Topic models for conference session assignment: Organising PRASA 2014(5)

Michael Burke
Mobile Intelligent Autonomous Systems
Modelling and Digital Science
Council for Scientific and Industrial Research
South Africa
Email: mburke@csir.co.za

Deon Sabatta
Mobile Intelligent Autonomous Systems
Modelling and Digital Science
Council for Scientific and Industrial Research
South Africa
Email: dsabatta@csir.co.za

***Abstract*—Conference scheduling and organisation is a particularly laborious task and can be extremely time consuming. While many online conference platforms allow manual topic selection, these can be expensive and typically still require that individual papers be scanned and labelled appropriately before being assigned to reviewers and relevant conference tracks or sessions. This paper shows how the bulk of this process can be automated using topic models. Latent Dirichlet allocation is applied to learn conference topics directly from documents, and a clustering algorithm introduced to separate these into suitably sized conference sessions, determining an appropriate session topic in the process. Conference tracks can then be scheduled by maximising the distance between these session topics, thereby avoiding potential topic conflicts in parallel tracks.**

I. Introduction and related Work

Conference organisation is particularly time consuming, and the process of allocating papers to sessions can be exhausting, particularly in extremely large, multi-track conferences. Here, conference organisers typically rely on multiple area chairs, each overseeing a particular topic, but this introduces numerous coordination problems, and a reliance on notoriously disorganised academics. Commercial conference scheduling aids like shdlr.com [1] can speed up this process, but generally operate on a drag and drop basis, requiring manual session and track entry.

In an attempt to avoid this laborious task for PRASA, this paper describes an automatic paper allocation approach, which relies on topic models to assign papers to sessions. Our approach operates directly on papers, learns appropriate topics using latent Dirichlet allocation (LDA) [2], and then assigns papers to topics so as to minimise the distance between paper topic distributions within sessions.

Topic modelling is a well established approach to natural language processing that aims to discover themes in large collections of documents. Here, documents are typically modelled as mixtures of topics, each of which contains a vocabulary of words (Figure 1). A number of effective topic modelling techniques have been introduced, most of which extend LDA, the most popular approach to topic modelling. For example, Ramage et al. [3] have proposed a semi-supervised LDA to characterise microblogs and Li and McCallum [4] introduced the Pachinko allocation model, which finds correlations between topics in documents using a directed acyclic graph, while Teh et al. [5] propose a hierachical Bayesian model that allows groups of data to be described by coupled Dirichlet processes. Buntine and Mishra [6] have shown that topic modelling using these approaches is both rapid and efficient and can be implemented in parallel. However, despite the numerous extensions to LDA topic modelling, LDA remains ubiquitous across many applications.

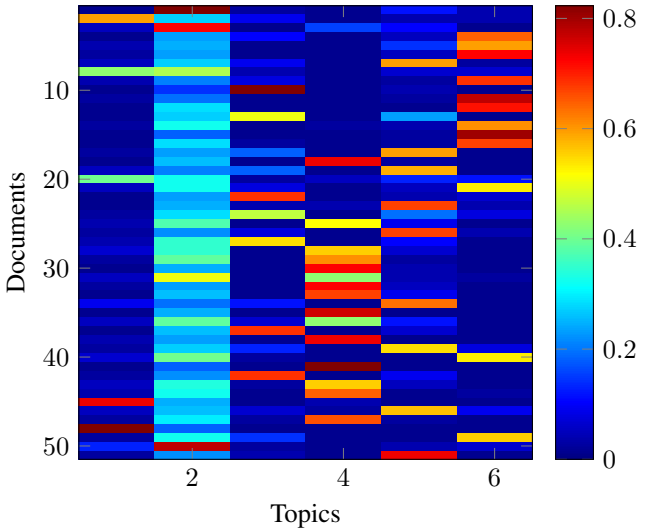

Fig. 1. Documents are formed by a combination of topics and topics are represented with varying probability in each document.

Topic modelling techniques are often tested on academic articles or proceedings [7, 8, 9] and provide excellent results suitable for end user applications. For example, JSTOR [10] uses optical character recognition and topic modelling to index documents [11]. However, to the best of our knowledge, topic modelling has yet to be used for conference scheduling, presumably due to the disconnect between the topics found and conference session allocation.

This paper is organised as follows. Section II introduces latent Dirichlet allocation, while Section III describes our approach to conference session assignment along with results when the algorithms are applied to the PRASA 2014 proceed-

ings. Section IV discusses some of the benefits of the proposed approach, and finally, conclusions are presented in Section V.

II. Latent Dirichlet allocation

Latent Dirichlet allocation [2] is a generative model of documents that is frequently used to find topics from a corpus of documents. Documents are treated as bags of words, drawn from a variety of topics. Let θ_i be the distribution of topics in document i and ϕ_k the distribution of words for topic k. LDA assumes that the M documents, each containing N_i words, in a given collection can be formed using random mixtures over K latent topics, with each topic described by a distribution over words.

Topic and word distributions are assumed to arise from Dirichlet distributions parametrised by α and β respectively,

$$\theta_i \sim \text{Dir}(\alpha), \tag{1}$$

$$\phi_k \sim \text{Dir}(\beta). \tag{2}$$

The j-th word in a corpus is drawn by first choosing a document topic using a categorical distribution with respective event probabilities described by θ_i,

$$Z_{i,j} \sim Cat(\theta_i), \tag{3}$$

and then drawing a word from a second categorical distribution, with event probabilities corresponding to the relevant word distribution for the sampled document topic $\phi_{z_{i,j}}$,

$$W_{i,j} \sim Cat(\phi_{z_{i,j}}). \tag{4}$$

The joint probability of this generative process is

$$p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \alpha, \beta) = \prod_{k=1}^{K} p(\phi_k; \alpha) \prod_{i=1}^{M} p(\theta_i; \beta) \prod_{j=1}^{N_i} p(Z_{i,j}|\theta_i) p(W_{i,j}|\phi_{z_{i,j}},) \tag{5}$$

and the various distributions forming this can be learned using Bayesian inference. Here, $\mathbf{W}$ is an $M \times N$ matrix of word identities, with N the total number of words in the corpus vocabulary, and $\mathbf{Z}$ an N dimensional vector of topics corresponding to each word in the vocabulary. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ refer to a matrix of document topic distributions and word topic distributions respectively.

Unfortunately, the distributions in (5) cannot be determined in closed form, and a numerical approximation is required. A variational Bayes approximation was used in [12], while Minka and Lafferty [13] apply expectation propagation and Griffiths and Steyvers [14] apply collapsed Gibbs sampling. The latter is used here, as this provides an unbiased estimate of the distributions of interest after an initial burn in period.

III. Assigning documents to sessions

The goal of this work is to learn topics directly from papers submitted to a conference and then use these to assign papers to conference sessions. Unfortunately, while LDA is extremely effective at finding topics in a corpus of documents, these topics are not immediately of use in conference session assignment.

TABLE I
Topics found for PRASA/RobMech/AfLat proceedings

Topic 1	Topic 2	Topic 3
learn	based	image
training	time	method
programming	model	camera
network	system	pixel
genetic	process	filter

Topic 4	Topic 5	Topic 6
language	feature	robot
speech	hand	control
word	classification	manufacturing
translation	face	system
model	recognition	design

Table I shows the top ranked words that were found for each topic (Somewhat arbitrarily, 6 were used here) when LDA [15] was applied to the PRASA/RobMech/AfLat 2014 proceedings. Training took place on the full corpus of submitted papers (103 papers), but only accepted papers were assigned to sessions. Common English words were removed from the corpus using the collection of stopwords available at [16]. It is clear that the LDA topic model managed to discover the main conference themes; robotics, machine learning, speech and machine translation, computer vision and image processing.

However, papers are mixtures of these topics and as result cannot be assigned directly to a single topic. This is illustrated in Figure 1, which shows the distribution over topics for each paper in the proceedings. While papers could be assigned to the most probable topic for a given document, this could cause errors in the case of application papers, which are typically distributed across multiple topics. In addition, certain topics may contain a larger number of papers. This is common at PRASA, where speech and machine translation papers typically outnumber those in other fields.

Furthermore, there is no guarantee that the number of papers assigned to a given topic in this way will correspond to the required number of conference sessions or tracks. For example, a typical conference consists of three of four 90 minute sessions per day, each comprising 6 papers, and the best papers assigned to each of the 6 topics listed above are highly unlikely to be neatly divisible in this way.

Instead, we allocate papers to sessions so as to minimise the mean of the average sum of distances between the topic distributions of papers assigned to a given session across all sessions,

$$C = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{G_k^2} \sum_{i=1}^{G_k} \sum_{j=1}^{G_k} D(i,j). \tag{6}$$

Here, K denotes the number of conference sessions, G_k the number of papers to be presented in session k, and $D(i,j)$ the Euclidean distance between topic probability distributions for papers i and j in session k.

Referring back to Figure 1, it is clear that topic 2 is something of a catch-all topic, describing words that are common across all papers. We prefer to remove this topic when allocating documents to sessions, as it provides very

little information that aids in paper separation, and can result in an undesirable catch-all session. Catch-all topics like this are inevitable in topic modelling, and are easily detected by selecting the topic with the largest sum probability over all documents.

The cost C is minimised in a brute force manner, by initially assigning papers to sessions at random without replacement, and then iterating over all sessions, testing whether swapping a paper in a given session with a paper in another session will result in a reduced cost and exchanging papers if this is the case. This process is repeated a number of times, with different starting points. Algorithm 1 illustrates this more clearly.

Algorithm 1 Paper allocation

```
C_best = Inf
C = C_best
for iter = 1 : MaxIter do
  for k = 1 : K do
    m_k = G_k papers drawn without replacement
  end for
  loop
    C_p = C
    for a = randperm(N) do
      for b = 1 to setdiff(1 : N, a) do
        swap m(a) and m(b)
        C_t = (1/K) Σ_{k=1}^{K} (1/G_k^2) Σ_{i=1}^{G_k} Σ_{j=1}^{G_k} D(i, j)
        if C_t < C then
          C = C_t
        else
          swap m(a) and m(b)
        end if
      end for
    end for
    if C_p = C then
      break loop
    end if
  end loop
  if C < C_best then
    C_best = C
    m_best = m
  end if
end for
```

Figure 2 shows the topic distributions when the 2014 PRASA papers are grouped in this manner. Here, we used a schedule of 7 sessions comprising 5 papers, 4 sessions containing 3 papers and a single session of 4 papers, in line with the 2014 PRASA schedule. It is clear that the papers in Figure 2 are successfully clustered by similarity.

After papers have been assigned to sessions, a set of words that best describe each session is desired. Let $p(T|D)$ be the probability of a topic being present given document D, and $p(W|T)$ be the probability of a given word being observed given topic T, both obtained using LDA. The probability of words being observed in a given session can be obtained by first marginalising to find the average distribution over topics for session k,

$$p(T^k) = \sum_{i=1}^{G_k} p(T_i^k|D_i^k)p(D_i^k) = \frac{1}{G_k}\sum_{i=1}^{G_k} p(T_i^k|D_i^k), \quad (7)$$

and then using this to determine an appropriate distribution over words for session k,

$$p(W^k) = \sum_{j=1}^{K} p(W^k|T_j^k)p(T_j^k) = \sum_{j=1}^{K} p(W^k|T_j^k)\frac{1}{G_k}\sum_{i=1}^{G_k} p(T_i^k|D_i^k). \quad (8)$$

Selecting a subset of words for which $p(W^k)$ is greatest provides a set of keywords to describe the session, which can be used to guide the selection of the session title.

While the evaluation of topic model performance is often subjective [17], a comparison between the session assignment found by the proposed approach and that actually used at the conference is useful for performance evaluation. Table II shows paper titles and PRASA 2014 session titles for each session allocation the proposed approach made. It is clear that papers are clustered into suitably similar groupings, but, as expected, the clusters do not quite match those assigned manually. The conference session assignment that was used at the conference is shown in the second column and colour coded according to the manually selected topic a paper was allocated to. Of the automatic paper allocations, almost all seem acceptable, and in many cases groupings appear more sensible that those assigned by hand (Session 4).

The session of most concern is Session 3, which appears to contain somewhat loosely related papers. This grouping appears to have been made because none of these papers seem to fit in a specific topic, and the content appears to be well distributed across all topics. This is clearly exhibited in Figure 2, which shows that the topic probabilities for many of the papers assigned to Session 3 are relatively low. As a result, the algorithm appears to have created an applications session of its own, and assigned 'left-over' papers to this.

Figure 3 shows a trace of the average assignment cost over 30 execution iterations, roughly equivalent to 5 minutes of execution time. Shaded areas indicate standard deviations in cost (the experiment was repeated 100 times). Empirical analysis shows that a cost of less than 0.105 provided reasonable session assignments for this corpus. The results presented above corresponded to an overall assignment cost of 0.094, which was obtained after approximately 8 hours of execution time.

IV. Useful algorithm by-products

The proposed approach to paper session allocation has some useful by-products. These are briefly discussed below.

TABLE II
PAPERS ASSIGNED TO SESSIONS

Paper	Session	Title
	Session 1: language speech word translation model	
29	3B: Speech Processing	Number pronunciation in a multilingual environment ...
32	4B: Natural Language Processing	Unsupervised Topic Modelling on South African Parl ...
33	5A: Natural Language Processing	An investigation into Spoken Audio Topic Identific ...
44	5A: Natural Language Processing	Experiments with syllable-based Zulu-English machi ...
47	4B: Natural Language Processing	Exploring unsupervised word segmentation for machi ...
	Session 2: robot control manufacturing system design	
5	2B: Robotic Case Studies and Applications	Improvements on a Prosthetic Hand - The UKZN Touch ...
14	2B: Robotic Case Studies and Applications	The Case for a General Purpose First Response Resc ...
21	4A: Robotics Sensing and Design	Development of a Two-Wheel Balancing Robot using t ...
40	4A: Robotics Sensing and Design	Development of an Educational Process Control and ...
49	4A: Robotics Sensing and Design	Kinematics Analysis and Workspace Investigation of ...
	Session 3: image feature method detection camera	
1	5B: End User Applications and Systems	Interactive Energy Consumption Visualization ...
50	5B: End User Applications and Systems	Application of Multi-Objective Local Search to Har ...
3	4B: Natural Language Processing	SVM Classification of Dr Math Microtext ...
13	2A: Classifiers AI Machine Learning and Related Topics	Comparison of two detection algorithms for spot tr ...
24	4A: Robotics Sensing and Design	A vision-based error metric for path following con ...
	Session 4: learn training programming network genetic	
2	5B: End User Applications and Systems	A Novel Approach to Visual Password Schemes Using ...
8	1A: Natural Language Processing	Context-based Online Policy Instantiation for Mult ...
20	5A: Natural Language Processing	South African Sign Language Dataset Development an ...
45	4B: Natural Language Processing	Genetic Programming for Password Cracking. Phase O ...
48	3A: Image Processing Classifiers and Related Topics	A Comparative Study of Genetic Programming and Gra ...
	Session 5: image method camera pixel filter	
42	6B: Image Processing	Generation of Super-Resolution Stills from Video ...
22	3A: Image Processing Classifiers and Related Topics	A study on the sensitivity of photogrammetric came ...
10	2A: Classifiers AI Machine Learning and Related Topics	Retinal Vessel Segmentation Based on Difference Im ...
27	2A: Classifiers AI Machine Learning and Related Topics	Automatic infarct planimetry by means of swarm-bas ...
37	2A: Classifiers AI Machine Learning and Related Topics	A two-Stage Fuzzy c-Means Clustering Algorithm for ...
	Session 6: language speech word translation model	
25	4B: Natural Language Processing	Comparing Support Vector Machine and Multinomial N ...
28	5A: Natural Language Processing	Phrase chunking for South African languages: an in ...
43	1A: Natural Language Processing	Topic Models for Short Text ...
31	5B: End User Applications and Systems	Performance analysis of a multilingual directory e ...
36	3B: Speech Processing	Automatic segmentation and clustering of speech us ...
	Session 7: language speech word translation model	
18	1A: Natural Language Processing	Semi-Supervised Training for Lecture Transcription ...
35	5A: Natural Language Processing	An English to Xitsonga statistical machine transla ...
30	3B: Speech Processing	Aligning Audio Samples from the South African Parl ...
38	3B: Speech Processing	Investigating The Use Of Syllable Acoustic Units F ...
41	3B: Speech Processing	Effect of Language Resources on Automatic Speech R ...
	Session 8: robot control manufacturing system design	
4	6A: Robot Design	Towards a Mobile Mapping Robot for Underground Min ...
9	2B: Robotic Case Studies and Applications	Development of the Electronics Pod for an Underwat ...
16	2B: Robotic Case Studies and Applications	The Design of a Rugged Low-Cost Man-Packable Urban ...
	Session 9: feature hand face classification tracking	
7	2A: Classifiers AI Machine Learning and Related Topics	Long-term tracking of multiple interacting pedestr ...
39	4A: Robotics Sensing and Design	Visual Features in Varying Illumination for Enhanc ...
46	1B: Image Processing	Single-trial EEG Discrimination between Five Hand ...
	Session 10: feature image hand face classification	
17	6B: Image Processing	Comparison of background subtraction techniques un ...
19	1B: Image Processing	Hybrid Age Estimation using Facial Images ...
34	3A: Image Processing Classifiers and Related Topics	Temporal Classification of FACS AUs using SURF Des ...
	Session 11: feature hand face classification recognition	
23	3A: Image Processing Classifiers and Related Topics	Automatic classification of sheep behaviour using ...
26	1B: Image Processing	Vision-based hand motion detection for intent reco ...
51	6B: Image Processing	Augmenting the L1 Tracker with appearance based tr ...
	Session 12: robot control manufacturing system design	
6	5B: End User Applications and Systems	CHAMP: a Bespoke Integrated System for Mobile Mani ...
12	2B: Robotic Case Studies and Applications	Development of a mechatronic transmission control ...
11	6A: Robot Design	Programmable Logic Control of an Electro- hydrauli ...
15	6A: Robot Design	Development of a docking mechanism for self-reconf ...

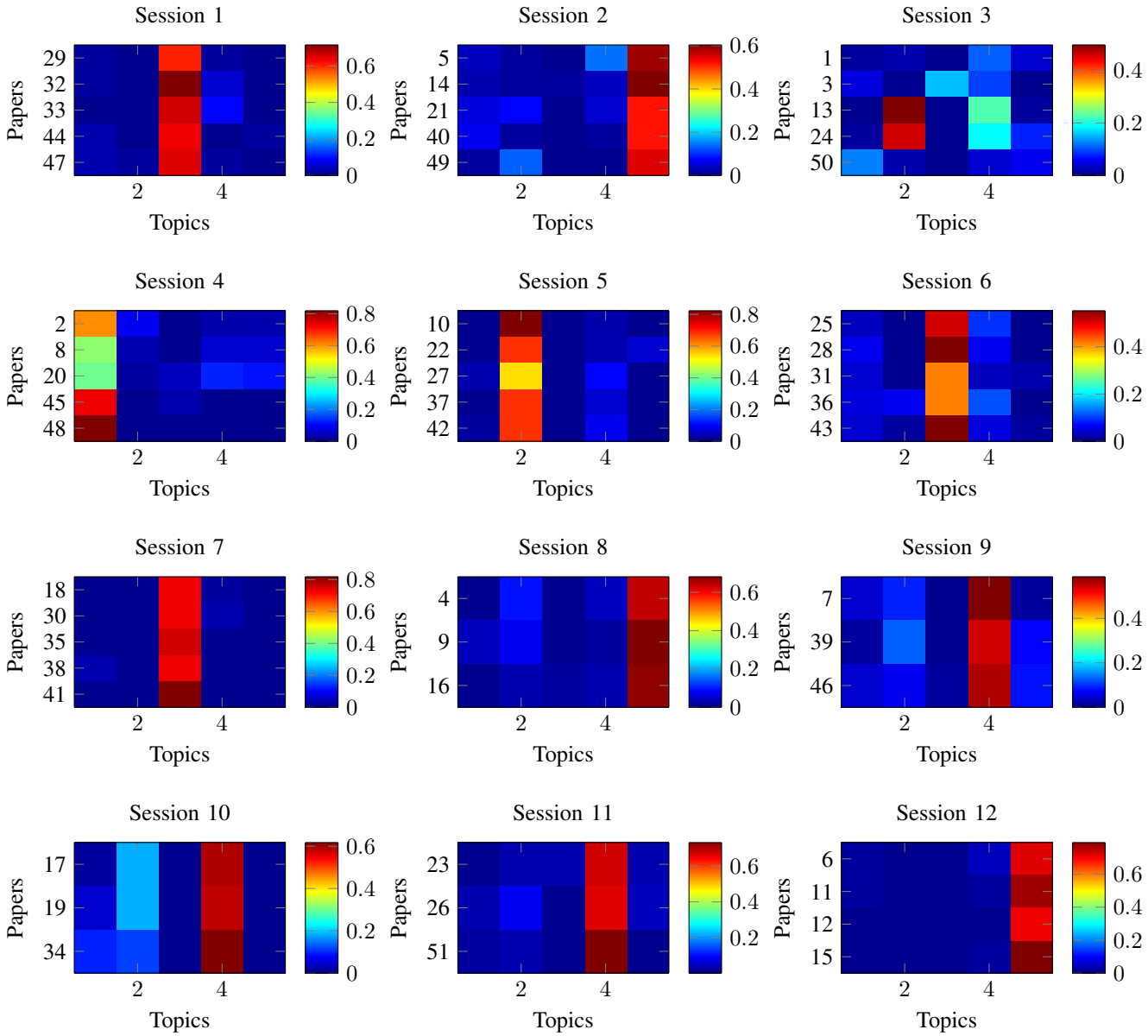

Fig. 2. Topic probabilities are similar for documents assigned to respective sessions. Note that the catch-all topic in Figure 1 has been removed here.

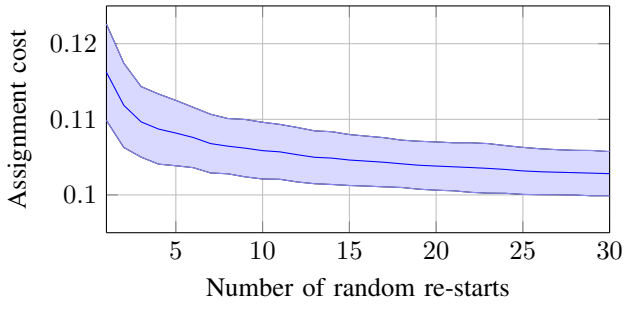

Fig. 3. Mean and standard deviation costs for session assignment iterations show typical convergence rates over a 5 minute period.

A. *Scheduling conference tracks*

Conference tracks are frequently parallel, and it can be particularly annoying when tracks are poorly scheduled, with talks on similar topics occurring simultaneously. The proposed approach to session allocation can be used to avoid this problem, by scheduling tracks such that distance between session topic distributions, $P(T^k)$, is maximised. Figure 4 shows a distance matrix for the sessions selected for PRASA 2014. The figure provides a simple visual aid for track scheduling and allows scheduling conflicts to be avoided. For example, the distance between sessions 1 and 6 (speech and language) is low so these should be scheduled further apart. The session assignment algorithm described in Section III can be used

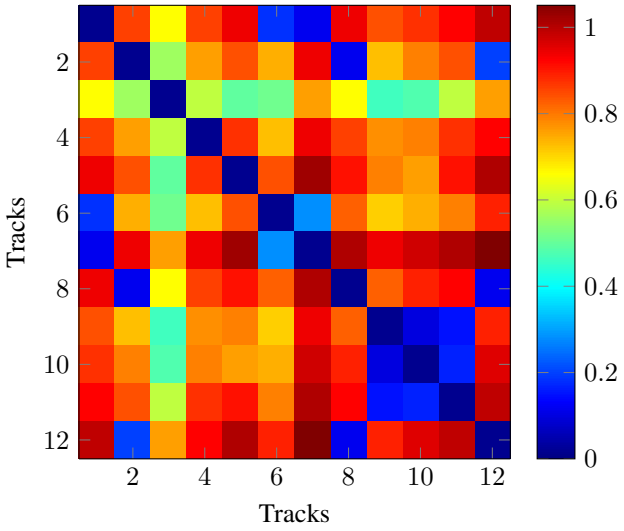

Fig. 4. A Euclidean distance matrix of the average topic distribution for each session can be used to ensure that session conflicts are avoided.

to do this automatically, in this case by maximising the assignment cost.

B. Relevance detection

The proposed approach can also be used for relevance detection in the review stages of a conference. Papers that are off topic typically result in a separate latent topic appearing, as they tend to use different vocabularies to relevant topics. As a result, topic distributions for these less relevant papers tend to be peaky, with topic probability mass biased towards a single topic. Thresholding the LDA document topic distributions can flag papers of this type.

C. Duplicate or similar submission detection

By allocating papers to topics in the proposed manner, papers with similar content tend to be grouped together, making duplicate submission detection easy. This is a particularly useful property for extremely large conferences, where fraudulent submissions can escape the notice of conference organisers.

V. Conclusions

This paper has shown how topic distributions learned using latent topic models can be used to assign conference papers to sessions or tracks automatically. Latent Dirichlet allocation was used to find topics in PRASA 2014 proceedings, and the resultant topic distributions used to group papers into sessions. Potential scheduling conflicts are avoided by maxmising the distance between session topic distributions, and the proposed approach allows for relevance and duplicate submission detection. Future work involves scheduling PRASA 2015.

Acknowledgment

This work was supported by funding from the Council for Scientific and Industrial Research, CSIR, South Africa.

References

[1] (2015) shdlr. [Online]. Available: shdlr.com

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[3] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models." *ICWSM*, vol. 10, pp. 1–1, 2010.

[4] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 577–584.

[5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, 2006.

[6] W. Buntine and S. Mishra, "Experiments with non-parametric topic models," in *20th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 2014.

[7] (2014) Nips 2014 papers. [Online]. Available: http://cs.stanford.edu/people/karpathy/nips2014/

[8] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, ser. UAI '04. Arlington, Virginia, United States: AUAI Press, 2004, pp. 487–494.

[9] G. E. Hinton and R. R. Salakhutdinov, "Replicated softmax: an undirected topic model," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 1607–1614.

[10] (2015) JSTOR. [Online]. Available: www.jstor.org

[11] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *Advances in neural information processing systems*, 2001, pp. 601–608.

[13] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[15] A. Riddell, "lda: 0.3.2," 2014. [Online]. Available: http://dx.doi.org/10.5281/zenodo.12737

[16] Stopwords. [Online]. Available: https://code.google.com/p/stop-words/

[17] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.