Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Sketch the Storyline with CHARCOAL: A Non-Parametric Approach

Siliang Tang, Fei Wu, Si Li, Weiming Lu*, Zhongfei Zhang, Yueting Zhuang
College of Computer Science, Zhejiang University
Hangzhou, China
{siliang, wufei, lisi_zzz, luwm, zhongfei, yzhuang}@zju.edu.cn

Abstract

Generating a coherent synopsis and revealing the development threads for news stories from the increasing amounts of news content remains a formidable challenge. In this paper, we proposed a hdd***C***RP (hybird distant-dependent Chinese Restaurant Process) based ***H***ier***ARC***hical t***O***pic model for news ***A***rticle c***L***ustering, abbreviated as ***CHARCOAL***. Given a bunch of news articles, the outcome of CHARCOAL is threefold: 1) it aggregates relevant new articles into clusters (i.e., *stories*); 2) it disentangles the chain links (i.e., *storyline*) between articles in their describing story; 3) it discerns the topics that each story is assigned (e.g.,Malaysia Airlines Flight 370 story belongs to the aircraft accident topic and U.S presidential election stories belong to the politics topic). CHARCOAL completes this task by utilizing a hddCRP as prior, and the entities (e.g., names of persons, organizations, or locations) that appear in news articles as clues. Moveover, the adaptation of non-parametric nature in CHARCOAL makes our model can adaptively learn the appropriate number of stories and topics from news corpus. The experimental analysis and results demonstrate both interpretability and superiority of the proposed approach.

1 Introduction

Massive amounts of information about news stories are published on the Internet every day. When retrieve news articles from search engines such as Google by keywords, the large volumes of unstructured search results returned by search engines make it hard to track the evolution and narrative threads of each news story. A *storyline* is a development path that reveal the hidden relationships of a set of relevant news articles in a given story. For examples, the news articles about Malaysia Airlines Flight 370 story are developed around the "search thread" and the "investigation thread", which are two separate but parallel storylines. The description of a given story in terms of appropriate storylines illustrates a big picture to help readers grasp the evolution and the development of the given story. Only a few studies [Kumar *et al.*, 2004; Lin *et al.*, 2012] have been reported on storyline generation, they are focusing on developing systems or frameworks for storyline extraction.

Here we are particularly interested in creating a probabilistic model to reveal and exploit the chain link between news articles about a news story, and simultaneous detect the topic that each news story is assigned to, e.g., MH370 story belongs to the aircraft accident topic.

In this paper, we propose the ***CHARCOAL*** model (abbreviated for "a hybird dd***C***RP based ***H***ier***ARC***hical t***O***pic model for news ***A***rticle c***L***ustering") from the perspective of unsupervised non-parametric modeling. Benefits from the proposed new hybrid distant-dependent CRP (hddCRP) prior, CHARCOAL adaptively achieves three goals simultaneously: 1) aggregates the news articles into coherent clusters (i.e., stories); 2) reveal the directed hidden connections between news articles about the same news story (i.e., the storylines); 3) assign each new story to its belonging news topic.

2 the CHARCOAL Model

Before introducing our CHARCOAL model, we first make a few assumptions. We assume that each news article merely remarks the progress of a news story, which represents a cell in a storyboard. We can trace the evolution of each story on the storyboard by linking article to its most relevant progress (i.e., another article), and these links are defined as *storylines*. Therefore, each news *story* in our model is defined as collections of linked articles, and each news *topic* is defined as the collection of similar *stories*. *Storylines*, *stories* and *topics* are three kinds of latent variables, which will be discerned by the proposed CHARCOAL model.

In CHARCOAL, we first decompose entities from news articles and utilize them as well as news contents as clues to reveal the storylines between articles. Then, these lines (storylines) and their connected dots (articles) gradually form a Directed Acyclic Graph during the model estimation, which from another perspective composes a story sketch for us. Finally our model creates story collections (i.e., topics) to enclose similar stories. We will present the details of each step in the next subsections, and show how we combine them

*Corresponding author

into a flexible and unified model. Some useful notations are shown in Table 1.

Symbols	Descriptions
$D/M/K$	# of observed articles/stories/topics
V	vocabulary size
dS/dT	shared entity appearance/relative entropy based distance measure
$\mathrm{F}_s/\mathrm{F}_T$	decay function for story/topic level distance measure
x_i	a V dimensional vector, the bag of words representation of document i
c_i	storyline, $c_i = j$ indicates document i is connecting to document j
z_m	story level connection, $z_m = n$ indicates that story S_m is connected to story S_n
S_m/T_k	a V dimensional vector, represents story m/topic k
α/π	story/topic level concentration parameter, scaling between 0 and 1, which controls the number of observed stories/topics.
$\mathbf{X}_C$	a group of data generated from a multinomial distribution C.
$f_c(l)$	a function returns the cluster index of customer assignment l.
$f_s(c)$	a function returns the self-link of cluster c.

Table 1: The notations in CHARCOAL

2.1 hddCRP and its relationship to CRP and ddCRP

The goal of CHARCOAL could be formally defined as follows: Given a bunch of news articles $\mathbf{X} = \{x_{1:D}\}$, we are going to jointly model three types of unobserved random variables, namely the discrete scalars c_i indicating the hidden connections from article i to its most related article j in the past, underlying news story S_m, and potential topic T_k. It should be noted here that the three kinds of unobserved random variables are in a hierarchy (i.e., news articles, stories, and topics). In order to make this feasible, we proposed a novel prior distribution over hierarchical document partitions, which is named as hybrid distant-dependent CRP (hddCRP) here.

hddCRP is a stochastic process which has strong relationships to CRP [Pitman and others, 2002] and distance dependent CRP [Blei and Frazier, 2011]. The CRP exhibits the clustering property of the Dirichlet Process. It was named under a metaphor of a table assignment strategy for a sequence of customers that arrive in a Chinese restaurant with infinite number of tables. Each customer (data point) x_i in $\mathbf{X}$ (dataset) sits at a previously occupied table (cluster) k with a probability proportional to the number of the customers N_k already sitting there, and at a new table with a probability proportional to a concentration parameter α. This can be interpreted as that each customer $x_{j(j\neq i)}$ in table k contributes 1 weight to N_k.

ddCRP generalizes CRP by introducing a changeable weight on x_j, and the weight is decided by two things: a distance measure d_{x_i,x_j} between x_i and x_j; and a decay function $\mathrm{F}(d)$, which satisfies $\mathrm{F}(\infty) = 0$. Moreover, instead of assigning table k to x_i directly, ddCRP assigns x_j to x_i. Such assignments create directed links between customers, and customers that are reachable from each other through customers assigned to the same table. This yields a *non-exchangeable* distribution over partitions, since table assignments may change when a new customer assignment is observed.

Although ddCRP allows to model the (directed) connections between news articles, it does not provide a nature of the hierarchy between news articles. This limitation makes ddCRP inappropriate for our initial goal (i.e., the hierarchical modeling of three kinds of unobserved random variables); therefore, we attempt to extend ddCRP hierarchically. There are two types of extensions. Hierarchical dirichlet process (HDP) [Teh *et al.*, 2006] demonstrates the first type. HDP groups data into a pre-defined hierarchical structure, where each level is associated with a Dirichlet process, whose base measure is sampled from a higher-level Dirichlet process. Another type is to create nested priors, such as nested CRP [Blei *et al.*, 2010], nDP [Rodriguez *et al.*, 2008], and TSSB [Ghahramani *et al.*, 2010]. They introduce nested DP or nested CRP prior on topics, which organize the topics according to a tree structure, in which more abstract topics are near the root and more concrete topics are near the leaves.. They can be described in a fragmentation process, where each group in a higher-level is further divided into small groups, and each fragmentation is governed by CRPs/DPs with a shared concentration parameter. However, due to the non-exchangeable property of ddCRP, in most cases, it is not feasible to build nested ddCRP priors, nor is associated each level with a ddCRP prior as HDP does. (However, creating a hierarchical model with a mixture of a single ddCRP layer and CRP/DP layers may be feasible.)

Our proposed hddCRP hierarchically extend ddCRP in a different way. hddCRP not only creates a hierarchy of temporal data (e.g., news articles), but also embeds knowledge of connections from different aspects to each hierarchy, which is crucial to many applications of practical interest. As we observe, ddCRP only allows one single self customer assignment (self-link) for each table (cluster). Under a given ddCRP prior, if a customer assignment is not a self-link, two tables are merged. Only a self-link assignment can stop the table merging, and completes a cluster. However, the discovered clusters may continue to merge given other ddCRP priors. Introducing a new ddCRP prior with a new distance measure and a decay function may lead clustering of tables by assigning the current table to other tables. Every time when we introduce a new prior, a new level of hierarchy is created. By combining different ddCRP priors, we may assign probability distributions to ensembles of forests, where each tree in the forests has the same depth, but an unbounded branching. If we consider a hddCRP with 3 depths, its structure can be interpreted as follows: the leaf nodes of all the trees in a forest are news articles (customers); the second level nodes are stories (tables); and the top level nodes are topics (restaurants). We need to specify d_{x_i,x_j} (distance measure) and $\mathrm{F}(d)$

(decay function) for each non-leaf level to complete this hddCRP prior.

2.2 Distance and Decay Functions

As described above, to complete the hddCRP prior for our CHARCOAL model, we need to specify distance measures and decay functions for both article link assignment and story link assignment.

For article link assignments, we adopt a shared entity appearance based distance $dS_{i,j}$. We define entities as names of persons, organizations, or locations, which can be extracted by name entity recognition (NER) tools in OpenNLP[Baldridge, 2005], or the Stanford Named Entity Recognizer [1]. More recent developed Named Entity Extraction (NEE) services, such as AlchemyAPI [2], and DBpedia-spotlight [Mendes *et al.*, 2011], offered even better solutions, since they not only identify entities, but also link them to related semantic resources.

With extracted entities, the distance between document i and j is defined as follows: $dS_{i,j} \triangleq \mathrm{F}_S(i,j) = \frac{1}{2}\left(\frac{e_{ij}}{|i|_e} + \frac{e_{ji}}{|j|_e}\right)$ and e_{ij} represents the number of times when the shared entities have appeared in document i, and $|i|_e$ represents the number of the discovered entities in total in document i. The decay function we use for $dS_{i,j}$ is a weighted window decay, which is specified as follows. $\mathrm{F}_S(i,j,dS_{i,j}) = 1[0 < t_i - t_j < a]dS_{i,j}$, where $1[\,]$ is an indicator function, t_i is the time-stamp of document i, and a is an integer indicating a window size.

For story link assignments, since stories with similar contents belong to the same topic, and all the stories $\mathbf{S}$ in CHARCOAL are represented as probabilistic distributions, we may adopt the relative entropy based approaches to measure the story similarity. Therefore, their distance $dT_{m,n} \triangleq \mathrm{F}_T(m,n)$ may be defined by any symmetric relative entropy based distance measures, such as symmetric KL divergence, i.e., $\frac{\mathrm{KL}(S_m,S_n)+\mathrm{KL}(S_m,S_n)}{2}$, or Jensen-Shannon divergence. For these measurements higher similarities actually yield shorter distances; therefore, a new decay function $\mathrm{F}_T(d)$ is defined to ensure the positive correlation between similarities and distances, where $\mathrm{F}_T(m,n,dT_{m,n}) = \begin{cases} 1 & (m=n) \\ \frac{(\max(dT_{m,n\in\{n|n\neq m\}})-dT_{m,n})(1-\pi)}{\sum_{m\neq n}(\max(dT_{m,n\in\{n|n\neq m\}})-dT_{m,n})} & (m\neq n) \end{cases}$, and the decay function satisfies a normalization requirement that $\sum_{m\neq n}\mathrm{F}_T(m,n,dT_{m,n}) = 1-\pi$. ($\pi$ is the concentration parameter of the ddCRP at the current level.)

2.3 Generative Process

With the help of the fully specified hddCRP prior, we now build our CHARCOAL model and describe the generative process of the proposed CHARCOAL model.

In CHARCOAL, the topic hierarchy is modeled sightly differently from the most traditional topic models. First, article i in text corpus is represented as a V dimensional vector x_i. x_i is a collection of words from the same article, where each dimension of x_i indicates the frequency of a particular word in article i. Therefore, x_i is viewed as a collection of word samples, and we assume that these word samples are generated from a Multinomial distribution $S_{m(m=1:M)}$. S_m is also a V dimensional vector, and can be viewed as the underlying story representation of a article collection $\mathbf{X}_{S_m}$. We then further assume that $\mathbf{S} = \{S_{1:M}\}$ is the collection of the samples from higher K Dirichlet distributions with their corresponding base measures $\mathbf{T} = \{T_{1:K}\}$, and each $T_{k(k=1:K)}$ in $\mathbf{T}$ is also a V dimensional Multinomial distribution which represents the underlying topic of a story collection $\mathbf{S}_{T_k}$. In this model, the story/topic assignment of each article/story is controlled by both their similarities and our defined hddCRP prior.

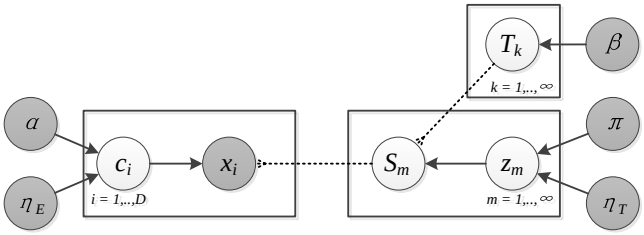

Figure 1: The pseudo graphic model representation of CHARCOAL.

The pseudo graphic model representation of CHARCOAL is shown in Figure 1. The representation is pseudo because the underlying story assignment of article x_i depends on the story assignments of other document $\mathbf{X}_{-i}$, and the topic assignment of story S_m depends on the topic assignments of other stories $\mathbf{S}_{-m}$. It is impossible to present such irregular and unobserved dependencies by the traditional plate notation. Therefore, we make a small extension to the traditional plate notation, where one-to-many dashed links are used to represent such dependencies.

If we denote η_S as a collection of story level distances and decay function $\eta_S = \{\mathrm{F}_S, dS\}$, and η_T as a collection of topic level distances and decay function $\eta_T = \{\mathrm{F}_T, dT\}$. The generative process of CHARCOAL is described as follows:

1. For each $i \in \{1,...,D\}$, draw a customer assignment $c_i \sim ddCRP(\eta_S;\alpha)$ for article x_i. If c_i is a self link, then continue to draw a story assignment $z_{f_c(c_i)} \sim ddCRP(\eta_T;\pi)$, otherwise assign $z_{f_c(c_i)} = z_{f_s(f_c(c_i))}$
2. For each $k \in \{1,...\}$, draw a topic $T_k^* \sim Dir(\beta)$
3. Draw $S_{f_c(c_i)} \sim Dir(T_{f_c(z_i)})$
4. For each $i \in \{1,...,D\}$, draw article $x_i \sim Dir(S_{f_c(c_i)})$
5. For each $n \in \{1,...,|x_i|\}$, draw word $w_n \sim Multi(x_i)$

where Function $f_c(l)$ returns the cluster index of customer assignment l, and $f_s(c)$ returns the self assignment link of cluster c. $Multi()$ represents a Multinomial distribution; $Dir(\delta)$ represents a Dirichlet distribution with a base measure δ; $ddCRP(\eta_S;\alpha)$ and $ddCRP(\eta_T;\pi)$ represent story and topic level ddCRP priors, respectively; $|x_i|$ represents the

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

[2] http://alchemyapi.com

total number of words in article x_i. Other notations may be found in Table 1.

3 Approximate Inference

An exact inference in topic models is often intractable and many approximate inference methods are adopted in practice. In this paper, we resort to collapsed Gibbs sampling for model learning and hyper-parameter inference.

3.1 Posterior Sampling

Collapse S and T: Given a collection of data points $\mathbf{X}$, if we define $\mathbf{X}_S$ as a group of data governed by a Multinomial distribution S, where S is a latent variable generated from a Dirichlet distributed with a base measure β, then the variable S could be collapsed due to the Dirichlet-multinomial conjugacy, and the marginal likelihood of $\mathbf{X}_S$ given β, could be written as:

$$\begin{aligned} p(\mathbf{X}_S|\beta) &= \int p(\mathbf{X}_S|S,\beta)p(S|\beta)dS \\ &= \left(\frac{\Gamma(|\beta|\beta)}{\Gamma(\beta)^{|S|}}\right)\left(\frac{\prod_v \Gamma(n_v+\beta)}{\Gamma(n_{(.)}+|\beta|\beta)}\right) \end{aligned} \tag{1}$$

where $n_{(.)}$ is the total number of the terms in $\mathbf{X}_S$, n_v is the number of the times term v appeared in $\mathbf{X}_S$, $|\delta|$ indicates the number of dimensions of δ, and $\Gamma()$ is the gamma function.

Similarly, in CHARCOAL latent variables $\mathbf{S} = \{S_{1:M}\}$ and $\mathbf{T} = \{T_{1:K}\}$ are in multinomial distributions, which represent topic and story, respectively. They can be integrated out according to the Eq. 1. Therefore, we only need to sample story link $\mathbf{z} = \{z_{1:M}\}$ and article link $\mathbf{c} = \{c_{1:D}\}$.

Sample c: For each article link assignment (i.e., storyline) c_i in $\mathbf{c}$, its marginal distribution on given all documents $\mathbf{X} = \{x_{1:D}\}$, and the rest document assignments without c_i, may be factorized into two parts:

$$\begin{aligned} & p(c_i = j|\mathbf{c}_{-i}, \mathbf{X}, T_{f_c(z_{f_c(c_i)})}, \alpha, \eta_S) \\ \propto\ & p(c_i = j|\alpha, \eta_S) p(\mathbf{X}|c_i = j, \mathbf{c}_{-i}, T_{f_c(z_{f_s(f_c(j))})}) \end{aligned}$$

The first part is the prior distribution. Since F_S yields temporally sequential distances in order, according to the hdd-CRP definition, this distribution may be written as:

$$p(c_i = j|\alpha, \eta_S) \propto \begin{cases} \alpha & (\text{ if } i = j) \\ \mathrm{F}_S(i, j, dS_{i,j}) & (\text{ if } i \neq j) \end{cases}$$

While the second part is empirical likelihood, if we denote $\theta = T_{f_c(z_{f_s(f_c(j))})}$, it can be written as:

$$p(\mathbf{X}|c_i = j, \mathbf{c}_{-j}, \theta) = \prod_{m=1}^{M} p\left(\mathbf{X}_{S_m}|c_i = j, \theta\right)$$

$$= \prod_{m=1}^{M} p(\mathbf{X}_{S_m}|\theta) \times \begin{cases} 1 & (\text{ if } i = j) \\ \frac{p(\mathbf{X}_{(S_{f_c(c_i)} \cup S_{f_c(j)})|\theta)}}{p(\mathbf{X}_{S_{f_c(c_i)}}|\theta) p(\mathbf{X}_{S_{f_c(j)}}|\theta)} & (\text{ if } i \neq j) \end{cases}$$

Since any $p(\mathbf{X}_{S_{(.)}}|\theta)$ may be computed easily according to Eq.(1), we now sample c_i according to following equations:

$$\begin{aligned} & p(c_i = j|\mathbf{c}_{-i}, \mathbf{X}, \theta, \alpha, \eta_S) \\ \propto & \begin{cases} \alpha & (\text{ if } i = j) \\ \mathrm{F}_S(i, j, dS_{i,j}) \frac{p(\mathbf{X}_{(S_{f_c(i)} \cup S_{f_c(j)})|\theta)}}{p(\mathbf{X}_{S_{f_c(i)}}|\theta) p(\mathbf{X}_{S_{f_c(j)}}|\theta)} & (\text{ if } i \neq j) \end{cases} \end{aligned} \tag{2}$$

Sample z: Story level link assignments $\mathbf{z} = \{z_{1:M}\}$, which connects stories to topics, may be sampled similarly; their marginal distribution is given as:

$$\begin{aligned} & p(z_m = n|\mathbf{z}_{-m}, \mathbf{c}, \mathbf{S}, \beta, \pi, \eta_T) \\ \propto & \begin{cases} \pi & (\text{ if } m = n) \\ \mathrm{F}_T(m, n, dT_{m,n}) \frac{p(\mathbf{S}_{(T_{f_c(m)} \cup T_{f_c(n)})|\beta)}}{p(\mathbf{S}_{T_{f_c(m)}}|\beta) p(\mathbf{S}_{T_{f_c(n)}}|\beta)} & (\text{ if } m \neq n) \end{cases} \end{aligned} \tag{3}$$

where β is the Dirichlet smooth of topics $\mathbf{T}$

3.2 Sampling Hyper-parameters

Sample α: Given all the article links (i.e., stroylines) $\mathbf{c}$, the posterior of α, $p(\alpha|\mathbf{c}) = p(\mathbf{c}|\alpha)p(\alpha)$. Its likelihood part is calculated as follows:

$$\begin{aligned} p\left(\mathbf{c}|\alpha\right) &= \prod_{i=1}^{D} \frac{1\left[c_i = i\right]\alpha + 1\left[c_i \neq i\right]\mathrm{F}_S\left(i, c_i, dS_{i,c_i}\right)}{\alpha + \sum_{j \neq i} \mathrm{F}_S\left(i, j, dS_{i,j}\right)} \\ &\propto \alpha^M \left[\prod_{i=1}^{D}\left(\alpha + \sum_{j \in \{j|0<t_i-t_j<a\}} dS_{i,j}\right)\right]^{-1} \end{aligned}$$

since the prior α is continuous, it is difficult to sample exactly from the posterior distribution $p(\alpha|\mathbf{c})$; as mentioned in the literature [Blei and Frazier, 2011], we may use the Griddy-Gibbs approach for sampling approximately.

Sample π: Similarly, according to the posterior of π, $p(\pi|\mathbf{z}) = p(\mathbf{z}|\pi)p(\pi)$. Its likelihood part $p(\mathbf{z}|\pi)$ is actually quite simple to calculate, due to the constraint that $\sum_{n \neq m} \mathrm{F}_T\left(m, n, dT_{m,n}\right) = 1 - \pi$.

$$\begin{aligned} p\left(\mathbf{z}|\pi\right) &\propto \pi^K \left[\prod_{m=1}^{M}\left(\pi + \sum_{n \neq m} \mathrm{F}_T\left(m, n, dT_{m,n}\right)\right)\right]^{-1} \\ &= \pi^K \end{aligned}$$

If we make π beta distributed, that is $p(\pi|a_0, b_0) = \mathrm{Beta}(a_0, b_0)$, then the posterior of π, $p(\pi|z, a_0, b_0) \propto \frac{B(a_0+K, b_0)}{B(a_0, b_0)} \mathrm{Beta}(a_0 + K, b_0) = E(\pi^K)\,\mathrm{Beta}(a_0 + K, b_0)$, where K is the number of the topics, $B(a, b)$ is the beta function $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a,b)}$, $E(\pi^K)$ is the K^{th} raw moments of π, and $E(\pi^K) = \frac{a_0^{(1,K)}}{(a_0+b_0)^{(1,K)}}$; note that $a^{(1,K)} = a(a+1)\cdots[a+(K-1)]$ is the ascending power of base a and order K.

3.3 Model Estimation

The model estimation procedure is shown in Algorithm 1.

Algorithm 1 Model Estimation of CHARCOAL

Input: a set of documents $\mathbf{X} = \{x_{1:D}\}$;
Output: article link assignments $\mathbf{c}$, story link assignments $\mathbf{z}$;

1: For each article x_i, we set its article/story link assignment to itself, i.e., $c_i = i$, $z_{f_c(c_i)} = f_c(c_i)$;
2: **for** each i in $\{1, ..., D\}$ **do**
3: Detach article i from the state by setting $c_i = i$;
4: **for** each j in $\{1, ..., D\}$ **do**
5: Compute distance $dS_{i,j}$, and then the posterior $p(c_i = j|\mathbf{c}_{-i}, \mathbf{X}, \theta, \alpha, \eta_S)$ according to Eq.(2);
6: **end for**
7: Sample a new c_i from the posterior distribution;
8: **if** $c_i = i$ **then**
9: **for** each n in $\{1, ..., M\}$ **do**
10: Compute distance $dT_{c_i,n}$, and then the posterior $p(z_{f_c(c_i)} = n|\mathbf{z}_{-f_c(c_i)}, \mathbf{c}, \mathbf{S}, \beta, \pi, \eta_T)$ according to Eq.(3);
11: **end for**
12: Sample a new $z_{f_c(c_i)}$ from the posterior distribution;
13: **else**
14: Assign $z_{f_c(c_i)} = z_{f_s(f_c(c_i))}$;
15: **end if**
16: **end for**
17: Update hyper-parameters every 50 iterations;
18: Repeats step 2 until convergence;

hddCRP can be decomposed into multiple ddCRPs, resulting in a more complex sampling process. The computational complexity of CHARCOAL is highly dependent on the data and distance/decay functions defined at each hddCRP level. In the worst case, each data point has non-zero distance with other data points (fully connected), but its table assignment is a self link; the complexity of our algorithm in such case is $O(n^2) * O(Eq.(1))$ for one iteration. In the best case, in which all the data points are unreachable from each other, the complexity is $O(n)$ for each iteration. While in the most cases, due to the existence of decay function, the connections among data points are quite sparse, and we only have to calculate $Eq.(1)$ when table assignment triggered table splitting or merging. Therefore, the computational cost of CHARCOAL is affordable in practice.

4 Model Evaluation

In this section, we employ CHARCOAL to analyze some challenging text corpra of different news categories, and analysis its clustering performance with two comparison studies. We also demonstrate its effectiveness by visualizing a few CHARCOAL topics.

4.1 Compare CHARCOAL with the Two Stage Clustering Methods

CHARCOAL jointly models the stories and topics as well as their inner connections. In this section, we verify that this joint modelling yields a better result than the two stage clustering by evaluating the document clustering results on a real world news corpus.

The data corpus we use to diagnose our model contains 4,908 news articles from the New York Times International News section and covers a time period from Jan 2014 to Aug 2014. We have removed the low frequency words and stop words, yielding a corpus of 4,780 unique terms and 1.63M observed words. Each article has a region tag, which is extracted through an analysis of its URL and is served as the ground truth during the diagnosis. Entities are extracted by AlchemyAPI. Since location tags are used as the ground truth, we therefore only employ the shared person names to calculate article distances dS. After removing non-shared entities, 45,868 entities are identified, and they belong to 3,326 unique (but may ambiguous) names.

In this experiment, we first degrade CHARCOAL to a single level model, which only contains one document-story clustering (this could be simply achieved by setting π=1). After the model convergence, we then apply complete-linkage clustering (CLC) [Brian S. Everitt, 2001] and k-means on the generated stories to further create hierarchical topics. CLC and k-means require a distance measure between pair of stories; we adopts the same distance measure and decay function that we apply to CHARCOAL for the story level clustering (for distance measure, Symmetric KL (SKL) and JSD are used respectively). On the other hand, we control the number of the topics that are estimated by CHARCOAL indirectly by tuning the topic level concentration parameter π (to achieve this the update of π is prohibited). CHARCOAL and the two stage clustering models are initialised with story level concentration parameter $\alpha = 0.5$, topic smooth $\beta = 0.05$, and time window size α = 8 days. For CHARCOAL, π varies from 0.05 to 0.95 with an interval of 0.05, and its topic level distance dT is measured by JSD.

The clustering results are evaluated under three well known metrics [Manning *et al.*, 2008]; they are the percentage of correct classification (Purity), the normalized mutual information (NMI), and the Rand Index. Figure 2 illustrates the model performances as a function of the number of the topics. As shown in Figure 2, by tuning π, CHARCOAL produces the number of the topics from 176 to 476, which outperforms all the three other approaches in all the three metrics all the time.

4.2 Compare CHARCOAL with other Non-parametric Hierarchical Topic Models

As mentioned in Section 2.1, there are two major ways of building the hierarchy on CRP; they are HDP and hLDA. It is important to make a quantitative comparison with these models, since they are the state-of-the-art topic models, and they are both non-parametric and hierarchical. The experiments are performed on collections of annotated news articles from New York Times with different major categories. Their characteristics are documented in Table 2 below:

The experiment setup is as follows:

1) the HDP inference is performed through a collapsed Gibbs sampler, with a burn-in of 500 samples; its hyper-parameters are updated every 50 iterations. For evaluation

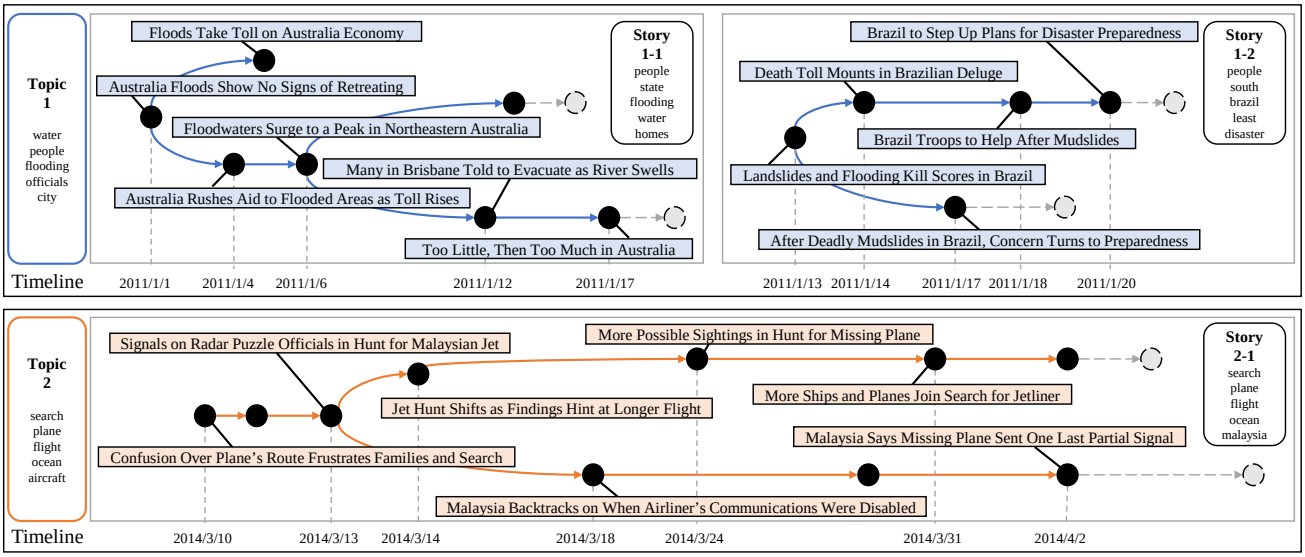

Figure 3: a close look at some topic fragments (articles are presented as dots, and aligned chronologically along the timeline with their titles. The arrows between two dots are the storylines, which reveal the diffusion paths of the CHARCOAL stories)

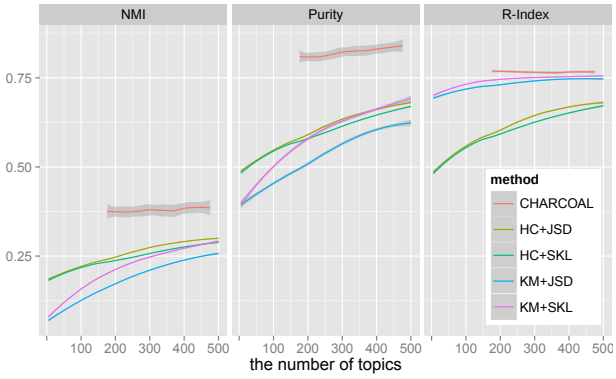

Figure 2: The clustering performances as a function of the number of the topics. (a LOESS[Fox, 2002] smoother is fitted to the data, and 95% confidence bands (gray areas) are applied to results)

	ARTS	BIZ	SPORTS	WORLD
time period	from 2012.1.1 to 2014.8.31			
# of articles	11,276	8,708	21,585	19,741
categories	9	10	18	5
total words	4.1M	3.1M	6.1M	5.9M
vocab. size	6,250	4,680	3,998	4,573
# of entities	224K	66K	558K	204K
unique entities	19K	6K	18K	12K

Table 2: The Description of the Datasets

purpose, articles are assigned to the topics that occupy the largest proportion in their topic mixture. 2) the hLDA inference is also performed by Gibbs sampling with a burn-in period of 500 samples; its hyper-parameters are fixed and initialized as follows: topic smooth $\alpha = 0.1$, word smooth $\eta = 0.1$, and nCRP concentration parameter $\lambda = 1$. Similar to CHARCOAL, hLDA creates hierarchical partitions on article collections. If we set the depth of nCRP prior to 4, hLDA generates trees with 4 depths, which are root, topic clusters, story clusters, and articles, respectively. In this experiment, topic clusters are used to evaluate hLDA performance. 3) CHARCOAL is initialized with story level concentration parameter $\alpha = 0.5$, topic smooth $\beta = 0.05$, and time window size $a = 8$ days. We use JSD to measure dT, and hyper-parameter π is updated every 50 iterations. Similarly, topic clusters are used to evaluate model performance.

metrics	Datasets	HDP	hLDA	CHARCOAL
Purity	ARTS	0.6711	0.4164	**0.8882**
	BIZ	0.7387	0.4528	**0.7997**
	SPORTS	0.7256	0.3637	**0.8934**
	WORLD	0.6604	0.3598	**0.8102**
NMI	ARTS	0.3240	0.0721	**0.4926**
	BIZ	0.3431	0.0837	**0.3946**
	SPORTS	0.5102	0.2817	**0.6244**
	WORLD	0.2373	0.0317	**0.3299**
R-Index	ARTS	0.6485	0.5547	**0.7828**
	BIZ	0.7482	0.6648	**0.7528**
	SPORTS	0.9006	0.8092	**0.9081**
	WORLD	0.7607	0.6989	**0.7616**

Table 3: Performance Comparisons

For each experiment samples for the evaluation are collected a few iterations apart after the model convergence (af-

ter the burn-in period). The experiments' results are presented in Table 3. According to Table 3, CHARCOAL outperforms HDP and hLDA in all datasets, which indicated that our method could effectively improve clustering performance for many different news categories. This may be partially due to the storyline prior (i.e., shared entity appearance) serves as a useful feature on characterizing the development of a real world news story, which creates a solid foundation for topic level clustering.

4.3 Qualitative analysis with data visualization

In order to understand the CHARCOAL outputs more intuitively, we visualized a few CHARCOAL topics detected from NYT data corpus in Figure 3. For a close look at these discovered topics, we presented their story fragments, storylines as well as the top five words in each topic/story.

The upper part of Figure 3 presents two stories from a flood topic. They are two different floods happened in Australia and Brazil. CHARCOAL splits them into separated stories due to the weak/no evidences from the appearances of name entities, but the stories later are grouped into the same topic because of their similarities in contents.

The lower part of Figure 3 presents another story from topic "search, plane, flight". It is a story that records the progress of searching the missing flight MH370. Its storyline revealed the parallel development of the MH370 story. As shown in Figure 3, the main thread of MH370 splits up into two separated threads, one is concentrated on searching the lost fight, while the other one is focused on the investigation of the possible causes of the disappearance.

5 Conclusion and Future work

In this paper, we have presented CHARCOAL, a novel nonparametric probabilistic model via a hddCRP prior, which jointly models news storylines, stories, and topics. CHARCOAL produces a structural output of a set of news articles, which may help readers who are overwhelmed by information overload, to explore the big picture of popular news topics. The improvement of our proposed hddCRP prior is nontrivial because hddCRP gives the possibility to model document hierarchy from a new perspective, where ddCRPs are "stacked" to form a multi-layer model. In such model, the higher-level clusters server as priors to its lower-level offspring clusters. By redefining appropriate distance measures, it is easy to apply our hddCRP prior to other applications such as document summarization or evolution analysis[Tang *et al.*, 2013b]. Some other extensions such as introducing domain information[Gao *et al.*, 2012], or leveraging the importance between different types of measures[Tang *et al.*, 2013a] may be considered as our future work.

Acknowledgments

This work was supported in part by the 973 Program (No. 2012CB316400), the NSFC (No. 61402401), the 863 Program (No. 2012AA012505), the China Knowledge Centre for Engineering Sciences and Technology (CKCEST), and the Zhejiang Provincial NSFC (No. LQ14F010004).

References

[Baldridge, 2005] Jason Baldridge. The opennlp project. [EB/OL], 2005. http://opennlp.apache.org/index.html.

[Blei and Frazier, 2011] David M Blei and Peter I Frazier. Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488, 2011.

[Blei *et al.*, 2010] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.

[Brian S. Everitt, 2001] Morven Leese Brian S. Everitt, Sabine Landau. *Cluster Analysis.(Fourth ed.).* 2001.

[Fox, 2002] John Fox. Nonparametric regression. *CRAN R Project. January*, 2002.

[Gao *et al.*, 2012] Haidong Gao, Siliang Tang, Yin Zhang, Dapeng Jiang, Fei Wu, and Yueting Zhuang. Supervised cross-collection topic modeling. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 957–960. ACM, 2012.

[Ghahramani *et al.*, 2010] Zoubin Ghahramani, Michael I Jordan, and Ryan P Adams. Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems*, pages 19–27, 2010.

[Kumar *et al.*, 2004] Ravi Kumar, Uma Mahadevan, and D Sivakumar. A graph-theoretic approach to extract storylines from search results. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 216–225. ACM, 2004.

[Lin *et al.*, 2012] Chen Lin, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, and Tao Li. Generating event storylines from microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 175–184. ACM, 2012.

[Manning *et al.*, 2008] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[Mendes *et al.*, 2011] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[Pitman and others, 2002] Jim Pitman et al. Combinatorial stochastic processes. Technical report, Springer, 2002.

[Rodriguez *et al.*, 2008] Abel Rodriguez, David B Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483), 2008.

[Tang *et al.*, 2013a] Siliang Tang, Hanqi Wang, Jian Shao, Fei Wu, Ming Chen, and Yueting Zhuang. πlda: document clustering with selective structural constraints. In *Proceedings of the 21st ACM international conference on multimedia*, pages 753–756. ACM, 2013.

[Tang *et al.*, 2013b] Siliang Tang, Yin Zhang, Hanqi Wang, Ming Chen, Fei Wu, and Yueting Zhuang. The discovery of burst topic and its intermittent evolution in our real world. *Communications, China*, 10(3):1–12, 2013.

[Teh *et al.*, 2006] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.