文章编号: 1000-5641(2018)05-0183-12

基于知识图谱和 LDA 模型的社会媒体数据抽取

麻 友[1], 岳 昆[1], 张子辰[1], 王笑一[2], 郭建斌[2]

(1. 云南大学 信息学院, 昆明 650500; 2. 云南大学 民族学与社会学学院, 昆明 650500)

摘要: 社会媒体数据的抽取, 是社会舆论集散、新闻信息传播、企业品牌推广、商业营销拓展等研究和应用的基础, 准确的抽取结果是数据分析有效性的重要保证. 本文针对社会媒体数据的非结构、多主题特征, 基于 LDA(Latent Dirichlet Allocation) 模型挖掘数据中的隐含主题, 利用数据特征词序列和知识图谱描述的实体及实体间的关联关系, 实现对特定领域数据的抽取. 建立在"今日头条"新闻数据和新浪微博数据之上的实验结果表明, 本文提出的方法能有效地实现社会媒体数据的抽取.

关键词: 社会媒体数据; 数据抽取; 隐含狄利克雷分配; 知识图谱

中图分类号: TP311 **文献标志码**: A **DOI:** 10.3969/j.issn.1000-5641.2018.05.016

Extraction of social media data based on the knowledge graph and LDA model

MA You[1], YUE Kun[1], ZHANG Zi-chen[1], WANG Xiao-yi[2], GUO Jian-bin[2]

(1. *School of Information Science and Engineering, Yunnan University, Kunming* 650500*, China;*
2. *School of Ethnology and Sociology, Yunnan University, Kunming* 650500*, China*)

Abstract: Social media data extraction forms the basis of research and applications related to public opinion, news dissemination, corporate brand promotion, commercial marketing development, etc. Accurate extraction results are critical to guarantee the effectiveness of the data analysis. In this paper, we analyze the underlying topics in data based on the LDA (Latent Dirichlet Allocation) model; we further implement data extraction in specific domains by adopting featured word sequences and knowledge graphs that describe entities and relevant relationships. Experimental results using "Headline Today" news and Sina Weibo data show that our proposed method can be used to extract social media data effectively.

Keywords: social media; data extraction; LDA (Latent Dirichlet Allocation);

收稿日期: 2018-07-10

基金项目: 国家自然科学基金(61472345); 云南大学青年英才培育计划(WX173602); 云南大学科研基金(2017YDJQ06); 云南大学研究生科研创新基金(Y2000211)

第一作者: 麻 友, 男, 硕士研究生, 研究方向为海量数据处理与知识发现.
E-mail: 1172880152@qq.com.

通信作者: 岳 昆, 男, 教授, 博士生导师, 研究方向为海量数据处理与知识发现.
E-mail: kyue@ynu.edu.cn.

knowledge graph

0 引 言

近年来, 随着社会媒体在信息传播中发挥日益重要的作用, 社会媒体数据的抽取与分析也受到国内外学者的高度关注[1]. 从海量、异构的社会媒体数据中获取特定主题的数据, 并根据领域不同进行主题分析、内容筛选和过滤, 是社会媒体数据抽取的重要研究内容, 也是决策支持、影响预测、知识库构建和舆情分析等工作的重要基础[2].

与传统结构化数据不同, 社会媒体数据多为非结构化或半结构化, 具有规模大、短文本、多主题、数据稀疏和随意性强等特点, 为数据抽取与分析带来了挑战. 具体包括: ① 社会媒体数据非结构、多主题的特点, 需要一种扩展性强、支持多主题的方法进行主题特征抽取; ② 特定领域的数据通常词源生僻、专业性强, 难以快速把握, 需要特定领域的先验知识作为基础, 利用领域知识丰富短文本信息量, 以提高针对特定领域数据抽取结果的准确性.

从对社会媒体数据处理和分析利用的方便性来看, 往往需要对数据大致内容按照主题识别和分类. 目前主题分类方法有支持向量机(Support Vector Machine, SVM)[3]、人工神经网络(Artificial Neural Network, ANN)[4]和随机森林(Random Forests Algorithm, RFA)[5]等. 然而, 由于 SVM 缺少对非线性问题的通用性, ANN 学习过程较复杂, RFA 在噪音较大时会产生过拟合结果, 因此, 在社会媒体数据的抽取问题上具有一定的局限性.

在研究文本集的隐含主题问题时, Blei 等[6]提出了 LDA(Latent Dirichlet Allocation)模型. LDA 是一种经典贝叶斯层次模型, 其基本思想是将每个文本表示为主题的多项分布, 每个主题表示为词汇的多项分布, 进而得到文本潜在的主题结构. LDA 模型被广泛用于信息抽取、文本分析、社交网络和自然语言处理等领域. Jaradat 等[7]将 LDA 应用于 Twitter 数据主题分析; 文献 [8] 提出了非参数贝叶斯模型 MB-HDP, 有效地实现主题分析; 文献 [9] 则使用 LDA 对海量的电影评论数据进行了定性和描述性的主题提取; 文献 [10] 在 LDA 的基础上根据情感层次扩展了在线模型, 实现更准确的微博主题和情感信息挖掘与分析; 文献 [11] 研究海量的电商评论数据, 实现了基于语义约束 LDA 的商品特征和情感词提取. 但是, 针对特定领域的社会媒体数据抽取, 上述基于 LDA 的主题分析和信息挖掘方法仍有待进一步扩展. 因此, 针对挑战(1), 本文基于 LDA 模型挖掘社会媒体数据中的隐含主题, 实现数据从“文本-词汇”到“文本-主题”的降维, 对于数据的多个主题, 获取各个主题的高频词, 对每条数据进行主题分析, 并得到数据的特征词序列, 实现对每条数据的特征抽取, 以简化数据的抽取问题.

知识图谱(Knowledge Graph, KG)是一种语义网络, 表达了实体、概念及其之间的语义关系, 广泛用于个性化推荐、智能搜索、知识发现、内容分发等领域. KG 被用于医疗[12-13]、音乐[14]、建筑[15]等领域的信息抽取研究. Meij 等[16]利用 KG 实现以文本为中心的信息检索. 文献 [17] 研究关系抽取时提出一种面向中文维基百科领域知识的演化方法. Marin 等[18]结合 KG 与无标签数据实现文本短语的分类. Kliegr 等[19]将知识图谱应用于文本的信息分析中, 为海量数据中的信息分析和抽取提供了先验知识. 文献 [20] 提出了一种将语义相似性聚类算法与 KG 结构相结合的 KGRank 方法, 以发现隐藏在文档中的语义关系, 实现关键词抽取的目的. KG 同时有助于解决短文本数据稀疏的问题, 如 Chen 等[21]利用领域知识作为先验知识指导模型推理. 针对挑战 ② 特定领域的约束或知识, 本文首先将领域

知识表示成 KG, 然后利用 KG 丰富的实体信息获取特定领域的生僻词, 进而根据 KG 实体间的关联、同名实体的不同属性和异名实体可能指代同一事物的特征, 来判别一词多义、异词同义等情况. 同时, 对于短文本数据稀疏、信息量少的特点, 我们利用 KG 的实体间的关联得到相似数据内容间的联系, 从而丰富短文本的信息、提高数据抽取的准确性.

本文第 1 节给出基于 LDA 的社会媒体数据主题分析; 第 2 节给出知识图谱引导下的数据抽取方法; 第 3 节给出实验结果; 第 4 节总结全文, 并展望将来的研究工作.

1　基于 LDA 的社会媒体数据主题分析

本节中首先给出了社会媒体数据的表示, 再获取各个主题的高频词与数据特征词, 以此进行社会媒体数据的主题分析. 为了便于阅读, 给出相关符号及其含义对照, 见表 1.

表 1　符号及含义

Tab. 1　Notations

符号	含义
I_i	第 i 条数据
z_k	第 k 个主题
M	社会媒体数据总条数
$w_{i,j}$	第i条数据第 j 个词
$z_{i,j}$	词 $w_{i,j}$ 所属的主题
Λ_i	第 i 条数据的主题向量
$\lambda_{k,i}$	I_i 中词汇属于主题 z_k 的概率
Δ_k	主题 z_k 的高频词向量
$\delta_{t,k}$	主题 z_k 总词汇中的词 w_t 的概率
χ_i	数据 I_i 的主题的高频词向量
d_i	数据 I_i 的特征词序列

1.1　社会媒体数据的表示

我们用 $I=\{I_1, I_2, \cdots, I_M\}$ 表示社会媒体数据集, 下面给出社会媒体数据的定义.

定义 1　社会媒体数据的一个实例 I_i 表示为一个三元组(id, T_i, A_i), 其中

(1) I_i 表示第 i 条数据, $1 \leqslant i \leqslant M$, M 为社会媒体数据总条数;

(2) id 表示数据实例标识, 唯一标识每条数据;

(3) T_i 表示第 i 条数据的文字内容, 用词序列 $T_i =< w_{i,1}, w_{i,2}, \cdots, w_{i,n} >$ 表示, $w_{i,j}$ 表示 T_i 的第 j 个词, n 为 T_i 包含的词数;

(4) $A_i = \{A_{i,u}, A_{i,p}, A_{i,l}, A_{i,v}, A_{i,f}, A_{i,q}, A_{i,c}, A_{i,r}\}$, 表示附加信息, 分别表示数据发布者 $A_{i,u}$、发布时间 $A_{i,p}$、发布地点 $A_{i,l}$、发布源 $A_{i,v}$、转发量 $A_{i,f}$、点赞量 $A_{i,q}$、评论数 $A_{i,c}$ 和数据的读取时间 $A_{i,r}$.

例 1　表 2 给出社会媒体数据的例子. $id = a001$ 是该数据的唯一标识, 发布内容为文本内容 T_i, 附加信息包括发布者“国家摄影 unpcn”, 发布时间“2012-9-4 17:14”, 发布地点为空, 数据来源“微博 weibo.com”, 转发量 0, 点赞量 22, 评论数 12, 数据的读取时间为“2016-8-24”.

表 2　社会媒体数据示例

Tab. 2　Examples of social media data

id	T_i	A_i
a001	“再赴西藏之—— 青藏高原的风光 http://t.cn/zWgYG2J”	{$A_{i,u}$=“国家摄影 unpcn”, $A_{i,p}$=“2012-9-5 17:14”, $A_{i,l}$=“”, $A_{i,v}$=“微博weibo.com”, $A_{i,f}$=0, $A_{i,q}$=22, $A_{i,c}$=12, $A_{i,r=}$“2016-8-24”}

数据由词汇组成, 词汇的集合存储在词典中, 定义如下.

定义 2 词典用 W 表示, $W=\{w_1,w_2,\cdots,w_S\}$, S 为词典的词汇总数, $w_i\neq w_j(1\leqslant i, j\leqslant S, i\neq j)$.

词典 W 存储数据包含的全部词汇, 在本文主题分析过程中, 为每个词汇分配唯一确定的主题, 记录每个词汇所属主题, 应用于主题分析和特征抽取过程中.

1.2 主题高频词与数据特征词的获取

本文从"数据-词汇"中挖掘数据隐含的"主题"维度, 转化为低维的"数据-主题"问题. LDA 用概率图模型表示[22], 如图1所示.

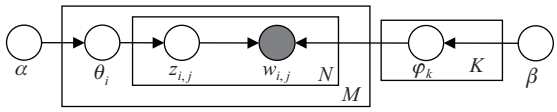

图1 LDA 图模型

Fig. 1 LDA graphic models

其中, φ_k 表示主题 k 中所有词汇的概率分布, θ_i 表示第 i 条数据的所有主题概率分布, θ_i 和 φ_k分别服从超参数 α 和 β 的 Dirichlet 先验分布. 该图模型主要包含两个物理过程:

(1) $\alpha\rightarrow\theta_i\rightarrow z_{i,j}$, 生成第 i 条数据第 j 个词的主题 $z_{i,j}$ 的过程;

(2) $\beta\rightarrow\phi_k\rightarrow w_{i,j}|k=z_{i,j}$, 生成第 i 条数据的第 j 个特征词 $w_{i,j}$ 的过程.

公式(1)给出了第 i 条数据 I_i 中第 j 个词 $w_{i,j}$ 的生成概率求解过程.

$$p(w_{i,j}|I_i)=\sum_{k=1}^{K}p(w_{i,j}|z_{i,j}=k)p(z_{i,j}=k). \tag{1}$$

其中, $p(w_{i,j}|z_{i,j}=k)$ 表示词 $w_{i,j}$ 出自主题 k 的概率, $p(z_{i,j}=k)$ 是数据包含主题 k 的概率.

根据式(1)的结果来更新该词对应的主题, 如果更新后其主题发生变化, 反过来也会影响 θ_i 和 φ_k, 迭代直到结果收敛.

本文用主题向量表示社会媒体数据, 用高频词向量描述各个主题. 设 $I_i(1\leqslant i\leqslant M)$ 为任意一条社会媒体数据, I_i 的主题向量定义为 $\Lambda_i=(\lambda_{1,i},\lambda_{2,i},\cdots,\lambda_{K,i})$. $\lambda_{k,i}$ 是 I_i 中词汇属于主题 z_k 的概率, $0\leqslant\lambda_{k,i}\leqslant1$. 其中, 主题 z_k 用高频词向量 $\Delta_k=((w_1,\delta_{1,k}),(w_2,\delta_{2,k}),\cdots,(w_{S_k},\delta_{S_k,k}))$表示, S_k 为 z_k 的总词数. $\delta_{t,k}$ 是 z_k 总词汇中的词 w_t 的概率, $0\leqslant\delta_{t,k}\leqslant1$, $\delta_{t,k}$ 和 $\lambda_{k,i}$ 分别由公式(2)和(3)计算.

$$\delta_{t,k}=\frac{n_k^{(t)}+\beta_t}{\sum_{r=1}^{S}n_k^{(r)}+S\beta_t}, \tag{2}$$

$$\lambda_{k,i}=\frac{n_i^{(k)}+\alpha_k}{\sum_{r=1}^{K}n_i^{(r)}+K\alpha_k}. \tag{3}$$

其中, $n_k^{(t)}$ 表示主题 z_k 的词汇 w_t 的总数, $n_i^{(k)}$ 表示 I_i 中包含主题 z_k 中词汇的数量.

每条数据 I_i 和每个主题 z_k 分别用主题向量和高频词向量表示. 按照 $\lambda_{k,i}$ 排序得到数据 I_i 的主题的高频词向量 χ_i, 将高频词向量与 T_i 进行匹配, χ_i 和 T_i 同时包含的词汇序列, 称为特征词序列, 记为 $d_i=<w_{i,1},w_{i,2},\cdots,w_{i,m_i}>$ 表示, m_i 为 d_i 的特征词的个数. 上述思路见算法 1.

算法 1 主题高频词与数据特征词的获取

输入: 数据集 I, 迭代数 N_{iter}, 主题总数 K, 参数 α, β, κ

输出: 数据主题向量 Λ_i, 主题高频词向量 Δ_k, 特征词序列 d_i

1: for $k = 1$ to K do
　　采样参数 φ_k∼Dir(β)
　end for
2: $n_i^{(k)} \leftarrow 0$; $n_k^{(t)} \leftarrow 0$
　for each I_i in I do
　　采样参数 θ_i∼Dir(α)
　　for each $w_{i,j}$ in I_i do
　　　采样主题 $z_{i,j}$∼Mult(θ_i); $w_{i,j}$∼Mult($\varphi_{z_{i,j}}$)
　　　$n_i^{(k)} \leftarrow n_i^{(k)} + 1$; $n_k^{(t)} \leftarrow n_k^{(t)} + 1$
　　end for
　end for
3: for $x = 1$ to N_{iter} do
　　for each I_i in I do
　　　for each $w_{i,j}$ in I_i do
　　　　if $z_{i,j} = k$ then
　　　　　$n_i^{(k)} \leftarrow n_i^{(k)} + 1$
　　　　　$n_k^{(t)} \leftarrow n_k^{(t)} + 1$
　　　　end if
　　　end for
　　$\delta_{t,k}$ ←根据公式 (2) 得
　　Δ_k ←降序排列 $\delta_{t,k}$, 得到 z_k 的高频词向量
　　$\lambda_{k,i}$ ←根据公式 (3) 得
　　Λ_i ←降序排列 $\lambda_{k,i}$, 得到 I_i 的主题向量
　　end for
　end for
4: for each I_i in I do
　　获取 I_i 的高频词向量 Λ_i 按 $\lambda_{k,i}$ 降序的 top-κ个主题
　　$d_i \leftarrow$ (T_i 中的词汇)∩(top-κ个主题的高频词向量 Δ_k 的词汇)
　　end for
　return $(\Lambda_i, \Delta_k, d_i)$

根据文献 [23] 的结论, 超参数 α 和 β 可以取合理的默认值, 即 $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 50/K, \beta = 0.01$. 算法 1 单次迭代复杂度为 $O(K \cdot S)$, K 和 S 分别为主题数和词的总数, S 与数据总条数 M 成正比. 因此, N_{iter} 次数迭代的时间复杂度为 $O(N_{iter} \cdot K \cdot M)$, 其中 N_{iter} 和 K 为算法初始时的常数, M 是影响算法效率的主要因素, 当 M 较小时算法 1 时间复杂度较高, 当$M \gg N_{iter} \cdot K$ 时算法的时间复杂度为 $O(M)$.

2　知识图谱引导的数据抽取

第 1 节通过特征抽取获得了数据特征词序列, 本节进一步实现特定领域数据的抽取. 首先

将领域知识表示为 KG, 进而利用领域 KG 引导从社会媒体数据中抽取出特定领域的数据.

2.1 领域知识图谱的表示

本节将领域知识表示为领域 KG, 记为 G_k, 定义如下.

定义 3 用 G_k=(V, E)表示 KG, 其中 V=$\{v_1, v_2,\cdots,v_n\}$ 表示 KG 中实体对应节点的集合, E=$\{e_1, e_2,\cdots, e_m\}$表示实体之间边的集合. 任意一条边对应一个节点三元组 e_x=(v_i, v_j, $label$), 节点 v_i 称为始点, 节点 v_j 称为终点, $label$ 为始点与终点的关系标签.

领域知识由该领域学者研究总结得出, 用 $Z =< term, attributes, addition >$ 表示, $term$ 为实体名, $attributes$ 为实体属性, $addition$ 为词条附加说明. 领域 KG 的表示, 首先依次取 Z 的元素实体名 v_i 与本领域名称 v_0 表示为三元组(v_0, v_i, $label$), $label$ 取 v_i 的属性作为 $v_0 \to v_i$ 的关系标签, 再依次建立每个元素 v_i 与 v_j 的三元组(v_i, v_j, $label$), 此时 $label$ 由节点的 $addition$ 得到 $v_i \to v_j$ 的关系标签. 如 v_i 与 v_j 无关系, 则相应的边也不存在, 所有的三元组共同构成领域知识图谱 G_k.

例 2 图 2 为领域 KG 的示例, 有向边表示节点间的关系, KG 中节点名称相同不一定表示同一实体, 不同名的实体也可能表示同一对象. 如两个“青藏高原”同名, 但表示不同属性的实体, 而“央金卓玛”与“韩红”表示同一人物.

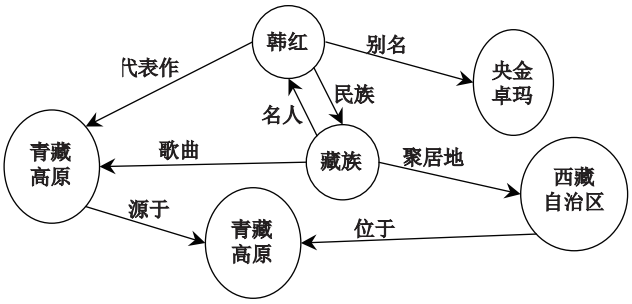

图 2 领域 KG 示例

Fig. 2 Example of field KG

我们对特定领域之外的数据, 如旅游、广告、电商以及歧义词汇等影响数据抽取的知识也表示为知识图谱, 记为 $\neg G_k$, 从而以 G_k 和 $\neg G_k$ 作为先验知识进行领域的数据抽取.

2.2 社会媒体数据的抽取

我们将 G_k 中的实体与数据的特征词匹配, 利用 G_k 实体间的关联, 找到数据中特定领域的全部词汇, 完成过滤匹配. 同时, 针对社会媒体数据的随意性, 我们利用 $\neg G_k$ 判别无关数据, 在数据抽取时进行去除. 用公式(4)判定数据 I_i 属于特定领域 G_k.

$$T = \begin{cases} 1, & (p > \tau)\text{且}(p' < \tau) \\ 0, & \text{其他} \end{cases}. \tag{4}$$

其中, p 表示数据 I_i 在领域 G_k 的概率, $p = \frac{n}{m_i}$, m_i 为特征词序列 d_i 的长度, n为 d_i 中包含在领域 G_k 中实体的个数. p' 表示 I_i 在领域 $\neg G_k$ 中的概率. τ 为给定的参数. 若 $T = 1$, 则 I_i 属于 G_k, 否则 I_i 不属于 G_k.

用 $D = \{I_1, I_2, \cdots, I_{M'}\}$ 表示抽取得到特定领域的数据集, M' 为 D 中社会媒体数据总条数. 从社会媒体数据的特征词序列中抽取特定领域数据的方法见算法 2.

算法 2 特定领域数据抽取

输入: 数据特征词序列 d_i, 领域 G_k, 领域 $\neg G_k$, 参数 τ

输出: 特定领域的数据集 D

for each $w_{i,j}$ in d_i do
  $m_i \leftarrow d_i$ 的长度
  for $v_x = v_0$ in G_k then
    if $w_{i,j} = v_x$ then
      $n \leftarrow n$+1
    else $v_x \leftarrow v_{x+1}$, 当边(v_x, v_{x+1}, *label*)存在于 G_k 中
    end if
  end for
  for $v_x = v_0$ in $\neg G_k$ then
    if $w_{i,j} = v_x$ then
      $n' \leftarrow n'$+1
    else $v_x \leftarrow v_{x+1}$, 当边(v_x, v_{x+1}, *label*)存在于 $\neg G_k$中
    end if
  end for
  $p \leftarrow n/m_i$
  $p' \leftarrow n'/m_i$
  if $p > \tau$ and $p' < \tau$ then
    $D \leftarrow D \cup \{I_i\}$
  end if
end for
return D

结合第 1 节主题分析获得的数据主题高频词向量和特征词序列, 最终抽取得到的领域数据表示为五元组(id, T_i, χ_i, d_i, A_i), χ_i 和 d_i 分别为数据的主题高频词向量和数据特征词序列.

例 3　如表 3 所示, 表中为一条以藏族领域为例的最终抽取的旅游主题的结果. χ_i 记录了旅游主题的高频词向量, 包括词汇和该词在旅游主题上的概率. d_i 记录了该条数据的特征词序列.

表 3　数据抽取结果示例

Tab. 3　Example of data extraction results

id	T_i	χ_i	d_i	A_i
ua6r54	#旅行攻略#【西藏】假期西藏游成为很多游客的旅游目的地, 去拉萨通常飞贡嘎机场, 离拉萨 67 km, 不过推荐坐高铁直达市区酒店.	((酒店, 0.009 1), (城市, 0.007 2), (旅游, 0.005 2), (旅行, 0.004 6), (假期, 0.004 5), (建筑, 0.004 3), (攻略, 0.003 8), (文化, 0.003 7), (公园, 0.003 2), (机场, 0.003 1), (推荐, 0.002 8), (km, 0.002 6), (游客, 0.002 5), (特色, 0.002 4), (高铁, 0.002 4))	<旅行, 攻略, 假期, 游客, 旅游, 机场, km, 推荐, 高铁, 酒店>	{$A_{i,u}$=“穷游网”$A_{i,p}$=“2016-7-31 7:44”,$A_{i,l}$=“”,$A_{i,v}$=“微博”,$A_{i,f}$=9, $A_{i,q}$=7, $A_{i,c}$=20, $A_{i,r}$=“2016-8-24”}

3　实验结果

3.1　实验设置

为了测试本文数据抽取方法的效率和有效性, 本文获取了今日头条平台[24] 2017 年的

21 567 条数据, 以及新浪微博平台 2017 年某月中 4 906 327 条数据作为实验测试数据. 数据包括 id、T_i 和 A_i 三个部分, 其中今日头条数据的附加信息 A_i 另包含了 $title$ 和 $category$, 分别表示文章的标题和分类. 实验环境如下: Intel® Core™i3-3240 CPU 3.40 GHz 处理器, 4.00 GB 内存, Windows7 64b 操作系统, Pycharm 开发平台, Python 语言编程实现.

本文的实验以藏族领域为例测试领域数据的抽取方法, 首先, 将藏族领域知识表示为藏族 KG, 用该藏族 KG 引导数据的抽取. 由于实验的新浪微博数据的随意性, 无自带的分类标签, 为了得到准确可靠的测试数据, 我们对微博一周内 906 327 条数据做人工标记, 标记的标准为该条数据内容是否与本文测试的藏族领域相关.

利用已知分类和人工标记相结合, 实验测试了数据抽取方法的有效性, 以及算法 1 和算法 2 的效率. 为了定量描述算法的有效性, 我们以准确率(Precision)、召回率(Recall)和 F 值(F-Measure)作为标准对本文的方法进行测试. 其中, 准确率 P 为抽取的准确数据总数占抽取总数据量的比例, 召回率 R 为抽取的准确数据总数与实际该领域总数据量的比例, F 值由公式(5)计算得到.

$$F = \frac{2 \cdot P \cdot R}{P + R}. \tag{5}$$

3.2 有效性测试

本文以“今日头条”新闻数据中的主题分类标签 $category$ 为依据, 测试算法 1 的有效性. 由于“今日头条”数据中不同 $category$ 的新闻条数不同, 本文选取数据量排前四的“娱乐”、“动漫”、“文化”和“科技”类别作为检测目标, 并在设置不同的主题数目情况下, 测试每个主题分类的准确率 P、召回率 R 和 F 值, 结果分别如图 3、图 4 和图 5 所示.

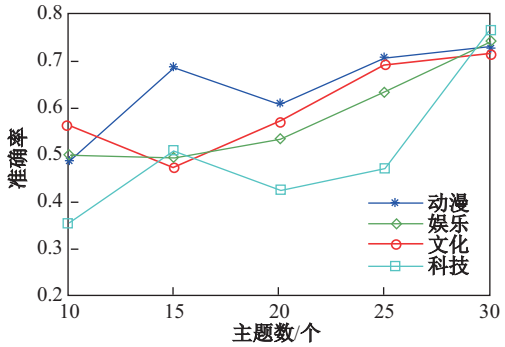

图 3 主题分析准确率

Fig. 3 Precision of topic analysis

从图 3 和图 4 可以看出, 随着设置主题数的增加, 各个主题检测的召回率下降, 而准确率上升. 即随着主题划分的细化, 每个主题下的数据进一步减少带来的变化, 对于如“科技”、“动漫”等区分性较强的主题, 其词汇与其他主题的词汇之间具有较为明显的区别, 设置主题数增多, 召回率依然较高. 此外, 由于“娱乐”、“动漫”、“文化”和“科技”四个主题的数据分别占总数据的 20.51%、7.85%、5.11% 和 3.35%, 不同主题的数据量也随着设置主题数的变化影响算法 1 的有效性. 从图 5 看出, 随着数据量的增加, F 值稳定在 0.55 上下, 说明算法 1 的有效性随着数据

量的增大仍能得到保证.

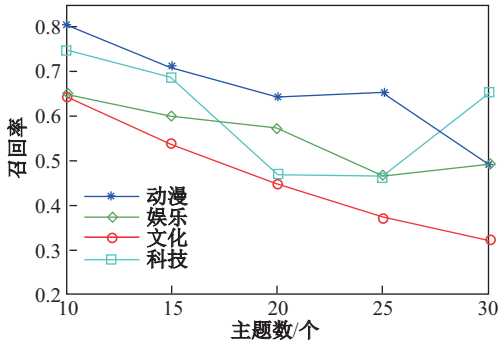

图 4　主题分析召回率

Fig. 4　Recall of topic analysis

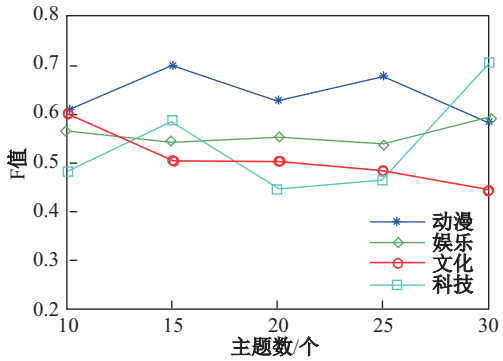

图 5　主题分析 F 值

Fig. 5　F-Measure of topic analysis

针对算法 2 特定领域的数据抽取, 本文以抽取藏族领域数据为例, 利用人工标记的 906 327 条微博数据进行测试, 并与 KMP 算法[25]为基础的关键词匹配筛选方法, 以及 KGRank 算法[20]为基础的数据抽取方法进行对比. KMP 算法是一种高效的字符串匹配算法, 以 KMP 算法为基础的数据过滤方法是数据抽取的最基本形式. 而 KGRank 是一种将语义相似性聚类算法与知识图谱结构相结合的关键词抽取方法, 以抽取的关键词为基础同样能实现数据的抽取. 实验测试了三种方法的有效性, 结果分别如图 6、图 7 和图 8.

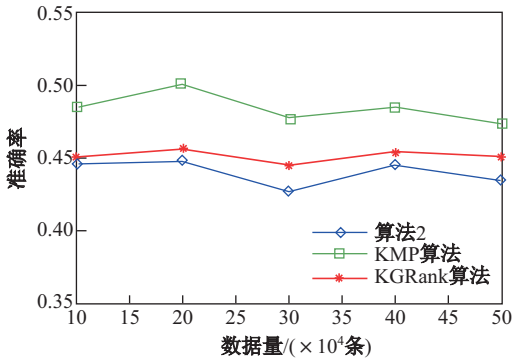

图 6　数据抽取准确率

Fig. 6　Precision of data extraction

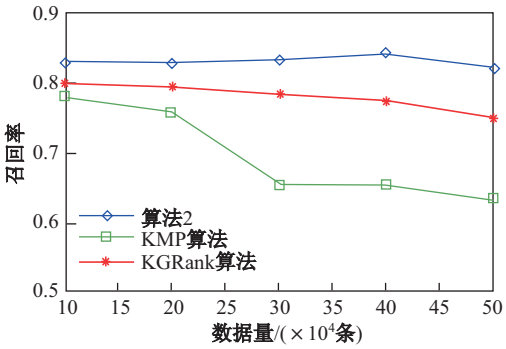

图 7 数据抽取召回率

Fig. 7 Recall of data extraction

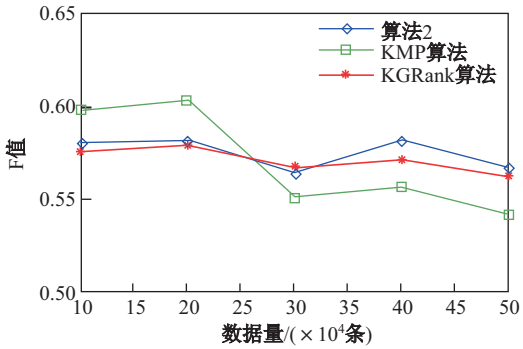

图 8 数据抽取 F 值

Fig. 8 F-Measure of data extraction

由图 6、图 7 和图 8 可以看出, 本文抽取结果的准确性稳定在 44%, 召回率在 82% 以上, 综合二者的 F 值在 57% 左右. 对比可知, 本文方法在准确率上略低于 KMP 算法和 KGRank 算法, 这是由于这两种算法为基础的数据抽取从词汇出发. 如 KMP 算法, 使用未经扩展的领域相关词汇进行字符匹配, 这些词汇与特定领域的关联性较高, 所抽取得到的数据准确率自然也较高. 而本文方法使用覆盖范围更全面的领域 KG 引导数据抽取, 在召回率上优于这两种方法, F 值也随着数据量增加优于 KMP 算法和 KGRank 算法. 结果说明本文方法能有效地实现特定领域数据的抽取.

3.3 效率测试

实验以新浪微博一个月的 4 906 327 条数据为依据, 对算法 1、算法 2 的效率进行了测试. 测试了算法 1 在不同迭代次数情形下, 随着数据量的增加执行时间的变化情况, 结果如图 9. 实验还测试了算法 2 与 KMP 算法和 KGRank 算法为基础的数据抽取方法的效率对比, 结果如图 10 所示.

由图 9 可看出, 算法 1 的执行时间随着数据量增加而呈线性增长, 而随着迭代次数的增加, 执行时间呈现非线性增长. 这是由于在一定的迭代次数内, 当结果已经达到收敛, 其后的迭代时间会减少. 本文实验综合考虑算法效率与有效性, 设定 100 次迭代. 由图 10 可看出, 三个算法的执行时间都与数据量呈线性正相关, 算法 2 在相同条件下比 KGRank 算法省时, 而比 KMP 算法耗时. 这是由于 KMP 算法的关键词匹配过程具有良好的性能, 其时间复杂度为 $O(m+n)$, m

和 n 分别表示目标文档长度和关键词长度, 在实现数据抽取时能发挥其性能的优势. 而本文的方法需要查找 KG 关联节点, KGRank 需要先实现关键词抽取进而抽取领域数据, 耗时增多.

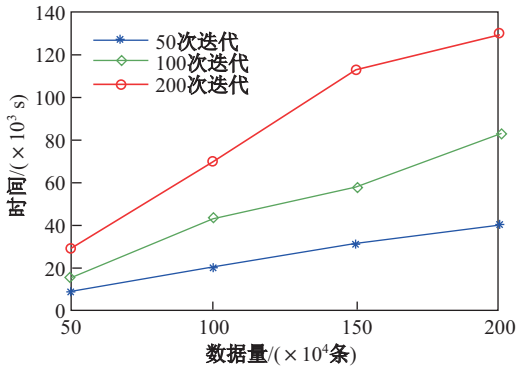

图 9　算法 1 执行时间

Fig. 9　Execution time of Algorithm 1

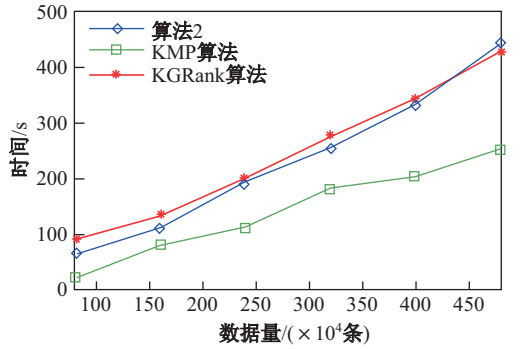

图 10　执行时间对比

Fig. 10　Comparison of execution time

4　总结与展望

本文研究社会媒体数据的抽取, 给出了基于 LDA 模型的社会媒体数据主题分析方法, 该方法能够挖掘数据中隐含的主题, 实现特征抽取. 进而引入领域 KG, 提供了一种特定领域的数据抽取方法. 但是, 实验数据来源还较为单一, 缺乏多平台、多维度的数据; LDA 模型是非监督学习, 难以准确地、有目的性地进行主题划分; 且本文用领域知识表示 KG, 领域知识的不完备性使得 KG 也不够全面.

我们接下来将进一步扩展数据来源, 获取包括网易新闻、Facebook、Twitter 等多平台、多维度的数据, 针对更大规模的社会媒体数据对所提出的方法进行进一步测试. 在特征抽取中实现监督或半监督学习, 提高准确性. 而针对领域知识的不完备, 接下来将研究以数据驱动的领域 KG 增量化补全. 针对海量数据规模的高效方法, 以及算法的进一步优化也是我们今后将要开展的工作.

[参　考　文　献]

[1] OUYANG Y, GUO B, ZHANG J, et al. SentiStory: Multi-grained sentiment analysis and event summarization with crowdsourced social media data[J]. Personal & Ubiquitous Computing, 2017, 21(1): 97-111.

[2] HE W, WANG F K, AKULA V. Managing extracted knowledge from big social media data for business decision making[J]. Journal of Knowledge Management, 2017, 21(2): 275-294.

[3] ZHOU X, GUO L, LIU P, et al. Latent factor SVM for text categorization[C]// IEEE International Conference on Data Mining Workshop. IEEE, 2015: 105-110.

[4] WAJEED M A, ADILAKSHMI T. Supervised and semi-supervised learning in text classification using enhanced KNN algorithm: A comparative study of supervised and semi-supervised classification in text categorization[J]. International Journal of Intelligent Systems Technologies & Applications, 2012, 11(3/4): 179-195.

[5] RISTIN M, GUILLAUMIN M, GALL J, et al. Incremental learning of random forests for Large-Scale image classification[J]. IEEE Trans Pattern Anal Mach Intell, 2016, 38(3): 490-503.

[6] BLEI D, NG A, JORDAN M. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[7] JARADAT S, DOKOOHAKI N, MATSKIN M. OLLDA: A supervised and dynamic topic mining framework in twitter[C]//2015 IEEE 15th International Conference on Data Mining Workshop. IEEE, 2016: 1354-1359.

[8] 刘少鹏, 印鉴, 欧阳佳, 等. 基于 MB-HDP 模型的微博主题挖掘[J]. 计算机学报, 2015, 38(7): 1408-1419.

[9] DUPUY C, BACH F, DIOT C. Qualitative and descriptive topic extraction from movie reviews using LDA[C]//Machine Learning and Data Mining in Pattern Recognition. Springer, 2017: 91-106.

[10] MA J, YAO Z, SUN M. WSO-LDA: An online "Sentiment+Topic" weibo topic mining algorithm[C/OL]//Pacific Asia Conference on Information Systems. [2018-07-01].http://aisel.aisnet.org/pacis2017/223.

[11] 刘冰玉, 王翠荣, 王聪, 等. 基于动态主题模型融合多维数据的微博社区发现算法[J]. 软件学报, 2017, 28(2): 246-261.

[12] KHOLGHI M, SITBON L, ZUCCON G, et al. External knowledge and query strategies in active learning: A study in clinical information extraction[C]//24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 143-152.

[13] 陈德华, 殷苏娜, 乐嘉锦, 等. 一种面向临床领域时序知识图谱的链接预测模型[J]. 计算机研究与发展, 2017, 54(12): 2687-2697.

[14] ORAMAS S, ESPINOSA-ANKE L, SORDO M, et al. Information extraction for knowledge base construction in the music domain[J]. Data & Knowledge Engineering, 2016, 106: 70-83.

[15] VELASCO-ELIZONDO P, MARÍN-PIÑA R, VAZQUEZ-REYES S, et al. Knowledge representation and information extraction for analysing architectural patterns[J]. Science of Computer Programming, 2016, 121: 176-189.

[16] DIETZ L, KOTOV A, MEIJ E. Utilizing knowledge graphs in text-centric information retrieval[C]//Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017: 815-816.

[17] 高俊平, 张晖, 赵旭剑, 等. 面向维基百科的领域知识演化关系抽取[J]. 计算机学报, 2016, 39(10): 2088-2101.

[18] MARIN A, HOLENSTEIN R, SARIKAYA R, et al. Learning phrase patterns for text classification using a knowledge graph and unlabeled data[J]. ISCA-International Speech Communication Association, 2014 (15): 253-257.

[19] KLIEGR T, ZAMAZAL O. LHD 2.0: A text mining approach to typing entities in knowledge graphs[J]. Web Semantics Science Services & Agents on the World Wide Web, 2016, 39: 47-61.

[20] SHI W, ZHENG W, YU J X, et al. Keyphrase extraction using knowledge graphs[J]. Data Science & Engineering, 2017, 2(4): 275-288.

[21] CHEN Z, LIU B. Mining topics in documents: Standing on the shoulders of big data[C]//20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 1116-1125.

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

[23] LU Y, MEI Q, ZHAI C. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA[J]. Information Retrieval, 2011, 14(2): 178-203.

[24] 北京字节跳动科技有限公司. 今日头条媒体平台[EB/OL]. [2017-12-31]. https://www.toutiao.com/.

[25] KNUTH D E, MORRIS J H, PRATT V R, et al. Fast pattern matching in strings[J]. SIAM Journal on Computing, 1977, 6(2): 323-350.

(责任编辑: 李万会)