

基于主题模型的法院文本典型案例推荐

吕 宾, 侯伟亮

(北京华宇信息技术有限公司, 北京 100084)

摘 要: 考虑到案件不能量化、业务法官经验局限以及相似案件的认定缺乏统一的标准和尺度等因素, 提出一种基于主题模型的法院文本典型案例推荐模型, 基于本模型从相似度最大的文本中利用正则和加权的方法根据法院文本中判决年限和对应刑法给出推荐. 同时对案件进行深度挖掘和分析, 将相似度最高的案件定为最优推荐案件推荐给用户. 实验结果表明, 提出的基于主题模型的文本相似度计算方法跟传统的方法相比能够取得更好的 F 值.

关键词: 主题模型 LDA; 多粒度文本特征提取; 文本相似度; 法院判决文本

中图分类号: **文献标识码:** A **文章编号:** 1000—7180(2018)02—0128—05

Typical Case Recommendation of Court Texts Based on Topic Model

LV Bin, HOU Wei-liang

(Beijing Huayu Information Technologys Co., Ltd, Beijing 100084, China)

Abstract: Taking into account that the case cannot be quantified, the limited experience of business judge, the lack of unified standards for the identification of similar cases and other factors, a typical model of court texts based on topic model is proposed in this paper. The recommendation is given by regular and weighted method according to the number of years of judgment and corresponding criminal law from the maximum similarity text in court texts based in this model. While the highest similarity of the case as the best recommendation case to the user. Experimental results show that the method based on topic model proposed in this paper is superior to the traditional text similarity measure in F-metrics.

Key words: topic model LDA; multi-granularity text feature extraction; text similarity; court texts

1 引言

伴随着时下互联网技术的快速发展, 现在的社会信息化程度越来越高, 信息量也正以前所未有的速度增长. 在管理和组织海量信息方面, 文本相似性分析作为基础性的工作在文本挖掘、信息检索、敏感信息过滤机器翻译等领域有着较为广泛的应用, 文本相似性分析结果的优劣直接影响到文本分类和聚类的结果.

作为文本分析应用的一种, 法院文本案例相似性分析受到越来越多的关注. 目前在法院审判过程中, 对审判案件中相似事实的认定法院内部没有形成统一的标准, 而法律又赋予了法官在法律许可的

范围内具有一定的自由裁量权, 这样当法官受不同知识框架、认知背景、经验和价值评价等因素的影响时, 容易造成同案不同判的情况. 当前社会案件数量持续增长, 复杂的新类型案件逐渐增多, 加上新时期社会公众的法治意识不断增强, 解决公众纠纷的渠道也趋于多元化, 但是往往因为不了解和没有比较, 绝大多数解决纠纷所采取的方式是法院诉讼. 从全局看, 这在一定程度上造成了社会经济资源的浪费, 包括当事人的时间成本的花费经济效益的受损, 法院诉讼资源的占用, 以及其他纠纷渠道资源的闲置等. 因此研究典型案例推荐模型, 可以为法官、当事人、律师等提供经典案例推荐, 在案例推荐模型中基于裁判文书和电子卷宗等文书信息, 从中自动提取

与本案案情有关的案情要素内容,并通过自然语言理解和分析算法,对采集所有案件的信息进行全方位分析,根据案情、案件量刑进行综合分析挖掘,智能识别相似案件。同时对案件进行深度挖掘分析将相似度最高的案件定为最优推荐案件推荐给用户。用户针对最优推荐案件与当前案件的匹配度情况进行评价。

综上,本文提出一种基于主题模型的法院文本典型案例推荐算法,从相似度最大的文本中利用基于正则和加权的方法根据法院文本中判决年限和对应刑法给出推荐。同时对案件进行深度挖掘分析将相似度最高的案件定为最优推荐案件推荐给用户。

2 相关工作

近些年随着数据挖掘技术的快速发展和日益成熟,以及大量的结构化、非结构化和半结构化数据的积累,许多研究者在文本挖掘领域做了大量研究,尤其是在文本相似性分析方面做了不少工作。刘群、李素建等人通过研究 HowNet 的多维知识表示模型提出构建一种相似语义树,词项之间的语义相似度可以通过计算树节点之间的距离得到^[1]。詹志建等人针对已有的一些基于语义和统计方法上的不足,在基于词项语义计算和语言网络的基础上,提出一种新的文本相似度计算方法,该方法跟传统的 TF-IDF 相比能得到更好的 F-度量值^[2]。张永安等人研究了一种基于历史案例本体知识库对判决结果进行推理的算法,通过基于案例特征将决策树、神经网络等算法相结合训练推力模型,预测目标案例的判决结果^[3]。Blei D M 等提出的 LDA 模型目前被广泛应用到实际项目中,并且取得了不错的效果,该模型不仅能有效降低数据维度,而且还能够识别出文本隐藏的主题信息^[4],不过在标准 LDA 模型中,每个特征词被认为是同等重要的,这种假设用在法院文本上是不适用的,因为在法院文本中,最能表示其特征的词项主要来源于摘要、时间、地点、标题、判决年限、对应刑法等要素中的词集,显然,不同的特征词需要对应不同的权重,结合某些加权方法来调整不同特征词的权重能有效提高文本相似性分析的结果。Si X 等已经通过实验证明了对于表示文本的特征词,其粒度不一样,对文本的描述能力不同^[5]。Wilson 等通过实验证明了基于 LDA 模型合理的加权策略会使高频词汇自然分配,而不会被强制分配或随机分散到各主题中^[6],文献^[7]中 Iwata T 等人指出如果仅使用基于词项来表示文本会造成一定的

稀疏性,而如果仅使用基于 LDA 模型得到的特征词会因为主题比较泛化,也就是粒度较粗而无法对文本进行细节上的表示。

综合以上对文本特征表示以及文本相似性计算相关文献方法的分析,本文将加权 LDA 模型结合多粒度特征来表示文本,然后使用语义分析结合词频统计,通过计算重要特征词的语义相似度来表示文本相似度。

3 模型构建

3.1 文本预处理

对法院文本进行特征提取、文本相似分析以及最后的推荐之前,需要对文本进行预处理,在这个过程中首先需要对文本进行分词,形成文本对应的词序列。

然后过滤掉文本中携带信息量较少的词,包括停用词、虚词、连词、介词等没有特征意义的词,降低文本的特征维数,紧接着通过全文检索引擎 Lucene 建立索引并入库,通过 Lucene 框架进行全文检索得到跟待分析法院文本相关的 Top-k 文本。

3.2 基于 LDA 主题模型的特征词权重调整和多粒度文本特征选择

在标注 LDA 模型中,特征词 w_i 的主题 z_i 的后验分布一般是使用 Gibbs Sampling 采样方法来估计,最后得到的参数采样公式如(1)所示。

$$p(z_i = k | z_{-i}, w, \alpha, \beta) \propto$$
$$p(z_i = k, w_i = t | z_{-i}, w_{-i}, \alpha, \beta) \propto$$
$$\frac{n_{k, \rightarrow i}^{<t>} + \beta_t}{\sum_{i=1}^V (n_{k, \rightarrow i}^{<t>} + \beta_t)} \cdot \frac{n_{m, \rightarrow i}^{<k>} + \alpha_k}{\sum_{k=1}^K (n_{m, \rightarrow i}^{<k>} + \alpha_k)}$$

(1)

式中, α_k 和 β_t 分别代表了主题 k 和特征词 t 的狄利克雷先验分布。最终 θ 和 φ 的后验估计值分别如公式(2)和公式(3)所示:

$$\theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_{k=1}^K (n_m^k + \alpha_k)}$$

(2)

$$\varphi_{k,t} = \frac{n_k^t + \beta_t}{\sum_{t=1}^V (n_k^t + \beta_t)}$$

(3)

式中, $\theta_{m,k}$ 表示文本 m 中主题 k 出现的概率; $\varphi_{k,t}$ 表示主题 k 中特征词 t 的概率;它们分别表示对应的两个狄利克雷分布在贝叶斯框架下的参数估计值。

上述的 LDA 模型没有考虑每个特征词的权重,这种假设对于法院文本是不适合的,显然不同词性、不同位置的特征词对法院文本的表示能力是不一样的,因为在法院文本中,最能表示其特征

的词项主要是来源于摘要、时间、地点、标题、判决年限、对应刑法等要素中的词集,将合适的加权方法代入到基于 LDA 模型生成特征词的过程中,即不同的特征信息对应特征词有不同的权重,这样获得的特征词更具有表意性,从而提高文本相似性分析的结果。

互信息是文本特征选择中常用的一种方法,其时间复杂度较低,不过在互信息中,一些小众词或者低频词容易被选为特征来表示文本,显然这是不合适的,因此需要降低文本中这些特征词出现的概率,结合互信息,在本文的方法中对上述标准 LDA 模型做出一些调整和改进,针对那些出现频次较高但是表示文本能力不足的词,降低其生成过程中的概率分布,同理对出现频次较低不过具有较大表示文本特征意义的词,提高其生成过程中的概率分布。

特征词的生成如式(4)所示。

$$p(w|\theta, \varphi) = \frac{\text{weight}(t, d) \cdot \varphi_{k,t}}{\sum_{t=d} \text{weight}(t, d)} \quad (4)$$

不同于标准 LDA 模型中的特征词生成过程,本文对其作出一些调整,将权重引入到采样过程中,得到改变后的 Gibbs Sampling 采样如式(5)所示。

$$\begin{aligned} p(z_i = k | z_{-i}, w, \alpha, \beta) & \propto \\ p(z_i = k, w_i = t | z_{-i}, w_{-i}, \alpha, \beta) & \propto \\ \frac{\text{weight}(t, d) n_{k, \rightarrow i}^{(t)} + \beta_t}{\sum_{t=1}^V (\text{weight}(t, d) n_{k, \rightarrow i}^{(t)} + \beta_t)} \cdot & \\ \frac{\text{weight}(t, d) n_{m, \rightarrow i}^{(k)} + \alpha_k}{\sum_{k=1}^K (\text{weight}(t, d) n_{m, \rightarrow i}^{(k)} + \alpha_k)} & \end{aligned} \quad (5)$$

结合互信息,其中特征词 w_i 跟文本 d 的点互信息计算如式(6)所示。

$$PMI(w_i, d) = \log_2 \frac{p(w_i, d)}{p(w_i)p(d)} \quad (6)$$

在点互信息中调整每个特征词的权重,并将其扩展到 LDA 模型中,调整后的特征词 w_i 跟文本 d 的点互信息计算如式(7)所示。

$$\text{weight}(t = w_i, d) = \sum_{j=1}^l \lambda \cdot \mu_j \cdot PMI(w_i)_j \quad (7)$$

其中 μ_j 表示权重,并且 $\mu_1 > \mu_2 > \dots > \mu_l, \sum_{j=1}^l \mu_j = 1, \lambda$ 为调节因子。

对于文本 d_i ,通过上述调整后的 LDA 模型生成其特征词,得到特征向量 $d_i = (t_{i1}, t_{i2}, \dots, t_{im})$,考虑到仅使用 LDA 模型得到的特征词主题比较泛

化,在文本细节上的表示能力不足,因此本文引入了一些能够有效描述文本细节方面的特征词.首先计算文本中每个特征词的 TF-IDF 值,对其进行排序,取一定阈值 δ 以上的词来表示文本,得到特征向量 $d_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{in})$,结合上述两种类型的特征,最终可以得到文本的特征向量为

$$d_i = (t_{i1}, t_{i2}, t_{i3}, \dots, t_{im}, w_{i1}, w_{i2}, w_{i3}, \dots, w_{in})$$

这里尽量保证每个法院文本特征向量具有相同的维度,其中如果同一个特征词被 w_m 与 t_{im} 同时表示,在本文的方法中会视为是不同的特征词。

3.3 基于语义分析结合词频统计的文本相似度计算

本文中使用文本特征向量中词项间的相似度来度量文本间的相似度计算.在此之前,通过改进的 HowNet 进行词语语义相似度计算,如式(8)所示。

$$\begin{aligned} \text{sim}(S_1, S_2) &= \alpha \times (\text{depth}(S_1) + \text{depth}(S_2)) / \\ &(\alpha \times (\text{depth}(S_1) + \text{depth}(S_2)) + \text{dist}(S_1, S_2)) \end{aligned} \quad (8)$$

设 v_i, v_j 分别表示两篇文本对应的特征词项向量, $v_i = (w_{i1}, w_{i2}, \dots, w_{in})$, $v_j = (w_{j1}, w_{j2}, \dots, w_{jn})$,由此定义文本相似度如式(9)所示。

$$\begin{aligned} \text{textSim}(v_i, v_j) &= wf \times \text{vecSim}(v_i, v_j) + \\ &(1 - wf) \cosSim(v_i, v_j) \end{aligned} \quad (9)$$

式中, $\text{vecSim}(v_i, v_j)$ 表示特征词项向量 v_i 与 v_j 之间的语义相似度,如式(12)所示。

并且特征词项向量 v_i 与 v_j 间的余弦相似度计算如式(10)所示(其中 λ 指的是 v_i 和 v_j 中出现的相同词项的个数)。

$$\begin{aligned} \cosSim(v_i, v_j) &= \\ \frac{\sum_{k=1}^{\lambda} TFIDF(w_{ik}) \times TFIDF(w_{jk}) /}{\sqrt{\sum_{k=1}^m (TFIDF(w_{ik}))^2 \times \sum_{l=1}^n (TFIDF(w_{jl}))^2}} & \end{aligned} \quad (10)$$

值得注意的是对于两篇文本而言,一般会存在一些语义相似度比较高的词项,这些词项相比其他剩余词项会拥有更高的 TF-IDF 值,通过计算它们之间的语义上的相似度能更好的反映文本的相似度情况,不过对于其他剩余的词项通过计算其在整个文本中的概率分布来反映文本相似度,因此在式(9)中需要对 $\text{vecSim}(v_i, v_j)$ 进行加权,其中加权因子 wf 的计算是通过关键词向量中满足相似度阈值条件的关键词的 TF-IDF 值在整个文本 TF-IDF 总和中所占的比例得到,如式(11)所示:

$$\omega f = \frac{1}{2} \left[\frac{\sum_{k \in \Lambda_i} TFIDF(\omega_{ik})}{\sum_{k=1}^m TFIDF(\omega_{ik})} + \frac{\sum_{l \in \Lambda_j} TFIDF(\omega_{jl})}{\sum_{l=1}^n TFIDF(\omega_{jl})} \right] \quad (11)$$

式中, $TFIDF(\omega_{ik})$ 表示特征词项 ω_{ik} 的 TF-IDF 值, 等式右边表示特征词项向量 v_i 中所有满足相似度阈值条件的特征词项 ω_{ik} ($k \in \Lambda_i$) 的 TF-IDF 值在 v_i 所有的词项 TF-IDF 值总和中所占的百分比. 这里定义两个集合 Λ_i 和 Λ_j , 如果特征词项向量 v_i 中的某个特征词项 ω_{ik} 与另一个特征词项向量 v_j 中的特征词项 ω_{jl} ($l = 1, 2, \dots, n$) 的相似度大于设定的阈值 μ , 则 ω_{ik} 属于 Λ_i . 同理对 v_j 中的特征词项进行选择得到 Λ_j .

$$\text{vecSim}(v_i, v_j) = \frac{1}{2} \left[\frac{1}{m} \sum_{k=1}^m \max_{1 \leq l \leq n} \{ \text{sim}(\omega_{ik}, \omega_{jl}) \} + \frac{1}{n} \sum_{l=1}^n \max_{1 \leq k \leq m} \{ \text{sim}(\omega_{jl}, \omega_{ik}) \} \right] \quad (12)$$

式中, $\text{sim}(\omega_{jl}, \omega_{ik})$ 表示特征词项 ω_{ik}, ω_{jl} 之间的语义相似度, 由式(8) 计算得到 $\text{vecSim}(v_i, v_j)$ 由特征词项向量 v_i, v_j 中所包含的词项语义相似度决定.

本文模型算法如下所示.

输入: 文本数据集和待分析文本

输出: 最优推荐案例文本

将文本进行预处理.

通过加权 LDA 模型和多粒度文本特征选择得到文本对应特征向量 v_i .

从向量 v_i 中的词项 ω_{i1} 开始, 利用式(8) 找到向量 v_j 中与 ω_{i1} 最相似的词项 ω_{jk} , 然后累加 v_i 中每个项的相似度, 除以向量 v_i 的维度, 以此来作为向量 v_i 和 v_j 的相似度 $\text{sim}(v_i, v_j)$. 重复步骤 1, 2 的过程, 得到向量 v_j 和 v_i 的相似度 $\text{sim}(v_j, v_i)$.

取 $\text{sim}(v_i, v_j)$ 和 $\text{sim}(v_j, v_i)$ 的算术平均值来表示向量 v_i 和 v_j 的语义相似度 $\text{vecSim}(v_i, v_j)$.

分别对向量 v_i 和 v_j 中的词项计算其 TF-IDF 权值, 并且 v_i 和 v_j 之间的余弦相似度通过公式(10) 计算得到.

经过上面的步骤得到 Λ_i 和 Λ_j 中的元素, 由此加权因子 ωf 通过公式(11) 计算得到.

综上, 利用式(9) 计算最终得到向量 v_i 和 v_j 之间的文本相似度.

从相似度最大的文本中利用基于正则和加权的方法根据法院文本中判决年限和对应刑法给出推荐.

4 实验

4.1 基于文本分类语料库数据集

本文实验数据部分先选用了其大学自然语言处理小组收集与整理的部分文本分类语料库. 该部分语料库分 10 个类别, 共 2 815 篇. 每个类别中根据文本内容又细分为不同子类别, 该语料库中各个类别的子类别可以作为各个文本集的标准聚类

结果^[8].

在基于词频统计与语义分析相结合相似度计算中, 首先需要知道词语相似度阈值 μ 的大小会对聚类结果产生的影响以及其具体表现分布, 在经过加权 LDA 模型和多粒度文本特征选择后, 本文通过利用经典的 DKM 聚类算法来进行聚类实验, 结果如图 1 所示, 从图中可以看出聚类的结果跟同一个聚类中的特征词项对应相似度阈值 μ 有关. 根据实验结果, 本文取词语相似度阈值为 0.7.

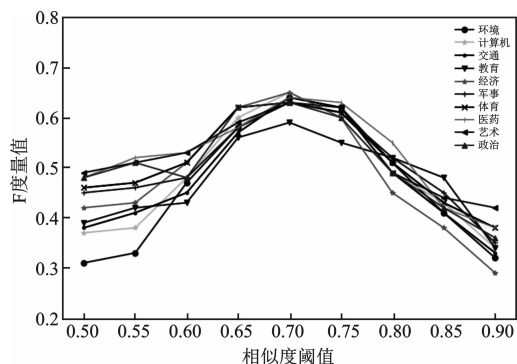


图 1 特征词相似度阈值 μ 对聚类结果的影响

将本文的算法跟传统 TF-IDF 算法和文献^[8]里的 TSemSim 算法分别应用在 DKM 聚类实验中, 得到如图 2 所示的实验结果, 从图 2 实验结果可以得到本文算法相比传统 TF-IDF 算法以及 TSemSim 算法有着更好的 F-度量值的结论.

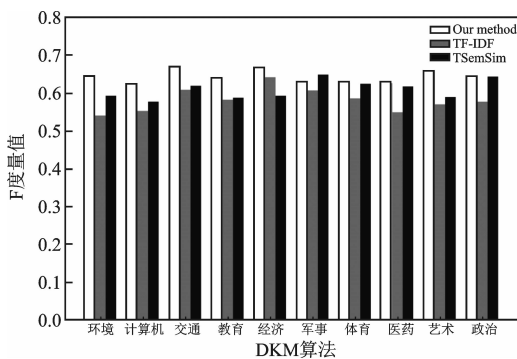


图 2 聚类结果在 F-度量值上比较

4.2 基于实际法院文本数据集

将本文算法应用到实际的法院文本数据中, 同样与传统 TF-IDF 算法以及 TSemSim 算法在通过经典的 DKM 聚类算法实验进行比较, 得到的结果如图 3 所示, 其中数据来源于法院判决文本刑事案件, 有交通事故、聚众斗殴、非法盗窃、合同诈骗 4 类案件, 共计 4 000 余份. 分析图 3 可以得到在应用于法院判决文本上本文算法比传统 TF-IDF 算法和

TSemSim 算法有更好的 F 值.

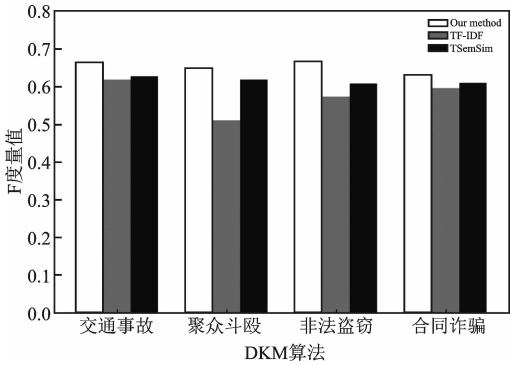


图 3 应用于法院判决文本上的
聚类结果在 F-度量值上比较

5 结束语

本文提出一种基于主题模型的法院文本典型案例推荐模型,在模型中不仅能有效降低文本模型的维度,还考虑到了文本隐含语义信息以及采用多粒度特征来表示文本,有效避免粒度太细造成的稀疏性.最后从相似度最大的文本中利用基于正则和加权的方法,根据法院文本中判决年限和对应刑法给出推荐.实验结果表明,本文提出的基于主题模型的文本相似度计算方法跟传统的方法相比能够取得更好的 F 值.

参考文献:

[1] 李素建. 基于语义计算的语句相关度研究[J]. 计算机工程与应用, 2002, 38(7): 75-76.

[2] 詹志建, 杨小平. 基于语言网络和语义信息的文本相似度计算[J]. 计算机工程与应用, 2014 (5): 33-38.

[3] 张永安. 审判案例自动抽取与标注模型研究[J]. 现代图书情报技术, 2013 (6): 23-29.

[4] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allo-

cation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[5] Si X, Liu Z, Li P, et al. Content-based and Graph-based Tag Suggestion [C]//DC @ PKDD/ECML. 2009.

[6] Wilson A T, Chew P A. Term weighting schemes for latent dirichlet allocation[C]//human language technologies: The 2010 annual conference of the North American Chapter of Association for Computational Linguistics. Association for Computational Linguistics, 2010: 465-473.

[7] Iwata T, Yamada T, Ueda N. Modeling social annotation data with content relevance using a topic model [C]//Advances in Neural Information Processing Systems. 2009: 835-843.

[8] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5): 856-864.

[9] Kashyap A, Han L, Yus R, et al. Robust semantic text similarity using LSA, machine learning, and linguistic resources[J]. Language Resources and Evaluation, 2016, 50(1): 125-161.

[10] Wang F, Xu T, Tang T, et al. Bilevel feature extraction-based text mining for fault diagnosis of railway systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(1): 49-58.

作者简介:

吕 斌 男,(1973-),硕士,副高级工程师,研究方向为信息法律知识认知技术和法律知识图谱构建.
E-mail: lb@thunisoft.com.

侯伟亮 男,(1984-),学士,研究方向为以大数据为中心的多元诉讼服务技术和产品开发.

(上接第 127 页)

[6] Song J, Wu Y, Xu Z, et al. Research on car-following model based on SUMO[C]// IEEE, InternationalConference on Advanced Infocomm Technology. Hangzhou, China, IEEE, 2014:47-55.

作者简介:

蒋华涛 男,(1991-),硕士研究生.研究方向为汽车电子和车联网.

常 琳 女,(1983-),博士,副研究员.研究方向为汽车电子和车联网.

娄玉强 男,(1991-),硕士研究生.研究方向为机器学习与图像处理.

李 庆(通讯作者) 男,(1972-),博士,研究员.研究方向为汽车电子和车联网. E-mail:liqing@ime.ac.cn.

陈大鹏 男,(1968-),博士,研究员.研究方向为汽车电子和车联网.