Noname manuscript No.
(will be inserted by the editor)

Robust Unsupervised Cluster Matching for Network Data

Tomoharu Iwata · Katsuhiko Ishiguro

Received: date / Accepted: date

Abstract Unsupervised cluster matching is a task to find matching between clusters of objects in different domains. Examples include matching word clusters in different languages without dictionaries or parallel sentences and matching user communities across different friendship networks. Existing methods assume that every object is assigned into a cluster. However, in real-world applications, some objects would not form clusters. These irrelevant objects deteriorate the cluster matching performance since mistakenly estimated matching affect on estimation of matching of other objects. In this paper, we propose a probabilistic model for robust unsupervised cluster matching that discovers relevance of objects and matching of object clusters, simultaneously, given multiple networks. The proposed method finds correspondence only for relevant objects, and keeps irrelevant objects unmatched, which enables us to improve the matching performance since the adverse impact of irrelevant objects is eliminated. With the proposed method, relevant objects in different networks are clustered into a shared set of clusters by assuming that different networks are generated from a common network probabilistic model, which is an extension of stochastic block models. Objects assigned into the same clusters are considered as matched. Edges for irrelevant objects are assumed to be generated from a noise distribution irrespective of cluster assignments. We present an efficient Bayesian inference procedure of the proposed model based on collapsed Gibbs sampling. In our experiments, we demonstrate the effectiveness of the proposed method using synthetic and real-world data sets, including multilingual corpora and movie ratings.

T. Iwata
NTT Communication Science Laboratories
2-4 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237, Japan
Tel.: +81-774-93-5161
Fax: +81-774-93-5155
E-mail: iwata.tomoharu@lab.ntt.co.jp

K. Ishiguro
NTT Communication Science Laboratories

Keywords unsupervised learning · object matching · network modeling · multilingual corpus analysis · stochastic block model

1 Introduction

Network analysis has become an important tool for a wide variety of fields, such as sociology, biology and information engineering (Albert and Barabási, 2002). Many networks from different fields exhibit common characteristics. For example, we can find heavy-tailed degree distributions, or scale-free properties, in collaboration networks, World Wide Web, power grid and protein-protein interaction networks (Barabási and Albert, 1999). Other examples of commonly found structures include the small world property (Watts and Strogatz, 1998), community structure (Girvan and Newman, 2002) and hierarchical structure (Clauset et al, 2008).

We consider situations in which different networks contain common latent clusters, where each cluster exhibits a particular interaction pattern to other clusters. For example, lexical networks from different languages would have similar synonym groups that have characteristic dependencies between the groups, social networks from different research labs would share similar relationship patterns among faculty, post-docs and students, and biological networks from different species would have some common components.

Our task is to discover common latent clusters of objects, or to find matching between clusters, in different networks without correspondence information, which we call *unsupervised cluster matching*. An example is to find common word clusters from document-word networks in English and German without dictionaries or parallel sentences. Other examples include matching user communities across different friendship networks, and matching genes groups in gene regulatory networks from different species. Some unsupervised cluster matching methods, such as ReMatch (Iwata et al, 2016), have been proposed, where every object is assumed to be assigned into a cluster. However, in real-world applications, some objects would not form clusters. We call these objects that do not form clusters with other objects as *irrelevant* objects, and objects that form clusters as *relevant* objects. The irrelevant objects deteriorate the cluster matching performance since mistakenly estimated matching affect on estimation of matching of other objects.

We propose a probabilistic model for robust unsupervised cluster matching that discovers relevance of objects and matching of object clusters, simultaneously, given multiple networks. The proposed method finds correspondence only for relevant objects, and keeps irrelevant objects unmatched, which enables us to improve the matching performance since the adverse impact of irrelevant objects is eliminated. With the proposed method, relevant objects in different networks are clustered into a shared set of clusters by assuming that different networks are generated from a common network probabilistic model, where interaction patterns between clusters are shared among different networks. We use infinite relational models (Kemp et al, 2006), which is an extension of stochastic block models (Wang and Wong, 1987; Nowicki and Snijders, 2001) with Bayesian nonparametrics, as a basic component for modeling relevant objects in a single network. Objects assigned into the same clusters are considered as

matched. Edges for irrelevant objects are assumed to be generated from a noise distribution irrespective of cluster assignments. We name the proposed method as *subset ReMatch* (subset relational matching).

Figure 1 shows an example of the output result with the proposed method when two bipartite networks are given as input. Each plot represents the adjacency matrix, where a dot at (i, j) indicates existence of an edge between the ith Type1 object and the jth Type2 object. Here, Type1 and Type2 objects are e.g. documents and words, respectively, in a document-word bipartite network. Note that input networks is not restricted to bipartite networks; they can be networks with a single type such as person-person networks, those with multiple types such as user-item-tag networks, or those with multiple relations such as 'like' and 'hate' relations. In addition, the proposed method can handle more than two networks. The number of objects can be different across different networks; the adjacency matrix size of Networks 1 and 2 is (120×140) and (90×110), respectively, in this example. The proposed method outputs relevance for each object and cluster assignments for relevant objects. In Figure 1(b), objects are sorted by their relevance and cluster assignments. The red lines indicates the border of clusters, and the lower right region that is not surrounded by red lines represents edges of irrelevant objects. The cluster assignments indicate their correspondence; for instance, with Type2, objects 1–20 in Network1 and objects 1–20 in Network2 are matched since they are assigned into cluster 'a', and objects 21–60 in Network1 and objects 21–40 in Network2 are matched since they are assigned into cluster 'b'. Clusters in different networks have similar connectivities; The first Type2 cluster 'a' has a high link probability to the first Type1 cluster 'A', but has a low link probability to the second Type1 cluster 'B', in both networks. Some clusters can appear only in a network; the fourth Type1 cluster 'D' does not appear in Network1, but appears in Network2. With the proposed method, the number of clusters is automatically inferred using Dirichlet processes. Irrelevant objects are connected irrespective of their partners' cluster assignments. The number of irrelevant objects can be different among networks; there are no irrelevant Type1 objects in Network2.

The paper is organized as follows: In Section 2, related work is briefly outlined. In Section 3, we propose a probabilistic model for robust unsupervised cluster matching that discovers relevance and cluster matching from multiple networks without correspondence. In Section 4, an efficient Bayesian inference procedure of the proposed model based on collapsed Gibbs sampling is presented. In Section 5, we describe connection between the proposed model and a closely related existing model. In Section 6, we experimentally demonstrate the effectiveness of the proposed model by using synthetic and real data sets, which include multilingual word clustering without dictionaries/aligned-texts. Finally, we present concluding remarks and a discussion of future work in Section 7.

2 Related work

Object matching is the task of finding correspondence between objects in different domains, such as images and annotations (Socher and Fei-Fei, 2010), user identifiers in different databases (Li et al, 2009), sentences written in different languages (Gale

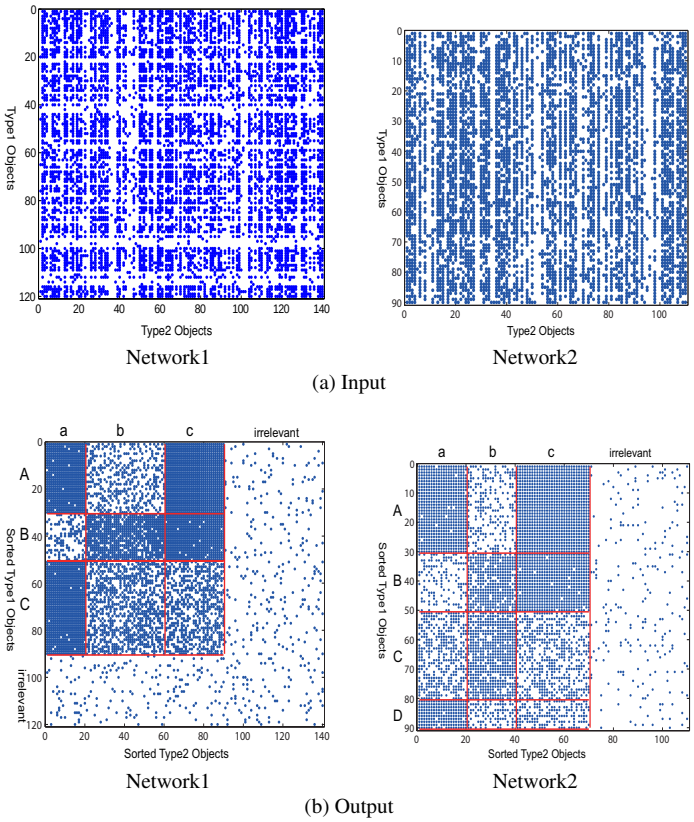


Fig. 1 Example of the input and output for the proposed method.

and Church, 1991; Rapp, 1999). Most object matching methods involve similarity or correspondence information. However, the information is unavailable in some applications because of its high cost or privacy protection.

For this situation, a number of unsupervised object matching methods, e.g. matching canonical correlation analysis (Haghighi et al, 2008), kernelized sorting (Quadrianto et al, 2010), convex kernelized sorting (Djuric et al, 2012), least square object matching (Yamada and Sugiyama, 2011) and Bayesian object matching (Klami, 2013, 2012), and unsupervised cluster matching methods, e.g. many-to-many matching latent variable models (Iwata et al, 2013) and ReMatch (Iwata et al, 2016), have been proposed. Intuitively, they find matching using relationship of objects within a domain, since distance between objects in different domains cannot be calculated with unsupervised setting, but distance between objects in the same domain can be calculated. When irrelevant objects exist, the relationship is distorted, and the performance of these methods would be deteriorated. On the other hand, with the proposed method, the relationship within a domain is maintained by introducing relevance. Another advantage of the proposed method over existing unsupervised object matching methods is the proposed method can handle different numbers of objects. The

Table 1 Notation.

Symbol	Description
D	number of networks
N_{dt}	number of Type-t objects in network d
K_t	number of realized clusters for Type-t
x_{dnm}	edge existence between the nth Type1 object and the mth Type2 object in network d, $x_{dnm} \in \{0, 1\}$
z_{dtn}	cluster assignment of the nth Type-t object in network d, $z_{dtn} \in \{0, 1, \cdots, \infty\}$
r_{dtn}	relevance of the nth Type-t object in network d, $r_{dtn} \in \{0, 1\}$

proposed method is an extension of the ReMatch (Iwata et al, 2016), which is an unsupervised cluster matching method for network data, for noisy data. By clustering only relevant objects, the proposed method can obtain intuitive cluster matching results compared with the ReMatch.

There have been proposed many methods for discovering latent clusters from a single network, such as the stochastic block model (Wang and Wong, 1987; Nowicki and Snijders, 2001), mixed membership stochastic block model (Airoldi et al, 2008), infinite relational model (IRM) (Kemp et al, 2006), and subset IRM (Ishiguro et al, 2012). However, these methods cannot discover shared groups from multiple networks. The proposed method corresponds to applying the subset IRM to a single large network that is constructed by combining all the networks, where edges between different networks are assumed to be missing.

3 Proposed model

Suppose that we are given D bipartite networks $\mathbf{X} = \{\mathbf{X}_d\}_{d=1}^{D}$, where $\mathbf{X}_d = \{\{x_{dnm}\}_{n=1}^{N_{d1}}\}_{m=1}^{N_{d2}}$ is the dth network, and $x_{dnm} \in \{0, 1\}$ indicates existence of an edge between the nth Type1 object and the mth Type2 object. Our notation is summarized in Table 1. Although we assume that given data are bipartite networks for simplicity, the proposed model is applicable to networks with a single type, those with more than two types, and those with multiple relations.

The model proposed for this task is a probabilistic generative model of multiple networks. We assume that there are two classes of objects: 1) relevant objects that have hidden common structure and are connected depending on their latent cluster assignments, and 2) irrelevant objects that are noisy and are randomly connected to other objects. A latent relevancy variable $r_{dtn} \in \{0, 1\}$ is associated with each object, where $r_{dtn} = 1$ if it is relevant and $r_{dtn} = 0$ otherwise. For modeling networks for relevant objects, the proposed model is based on an IRM (Kemp et al, 2006), but is extended for multiple networks by sharing clusters and connectivity parameters. The proposed model assumes that there are potentially a countably infinite number of clusters for each type, and clusters are shared across different networks. Each relevant object is assigned into a cluster $z_{dtn} \in \{1, \cdots, \infty\}$. An edge between two relevant objects is generated depending on their cluster assignments. In particular, the probability of connecting objects assigned into clusters k and ℓ is assumed to be $\theta_{k\ell}$, which is common across all networks. An edge for irrelevant objects are generated randomly with probability ϕ, which does not depend on the cluster assignment of

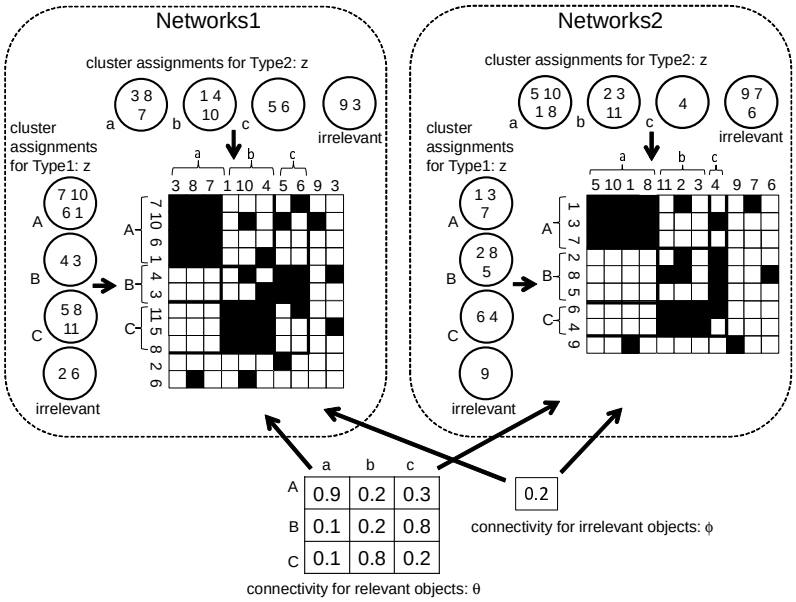


Fig. 2 Generative process of two bipartite networks with the proposed model.

the opponent object. Figure 2 shows an illustration of the generative process of two bipartite networks with the proposed model.

The proposed model generates multiple bipartite networks $\mathbf{X}$ according to the following process:

1. Draw connectivity for irrelevant objects $\phi \sim \mathrm{Beta}(a, b)$
2. For each cluster for Type1: $k = 1, \cdots, \infty$
   (a) For each cluster for Type2: $\ell = 1, \cdots, \infty$
      i. Draw connectivity for relevant objects $\theta_{k\ell} \sim \mathrm{Beta}(c, d)$
3. For each type: $t = 1, 2$
   (a) Draw relevance probability $\lambda_t \sim \mathrm{Beta}(e, f)$
   (b) For each network: $d = 1, \cdots, D$
      i. For each object: $n = 1, \cdots, N_{dt}$
         A. Draw relevance
            $r_{dtn} \sim \mathrm{Bernoulli}(\lambda_t)$
         B. Draw cluster assignment
            $z_{dtn} \sim \mathrm{CRP}(\alpha_t)$ if $r_{dtn} = 1$
            $z_{dtn} = 0$ otherwise

1. For each network: $d = 1, \cdots, D$
   (a) For each object of Type1: $n = 1, \cdots, N_{d1}$
      i. For each object of Type2: $m = 1, \cdots, N_{d2}$
         A. Draw relation
            $x_{dnm} \sim \mathrm{Bernoulli}(\theta^{r_{d1n} r_{d2m}}_{z_{d1n} z_{d2m}} \phi^{1 - r_{d1n} r_{d2m}})$

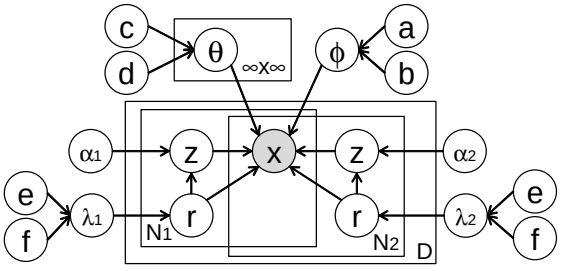


Fig. 3 Graphical model representation of the proposed model for bipartite networks.

Here, $\mathrm{CRP}(\alpha)$ indicates the Chinese restaurant process (Blackwell and MacQueen, 1973) with concentration parameter α, which is a representation of Dirichlet processes, and $\mathbb{I}(A) = 1$ if A is true, $\mathbb{I}(A) = 0$ otherwise. By using the CRP, the number of clusters can be automatically estimated from a given data. We use beta distributions for priors of Bernoulli parameters, ϕ, $\boldsymbol{\theta} = \{\theta_{k\ell}\}$, $\boldsymbol{\lambda} = \{\lambda_t\}$, since the beta distribution is the conjugate prior of the Bernoulli distribution, which enables us to develop an efficient Bayesian inference algorithm as describe in Section 4. At the last line of the generative process, an edge is generated using the Bernoulli distribution, where the probability is $\theta_{z_{d1n}z_{d2m}}$ when both objects are relevant because $r_{d1n}r_{d2m} = 1$, and it is ϕ when either object is irrelevant. Figure 3 shows graphical model representation of the proposed model, where shaded and unshaded nodes indicate observed and latent variables, respectively.

The joint distribution of networks $\mathbf{X}$, cluster assignments $\mathbf{Z} = \{z_{dtn}\}$ and relevance $\mathbf{R} = \{r_{dtn}\}$ given hyperparameters $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2\}, a, b, c, d, e, f$ is described as follows:

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{R}|\boldsymbol{\alpha}, a, b, c, d, e, f) = p(\mathbf{R}|e, f)p(\mathbf{Z}|\mathbf{R}, \boldsymbol{\alpha})p(\mathbf{X}|\mathbf{Z}, \mathbf{R}, a, b, c, d). \quad (1)$$

The first factor on the right hand side of (1) is calculated by

$$\begin{aligned} p(\mathbf{R}|e, f) &= \prod_{t=1}^{2} \int \prod_{d=1}^{D} \prod_{n=1}^{N_{dt}} p(r_{dtn}|\lambda_t)p(\lambda_t|e, f)d_{\lambda_t} \\ &= \prod_{t=1}^{2} \frac{B(e + L_{t1}, f + L_{t0})}{B(e, f)} \end{aligned} \quad (2)$$

where

$$L_{tr} = \sum_{d=1}^{D} \sum_{n=1}^{N_{dt}} r_{dtn}, \quad (3)$$

is the number of Type-t objects with relevance r, and $B(e, f) = \frac{\Gamma(e)\Gamma(f)}{\Gamma(e+f)}$ is the beta function. The Bernoulli parameters λ_t are analytically integrated out by using

conjugate Beta distribution priors $p(\lambda_t|e,f) = \mathrm{Beta}(e,f)$. The second factor of (1) is given by

$$p(\mathbf{Z}|\mathbf{R},\boldsymbol{\alpha}) = \prod_{t=1}^{2} \frac{\alpha^{K_t} \prod_{k=1}^{K_t}(M_{tk}-1)!}{\alpha_t(\alpha_t+1)\cdots(\alpha_t+M_t-1)}, \tag{4}$$

since cluster assignments are drawn from the Chinese restaurant process for relevant objects and deterministic functions for irrelevant objects. Here,

$$M_{tk} = \sum_{d=1}^{D}\sum_{n=1}^{N_{dt}} r_{dtn}\mathbb{I}(z_{dtn}=k), \tag{5}$$

is the number of Type-t relevant objects assigned to cluster k, $M_t = \sum_{k=1}^{K_t} M_{tk}$ is the total number of Type-t relevant objects, and K_t is the current number of realized clusters for Type-t. The third factor of (1) is given by

$$\begin{aligned} p(\mathbf{X}|\mathbf{Z},\mathbf{R},a,b,c,d) &= \int \prod_{d=1}^{D}\prod_{n=1}^{N_{d1}}\prod_{m=1}^{N_{d2}} (1-r_{d1n}r_{d2m}) p(x_{dnm}|\phi)p(\phi|a,b)d\phi \\ &\times \int \prod_{d=1}^{D}\prod_{n=1}^{N_{d1}}\prod_{m=1}^{N_{d2}} r_{d1n}r_{d2m} p(x_{dnm}|\theta_{z_{d1n}z_{d2m}})p(\boldsymbol{\theta}|c,d)d\boldsymbol{\theta} \\ &= \frac{B(a+Q,b+\bar{Q})}{B(a,b)} \prod_{k=1}^{K_1}\prod_{\ell=1}^{K_2} \frac{B(c+N_{k\ell},d+\bar{N}_{k\ell})}{B(c,d)}, \end{aligned} \tag{6}$$

where relations between relevant objects, $\{(d1n,d2m)|r_{d1n}r_{d2m}=1\}$, are drawn from a cluster-dependent distribution $\mathrm{Bernoulli}(\theta_{z_{d1n}z_{d2m}})$, and the other relations, $\{(d1n,d2m)|r_{d1n}r_{d2m}=0\}$, are drawn from $\mathrm{Bernoulli}(\phi)$. The Bernoulli parameters, $\boldsymbol{\theta}=\{\theta_{k\ell}\}$, ϕ, are analytically integrated out due to their conjugate priors. Here,

$$Q = \sum_{d=1}^{D}\sum_{n=1}^{N_{d1}}\sum_{m=1}^{N_{d2}} (1-r_{d1n}r_{d2m})x_{dnm}, \tag{7}$$

is the number of edges for irrelevant objects,

$$\bar{Q} = \sum_{d=1}^{D}\sum_{n=1}^{N_{d1}}\sum_{m=1}^{N_{d2}} (1-r_{d1n}r_{d2m})(1-x_{dnm}), \tag{8}$$

is the number of non-edges for irrelevant objects,

$$N_{k\ell} = \sum_{d=1}^{D}\sum_{n=1}^{N_{d1}}\sum_{m=1}^{N_{d2}} r_{d1n}r_{d2m}\mathbb{I}(z_{d1n}=k)\mathbb{I}(z_{d2m}=\ell)x_{dnm}, \tag{9}$$

is the number of edges between clusters k and ℓ for relevant objects, and

$$\bar{N}_{k\ell} = \sum_{d=1}^{D}\sum_{n=1}^{N_{d1}}\sum_{m=1}^{N_{d2}} r_{d1n}r_{d2m}\mathbb{I}(z_{d1n}=k)\mathbb{I}(z_{d2m}=\ell)(1-x_{dnm}), \tag{10}$$

is the number of non-edges between clusters k and ℓ for relevant objects.

4 Inference

We describe the inference procedures for the proposed model based on collapsed Gibbs sampling (Kemp et al, 2006). The parameters, ϕ, $\boldsymbol{\theta} = \{\theta_{k\ell}\}$ and $\boldsymbol{\lambda} = \{\lambda_t\}$, can be analytically marginalized out thanks to their conjugacy. The unknown latent variables to be inferred are cluster assignments $\mathbf{Z} = \{z_{dtn}\}$ and relevance $\mathbf{R} = \{r_{dtn}\}$. We sample z_{dtn} and r_{dtn} simultaneously. Given the current state of all but one latent assignment (z_i, r_i), where $i = (d, t, n)$, a new value for (z_i, r_i) is sampled from the following probability,

$$p(z_i = k, r_i | \mathbf{X}, \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}) \propto p(z_i = k, r_i | \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}) p(\mathbf{X}^{+i} | z_i = k, r_i, \mathbf{X}^{\setminus i}, \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}), \tag{11}$$

where the superscript $\setminus i$ denotes the set or value when excluding object i, $\mathbf{X}^{+i} = \{x_{dnm}\}_{m=1}^{N_{d\bar{t}}}$ is the set of edges for object i, $\bar{t} = 1$ if $t = 2$ and $\bar{t} = 2$ if $t = 1$, and this is derived from (1). Here, we omit the hyperparameters for simplicity. When the object is irrelevant $r_i = 0$, its cluster assignment is always $z_i = 0$. When the object is relevant $r_i = 1$, its cluster assignment is sampled from existing clusters $\{1, \cdots, K_t\}$ or a new cluster $K_t + 1$. Since we sample a cluster assignment for each object, there is no probability that z_i becomes greater than $K_t + 1$. The first factor is calculated by

$$\begin{aligned}
&p(z_i = k, r_i | \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}) \\
&= \frac{p(r_i, \mathbf{R}^{\setminus i}) p(z_i = k, \mathbf{Z}^{\setminus i} | r_i, \mathbf{R}^{\setminus i})}{p(\mathbf{R}^{\setminus i}) p(\mathbf{Z}^{\setminus i} | \mathbf{R}^{\setminus i})} \\
&\propto \begin{cases} f + L_{t0}^{\setminus i} & \text{if } r_i = 0, z_i = 0, \\ (e + L_{t1}^{\setminus i}) \frac{M_{tk}^{\setminus i}}{\alpha_t + \sum_{k'=1}^{K_t} M_{tk'}^{\setminus i}} & \text{if } r_i = 1, z_i = k \in \{1, \cdots, K_t\}, \\ (e + L_{t1}^{\setminus i}) \frac{\alpha_t}{\alpha_t + \sum_{k'=1}^{K} M_{tk'}^{\setminus i}} & \text{if } r_i = 1, z_i = K_t + 1, \end{cases}
\end{aligned} \tag{12}$$

using (2) and (4). The second factor for irrelevant objects is calculated by

$$\begin{aligned}
&p(\mathbf{X}^{+i} | z_i = 0, r_i = 0, \mathbf{X}^{\setminus i}, \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}) \\
&= \frac{p(\mathbf{X}^{+i}, \mathbf{X}^{\setminus i} | z_i = 0, \mathbf{Z}^{\setminus i}, r_i = 0, \mathbf{R}^{\setminus i})}{p(\mathbf{X}^{\setminus i} | \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i})} \\
&= \frac{B(a + Q^{\setminus i} + Q^{+i0}, b + \bar{Q}^{\setminus i} + \bar{Q}^{+i0})}{B(a + Q^{\setminus i}, b + \bar{Q}^{\setminus i})},
\end{aligned} \tag{13}$$

using (6), where the superscript $+i0$ denotes that the same statistics are computed on $\mathbf{X}^{+i}$ assuming $r_i = 0$. Similarly, the second factor for relevant objects is calculated as follows,

$$\begin{aligned}
&p(\mathbf{X}^{+i} | z_i = k, r_i = 1, \mathbf{X}^{\setminus i}, \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}) \\
&= \frac{B(a + Q^{\setminus i} + Q^{+i1}, b + \bar{Q}^{\setminus i} + \bar{Q}^{+i1})}{B(a + Q^{\setminus i}, b + \bar{Q}^{\setminus i})} \\
&\times \prod_{\ell=1}^{K_{\bar{t}}} \frac{B(c + N_{k\ell}^{\setminus i} + N_{k\ell}^{+ik}, d + \bar{N}_{k\ell}^{\setminus i} + \bar{N}_{k\ell}^{+ik})}{B(c + N_{k\ell}^{\setminus i}, d + \bar{N}_{k\ell}^{\setminus i})},
\end{aligned} \tag{14}$$

Algorithm 1 Inference procedure for the proposed model.

initialize cluster assignments $\mathbf{Z}$, relevance $\mathbf{R}$, and hyperparameters $\boldsymbol{\alpha}, a, b, c, d, e, f$
repeat
  for $d = 1$ to D **do**
    for $t = 1$ to 2 **do**
      for $n = 1$ to N_{dt} **do**
        sample cluster assignment z_{dtn} and relevance r_{dtn} by (11)
      end for
    end for
  end for
  sample hyperparameters $\boldsymbol{\alpha}, a, b, c, d, e, f$ by (15)
until end condition is met

where the superscript $+ik$ denotes that the same statistics are computed on $\mathbf{X}^{+i}$ assuming $r_i = 1$ and $z_i = k$.

We fit hyperparameters $\boldsymbol{\alpha}, a, b, c, d, e, f$ by posterior sampling assuming Gamma priors. In particular, we sample the value of hyperparameter a according to the following posterior probability:

$$p(a = \hat{a}|\mathbf{X}, \mathbf{Z}, \mathbf{R}, \boldsymbol{\alpha}, b, c, d, e, f) \propto p(a = \hat{a})p(\mathbf{X}, \mathbf{Z}, \mathbf{R}|\boldsymbol{\alpha}, a = \hat{a}, b, c, d, e, f), \tag{15}$$

In our experiments, we used $p(a) = \mathrm{Gamma}(5, 5)$, and ten candidates of $\hat{a}$ are generated for each sampling. The other hyperparameters $\boldsymbol{\alpha}, b, c, d, e, f$ are sampled in the same way.

By iterating sampling of cluster assignments and relevance variables for all objects and sampling of hyperparameters, we obtain an estimate of posteriors. Algorithm 1 summarizes the inference procedure for the proposed model. We initialize cluster assignments relevance and hyperparameters randomly. We use the last sample of the cluster assignments in the inference for matching. When initial assignments are different, different cluster assignments are obtained after the sampling. The computational complexity for each iteration of the collapsed Gibbs sampling is $O(\sum_{d=1}^{D}\sum_{t=1}^{T} N_{dt}K_t)$ since we calculate probabilities of $K_t + 1$ clusters for each object, and there are $\sum_{d=1}^{D}\sum_{t=1}^{T} N_{dt}$ objects in total.

Although all parameters are marginalized out during the inference, the posteriors of parameters are calculated using the samples as follows,

$$p(\phi|\mathbf{X}, \mathbf{Z}, \mathbf{R}) = \mathrm{Beta}(a + Q, b + \bar{Q}), \tag{16}$$

$$p(\theta_{k\ell}|\mathbf{X}, \mathbf{Z}, \mathbf{R}) = \mathrm{Beta}(c + N_{k\ell}, d + \bar{N}_{k\ell}), \tag{17}$$

$$p(\lambda_t|\mathbf{X}, \mathbf{Z}, \mathbf{R}) = \mathrm{Beta}(e + L_{t1}, f + L_{t0}). \tag{18}$$

5 Relationship with ReMatch

The ReMatch (Iwata et al, 2016) is an unsupervised cluster matching method for network data, which corresponds to the proposed model without relevance variables. In particular, the ReMatch assumes the following generative process:

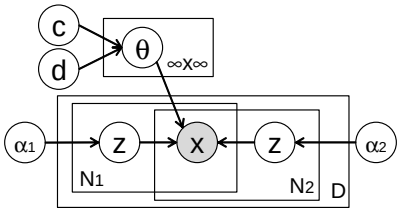


Fig. 4 Graphical model representation of the ReMatch for bipartite networks.

1. For each cluster for Type1: $k = 1, \cdots, \infty$
   (a) For each cluster for Type2: $\ell = 1, \cdots, \infty$
      i. Draw connectivity $\theta_{k\ell} \sim \mathrm{Beta}(c, d)$
2. For each type: $t = 1, 2$
   (a) For each network: $d = 1, \cdots, D$
      i. For each object: $n = 1, \cdots, N_{dt}$
         A. Draw cluster assignment $z_{dtn} \sim \mathrm{CRP}(\alpha_t)$

1. For each network: $d = 1, \cdots, D$
   (a) For each object of Type1: $n = 1, \cdots, N_{d1}$
      i. For each object of Type2: $m = 1, \cdots, N_{d2}$
         A. Draw relation
            $x_{dnm} \sim \mathrm{Bernoulli}(\theta_{z_{d1n} z_{d2m}})$

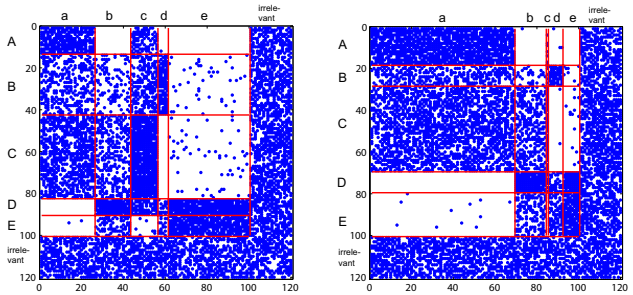
Figure 4 shows graphical model representation of the ReMatch.

The ReMatch assigns all objects into shared clusters even if they are irrelevant. In real-world applications, some objects would not form clusters. These irrelevant objects deteriorate the cluster matching performance since mistakenly estimated matching affect on estimation of matching of other objects. On the other hand, the proposed model assigns only relevant objects into shared clusters, which enables us to handle noisy observation, and improve the performance.

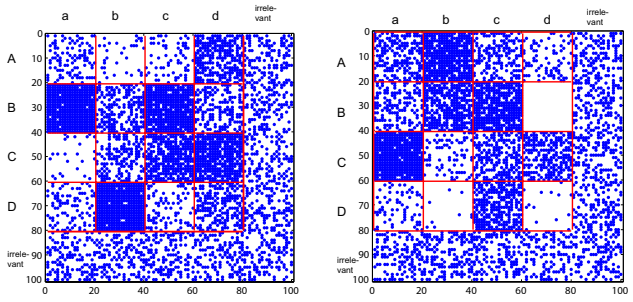
6 Experiments

6.1 Synthetic data

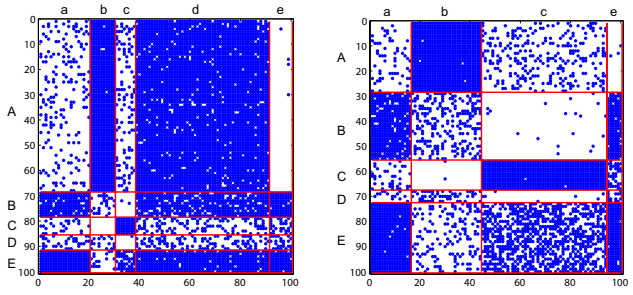
We evaluated the proposed subset ReMatch by using the following three types of synthetic data sets with two bipartite networks: Noisy-Dirichlet, Noisy-Partial and Dirichlet. With the Noisy-Dirichlet data, cluster proportions were generated from a symmetric Dirichlet distribution for each type and for each network, where there were 100 relevant objects, 20 irrelevant objects and five clusters for each type. With the Noisy-Partial data, some clusters appear in either network. In particular, there were five clusters, but the first/last cluster does not appear in Network2/Network1. There were 20 relevant objects for each cluster and 20 irrelevant objects for each type. The Dirichlet data is the same with the Noisy-Dirichlet data except that there were no irrelevant objects. With all data sets, the edge probability was generated from a beta



(a) Noisy-Dirichlet



(b) Noisy-Partial



(c) Dirichlet

Fig. 5 Examples of the synthetic data sets. Objects are aligned by their cluster assignments and the red lines indicate the border of clusters. The lower right region that is not surrounded by red lines represents edges of irrelevant objects in the Noisy-Dirichlet and Noisy-Partial data.

distribution, $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$, and edges were generated according to a Bernoulli distribution depending on the assigned clusters. Figure 5 shows examples of the synthetic data sets.

We compared the proposed method with the following four methods: ReMatch, MMLVM, IRM+KS and KS. The ReMatch is an unsupervised cluster matching method for network data (Iwata et al, 2016), and it corresponds to the proposed method without irrelevant objects. The MMLVM is the many-to-many matching latent variable

model (Iwata et al, 2013). It is an unsupervised cluster matching method for real-valued data, where Gaussian noise is assumed for observation. The KS is the convex kernelized sorting (Djuric et al, 2012), which is an unsupervised object matching method that finds one-to-one matching and does not cluster objects. The IRM+KS is the combination of the IRM and KS, where objects are clustered for each network individually, and then the obtained clusters are matched by the convex kernelized sorting.

For the evaluation measurement, we used matching adjusted Rand index (MARI) (Iwata et al, 2016). The MARI becomes high when pairs of matched/unmatched objects in different networks were correctly assigned into the same/different clusters, and it becomes low when they were incorrectly assigned. It takes the value from -1 to 1, and gives 0 for random cluster matching. In particular, the MARI for Type-t is calculated by

$$\mathrm{MARI}_t = \frac{h_{t1} + h_{t2} - \mu_t}{N_{t1}N_{t2} - \mu_t}, \tag{19}$$

where $h_{t1}(h_{t2})$ is the number of object pairs in different networks that are correctly assigned into the same (different) clusters both in the estimated and true assignments, and μ_t is the expected value of $h_{t1} + h_{t2}$, which is obtained by

$$\mu_t = \frac{(h_{t1} + h_{t3})(h_{t1} + h_{t4}) + (h_{t2} + h_{t3})(h_{t2} + h_{t4})}{N_{t1}N_{t2}}, \tag{20}$$

where h_{t3} (h_{t4}) is the number of object pairs in different networks that are incorrectly assigned into the same (different) clusters in the estimation but assigned into the different (same) clusters in the true assignments. With the proposed method, the set of irrelevant objects are considered as a cluster when the MARI was calculated.

Table 2 shows the MARI with the synthetic data sets. In all cases, the proposed method achieved the best MARI. With the Dirichlet data, which did not contain any irrelevant objects, the MARI of the proposed method and ReMatch were comparable. This result indicates that the performance of the proposed method is not deteriorated even when given networks consist of only relevant objects. The performance of the MMLVM was low since it assumes Gaussian noise, which is not appropriate for network data. On the other hand, the proposed model assumes Bernoulli noise, which is a natural assumption for network data. The MARI achieved by KS was low since it does not cluster objects. By clustering objects using IRM, IRM+KS improved the performance compared with KS. However, it did not outperform the proposed method, because errors accumulated in clustering by IRM cannot be corrected in the matching process with KS. In contrast, since the proposed method performs clustering and matching simultaneously, it can find clusters that are appropriate when matched.

6.2 Real-world data

Next, we evaluated the proposed method by using the following four real-world data sets: 20News, NIPS, Movie and Multilingual. The 20News data consisted of binary occurrence data for 100 words in documents obtained from 20 Newsgroups data

Table 2 Matching adjusted Rand indices with the synthetic data sets and their standard errors, which were averaged over 100 experiments for each data set. The values in bold are not significantly different from the best performing method in each row according to a paired t-test.

	Proposed	ReMatch	MMLVM	IRM+KS	KS
Noisy-Dirichlet	$\mathbf{0.579} \pm 0.029$	0.463 ± 0.030	0.228 ± 0.023	0.069 ± 0.021	0.008 ± 0.001
Noisy-Partial	$\mathbf{0.684} \pm 0.021$	0.579 ± 0.021	0.326 ± 0.018	0.062 ± 0.019	0.008 ± 0.001
Dirichlet	$\mathbf{0.454} \pm 0.036$	$\mathbf{0.438} \pm 0.036$	0.148 ± 0.026	0.124 ± 0.027	0.002 ± 0.001

set (Lang, 1995) [1]. The documents were categorized into the following four categories: Computers, Recreation, Science and Talk. We generated two disjoint sets of documents, where 250 documents were sampled for each category, and created two document-word networks with size (1000×100). The two networks contained the same set of 100 words although their correspondence were assumed to be unknown.

The NIPS data consist of binary word occurrence data for the NIPS conference papers from 2001 to 2003 [2]. There were 593 documents, and vocabulary size was 13,762. We split the documents and words into two disjoint sets, and created two document-word networks with size (296×6881) and (297×6881). The papers were categorized in 13 categories, such as Algorithms & Architectures, Applications and Neuroscience.

The Movie data consist of movie ratings obtained from the Eachmovie data set, where a movie and a user are linked when the user has rated the movie. We omit the users that rated less than 50 movies, and split the movies and users into two disjoint sets, and created two movie-user networks with size (507×7180). The movies were categorized into ten genres, such as Action, Comedy and Romance.

The Multilingual data were obtained from Wikipedia articles in English, German, Italian and Japanese in the following five categories: Nobel laureates in Physics, Nobel laureates in Chemistry, American basketball players, American composers and English footballers. We created two document-word networks for each pair of languages, where 50 documents were sampled for each category that appeared in both languages. We used 1,000 frequent words after removing stop-words for each language. The size of a created document-word network was (150×1000). Note that 20News, NIPS and Movie data were not originally multiple networks and therefore we split objects into two networks randomly, but the Multilingual data were originally multiple networks.

For comparing methods with real-world data, we used the ReMatch, MMLVM and IRM+KS. We excluded KS, since they were not effective and took too long computational time with large networks as shown Tables 2 and 4. The average MARI scores with the real-world data sets are shown in Table 3. The proposed method achieved the highest MARI with all data sets. This result implies that inferring relevance of objects is effective for find matching in noisy real-world network data. With the 20News data, there were no significant difference between the proposed method and ReMatch. It would be because words were completely shared across two different networks with the 20News data, and noise was small compared with the other data

[1] Available at `http://www.cs.nyu.edu/~roweis/data.html`

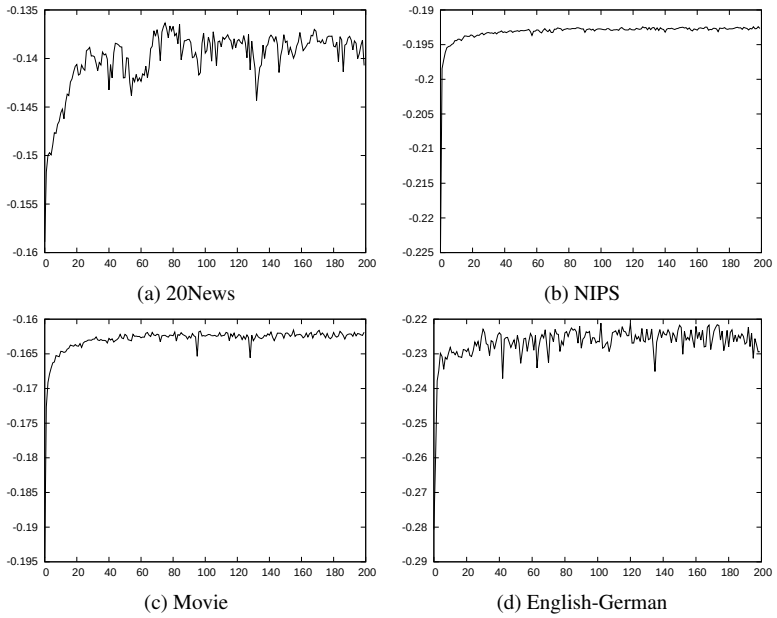
[2] Available at `http://ai.stanford.edu/~gal/`

Table 3 Matching adjusted Rand indices with the real-world data sets and their standard errors, which were averaged over 30 experiments for each data set. The values in bold are not significantly different from the best performing method in each row according to a paired t-test.

	Proposed	ReMatch	MMLVM	IRM+KS
20News	$\mathbf{0.178} \pm 0.011$	$\mathbf{0.153} \pm 0.015$	0.030 ± 0.001	-0.009 ± 0.010
NIPS	$\mathbf{0.221} \pm 0.004$	0.174 ± 0.005	NA	0.005 ± 0.011
Movie	$\mathbf{0.049} \pm 0.001$	0.045 ± 0.000	NA	0.001 ± 0.001
English-German	$\mathbf{0.203} \pm 0.026$	0.118 ± 0.021	0.098 ± 0.008	-0.004 ± 0.012
English-Italian	$\mathbf{0.263} \pm 0.026$	0.160 ± 0.021	0.102 ± 0.010	0.011 ± 0.011
English-Japanese	$\mathbf{0.244} \pm 0.022$	0.188 ± 0.014	0.069 ± 0.012	-0.001 ± 0.007
German-Italian	$\mathbf{0.208} \pm 0.023$	0.130 ± 0.017	0.087 ± 0.008	-0.007 ± 0.009
German-Japanese	$\mathbf{0.165} \pm 0.019$	0.090 ± 0.017	0.042 ± 0.009	0.013 ± 0.012
Italian-Japanese	$\mathbf{0.262} \pm 0.024$	$\mathbf{0.213} \pm 0.017$	0.078 ± 0.010	0.022 ± 0.011

Table 4 Computational time with the Noisy-Dirichlet data with 500 relevant objects for each type and for each network.

	Proposed	ReMatch	MMLVM	IRM+KS	KS
Noisy-Dirichlet	10 minutes	9 minutes	7 hours	4 minutes	13 hours
20News	9 minutes	4 minutes	10 hours	8 minutes	NA
NIPS	2 hours	2 hours	NA	4 hours	NA
Movie	38 hours	4 hours	NA	4 hours	NA
English-German	1 hour	13 minutes	25 hours	25 minutes	NA



(a) 20News (b) NIPS

(c) Movie (d) English-German

Fig. 6 Averaged training log likelihoods for each iteration in the inference.

sets. The results by the MMLVM with NIPS and Movie data sets were not available due to its high computational cost.

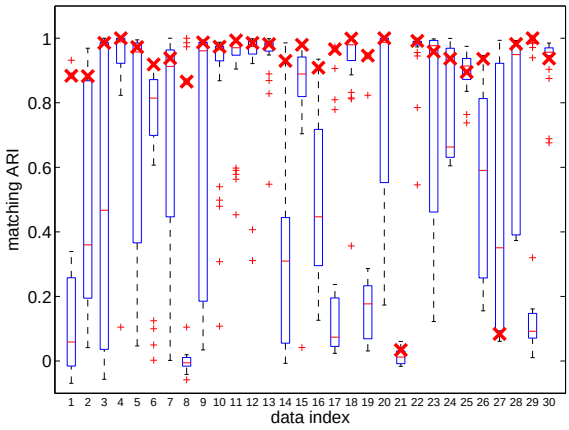
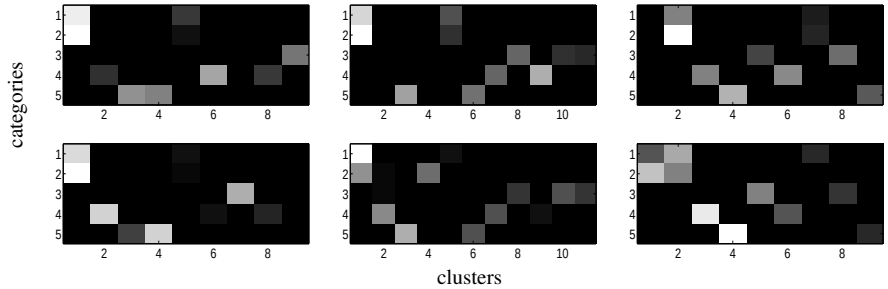


Fig. 7 Box plot of the MARI with different initializations using 30 Noisy-Dirichlet data sets. For each data set, we conducted 30 runs with different initializations. The bottom and top of the box are the first and third quartiles, the bar inside the box is the median, and red '×' indicates the MARI of the best log likelihood.



(a) English-German (b) English-Italian (c) English-Japanese

Fig. 8 Confusion matrices of the proposed method with the Multilingual data. The horizontal axis is the cluster index, and the vertical axis is the category index, 1: Nobel laureates in Physics, 2: Nobel laureates in Chemistry, 3: American basketball players, 4: American composers, and 5: English footballers. The brighter element indicates the larger number of documents in the category is assigned into the cluster. The top row shows the confusion matrix of English, and the bottom-most row shows that of German/Italian/Japanese.

Table 4 shows the computational time. The proposed method, ReMatch and IRM+KS are efficient. The KS requires long computational time since the computational complexity of KS is cubic in the number of objects. The IRM+KS does not take a long time since KS is applied not to objects but to clusters, where the number of clusters was much smaller than the number of objects. The MMLVM is inefficient because it requires matrix inversion for the inference. Figure 6 shows the averaged training log likelihoods of the proposed method for each iteration in the inference with the real-world data sets. The log likelihoods became high with about 100 iterations.

Figure 7 shows the sensitivity of the proposed model to different initialization. With some data sets, the variance is high, which indicates high sensitivity. However, when we selected a result according to the log likelihood (1), its MARI was high,

Table 5 Word clusters with the proposed method in the Multilingual data. Each pair of rows corresponds to a cluster, the upper row shows words from English, and the lower row shows words from German/Italy/Japanese in each pair of rows. The bottom-most pair of rows shows irrelevant words. Japanese words are translated in English.

(a) English-German

EN	composers musicians recording recordings songs composer walk concert piano orchestra album
DE	music life schrieb my jazz database movie is erschien at diskografie durchbruch komponisten
EN	basketball draft nba sportspeople averaged kg lb pick mvp rebounds rookie ncaa
DE	vereinigte basketballspieler nba draft basketball
EN	match goals cup clubs premier counted footballers friendly app gls fa uefa fifa manchester
DE	englischer verein trainer united nationalmannschaft fc league englische tore kader manchester
EN	prize laureates nobel
DE	preisverleihung nobelstiftung nobelpreis
EN	cooper eric felix dudley marcus phillips chairman receiving richardson fowler texas ramsey
DE	school leistungen treffer lee high bevor ernannt william anerkennung sonstiges abschluss

(b) English-Italian

EN	draft overall pro guard weight sportspeople thompson playoffs kg lb pick teammate assists
IT	media game high team basketball rookie scelto finali ala scelta titoli assoluta assist mvp star
EN	recording composers composer album piano musicians write recordings instrumental piece
IT	compositori compositore musicale tecniche album it me musicali band musicisti visita
EN	manchester honours transfer fifa euro correct draw soccer premiership substitute beckham
IT	manchester ferdinand presenze gol fifa cup premier neville reti centrocampista campbell ham
EN	prize nobel laureates
IT	premio premi vincito
EN	appointed moore mitchell sir aaron richardson elected marcus todd charge anthony heat gilbert
IT	van ottenne nominato wilson jack alan richardson londra membro national award britannici

(c) English-Japanese

EN	musicians composers album composer grammy
JP	performance piano classic history event instrument orchestra religion violin classic roman
EN	squad cup match goals premier clubs manchester counted footballers
JP	commentary england soccer club birthday foot relieve united
EN	fields alma mater professor institutions chemistry sciences physics faculty physicist scientists
JP	physics william scholar strike professor frederick chemical issue richard phd mali max
EN	johnson kevin retired jones round barry jordan teams basketball draft jackson thompson overall
JP	victory performance enters player basketball name
EN	fellow lawrence retirement mitchell glenn howard cooper confirmed anthony donald moore
JP	california london in chicago ray born source hertford al smith leave tony bra jack sony black

which is represented by ‘x’ in Figure 7. This result implies that the log likelihood is a good measurement to select a result from multiple runs.

Figure 8 shows the confusion matrices of the proposed method with the Multilingual data. The documents in the first and second categories, which were the Nobel laureates in Physics and Chemistry, were assigned into the same clusters, e.g. the first cluster in English-German. It is reasonable since they are closely related categories. Although the proposed method could not discriminate the Nobel laureates in Physics from those in Chemistry, it separated the Nobel laureates in Physics and Chemistry from the other categories. In addition, documents in the other categories were successfully assigned into different clusters. Some documents in the same category were assigned into different clusters, e.g. documents in the fifth category, which was En-

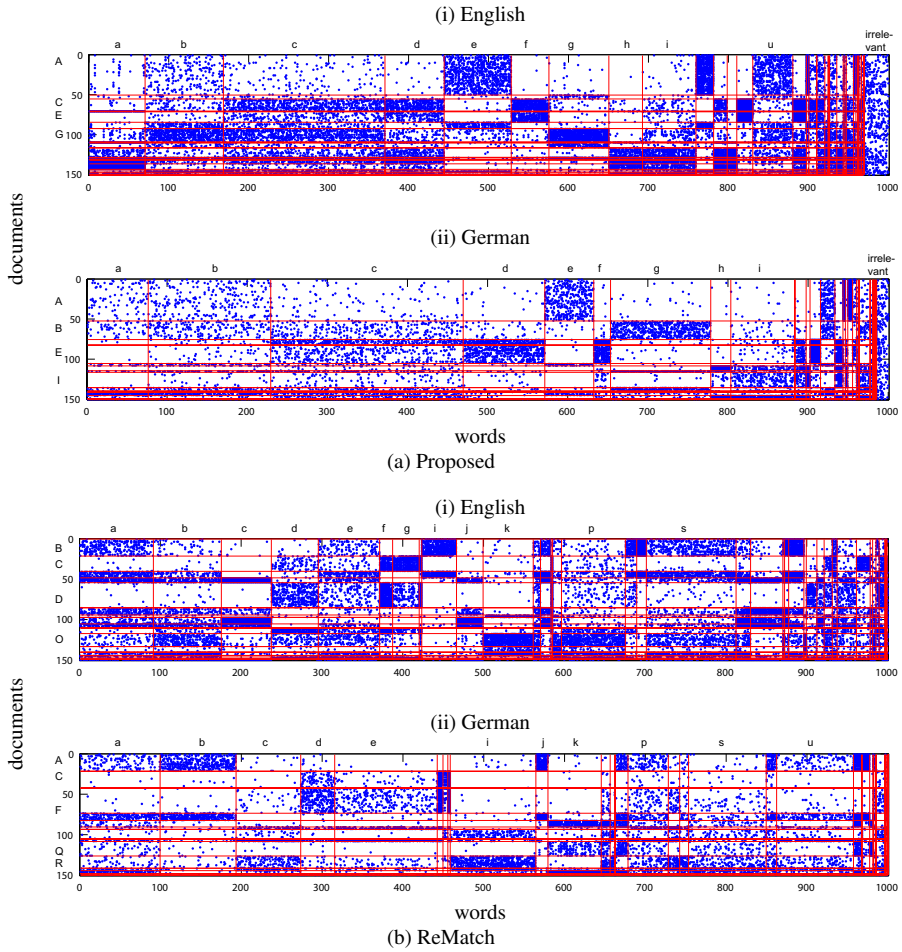


Fig. 9 Inferred shared latent clusters and relevance with the Multilingual English-German data (a) by the proposed method, and (b) by the ReMatch.

glish footballer, were assigned into the third and fourth clusters in English-German. It would be because the proposed method is an unsupervised method, and documents in the same category could use different vocabulary.

Table 5 shows word clusters inferred by the proposed method with the Multilingual data. Related words were appropriately assigned into the same clusters across different languages, e.g. in Table 5(a) the first, second, third and fourth cluster related to music, basketball, football and Nobel prize, respectively. The words that were not related to a specific category were inferred as irrelevant, which were shown in the bottom-most pair of rows in the table, such as common family name, city name, common noun and verb.

Figure 9(a) shows the inferred shared latent clusters and relevance by the proposed method with the Multilingual English-German data. There were few irrelevant

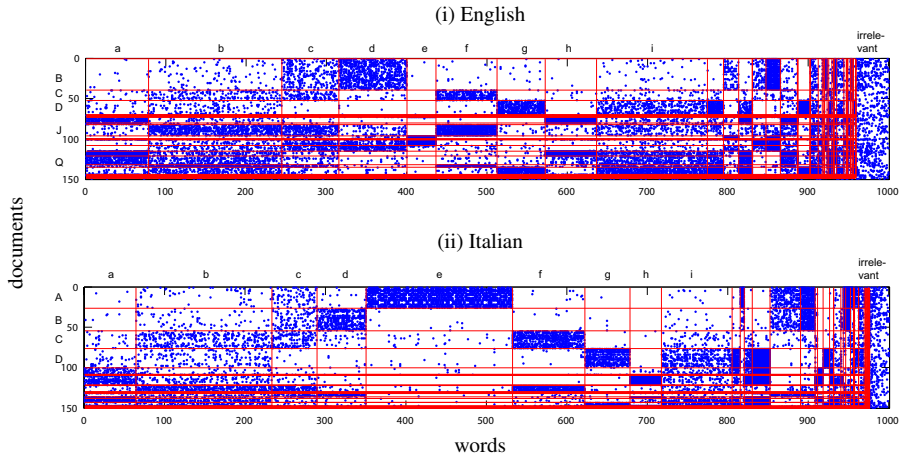


Fig. 10 Inferred shared latent clusters and relevance by the proposed method with the Multilingual English-Italian data.

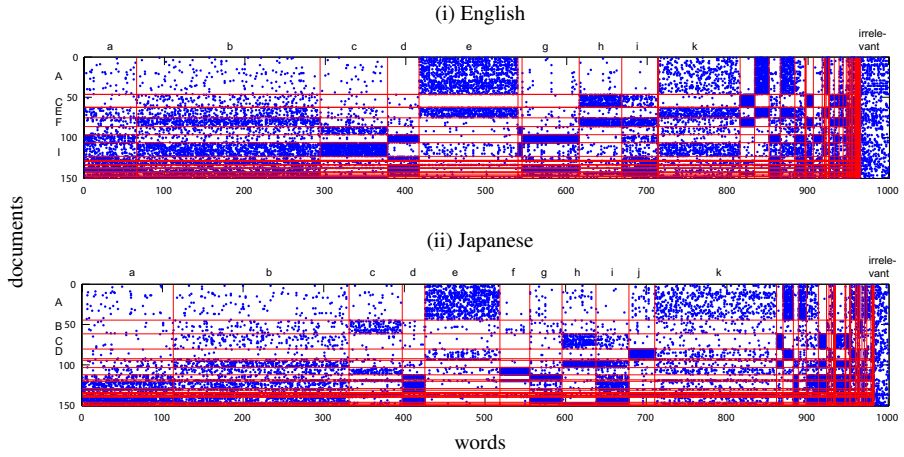


Fig. 11 Inferred shared latent clusters and relevance by the proposed method with the Multilingual English-Japanese data.

documents, but many irrelevant words. It is natural since all documents were sampled from the specific categories, but some words were not specific to a single category. The irrelevant words, which were shown in the right region, have edges to documents in all categories. Figure 9(b) shows the inferred shared latent clusters by the ReMatch. The size of clusters inferred by the ReMatch was smaller than that by the proposed method. By considering relevance, the proposed method could discover larger clusters which are easy to interpret. Figures 10 and 11 show the inferred shared latent clusters and relevance by the proposed method with the Multilingual English-Italian and English-Japanese data, respectively.

7 Conclusion

We proposed a probabilistic model for unsupervised cluster matching, which is a task to find matching between clusters of objects in different domains. Given multiple networks as inputs, the proposed method infers matching of clusters and relevance of objects simultaneously, which enables us to improve the matching performance by handling noisy observation. We have confirmed experimentally that the matching performance of the proposed method is higher than existing methods with synthetic and real-world data sets. With the proposed method, the ReMatch was extended by introducing relevance. This approach, i.e. introducing relevance, can be used for extending other existing object matching methods. The proposed method is based on the IRM for clustering. We plan to use other probabilistic models for clustering, such as dynamic IRM (Ishiguro et al, 2010) and latent feature models (Miller et al, 2009). The scalability of the inference can be improved by using stochastic variational inference (Hoffman et al, 2013) or parallel inference (Williamson et al, 2013).

References

Airoldi E, Blei D, Fienberg S, Xing E (2008) Mixed membership stochastic blockmodels. Journal of Machine Learning Research 9:1981–2014

Albert R, Barabási A (2002) Statistical mechanics of complex networks. Reviews of modern physics 74(1):47

Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Blackwell D, MacQueen JB (1973) Ferguson distributions via Pólya urn schemes. The Annals of Statistics pp 353–355

Clauset A, Moore C, Newman M (2008) Hierarchical structure and the prediction of missing links in networks. Nature 453(7191):98–101

Djuric N, Grbovic M, Vucetic S (2012) Convex kernelized sorting. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence

Gale WA, Church KW (1991) A program for aligning sentences in bilingual corpora. In: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, pp 177–184

Girvan M, Newman M (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12):7821–7826

Haghighi A, Liang P, Berg-Kirkpatrick T, Klein D (2008) Learning bilingual lexicons from monolingual corpora. In: Proceedings of ACL-08: HLT, pp 771–779

Hoffman MD, Blei DM, Wang C, Paisley JW (2013) Stochastic variational inference. Journal of Machine Learning Research 14(1):1303–1347

Ishiguro K, Iwata T, Ueda N, Tenenbaum J (2010) Dynamic infinite relational model for time-varying relational data analysis. Advances in Neural Information Processing Systems 23

Ishiguro K, Ueda N, Sawada H (2012) Subset infinite relational models. In: International Conference on Artificial Intelligence and Statistics, pp 547–555

Iwata T, Hirao T, Ueda N (2013) Unsupervised cluster matching via probabilistic latent variable models. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence

Iwata T, Lloyd J, Ghahramani Z (2016) Unsupervised many-to-many object matching for relational data. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(3):607–619

Kemp C, Tenenbaum J, Griffiths T, Yamada T, Ueda N (2006) Learning systems of concepts with an infinite relational model. In: Proceedings of the 20th AAAI Conference on Artificial Intelligence, vol 21, p 381

Klami A (2012) Variational Bayesian matching. In: Proceedings of the 4th Asian Conference on Machine Learning, pp 205–220

Klami A (2013) Bayesian object matching. Machine Learning 92:225–250

Lang K (1995) Newsweeder: Learning to filter netnews. In: Proceedings of the 12th International Conference on Machine Learning, pp 331–339

Li B, Yang Q, Xue X (2009) Transfer learning for collaborative filtering via a rating-matrix generative model. In: Proceedings of the 26th International Conference on Machine Learning, pp 617–624

Miller K, Griffiths T, Jordan M (2009) Nonparametric latent feature models for link prediction. Advances in Neural Information Processing Systems 22

Nowicki K, Snijders T (2001) Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96(455):1077–1087

Quadrianto N, Smola A, Song L, Tuytelaars T (2010) Kernelized sorting. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(10):1809–1821

Rapp R (1999) Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th Annual Meeting on Association for Computational Linguistics, pp 519–526

Socher R, Fei-Fei L (2010) Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 966–973

Wang Y, Wong G (1987) Stochastic blockmodels for directed graphs. Journal of the American Statistical Association 82(397):8–19

Watts D, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Williamson S, Dubey A, Xing EP (2013) Parallel Markov Chain Monte Carlo for nonparametric mixture models. In: Proceedings of the 30th International Conference on Machine Learning, pp 98–106

Yamada M, Sugiyama M (2011) Cross-domain object matching with model selection. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, pp 807–815