情报工程 ISSN: 2095-915X
第1卷 第3期 90-97 2015年6月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.3 90-97 Jun 2015

doi:10.3772/j.issn.2095-915x.2015.03.012

基于主题发现的专利发明人推荐方法

黎楠，杜永萍，何明

（北京工业大学计算机学院 北京 100124）

摘要：LDA 主题模型可用于识别大规模文档集中潜藏的主题信息，本文提出了一种基于 LDA 建立发明人兴趣主题模型的方法，合并每位发明人的专利数据，专利信息基于发明人进行划分，将标准的文档－主题－词的三层 LDA 模型变为专利数据中的发明人－主题－词的发明人兴趣模型，实现发明人的主题发现，并利用该模型中主题分布之间的相似性进行发明人的个性化推荐。在采集真实专利数据集上的实验结果表明该方法相比传统的向量空间模型方法和隐马尔科夫模型方法具有更高的准确率，推荐效果更优。

关键词：LDA 主题模型，专利，主题发现，推荐技术

Patent Inventor Recommendation Method Based on Topic Discovery

Li Nan, Du Yongping, He Ming

(College Of Computer Science And Technology, Beijing University Of Technology, Beijing, 100124)

Abstract: LDA model can be used for identifying topic information from large-scale document set. We propose the approach to build the inventor topic model based on LDA, which divides patent data based on inventor and represents each inventor with the ownedpatents. The standard three layers of document-topic-word in LDA(Latent Dirichlet Allocation) model becometheinventor topic model of inventor-topic-word. The topics of the inventor are discovered and recommendation is implementbased on thesimilarity of topic distribution. Experiment on real data set shows that the new approach has a better performance compared to the traditional Vector Space Model method and Hidden Markov Model method.

Keywords: LDA Topic Model, Patent Data, Topic Discovery, Recommendation Method

资助项目：国家科技支撑计划子课题（2013BAH21B02-01）；北京市自然科学基金资助项目 (4153058)；上海市智能信息处理重点实验室开放基金（IIPL-2014-004）。

作者简介：黎楠，（1989-），硕士，主要研究领域为自然语言处理、数据挖掘，email：maggie_lee10@163.com；杜永萍，（1977-），副教授，硕士生导师，主要研究方向为自然语言处理、信息检索等，email：ypdu@bjut.edu.cn；何明，（1975-），副教授，硕士生导师，主要研究方向为数据库理论与技术、数据挖掘、信息检索，email：heming@bjut.edu.cn。

1. 引言

专利信息是科学研究的产物，是重要的技术文献和知识宝库。近年，基于专利信息的分析和挖掘被广泛应用，机器学习技术的发展也为专利挖掘提供了有利的技术支持。评估专利发明人的研究成果，有效的挖掘发明人的兴趣主题，面向发明人的个性化推荐，逐渐发展成为一个研究热点。

传统的主题挖掘是采用文本聚类的算法[1]，通过向量空间模型（Vector Space Model）将文本里的非结构化数据映射到向量空间中的点，然后用传统的聚类算法，如基于划分的算法（如K-means 算法）、基于层次的算法（自顶向下和自底向上算法）、基于密度的算法等等[2]，实现文本聚类。

主题模型是一种运用于文本的典型主题发现统计模型[3]，已经在自然语言处理领域中被广泛应用。主题是对文本所包含的语义的概括和抽象[4]。在专利数据集上深入分析有影响力的发明人的兴趣主题，每一个发明者可以表示为基于该发明人专利的主题分布，而建立主题模型可以实现主题分布。

主题模型 PLSA（Probabilistic Latent Semantic Analysis）[5]，主题的概率分布是一个参数，而不是一个随机变量，但参数的增加可能会导致过度拟合的问题[6]。LDA（Latent Dirichlet Allocation）模型是一个更优的主题模型，是由 Blei 在 2003 年提出的[7]。Blei 介绍了 Dirichlet 先验分布的超参数的使用，以及转换概率分布的随机变量。参数随着语料库的容量增长，容易产生过拟合的问题，而 LDA 则很好的避免了过拟合问题。

本文是在专利数据的基础上，提出了一种基于 LDA 模型的发明人主题发现与推荐方法，对每一位发明人进行兴趣建模，并将相似度最高的发明人推荐给指定的发明人。我们将算法应用于真实的专利数据，并与传统的向量空间模型方法和经典的隐马尔科夫模型方法进行对比实验，结果表明，我们的算法能够更好的表现发明人的兴趣分布，在此基础上的推荐算法具有更高的准确率，在实际应用中具有更好的价值。

2 相关研究

PLSA (Probabilistic Latent Semantic Analysis) 是 Hofmann[5-6] 在研究 LSA 的基础上提出的基于最大似然法和产生式模型的概率模型。PLSA 沿用了 LSA 的降维思想：在常用的文本表达方式 (tf-idf) 下，文本是一种高维数据；主题的数量是有限的，对应低维的语义空间，主题挖掘就是通过“降维”将文档从高维空间投影到了语义空间。PLSA 通常运用 EM 算法对模型进行求解[8]。

LDA(Latent Dirichlet Allocation)[7,9] 在 PLSA 的基础上加入了 Dirichlet 先验分布，是 PLSA 的一个突破性的延伸。LDA 的创始者 Blei 等人指出，PLSA 在文档对应主题的概率计算上没有使用统一的概率模型，过多的参数会导致过拟合现象，并且很难对训练集以外的文档分配概率。基于这些缺陷，LDA 引入了超参数，形成了一个“文档－主题－单词”3 层的贝叶斯模型[9]，通过运用概率方法对模型进行推导，来寻找文本集的语义结构，挖掘文本的主题。

本文针对专利信息数据，提出了一种基于 LDA 主题模型对发明人进行兴趣主题分析的建模方法，并在发明人主题概率分布的基础上，提出了一种适用于专利发明人的推荐算法。

3 LDA 主题模型

LDA 模型是一个层次贝叶斯模型[7]，它有如下三层：

(1) 单词层：由单词集合 $V=\{w_1,w_2,\cdots,w_V\}$ 来表示。对于文档集中的所有文本，对其进行分词处

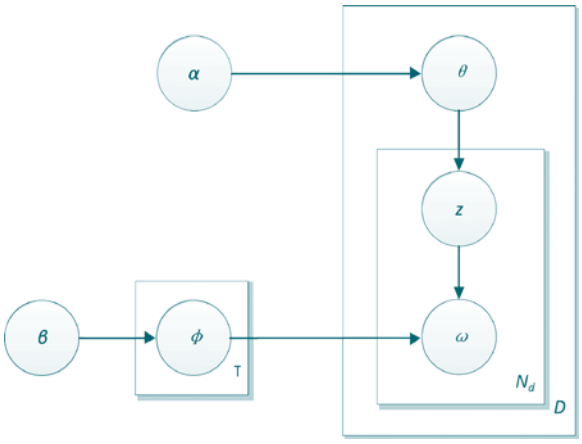
理及停用词过滤等处理后，即可得到单词集合 V，其中 w 便代表了一个单词。

(2) 主题层：主题层可以由一个主题集合 $\phi=\{z_1,z_2,\cdots,z_k\}$ 来表示。其中，z_i 代表某个主题。它可以表示成一个概率分布，其中，p_{kj} 表示的就是单词 w_j 属于主题 z_k 的概率。

(3) 文档层：文档层可以有两种表示方法，若从单词层的角度来考虑，可以利用向量空间模型的思想，使用基于单词的特征向量来表示文档 $d_i=<tf_{i,1},tf_{i,2},\cdots,tf_{i,V}>$，其中，$tf_{i,j}$ 是单词 i 的词频。若是换一个角度，从主题层面来考虑，可以将文档集表示为主题空间下的概率分布 $\theta=<\theta_1,\theta_2,\cdots,\theta_D>$，其中每一个向量 $\theta_d=<p_{d,1},p_{d,2},\cdots,p_{d,K}>$ 表示了一个文档的主题分布，$p_{d,z}$ 是主题 z 在该文档 d 中的生成概率。LDA 主题模型表示如图 1 所示。

LDA 模型采用 Dirichlet 分布作为概率主题模型中多项分布的先验分布。图中，D 为整个文档集，N_d 为文档 d 的单词集，α 和 β 分别是文档 – 主题概率分布 θ 和主题 – 单词概率分布 φ 的先验知识，其中 θ 和 φ 服从 Dirichlet 分布、α 和 β 是 Dirichlet 分布中的参数。

图 1. LDA 主题模型



4 发明人主题发现与推荐方法

4.1 发明人主题兴趣模型

标准的 LDA 模型是基于文档 – 主题 – 词的一个三层贝叶斯模型[7]，其矩阵表示如图 2 所示。

从上图中可以看出，LDA 模型将文档集合 D 中的每篇文档 d 表示成了在主题空间 Φ 下的概率分布 Θ。

在使用 LDA 主题模型构建发明人的兴趣模型时，我们将发明人对每一个主题的喜爱程度定义为该发明人的兴趣。因此，主题模型下发明人 – 主题生成概率多项分布表示了发明人对各个主题的兴趣。实验中，我们将每一个发明人的所有专利条目合并成一个文档进行主题建模，每一个发明人被表示为一个文档，从而得到发明人生成主题的概率多项分布，即发明人兴趣模型。该兴趣模型的发明人层就对应到了 LDA 模型中的文档层，即将文档 – 主题 – 词的三层关系变为了发明人 – 主题 – 词的关系，这即为发明人兴趣模型与标准 LDA 模型之间的区别。和其矩阵表示如图 3 所示。

在发明人层中，发明人集合 $U=\{u_1,u_2,\ldots,u_m\}$ 中的每一个发明人 u_i 由其对应的专利集 $\{f_{ui,1},f_{ui,2},\ldots,f_{ui,n}\}$

$$\underset{\text{词语}}{\begin{matrix}w_1\\w_2\\\cdots\\w_n\end{matrix}}\overbrace{\begin{Bmatrix}\\D\\\\\end{Bmatrix}}^{\text{文档}\ d_1\ d_2\ \ldots\ d_m}=\underset{\text{词语}}{\begin{matrix}w_1\\w_2\\\cdots\\w_n\end{matrix}}\overbrace{\begin{Bmatrix}\\\Phi\\\\\end{Bmatrix}}^{\text{主题}\ z_1\ z_2\ \ldots\ z_k}\times\underset{\text{主题}}{\begin{matrix}z_1\\z_2\\\cdots\\z_k\end{matrix}}\overbrace{\begin{Bmatrix}\\\Theta\\\\\end{Bmatrix}}^{\text{文档}\ d_1\ d_2\ \ldots\ d_m}$$

图 2 标准 LDA 模型矩阵

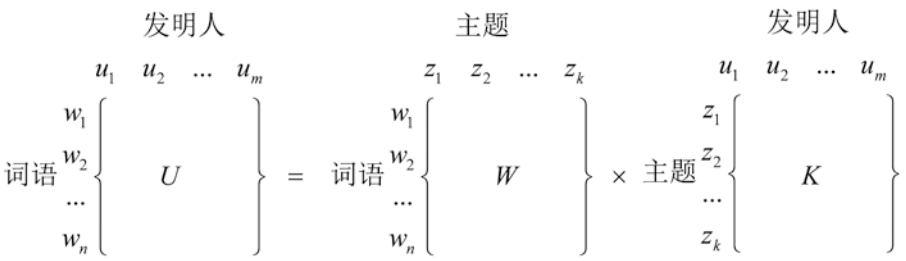
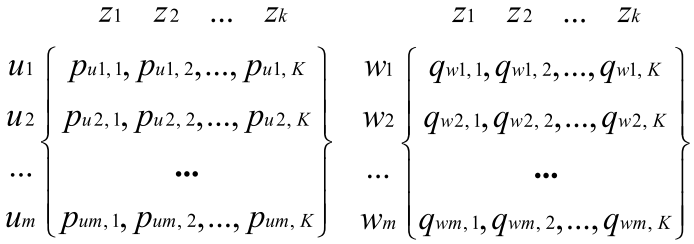


图 3 基于 LDA 的发明人兴趣模型矩阵

组成，专利集中的$f_{ui,j}$是发明人 所发明的所有专利组成的词频向量，如图 3 中的矩阵 U 所示。从主题层面而言，发明人 可以被表示成向量 ρ ui={pui,1,pui,2,⋯,pui,k}，其中 表示主题 z 在发明人 中的生成概率，用它来表示发明人 对主题 z 的感兴趣程度，从而，发明人层构成了发明人与主题的生成关系，从而生成主题用户模型，其矩阵表示如下图 4（a）所示，而词语 w_i 可以被表示成向量 $\theta_{wi}=\{q_{wi,1},q_{wi,2},...,q_{wi,k}\}$，其中 $q_{wi,z}$ 表示词语 w_i 属于主题 z 的生成概率，其矩阵表示如下图 4(b) 所示。



(a) 发明人 - 主题 (b) 词语 - 主题

图 4 发明人兴趣模型的矩阵表示

4.2 发明人相似度计算

KL(Kullback Leibler) 散度，也可以叫做 KL 距离，对于概率分布间的距离而言，它是最合适的计算方法 [10]，其计算公式如下式 (1) 所示：

$$D_{KL}(P\|Q)=\sum_i P(i)\ln\frac{P(i)}{Q(i)} \quad (1)$$

其中，$D_{KL}(P\|Q)$表示随机变量 $P=(p_1,p_2,...,p_M)$ 和 $Q=(q_1,q_2,...,q_M)$ 之间的 KL 距离，当两个概率分布完全相同时，它的值等于零，也就是他们的距离为零，其他情况下，这个 KL 距离始终都是大于零的。

由于 KL 散度是不对称的，即 $D_{KL}(P\|Q)\neq D_{KL}(Q\|P)$，因此可以将其转换为对称的形式，如下式 (2) 所示：

$$D(P,Q)=[D_{KL}(P\|Q)+D_{KL}(Q\|P)]/2 \quad (2)$$

这个距离值越小，则表明两个概率分布 P 和 Q 之间的差异越小，即相似度越高。

在基于 LDA 的发明人主题模型中，发明人的兴趣是由其主题空间的概率分布来表示的，衡量发明人之间的相似性也就是衡量发明人主题概率分布的相似性，因此发明人之间的相似程度就可以采用 KL 距离来表示，计算公式如下式 (3) 所示：

$$sim(U_i,U_j)=\frac{1}{D(U_i,U_j)}=\frac{2}{[D_{KL}(U_i\|U_j)+D_{KL}(U_j\|U_i)]} \quad (3)$$

$sim(U_i,U_j)$ 为发明人 u_i 和 u_j 的相似度，U_i 和 U_j 分别是他们的主题概率分布。该值越大，则表明参与计算的两发明人之间越相似。

4.3 推荐算法

算法：基于 LDA 主题模型的发明人推荐算法

(1) 建立发明人主题模型：

提取发明人的所有专利，将发明的主题内容合并到一起，并进行分词、去除停用词等预处理，得到代表每个发明人的专利词语的词频向量 ，对模型进行求解，得到每个发明人的主题概率分布；

(2) 相似度计算：

借助于概率分布之间的 KL 散度计算方法，发明人之间的相似度使用公式 (3) 来计算，该值越大则表示发明人之间的主题概率分布越相似，也即发明人之间的兴趣越相似，双方可以相互作为被推荐给对方的候选发明人；

(3) 用户推荐：

根据计算出的相似度，对发明人进行排序，从推荐列表中选取前 t 个发明人作为推荐列表。

4.4 评价指标

我们假设同一个技术领域下的发明人为兴趣相近的，并且他们的发明内容也是和自己感兴趣的方向相关的。

U 为发明人集合，对发明人 u_i 和发明人子集 U_i，其中，$u_i \in \boldsymbol{U}$，且 $\boldsymbol{U}_i=\boldsymbol{U}-u_i$。按照公式 (3)，对发明人集合 中的每个发明人分别与 u_i 计算相似度，然后对 U_i 中的所有发明人按照相似度值进行排列作为推荐列表，排在前面的用户就和用户 u_i 更相似。

选取 topt 个发明人作为对发明人 u_i 的推荐列表，$\boldsymbol{Rec}_{i,t}=\{u_1,u_2,\cdots,u_j,\cdots,u_t\}$。对推荐集合 $R_{eci,t}$ 中的每个发明人 u_j，分别判断其是否与发明人 u_i 属于同一领域，若属于同一领域，则认为将 u_j 推荐给发明人 u_i 是正确的。发明人 u_i 的推荐准确率计算如公式 (4)、(5) 所示：

$$Accuracy(u_i)=\frac{\sum_{u_j\in \boldsymbol{Rec}_{i,t}} f(u_i,u_j)}{t} \quad (4)$$

$$f(u_i,u_j)=\begin{cases}1 & u_i,u_j \text{ 属于同一领域}\\ 0 & u_i,u_j \text{ 不是同一领域}\end{cases} \quad (5)$$

其中，t 是推荐列表的大小并且 $t \leqslant N_i\text{-}1$，N_i 为发明人 u_i 所属领域下的发明人总数，t 的取值不超过该领域下的发明人总数减 1(除去用户 u_i 自身)。

某领域 p 下发明人的推荐准确率计算公式见下式 (6)：

$$Accuracy(p)=\frac{\sum_{i=1}^{N_p} Accuracy(u_i)}{N_p} \quad (6)$$

其中，N_p 为领域 p 下的发明人总数。

实验中所有发明人的平均推荐准确率计算公式如 (7) 所示：

$$\overline{Accuracy}=\frac{\sum_{i=1}^{N} Accuracy(u_i)}{N} \quad (7)$$

其中，N 为实验中的发明人总数。

5 实验与评价

5.1 数据集

实验中，我们采集了 2008 年 1 月至 2008 年 12 月的相关专利数据，解析并处理重要的专利信息，这些信息包括专利标题、技术类别、发明人、申请时间、公开号、所属地区和摘要等数据项。对这些数据进行分类整理，提取出所有发明人并根据技术类别进行分类。我们选取其中有代表性的 10 个技术类别中的发明人，根据专利数量进行排序，选取专利数量最多的前 100 个发明人作为实验数据集。最终，实验数据集的组成如表 1。

表 1 实验数据统计表

领域编号	领域	发明人数量
40	计算机和自动化技术	100
41	环境	100
38	电子元件；电子线路	100
32	交通运输	100
34	包装；容器；陈列	100
11	食品	100
16	医疗；卫生；消防	100
22	石油；燃料；能源	100
27	机械元件	100
12	农林；畜牧；水产	100
总计	/	1000

如表 1，实验数据集由 1000 个发明人及其专利组成，将数据集中每一个发明人的所有专利进行信息提取和处理。我们提取专利标题和摘要信息，对其进行分词、去除停用词等预处理，同时，我们认为标题和摘要数据对发明人兴趣提取的影响不同，因此对其设置不同的权重，将标题中的单词的出现次数设为 2，将摘要中的单词的出现次数设为 1，最后将同一个发明人的所有专利进行合并，每个发明人都可以由其专利组成的单词向量所表示。

5.2 实验设计

我们使用采集的专利数据作为实验数据集，对其中的每个发明人，使用 LDA 进行兴趣主题建模，并进行发明人的推荐。实验步骤如下：

（1）对每个发明人的合并专利数据进行分词、去除停用词等预处理，得到每个发明人单词向量；

（2）使用 LDA 对发明人进行兴趣主题建模，得到每个发明人对主题的概率分布；

（3）对每一位发明人，将数据集中其余的发明人作为其推荐候选者集合；

（4）使用 4.2 节介绍的 KL 散度方法，计算不同发明人之间的主题分布的相似度，也即兴趣的相近程度；

（5）根据主题分布的相似度排序，为发明人推荐最为相似的前 K 个用户；

（6）使用 4.4 节介绍的评价指标计算方法，计算推荐准确率。

上述实验步骤 2 中，对发明人进行 LDA 主题建模，其中模型的求解采用了 Gibbs 抽样方法。该方法需要设置两个主要的参数，实验中根据经验取值的方法，将参数 α 的值设置为 $50/T$，T 为主题数，将参数 β 的值设置为 0.01。主题数 T 取不同值时，主题模型的分布有差异。实验中我们令 T 取值从 8 至 20，观察不同取值下主题分布的情况，发现当 T 取值为 12 时主题分布结果更加符合实际情况。

为了进一步对比实验效果，把本文方法与下面 2 个方法进行比较：

· 基于向量空间模型 (VSM) 的算法

使用传统的 VSM 方法 [11] 建立发明人兴趣模型，同样对于发明人集合 $\boldsymbol{U}=\{u_1,u_2,\cdots,u_m\}$，将发明人 u_i 的所有发明专利数据进行预处理后得到其单词权重向量 $\boldsymbol{U}_i=<w_{i,1},w_{i,2},\cdots,w_{i,V}>$，其中，$w_{i,j}$ 表示单词 j 在发明人 u_i 的专利数据中的权重。这里的权重计算采用 TF−IDF 值。发明人间相似度的计算采用常规的向量夹角的余弦值来计算，公式如下所示：

$$Sim(u_i,u_j)=\frac{\boldsymbol{U}_i\cdot\boldsymbol{U}_j}{|\boldsymbol{U}_i^2||\boldsymbol{U}_j^2|} \tag{8}$$

· 基于隐马尔科夫模型 (HMM) 的算法

应用文献 [12] 中介绍的方法。我们将发明人按照主题来发明专利的行为看做一个隐马尔科夫过程，所以专利作为发明人的文档，可以用隐马尔科夫模型来描述发明人，模型描述如公式 (9) 所示：

$$\lambda=(A,B,\pi,N,M) \tag{9}$$

其中，$A=\{a_{ij}\}$是隐含状态之间的转移概率，a_{ij}代表从状态 i 到状态 j 的转移概率。$B=\{b_j(k)\}$是每个隐含状态下单词的分布概率，$b_j(K)$ 表示在当前状态 j 下输出单词状态为 V_k 的概率。$\pi=\{\pi_i\}$是隐藏状态的初始概率分布。N 是隐藏状态的数量，也即主题的数量。M 是可观测状态的数量，也即文档中不同单词的数量。

为发明人建立隐马尔科夫模型的过程属于隐马尔科夫模型要解决的第三类问题，即学习问题。专利集合中的每个单词是我们可以观察获取到的，可以看成是观测序列 。$O=O_1O_2...O_t$模型中的隐含状态即可以认为是发明人所感兴趣的主题，我们的目的就是根据观测到的专利中的单词序列获取模型中的未知参数：隐含状态转移概率矩阵 A 和

观测状态转移概率矩阵 B。

隐马尔科夫模型的学习问题使用 Baum-Welch 算法来解决，经过迭代训练后，得到如公式（9）所示的隐马尔科夫发明人模型，然后使用 KL 散度计算发明人间的相似度，计算公式为：

$$D_{KL}(\lambda_m,\lambda_n)=\frac{1}{V}(\ln P(w_s\mid\lambda_m)-\ln P(w_s\mid\lambda_n)) \quad (10)$$

以上 2 种方法的发明人推荐准确率的计算方法和 LDA 用户模型的计算方法相同，不再赘述。

5.3 实验结果与分析

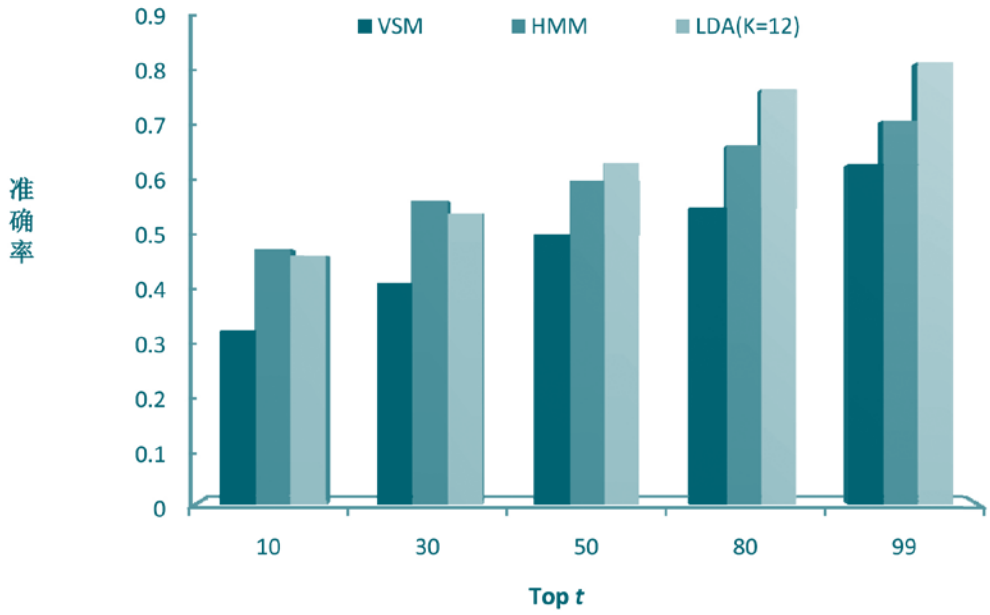
LDA 主题模型在不同主题数下有不同的性能，实验通过 LDA 对发明人进行主题兴趣建模并进行相应的发明人推荐，使用公式（7）所示的平均推荐准确率作为实验结果的评价指标。实验中，主题数 K 分别取值为 8 至 15，以及 20，平均准确率中的参数 t 分别取值 10、20、50、80 和 99，具体统计结果如表 2 所示：

表 2 实验结果统计表

Topt	LDA 用户模型 (主题数为 K)								
	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=20
10	0.340	0.335	0.380	0.326	0.490	0.380	0.336	0.377	0.411
30	0.390	0.415	0.445	0.470	0.540	0.516	0.460	0.435	0.434
50	0.540	0.488	0.536	0.535	0.633	0.546	0.490	0.485	0.492
80	0.531	0.560	0.575	0.632	0.770	0.680	0.595	0.626	0.550
99	0.590	0.672	0.754	0.786	0.820	0.770	0.745	0.706	0.711

当主题数 K 为 12 时，我们分别对使用 LDA 用户模型、VSM 方法以及 HMM 模型的推荐结果使用公式 (7) 进行平均推荐准确率计算，对在不同推荐数量情况下的计算结果进行对比，其中，参数 t 分别取值 10、20、50、80 和 99，结果如下图 5 所示：

图 5 三种推荐算法的实验性能



结合表 2 和图 5 对实验结果进行分析，得出以下结论：

（1）推荐性能与主题数相关。随着主题数的增加，推荐性能提高，在主题数为 12 时，性能最好，当主题数进一步增加时，效果基本保持稳定甚至略微有所回落。主题数越大，模型的计算量也越大，耗时越久，综合可虑，在主题数取 12 的时候，无论是推荐效果还是计算效率都有着不错的结果。

（2）对比主题模型和其他两种模型的实验结果可以看出，当主题数取最优值 12 的情况下，本文提出的基于 LDA 的用户兴趣模型的性能比 VSM 和 HMM 均有所提高。在推荐 Top10 的情况下，LDA 方法的推荐准确率要比传统的 VSM 方法高 0.15，与 HMM 则相差不多；在推荐 Top50 的情况下，三种方法的准确率都有所提升；当推荐 Top99 的情况下，LDA 比 HMM 和 VSM 的准确率均高，到达最高值，明显体现出了主题模型的优势。

6 结论

本文针对专利文本数据，结合 LDA 模型的文档 – 主题 – 词分层模型的特点，用发明人所拥有的专利数据集合来代表发明人，提出了发明人 – 主题 – 词的发明人兴趣模型，不仅能有效挖掘发明人所关注的主题，并基于相似度计算的方法对发明人进行了个性化推荐。

本文的研究工作还有一些不足之处，后续工作需要继续改进：在更大规模和更广泛领域的专利数据集上验证算法的效果；继续优化专利发明人兴趣模型的效果和效率；将算法应用于除中文专利外的其他类型数据中，进一步验证算法在实际应用中的效果。

参考文献

[1] Kang J H, Lerman K, Plangprasopchok A. Analyzing Microblogs with Affinity Propagation[C]// Proceedings of the First Workshop on Social Media Analytics. New York, USA: ACM Press, 2010: 67-70.

[2] Xu R, Wunsch D. Survey of Clustering Algorithms[J]. IEEE Trans on Neural Networks,2005,16(3):645-678.

[3] Deerwester S, Dumais S, LandauerT,et al. Latent Semantic Analysis for multiple-type interrelated data objects[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. New York, USA: ACM Press, 2006: 236-243.

[4] David Blei. Probabilistic topic models[J]. Communications of the ACM, 2012, 4(55): 77-84.

[5] Young-Min Kim. An extension of PLSA for document clustering[C]//Proceeding of the 17th ACM conference on Information and knowledge management. New York, USA: ACM Press, 2008: 1345-1346.

[6]Xuning Tang, Christopher C. Yang. TUT: A Statistical Model for Detecting Trends, Topics and user interests in Social Media[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. New York, USA: ACM Press, 2012: 972-981.

[7]Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research,2003,3(3):993-1022.

[8] Sarah Zelikovitz, Haym Hirsh. Using LSI for text classification in the presence of background text[C]// Proceedings of the tenth international conference on Information and knowledge management. New York, USA: ACM Press, 113-118.

[9] Wei X, Croft W B. LDA-based document models for ad hoc retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. New York, USA: ACM Press, 2006: 178-185.

[10]Sun C N,Zheng C,Xia Q S. Chinese Text Similarity Computing Based on LDA[J]. Computer Technology and Development, 2013(1)217-220.

[11] Fei H X,Jiang C,Xu L J.User profile based on dendriform vector space model.Computer Technology and Development,2009,19(5):79-81

[12] Jianping Zeng, Shiyong Zhang, Chengrong Wu. A Framework for WWW User Activity Analysis Based on User Interest[J]. Knowledge-Based Systems, 2008, 12(21):905-910.