多关系数据挖掘中的概率模型研究

李正洋

2014.01.15

研究背景

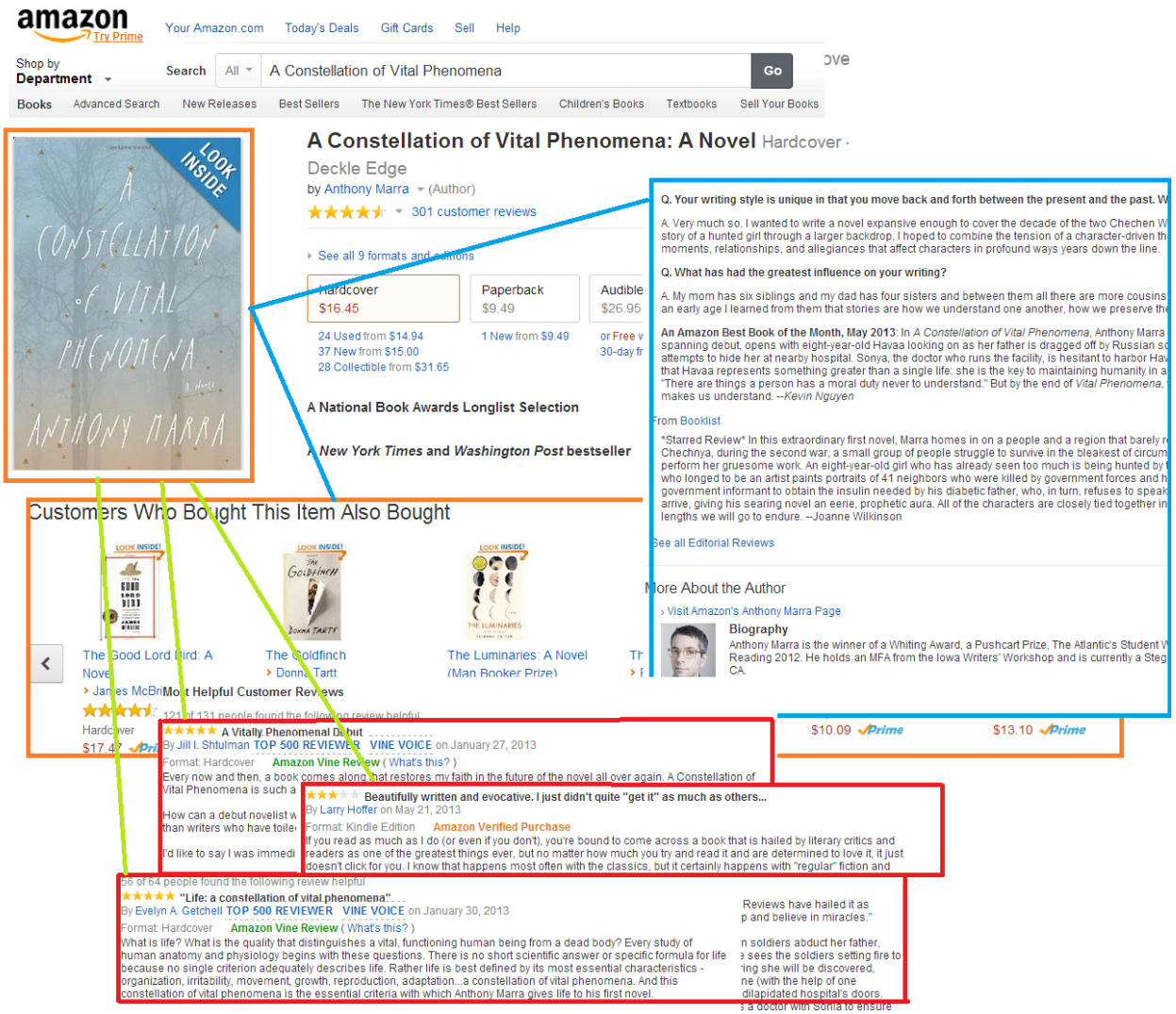


Figure: Amazon 网站的一本畅销书;

研究背景

Web2.0 时代的互联网有什么特点？

需要解决哪些问题？

如何解决这些问题？

本文的研究重点？

研究背景

Web2.0 时代的互联网有什么特点？
一个网站可能同时包含文字、图片、链接、音乐等信息，并为用户提供社交、购物、通讯等多种服务，在互联网公司的服务器中存储大量的复杂数据和历史信息。

需要解决哪些问题？
人工智能的观点，我们寄希望于计算机可以有效地提取出文本、图片的有效特征，同时依据已有特征的标注信息完成模型训练，进而对新的内容做预测以达到自动分析文本和图片的目的，提高搜索引擎的效率。我们也希望计算机可以依据用户的历史信息，为用户推荐新的朋友、商品、音乐、新闻、链接等。

如何解决这些问题？
这迫切地需要有效的建模方法来处理日益增长的多关系型数据。因此，对这些多关系数据做有效的建模分析和用户行为预测已成为计算机科学与统计学、管理科学最为重要的研究热点，这一学科就是**机器学习与数据挖掘**。而针对互联网应用问题，则具体到**推荐系统、自然语言处理**的研究范围。

本文的研究重点？
本文研究的多关系数据挖掘模型，主要基于推荐系统中的**协同过滤模型**与自然语言处理中的**文本分析模型**。

研究领域

- 统计机器学习

  统计机器学习是这一领域理论发展过程中最为主流的建模方法，概率图模型与贝叶斯学习是其中的一种，它通常假设一个概率模型来刻画数据的相互关系，通过最大化似然函数得到模型参数的估计值。概率模型建模方法已经在人工智能诸多领域得到了广泛的应用。

- 建模方法——概率图模型（有向图模型与无向图模型）

  概率图模型对结构化数据做建模，依据变量之间的依赖关系可分为有向图模型与无向图模型。有向图模型，如隐马尔科夫模型HMM、隐狄利克雷分配模型LDA、概率隐语义分析PLSA、高斯混合模型GMM 等。无向图模型，如高斯马尔科夫随机域GMRF、贝叶斯网BN、波尔兹曼机BM、深度学习网DL 等。当然变量之间的依赖关系也可以是两者都有。

- 求解方法——贝叶斯学习（近似推断与随机模拟）

  无论是有向图还是无向图模型，当得到联合似然函数后，贝叶斯学习的求解方法是一样的，一般有两种：1. 近似推断,即EM 算法或变分EM 算法2. 随机模拟——MCMC 方法、置信传播。

研究内容

文本分析模型（混合主题模型）有哪些？
- Probabilistic Latent Sematic Analysis(PLSA)
- Latent Dirichlet Allocation(LDA)
- Correlated Topic Model(CTM)
- Supervised Topic Model(sLDA)

协同过滤模型（因子分解模型）有哪些？
- Latent Factor Model(LFM)
- Probabilistic Matrix Factorization(PMF)
- Factorization Machine(FM)

Ideas? 融合文本和辅助信息的协同过滤模型！
- Collaborative Topic Regression(CTR)
- PMF-LDA
- PMF-CTM

文章框架

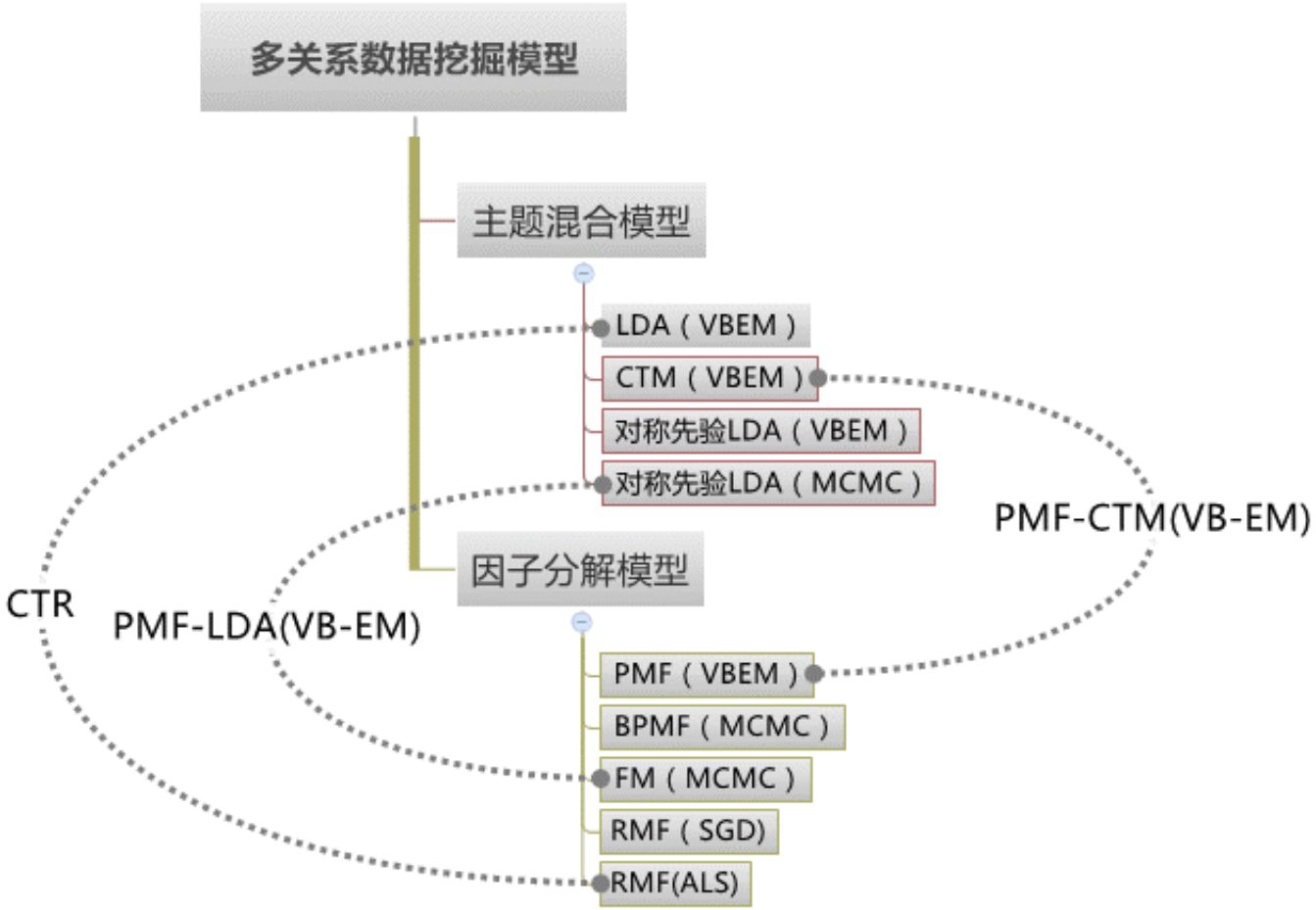


Figure: 模型关系图;

混合主题模型

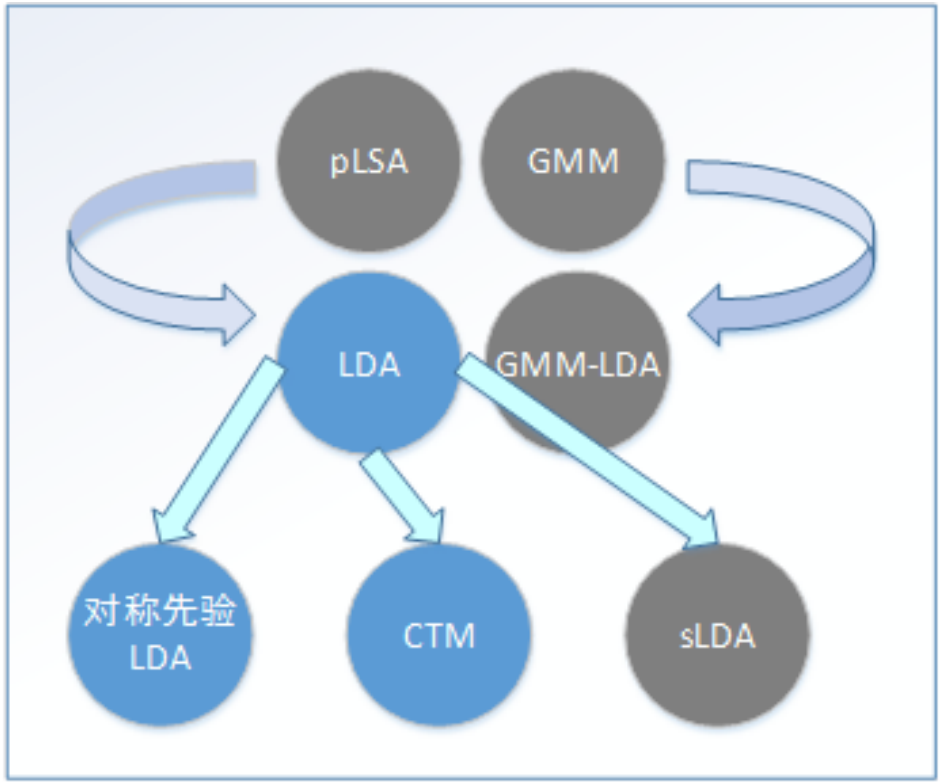


Figure: 混合主题模型衍生关系图;

PLSA

定义

PLSA 中主题的选择由 θ 确定话,题变量Z_i 服从 θ 确定的*Multinomial* 分布，记作$Z_i \sim Mult(\theta)$。文本中的单词$\{Y_1 \ldots\ldots Y_n\}$ 都各自对应一个主题$\{Z_1 \ldots\ldots Z_n\}$，这样文本中的每个Y_i 服从$\beta_{Z_i.}$ 确定的*Multinomial* 分布，记作$Y_i \sim Mult(\beta_{z_i.})$。

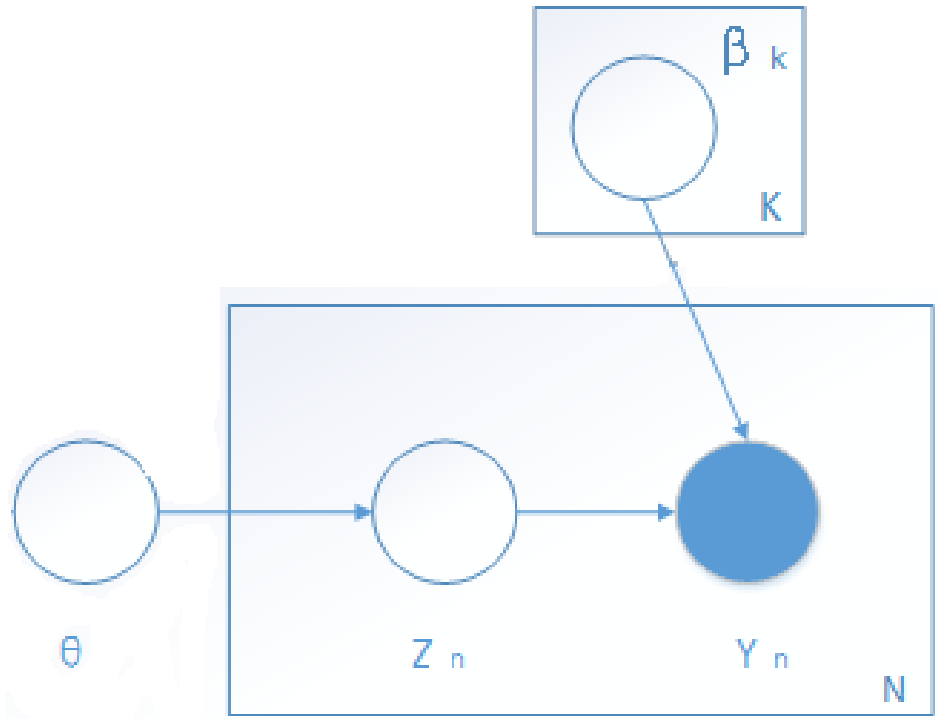


Figure: Probabilistic Latent Semantic Analysis;

LDA

定义

Latent Dirichlet Allocation 模型如下图所示，它是*pLSA* 的层级推广模型*(Dirichlet* 为*Multi* 的共轭先验*)*。详见论文*P12*。

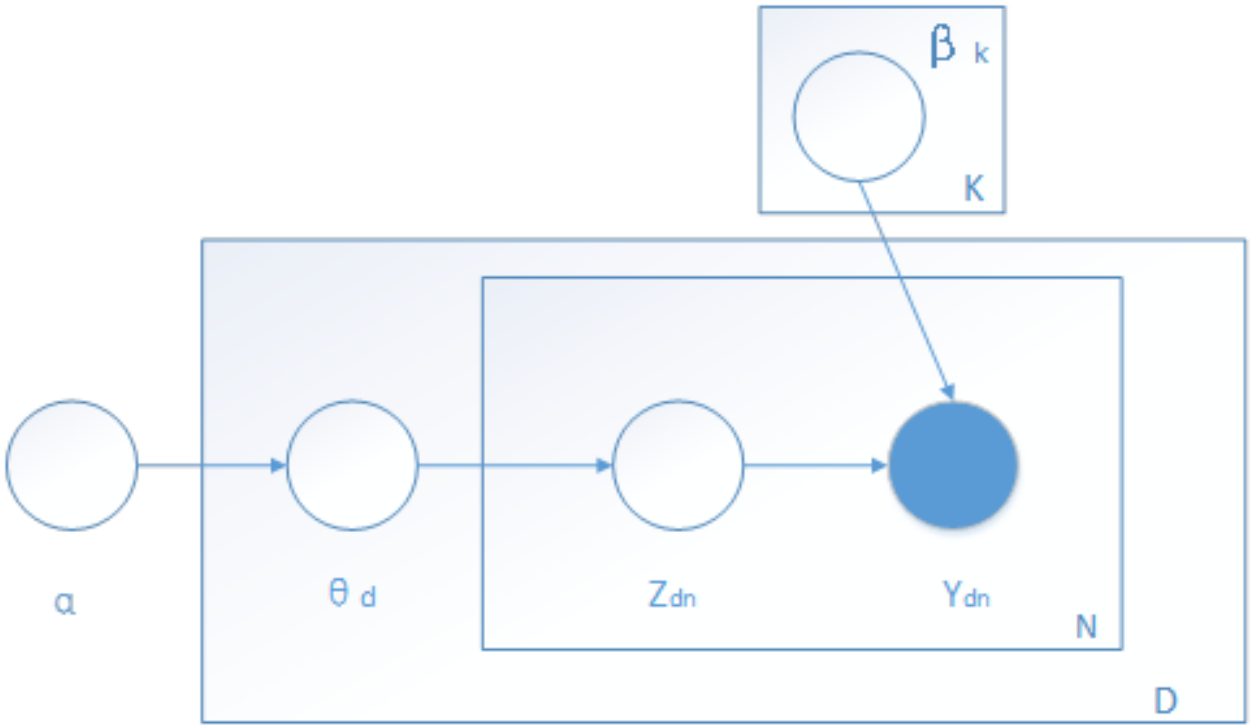


Figure: Latent Dirichlet Allocation;

LDA(VB-EM 算法)

详见论文P12-13, 附录P54-57。

算法 2-1 LDA 的 VB-EM 算法[2]

输出：$\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}$。

输入：文本单词$\mathbf{Y}$，收敛率ϵ，最大迭代次数 S，初始化$\{\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}\}$。

For t =1…S：

For d=1…D：

$\gamma_{di} = \alpha_{di} + \Sigma_{n=1}^{N_d} \phi_{dni}\ (i = 1 \ldots K)$

For n=1…N_d：

$\phi_{dni} \propto \beta_{iY_{dn}} \exp(\Psi(\gamma_{di}) - \Psi(\sum_{i=1}^{K} \gamma_{di}))$，$\Sigma_{k=1}^{K} \phi_{dnk} = 1$

Update $\beta_{kv} \propto \Sigma_{d=1}^{D} \Sigma_{n=1}^{Nd} \delta(Y_{dn} = v) \phi_{dnk}$

Update α

IF converge：break

Figure: LDA(VBEM);

CTM

定义

correlated topic model 是对*LDA* 的改进，改变*Dirichlet* 先验为*Multivariate Gaussian Logistic* 分布，这样可以刻画话题之间的关系，又由于非共轭关系，变分近似下界需要做放缩处理，详见正文*P15-16*, 附录*P57-58*。

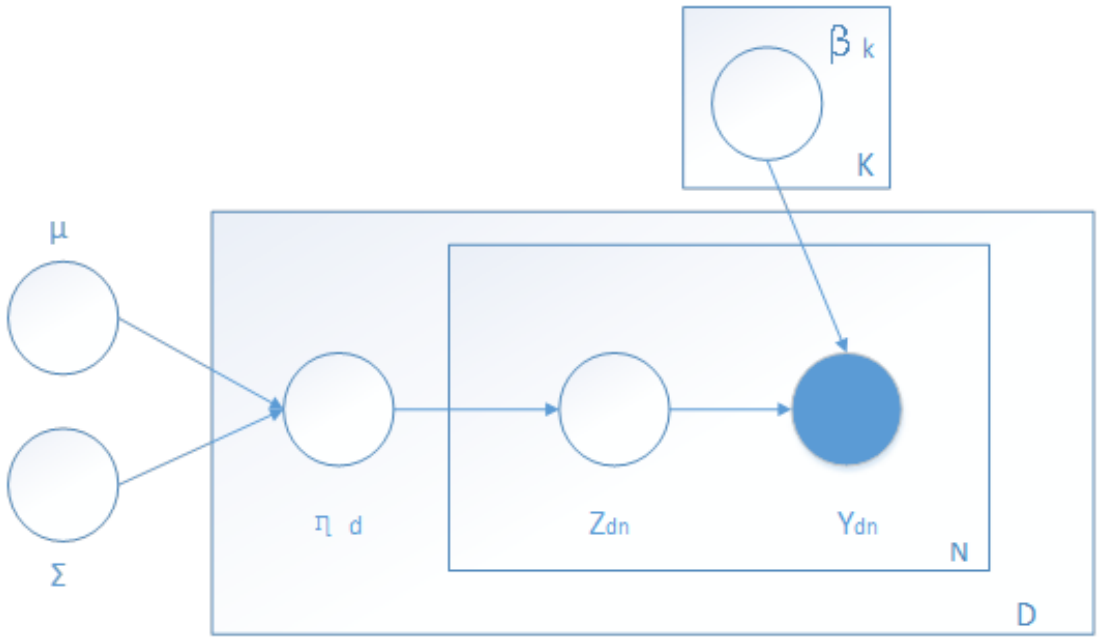


Figure: Correlated Topic Model;

CTM(VB-EM 算法)

详见正文P15-16, 附录P57-58。

算法 2-2 CTM 的 VB-EM 算法[3]

输出：$\mu, \Lambda, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\gamma}$。

输入：文本单词 $\mathbf{Y}$，收敛率 ϵ，最大迭代次数 S，初始化 $\{\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\gamma}\}$。

For t =1…S：

For d=1…D：

Update λ_d, γ_d^2

For n=1…N_d：

$\phi_{dni} \propto \beta_{iY_{dn}} \exp(\lambda_{di})$，$\Sigma_{k=1}^{K} \phi_{dnk} = 1$

Update $\beta_{kv} \propto \Sigma_{d=1}^{D} \Sigma_{n=1}^{Nd} \delta(Y_{dn} = v) \phi_{dnk}$

Update $\mu = \frac{1}{D} \Sigma_{d=1}^{D} \lambda_d$，$\Lambda^{-1} = \frac{1}{D} \Sigma_{d=1}^{D} (\gamma_d^2 + (\lambda_d - \mu)(\lambda_d - \mu)^T)$

IF converge：break

对称先验LDA

定义

为外层参数增加*Dirichlet* 先验，我们得到如下图的模型，叫作“对称先验*LDA*”模型。

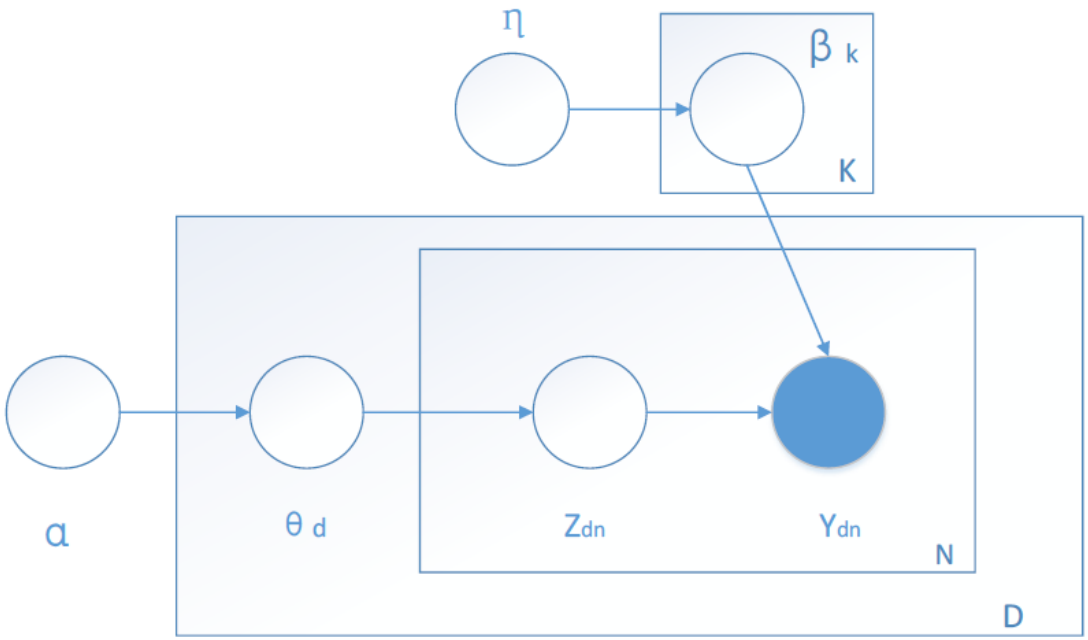


Figure: 对称先验LDA;

对称先验LDA VBEM 算法与MCMC 算法

输出：$\alpha, \eta, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}$。

输入：文本单词$\mathbf{Y}$，收敛率ϵ，最大迭代次数S，初始化$\{\alpha, \eta, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}\}$。

For t =1…S：

For d=1…D：

$\gamma_{di} = \alpha_{di} + \Sigma_{n=1}^{N_d} \phi_{dni}$ (i = 1 ... K)

For n=1…N_d：

$\phi_{dni} \propto \eta_{Y_{dn}} \exp(\Psi(\gamma_{di}) - \Psi(\sum_{i=1}^{K} \gamma_{di}))$，$\Sigma_{k=1}^{K} \phi_{dnk} = 1$

For k=1…K：$\lambda_{ki} = \eta_i + \Sigma_{d=1}^{D} \Sigma_{n=1}^{Nd} \phi_{dnk} \delta(Y_{dn} = i)$

Update α，η

IF converge：break

输出：$\mathbf{Z}$。

输入：文本单词$\mathbf{Y}$，最大迭代次数S，初始化$\{\alpha, \eta)\}$。

For t=1…S：

For d=1…D：

$$p(Z_{dn} = k|\boldsymbol{Z}^{\neg(d,n)}) \propto \frac{\eta_{Y_{dn}} + \#\{\cdot, Y_{dn}, k\}^{\neg(d,n)}}{\sum_{v=1}^{V} \eta_v + \#\{\cdot,\cdot,k\}^{\neg(d,n)}} \frac{\alpha_k + \#\{d,\cdot,k\}^{\neg(d,n)}}{\sum_{k=1}^{K} \alpha_k + \#\{d,\cdot,\cdot\}^{\neg(d,n)}}$$

Figure: LDA(VBEM);

文章框架

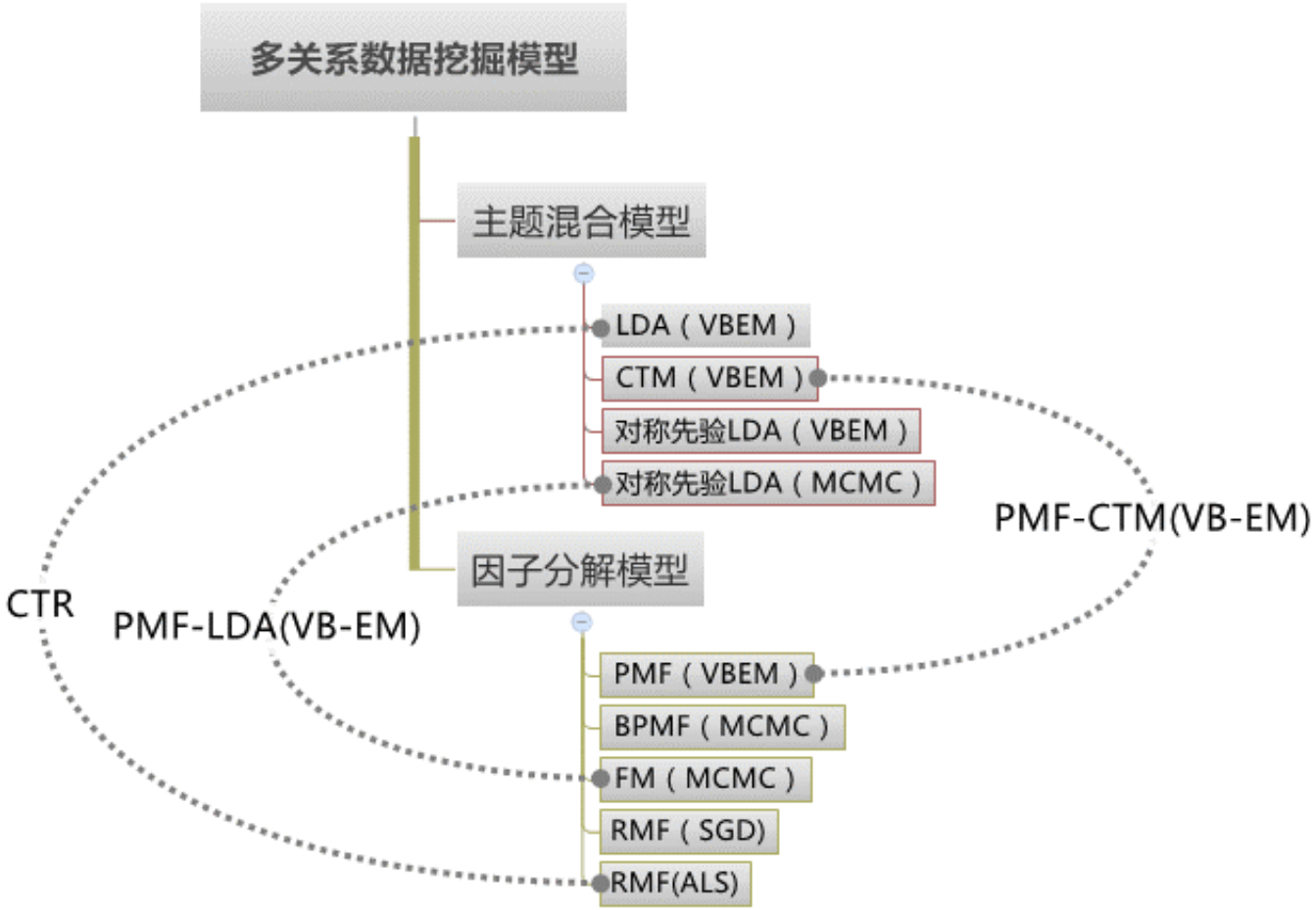


Figure: 模型关系图;

PMF

概率矩阵分解模型Probabilistic Matrix Factorization

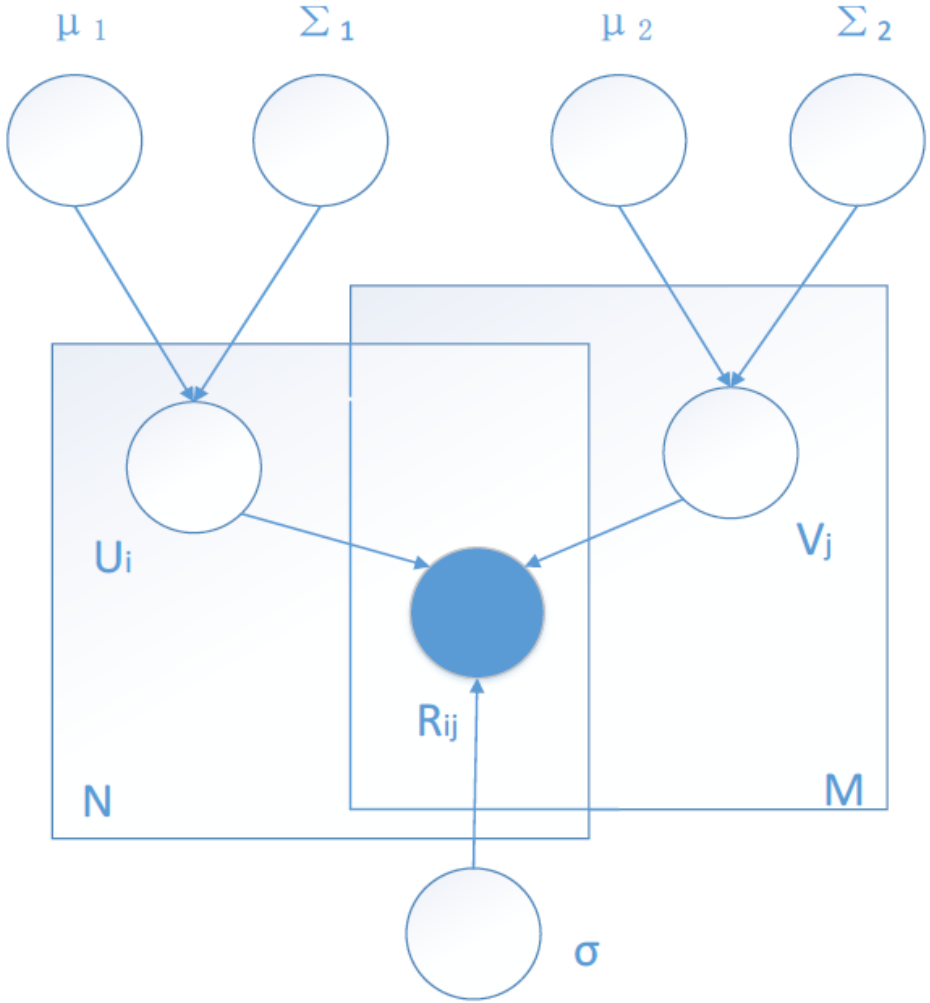


Figure: PMF;

PMF

PMF VBEM算法

算法 2-6 PMF(VB-EM 算法)

输出：U，V。

输入：指示矩阵$Y_{ij} = \delta(i,j)$，观察值矩阵 R，最大迭代次数 S，初始化$\mu_1, \Sigma_1, \mu_2, \Sigma_2, \sigma$以及$\Sigma^i$与 Φ^i(i=1…N)，$\Sigma^{\sim j}$与 $\Phi^{\sim j}$(j=1…M)。

For t =1…S：

For i from 1 to N, j from 1 to M：

update $\Sigma^i, \Phi^i, \Sigma^{\sim j}, \Phi^{\sim j}$

update $\mu_1, \Sigma_1, \mu_2, \Sigma_2, \sigma$。

$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

For i from 1 to N, j from 1 to M：

$U_i = \Phi^i, V_j = \Phi^{\sim j}$

Figure: PMF(VBEM);

正则矩阵分解

Regularized Matrix Factorization

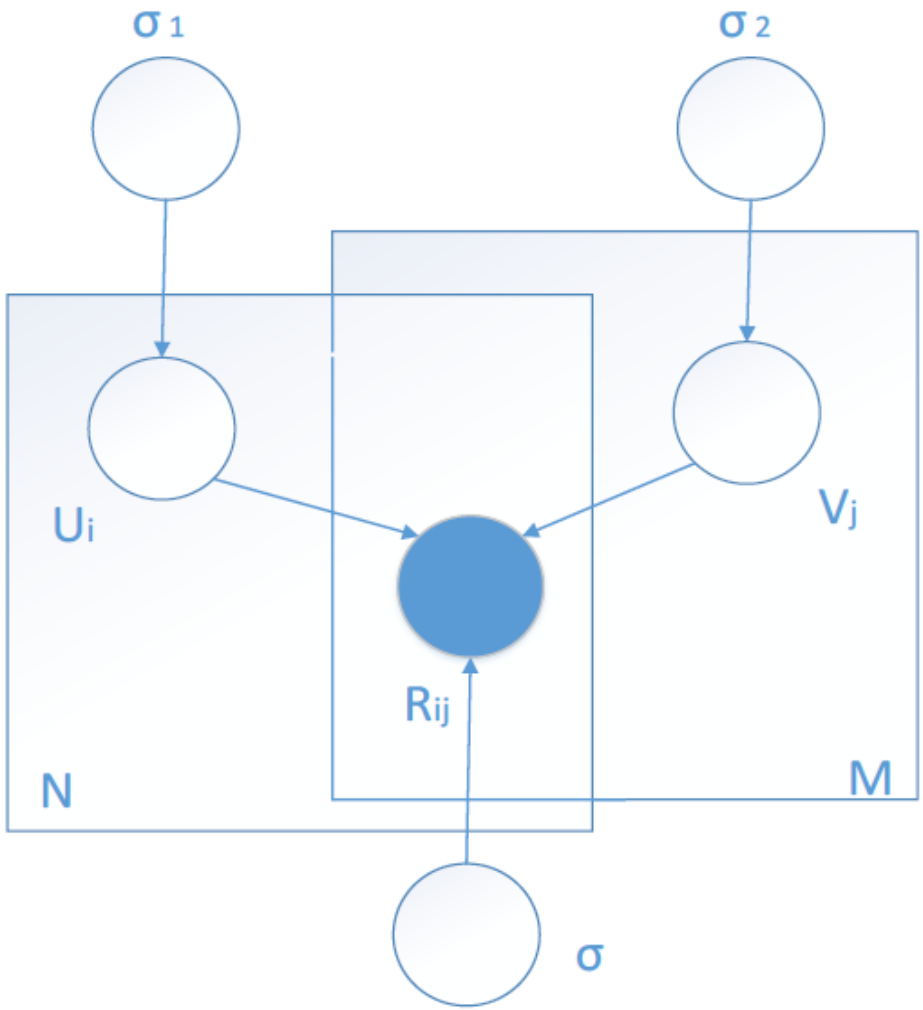


Figure: PMF;

RMF

SGD算法

算法 2-7 RMF(SGD 算法)

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，正则化参数$\lambda = \lambda_U = \lambda_V$，学习率$\eta$，收敛率$\epsilon$，最大迭代次数 S，初始化 U，V。

For t =1…S：

For each (i, j) with $Y_{ij} \neq 0$：

$$\Delta_{ij} = R_{ij} - U_i^T V_j$$

$$U_i = U_i + \eta(\Delta_{ij} V_j - \lambda U_i)$$

$$V_j = V_j + \eta(\Delta_{ij} U_i - \lambda V_j)$$

$$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

Figure: RMF(SGD);

RMF

ALS算法

算法 2-8 RMF(ALS 算法)

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，正则化参数$\lambda = \lambda_U = \lambda_V$，收敛率$\epsilon$，最大迭代次数 S，初始化 U，V。

For t =1…S：

For each i from 1 to N:

$$U_i = \left(\sum_j Y_{ij} V_j V_j^T + \lambda I \right)^{-1} \sum_j Y_{ij} R_{ij} V_j$$

For each i from 1 to M:

$$V_j = \left(\sum_i Y_{ij} U_i U_i^T + \lambda I \right)^{-1} \sum_i Y_{ij} R_{ij} U_i$$

$$\Delta^t = \| Y \odot (R - U^T V) \|_2^2$$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

Figure: RMF(ALS);

BPMF

贝叶斯概率矩阵分解Bayesian Probabilistic Matrix Factorization

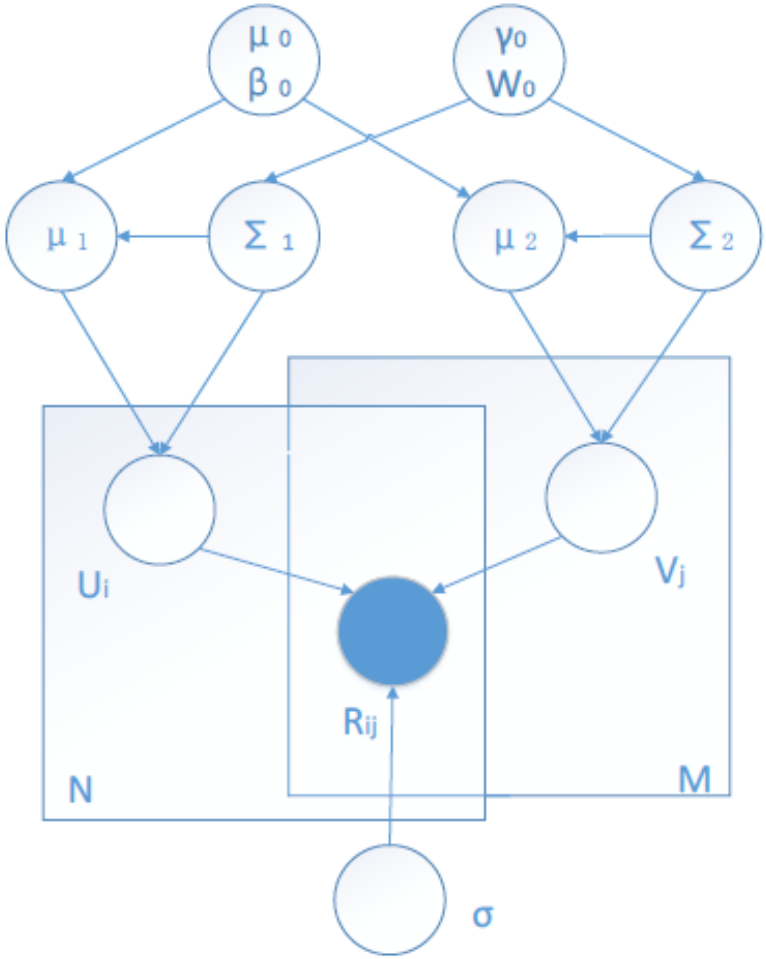


Figure: BPMF;

BPMF(MCMC)

BPMF(MCMC)算法

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，收敛率ϵ，最大迭代次数 S，固定值σ，初始化$\mu_0, \beta_0, \gamma_0, W_0, U, V$。

For t =1…S：

compute $\beta_0^*, \gamma_0^*, \mu_0^*, W_0^*$ with U

sample $\Sigma_1^{-1} \sim \text{Wishart}(W_0^*, \gamma_0^*)$

sample $\mu_1 \sim N(\mu_0^*, \beta_0^{*-1}\Sigma_1)$

For each i from 1 to N：

compute μ_i^*, Σ_i^*

sample $U_i \sim N(\mu_i^*, \Sigma_i^*)$

compute $\beta_0^*, \gamma_0^*, \mu_0^*, W_0^*$ with V

sample $\Sigma_2^{-1} \sim \text{Wishart}(W_0^*, \gamma_0^*)$

sample $\mu_2 \sim N(\mu_0^*, \beta_0^{*-1}\Sigma_2)$

For each i from 1 to M：

compute $\mu_{\sim j}^*, \Sigma_{\sim j}^*$

sample $V_j \sim N(\mu_{\sim j}^*, \Sigma_{\sim j}^*)$

$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

Figure: BPMF(MCMC);

FM

贝叶斯概率因子分解Bayesian Factorization Machine

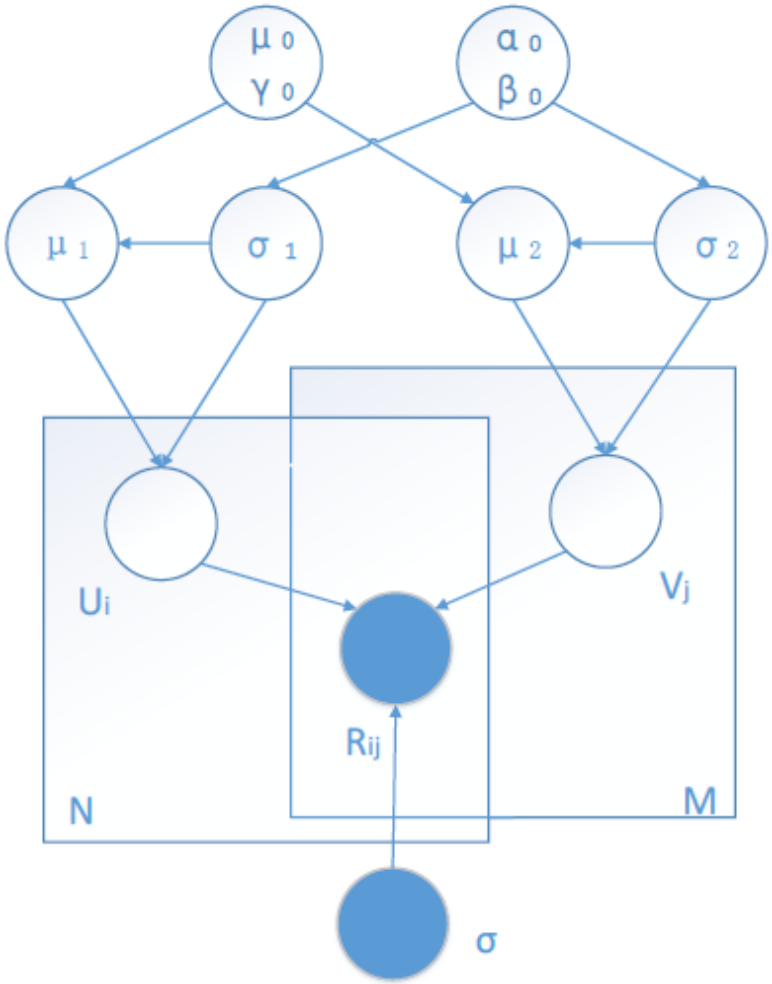


Figure: FM;

FM(MCMC)

FM(MCMC)算法

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，收敛率ϵ，最大迭代次数 S，固定值σ，初始化$\mu_0, \gamma_0, \alpha_0, \beta_0$, U, V。

For t =1…S：

compute $\mu_0^*, \gamma_0^*, \alpha_0^*, \beta_0^*$ with U

sample $\sigma_1^{-1} \sim \Gamma(\alpha_0^*, \beta_0^*)$

sample $\mu_1 \sim N(\mu_0^*, \gamma_0^{*-1}\sigma_1^2 I)$

For each i from 1 to N：

compute μ_i^*, Σ_i^*

sample $U_i \sim N(\mu_i^*, \Sigma_i^*)$

compute $\mu_0^*, \gamma_0^*, \alpha_0^*, \beta_0^*$ with V

sample $\sigma_2^{-1} \sim \Gamma(\alpha_0^*, \beta_0^*)$

sample $\mu_2 \sim N(\mu_0^*, \gamma_0^{*-1}\sigma_2^2 I)$

For each i from 1 to M：

compute $\mu_{\sim j}^*, \Sigma_{\sim j}^*$

sample $V_j \sim N(\mu_{\sim j}^*, \Sigma_{\sim j}^*)$

$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

Figure: FM(MCMC);

文章框架

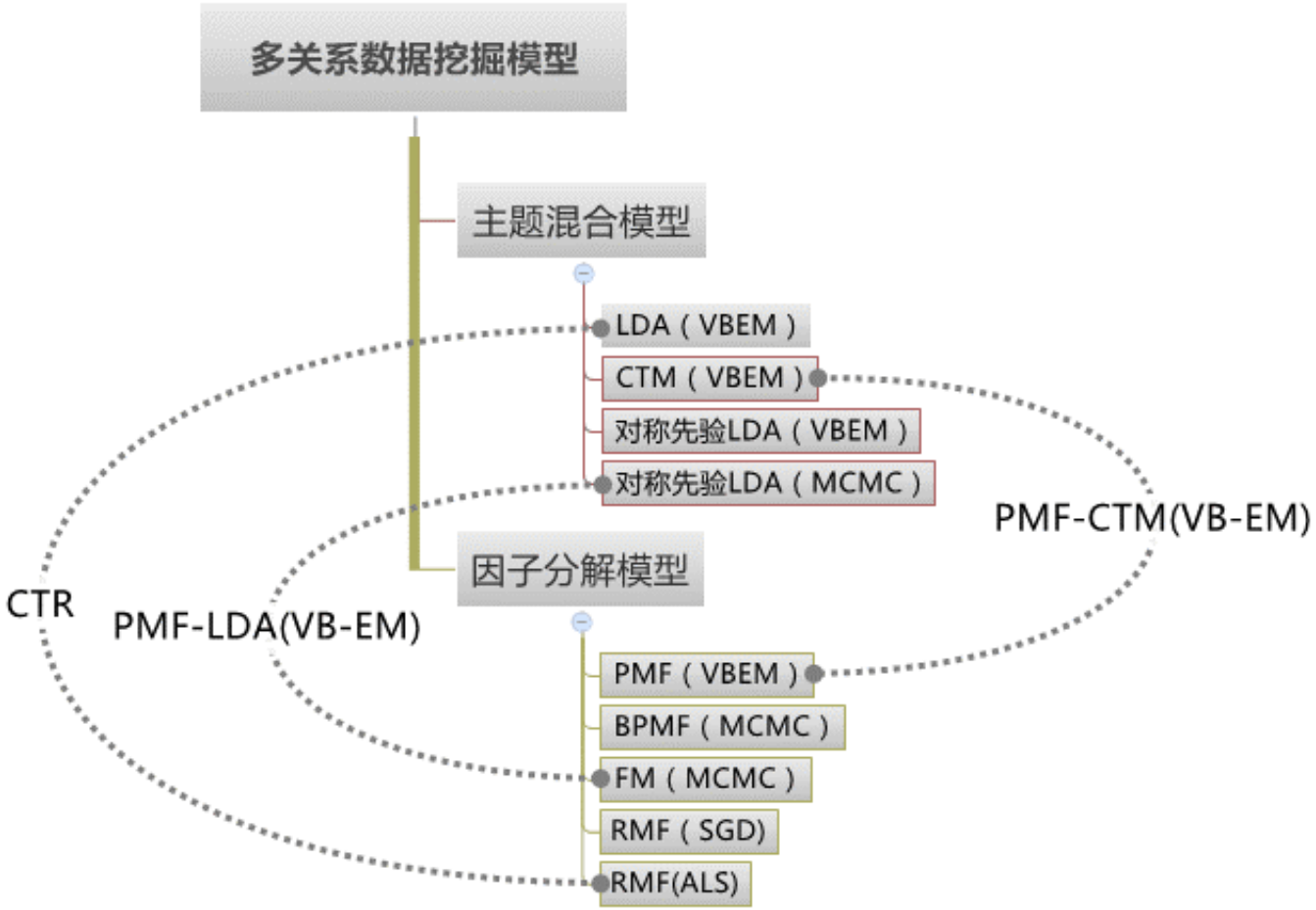
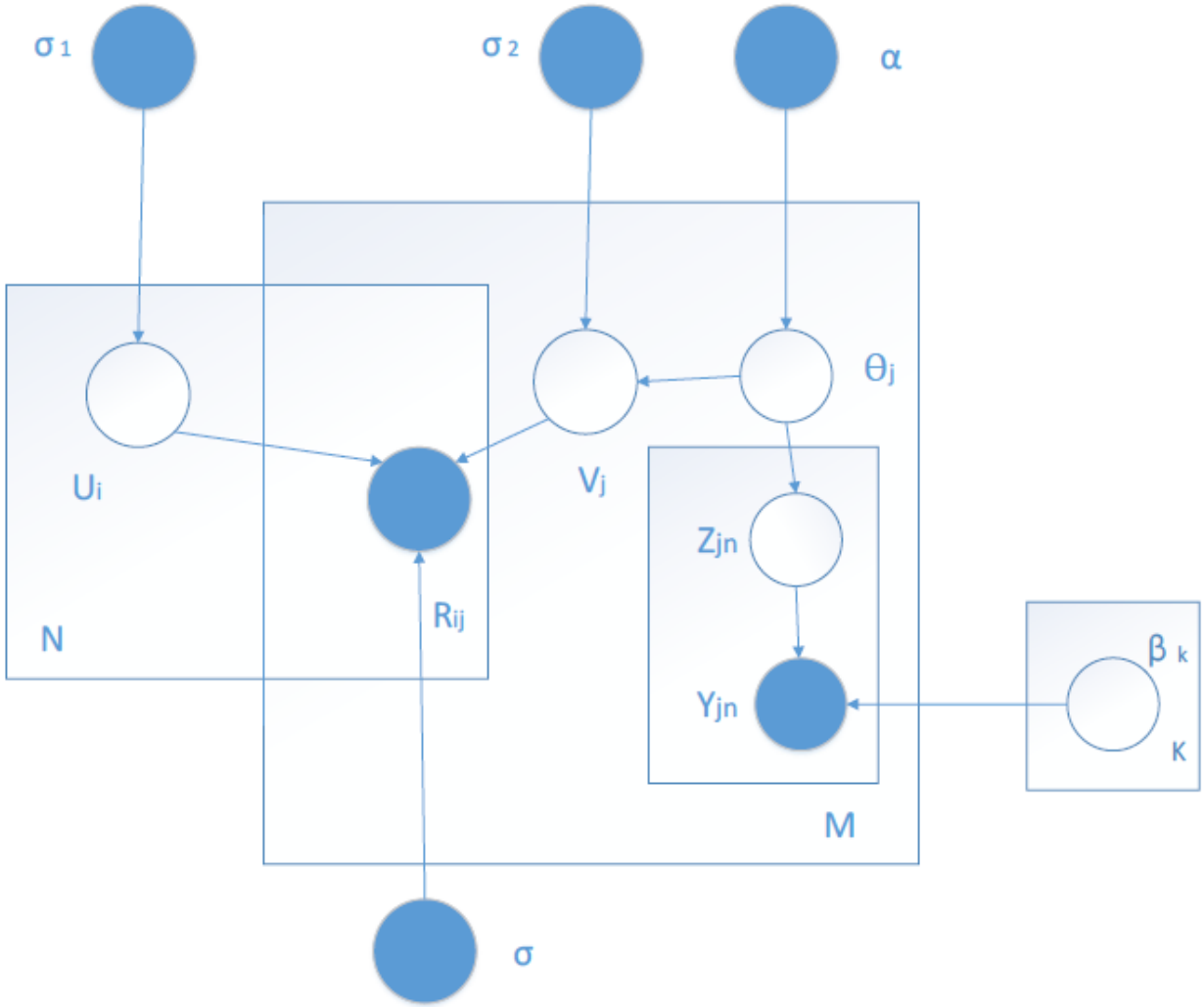


Figure: 模型关系图;

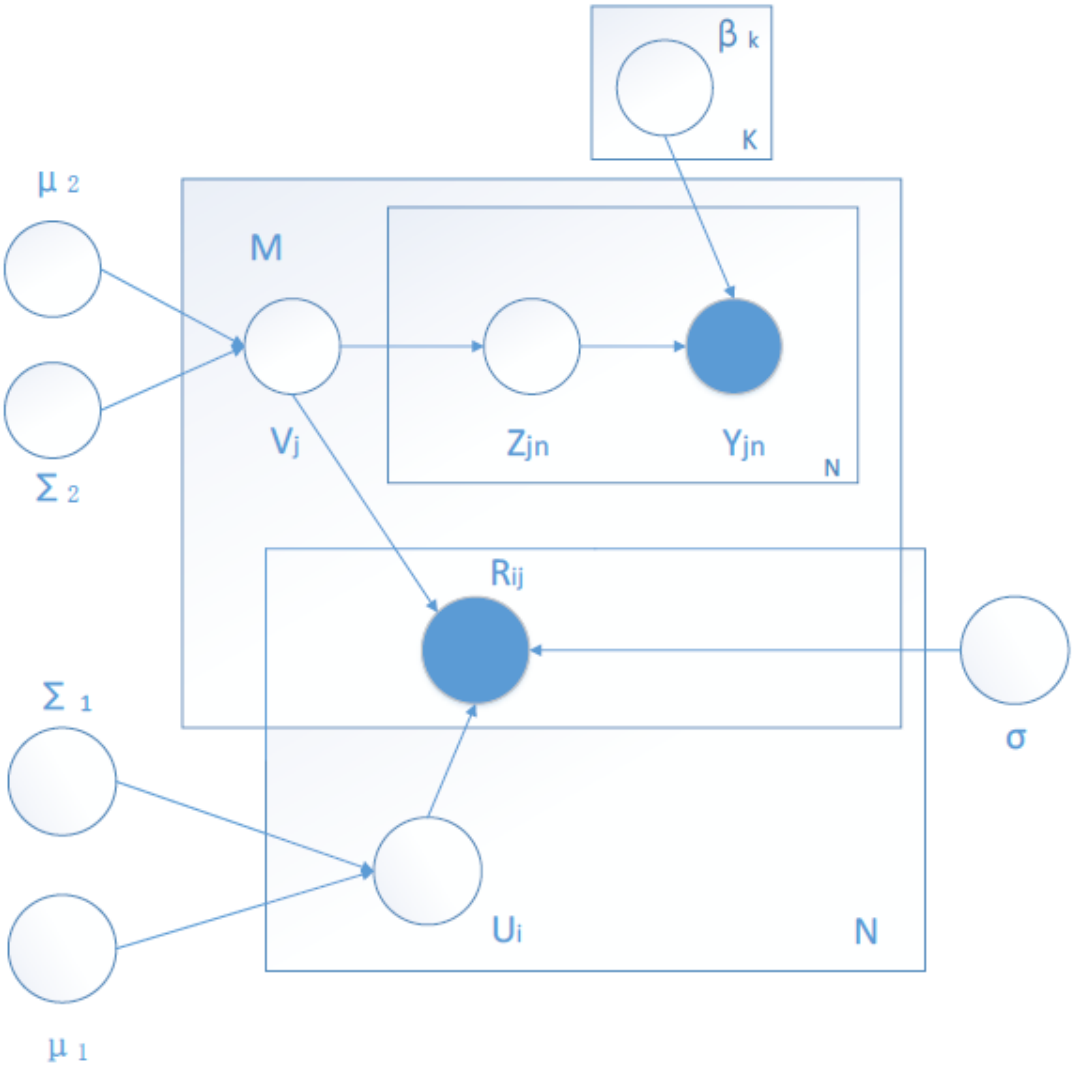
融入辅助信息协同主题回归模型CTR

融入辅助信息协同主题回归模型Collaborative Topic Regression Model, 实质是正则矩阵分解与层级概率隐语义分析的组合



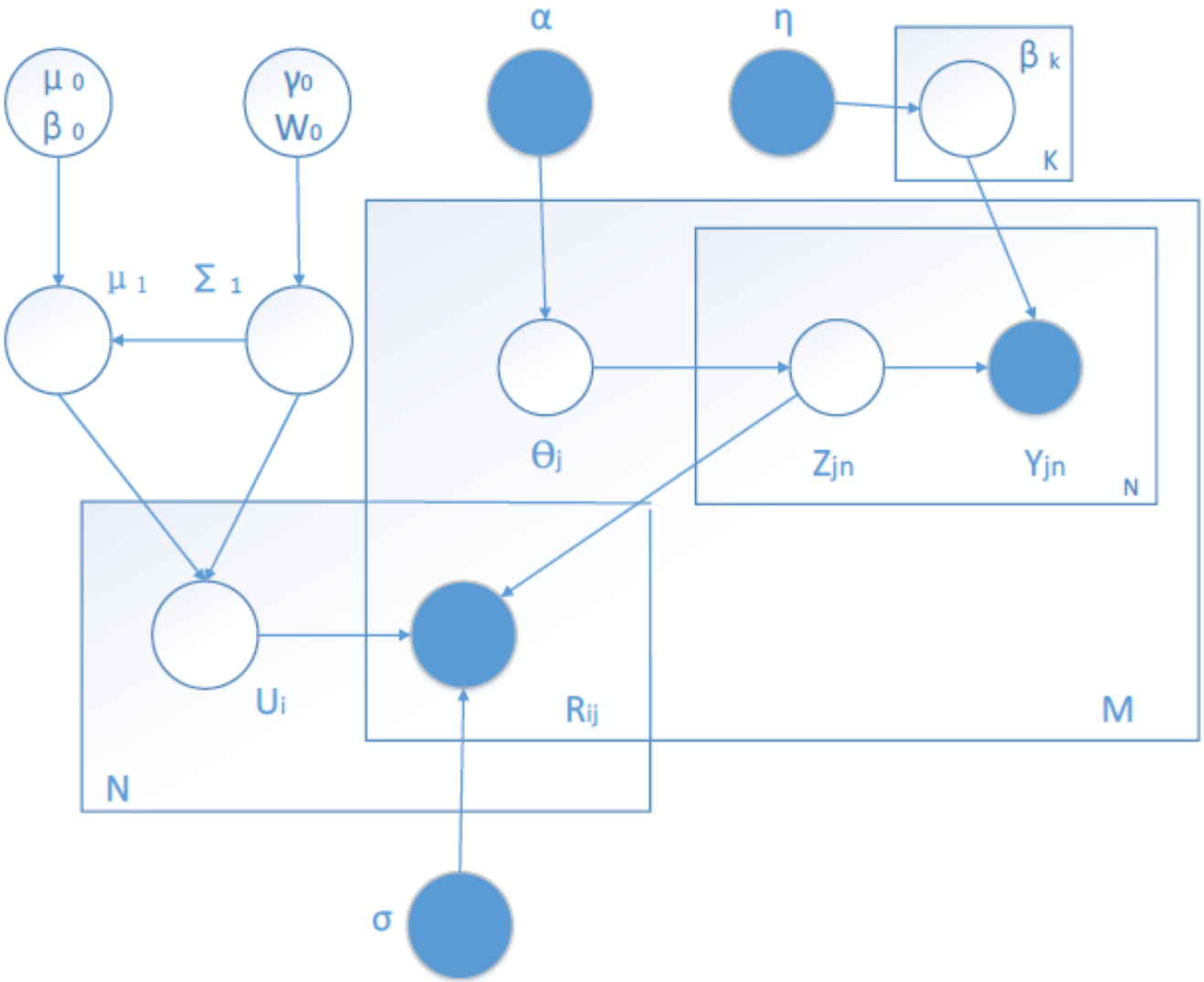
融入辅助信息概率矩阵分解PMF-CTM

Probabilistic Matrix Factorization VBEM & Correlated Topic Model VBEM



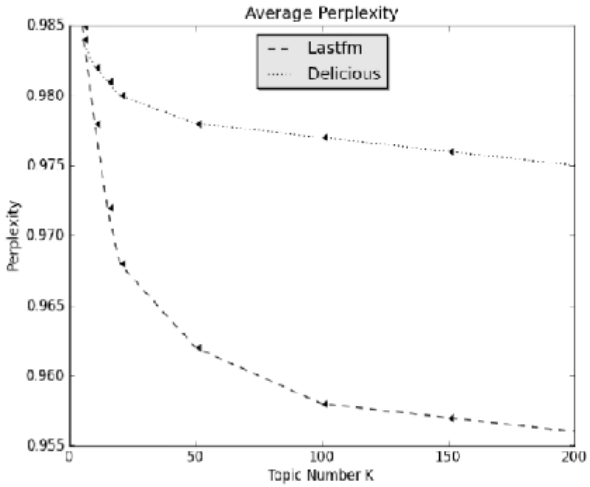
融入辅助信息概率矩阵分解PMF-LDA

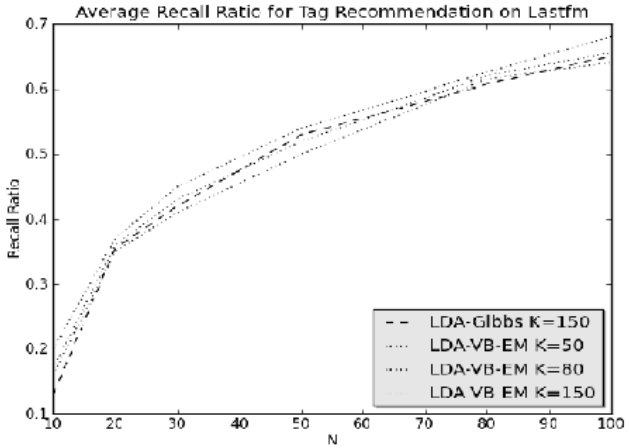
Bayesian Factorization Machine & Latent Dirichlet Allocation MCMC

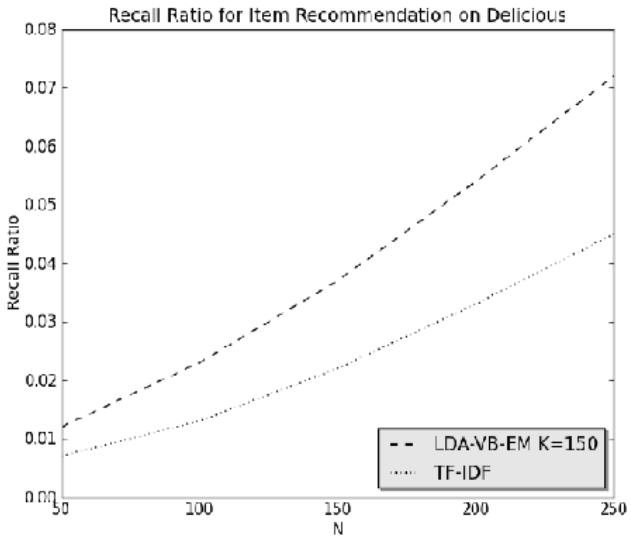


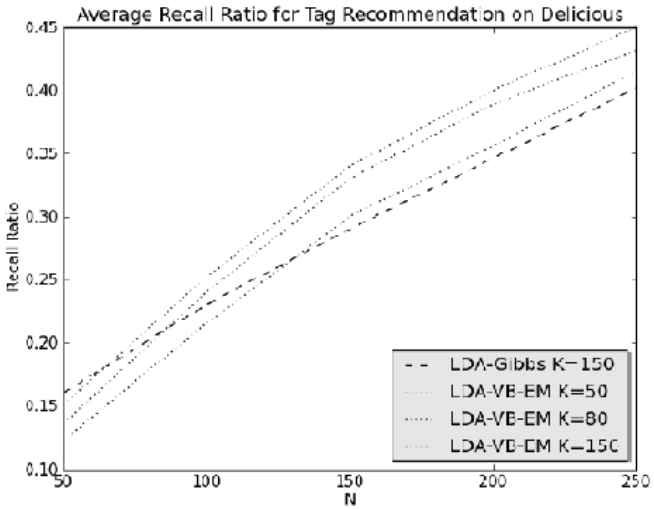
主题模型在推荐算法中的应用

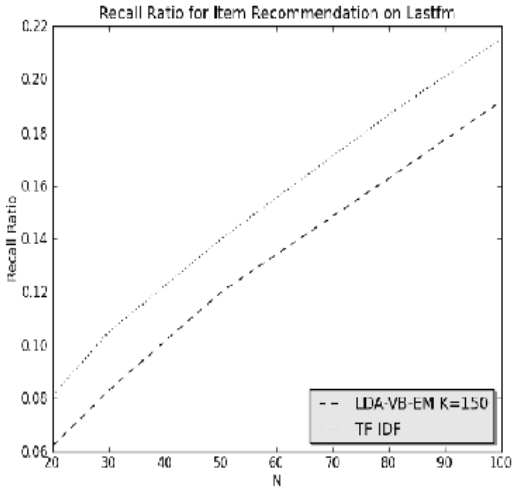
这部分内容详见小论文：Li, Zhengyang, and Congfu Xu. "Tag-based top-N recommendation using a pairwise topic model." Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on. IEEE, 2013.











数据集与实验方法

选用两个公开数据集Movielens与Netflix的一部分子集。Movielens电影评分数据由GroupLens Research 项目小组提供，它包含三个大小分别为100k，1M与10M 的数据集。电影评分数值范围是1 到5，间隔0.5分。此外，Movielens 数据提供了电影属性，这将作为辅助信息使用。Netflix 数据集来自Netflix Prize比赛，评分跨度为1到5的整数。在本文的实验中，我们选择的Movielens 数据集包含100,000 条评分有943个用户ID与1682 个电影ID，而Netflix数据集则收集其中的900,000条评分有6040 个用户ID与3950 个电影ID。

采用最小方均根误差 RMSE 刻画评分预测算法的准确程度，计算公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(r_n - \hat{r}_n)^2}{N}} \qquad (4.3)$$

在这里，所有的预测评分$\hat{r}_n$都需要经过后处理再得到相应的 RMSE 值。每次实验，随机划分其中的 80%作为训练集，20%作为测试集。为了方便起见在 Movielens 数据集上 SGD 学习率这里采用固定值$\eta = 0.001$而 Netflix 数据集上$\eta = 0.01$。每次实验将 80%数据集化分成四份，做四次交叉验证确定好最适当的正则化参数$\lambda = \lambda_V = \lambda_U \in \{0.1, 0.01, 0.001\}$，经验迭代次数则为在验证集上取得最好 RMSE 的四次迭代次数平均值，再利用数据集的 80%完整部分完成训练。

实验结果

方法 \ D	5	10	15	20	25	30
SGD	0.956	0.943	0.938	0.929	0.922	0.917
ALS	0.942	0.936	0.923	0.915	0.912	0.905
PMF	0.985	0.974	0.964	0.957	0.953	0.950
BPMF	0.933	0.925	0.917	0.912	0.904	0.893
FM	0.923	0.912	0.905	0.903	0.896	0.891

表格 4-10 Netflix 数据集上 RMSE 结果

方法 \ D	5	10	15	20	25	30
PMF	0.845	0.852	0.868	0.879	0.891	0.903
PMF-CTM	0.839	0.843	0.862	0.877	0.885	0.892
BPMF	0.821	0.835	0.847	0.859	0.864	0.879
PMF-LDA	0.817	0.824	0.838	0.852	0.859	0.871
ALS	0.802	0.815	0.824	0.825	0.829	0.833
CTR	0.796	0.808	0.812	0.818	0.821	0.829

表格 4-11 Movielens 数据集上 RMSE 结果

实验结论

从表格4-10中可以看到在Netflix数据集上，5种baseline算法随不同特征维数D变化的RMSE预测值。可见D值越大所得到的RMSE值越小，PMF、BPMF等算法的效果都要好于SGD算法，而ALS、PMF与BPMF、FM之间的差别不大。这里需要说明的是，由于单一数据集和实验初始值设置的影响，这不能说明各算法之间的优劣。

从表格4-11中可以看到在Movielens数据集上，多关系模型算法与baseline算法随不同特征维数D变化的RMSE预测值。可见D值变大所得到的RMSE变大，这说明特征维数并不是越大越好。PMF-CTM优于PMF，PMF-LDA优于BPMF算法，CTR与ALS相同参数设置一致时CTR优于ALS算法。依据图1-1所示的模型关系，融入辅助信息的协同过滤算法更为有效。

总结

本文对多关系数据挖掘模型的阐述，主要介绍了两种基本的概率模型建模方法即混合主题模型和因子分解模型，并给出基于变分近似推断和蒙特卡洛采样两种求解策略的对应算法。基于这两种模型的特点，引出了融合辅助信息的多关系模型，有效地结合了两种模型的功能，并为此提出一套完整的模型求解策略。具体研究内容如下：

1. 对现有主题模型算法的研究和应用。研究多项式混合模型从PLSA、GMM到层级贝叶斯推广模型LDA，以及改进模型CTM。总结几种重要的主题模型方法并推导其中关键步骤。应用LDA完成基于标签的推荐系统算法，比较了不同话题数量下，VBEM与MCMC采样算法的效果。

2. 对现有因子分解模型算法的研究和应用。研究概率矩阵分解模型，如PMF、BPMF、FM，以及它们与正则矩阵分解的联系。总结几种主要的矩阵分解建模方法并给出了VBEM、MCMC、SGD、ALS几种相应的解法，在公开数据集上比较了协同过滤效果

3. 提出基于混合主题模型和因子分解模型的多关系模型。给出三种建模方法，有效地利用了基本模型求解策略，扩展了因子模型和主题模型功能，给出相应的三种优化求解算法，有效地利用辅助信息提高了基本协同过滤方法。

展望

提高模型的准确率和求解速度以提高模型适应大数据的能力，这需要考虑主题模型和因子模型采用不同策略以更有效地结合，还需要利用贝叶斯在线学习方法来求解图模型。

考虑文本和辅助信息建模的其它模型，寻找引入辅助信息的新方法，以提高基本协同过滤系统的准确率。

将用户的社会化网络关系融入到矩阵分解中，将用户浏览历史的时序信息融入到矩阵分解中等。

无论是从混合主题模型与因子分解模型更有效结合的观点，还是从应用背景的角度，这方面的研究都是值得进一步深入和探索。

谢谢大家！