95, pp. 246-254.

Ongaro, L. (2013), JQCloud Homepage, http://www.lucaongaro.eu/demos/jqcloud/, Last date accessed 02.2013.

Pennacchiotti, M., (2011) Investigating Topic Models for Social Media User Recommendation, WWW 2011, Hyderabad, India, ACM 978-1-4503-0637-9/11/03.

Ramage, D. , Hall, D., Nallapati, R. , Manning, C. D. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled cor-pora, Proceedings of the 2009 Conference on Empirical Methods in Natu-ral Language Processing: Vol 1, pp. 248–256

Rennie, J.D.M. , Shih, L., Teevan, J. and Karger, D.R. (2003) Tackling the Poor Assumptions of Naive Bayes Text Classifiers, In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC.

Sekiya, T., Matsuda, Y., Yamaguchi, K. (2010) Development of a Curriculum Analysis Tool, ITHET 2010, 9th International Conference on Information Technology Based Higer Education and Training, pp. 413-418, Cappadocia, TURKEY.

Horanont, T., Shibaski, R. (2011) Nowcast of Urban Population Distribution using Mobile Phone Call Detail Recordsand Person Trip Data, 11th International Conference on Computers in Urban Planning and Urban Management (CUPUM), reference number 266.

UDMS (2013), Urban Data Management Society Homepage, http://www.udms.net/, Last date accessed 02.2013.

Wagner, C. (2010) Topic Models, Slide Share http://www.slideshare.net/clauwa/topic-models-5274169, Last date accessed 02.2013.

Wang, C. and Blei, D. M. (2011) Collaborative topic modeling for recommending scientific articles, Proceedings of the 17thACM SIGKDD international conference on Knowledge discovery and data mining, pp. 448–456.