基于关键短语和 LDA 模型的军事舆情热点话题发现及推荐方法

葛斌　何春辉 胡升泽 张翀

（国防科技大学科学信息系统工程重点实验室）

摘要：热点话题的发现和推荐是军事舆情领域的热点研究方向，但现有技术尚不成熟。在传统主题模型 LDA 基础上，使用关键短语作为特征项来构造“短语袋子”模型，提出了基于关键短语和 LDA 模型（KPLDA）的军事舆情热点话题发现及推荐方法。为验证方法性能，选取了三个军事舆情子领域的数据集进行实验验证，并根据训练模型耗时情况、困惑度、发现的热点话题质量这三个指标来评估方法的性能。实验结果表明，新方法可以准确发现热点话题并能给相关用户做出有效推荐。

关键词：热点话题发现；推荐；关键短语抽取；KPLDA；聚类

Military hot topic discovery and recommendation method based on key phrases and LDA model

GE Bin, HE Chun-hui, HU Sheng-ze, ZHANG Chong

(Science and Technology on Information Systems Engineering Laboratory , National University of Defense Technology,Changsha, Hunan, P.R. China, 410073)

Abstract: The discovery and recommendation of hot topics is a hot research direction in the field of military public opinion, But the existing technology also is not mature. Based on the traditional topic model LDA, the key phrase is used as the features to construct the "bag-of-phrase" model, proposed the hot topic discovery and recommendation method based on key phrase and LDA model (KPLDA). In order to evaluate the proposed method, Three data sets of military sub-areas were selected for experimental, and adopting: (1) time consumption of the training model, and (2) Perplexity value, and (3) quality of the discovery hot topic, three indexes to evaluate the performance of the method. The results show that the new method it can accurately discovery hot topics and make effective recommendations to relevant users.

Keywords: Hot topic discovery; Recommendation; Key phrase extract; KPLDA; Clustering

1 引言

随着互联网和媒体技术的迅速发展，大量与军事相关信息被发布在互联网上。网络军事舆情信息传播速度快，信息量大，已逐渐成为相关用户获取该类信息的主要途径。为了帮助相关用户快速获得有价值的军事舆情信息，热点话题发现技术已成为该领域的研究热点。

在传统的文本挖掘中，向量空间模型（VSM）[1]通常被用于文档表示。在此基础上，研究人员提出了基于词汇共现的潜在语义索引（LSI）[2, 3]方法来寻找词语之间的语义关系。潜在 Dirichlet 分配（LDA）[4]模型是一种无监督概率生成模型，使用多个潜在主题的概率分布来表示文档特征。

LDA 模型中的参数推理[5-6]通常采用 VEM 和 Gibbs 抽样方法等。Zou[7]提出了一种局部一致的潜在狄利克雷分配(LC-LDA)模型，通过使用词袋特征来学习参数空间的分布情况。Adams [8]分析了一种向量空间模型，将文档和查询表示成特征向量形式，通过计算向量之间的相似性来获取与会话内容相对应的信息。Lane[9]结合了多类主题模型将不同的语料自动归类到相应的主题。Blei[10]将时间序列信息引入 LDA 模型，实现了一种复杂的动态 LDA 模型（DLDA）。Zhao[11]利用 LDA 模型分析了 Twitter 上的数据集，并将结果与传统新闻媒体的内

容做了对比。 Lu[12]采用自适应 LDA 模型完成了文档的主题提取任务。综上所述，虽然 LDA 主题模型在大规模文本语料中的潜在主题发现上已经取得了一些成果，但它还存在一些不足之处，例如普遍存在发现主题的准确性低和可读性差等问题。为了改善这些问题，Huang[13]初步讨论了 LDA 模型发现主题的词分配效应。为了在军事舆情领域进一步解决以上问题，提出了基于关键短语和 LDA 模型的军事舆情热点话题发现及推荐方法。该方法具有以下特点：（1）在特征选取方面，使用关键短语代替独立词作为文档的特征；（2）因为传统的 LDA 模型发现话题的准确性和可读性太低，因此从语料库中抽取关键短语作为特征来训练 LDA 模型从而得到主题-短语分布，并通过热点话题发现和推荐方法来发现和推荐高质量的热点话题；（3）在话题聚类过程中，采用了字符串之间最长公共子序列的值作为相似度的距离度量。

2 TF-IDF 体重计算和关键词抽取

为了分析文档，首先我们通过分词和预处理来获得候选关键词。然后，使用 TF-IDF 算法计算每个候选词的权重。接下来，我们采用单词共现关系来构造节点和边的拓扑图。然后使用 TextRank 算法计算候选词的分数和相应的得分降序排列结果。最后根据候选单词的得分降序排列结果选择排名靠前的 N 个单词作为候选关键词，并将原文档中具有共现关系的候选关键词组装成关键短语。这样，提取的关键短语列表可以用来构造原始文档的向量矩阵作为 LDA 模型的特征。从文档中提取关键短语的流程图如下所示：

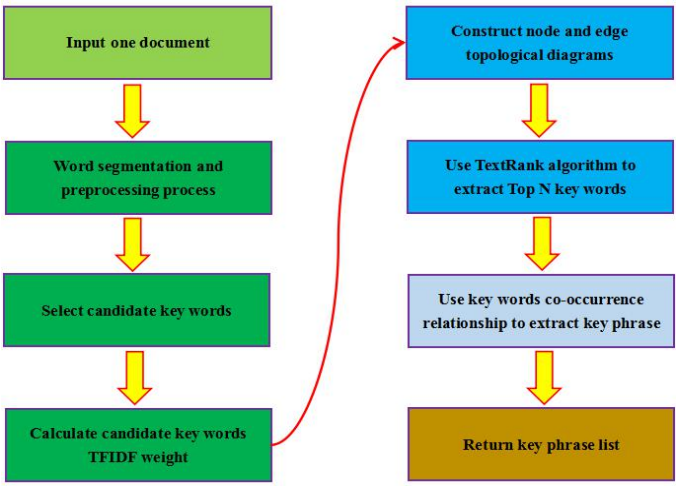

图 1：从文档中提取关键短语的流程图

2.1 TF-IDF 权重计算

TF-IDF [14]是一种统计方法，主要用来评估一个词语在语料库中某个文档的重要性。词语频率是指给定单词在文件中出现的次数。这个数字通常是经过归一化处理的。对于特定文档中的每个单词，TF 权重定义如下：

$$TF_{\mathrm{tj}} = \frac{n_{tj}}{\sum_{k} n_{kj}} \tag{1}$$

在公式（1）中，n_{tj}是词条 t 在文档 j 中出现的总次数，$\sum_{\mathrm{k}} n_{kj}$ 是出现在文档 j 中的所有词条 k 的总数。

逆文档频率（IDF）的主要思想是含有 t 的文档数越少，IDF 的值越高，这表明这些词具有良好的区分度。 一个特定单词的 IDF 值可以用语料库中所有文档的数量除以包含该词语 t 的文档总数来计算。 IDF 的定义如下：

$$IDF_{\mathrm{t}} = \log(\frac{D}{\sum D_t + 1}) \tag{2}$$

在公式（2）中，D 是语料库中的文档总数量，D_{t} 是出现词条 t 的文档总数。词条 t 的 TF 值和 IDF 值做乘积可以产生 TF-IDF 权重。利用 TF-IDF 权重，可以有效地过滤掉常用词并保留一些重要的词语。 因此，TF-IDF 的定义如下：

$$TF\text{-}IDF = TF * IDF \tag{3}$$

2.2 关键短语抽取

TextRank 是一种常用的关键词提取方法。它的计算公式如公式（4）[15]：

$$WS(V_{\mathrm{i}}) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_{\mathrm{k}} \in Out(V_j)} w_{jk}} WS(V_j) \tag{4}$$

这里 WS(V_j)是节点 V_{i} 的得分。常数 d 是阻尼因子，在上述方法中该值一般设置为 0.85。W_{ji} 是前一节点 V_{j} 到当前节点 V_{i} 之间边上的权重。$In(V_i)$ 是指向它的所有节点的集合(入链集合)。$Out(V_j)$ 是节点 V_{j} 指向的所有节点的集合（出链集合）。$\sum_{V_{\mathrm{k}} \in Out(V_j)} w_{jk}$ 是前一节点 V_{j} 中所有边的权重总和。 w_{ji} 为 V_{j} 和 V_{i} 在相关文档中指定窗口里出现的词频。

3 基于关键短语 LDA 模型

LDA [16]是一个三层贝叶斯模型，分别是文档层，话题层和关键短语层。该模型基于以下假设：

（1）所有文档中存在 k 个独立主题；

（2）每个话题与所有关键短语之间都服从多项式分布；

（3）每个文档与所有话题之间都服从多项式分布；

（4）每篇文档的先验分布服从 Dirichlet 分布；

（5）每个话题的先验概率分布服从 Dirichlet 分布。

生成一篇文档的过程如下：

（a）对文档集合 M，短语分布参数Ø是通过话题 Dirichlet 分布参数 β 采样生成。

（b）对来自 M 的一个文档 m，话题分布参数 θ 是通过文档 Dirichlet 分布参数α采样生成。

（c）对文献 m 中的短语 N，W_{mn}，首先根据 θ 分布提取文档 m 的隐藏话题 Z_{m}，然后根据Ø分布对话题 Z_{m} 中的关键短语 W_{mn} 进行抽样。

因此，模型的联合概率分布定义如下：

$$P(z,w,\theta,\phi \mid \alpha,\beta) = P(w \mid \phi,z) \cdot P(z \mid \theta) \cdot P(\theta) \cdot P(\phi) \tag{5}$$

移除一些隐藏变量后，联合分布的积分形式为：

$$P(z,w|\alpha,\beta)=\iint P(z,w,\theta,\phi|\alpha,\beta)d_\theta d_\phi$$
$$=\int P(w|\phi,z)\cdot P(\phi)d_\phi\int P(z|\theta)\cdot P(\theta)d_\theta \quad (6)$$
$$=\prod_{k=1}^{K}\frac{\Delta(n_k+\beta)}{\Delta(\beta)}\cdot\prod_{m=1}^{M}\frac{\Delta(n_m+\alpha)}{\Delta(\alpha)}$$

间接计算的转移概率可以消除中间参数 θ 和Ø，这样我们可以对每一轮迭代使用 Gibbs 采样，迭代过程：首先产生一个均匀的随机数，然后计算每个主题上转移概率，然后迭代更新参数矩阵，直到迭代收敛才结束参数更新。

关键短语 KPLDA 模型与传统的 LDA 模型之阿金存在一些差别。传统 LDA 模型的特征通常时是由单个词语组成。 但是 KPLDA 模型特征是通过每个文档中提取的关键短语构成。在 KPLDA 模型中，我们设定的参数α= 50 / K， β = 0.1。

最后，根据 KPLDA 模型，就可以对训练语料库生成文档-话题分布和话题关键短语分布。

KPLDA 模型的三层结构图如图 2 所示。

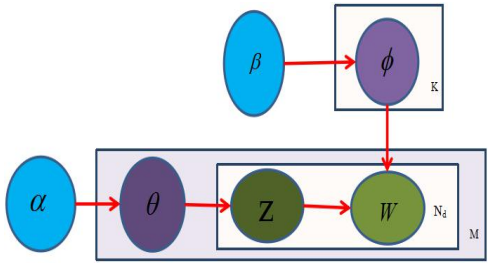

图 2：KPLDA 模型的参数结构图

图 2 中相应变量的定义如表 1 所示：

表 1：LDA 模型中相关变量的定义

Variable	Definition of Variables
M	Number of documents in corpus
K	Number of topics in all documents
W	Phrases Vocabulary
N_d	Number of phrases in one document
Z	Topic for a phrase
α	Parameter of the Dirichlet prior on θ
β	Parameter of the Dirichlet prior on Ø
θ	document topic probability distribution
Ø	topic phrase probability distribution

4 热点话题发现和推荐方法

根据第 3 节的方法，使用 KPLDA 模型训练语料库，可以得到话题与关键短语之间的分布。这些短语比单个词语[13]更能反映话题的含义。基于话题-关键短语分布，我们提出了一种话题发现和推荐方法。该方法分为以下几个步骤：

步骤 1：根据话题-关键短语分布，通过提取每个主题下权重最大的前 K 个关键短语来构成粗糙话题集合 T_1。

步骤 2：根据 T_1，执行集合并集和差集操作来移除重复元素，得到集合 T_2。

步骤 3：计算集合 T_2 中所有元素之间最长公共子序列值作为相似性距离 D。

步骤 4：根据相似性距离做聚类，得到话题列表为 T_3。

步骤 5：对 T_3 中每个话题的关键短语进行分词，并使用词性来过滤不包含名词或动词的话题，得到最终的热点话题列表 T_4。

步骤 6：结合用户个性化设置信息，将发现热点话题列表 T4 依次进行推荐。

在步骤 3 中，对于两个字符串 S =（s1，s2，...，si）和 T =（t1，t2，...，tj）的最长公用子序列（LCS）[17]递归计算公式定义如下：

$$LCS[\mathrm{i},j]=\begin{cases}0, if\,(i=0) or (j=0)\\ LCS[i-1,j-1]+1, if\,(i,j>0, \mathrm{s}_i=t_j)\\ \max\{LCS[i,j-1], LCS[i-1,j]\}, if\,(i,j>0, s_i\neq t_j)\end{cases} \tag{7}$$

LCS [i，j]是一个非负整数，表示 S =（s_1，s_2，...，s_i）和 T =（t_1，t_2，...，t_j）的最长公共子序列。 该方法中相似距离 D 的范围> 1。

热点话题发现和推荐方法的过程如图 3 所示：

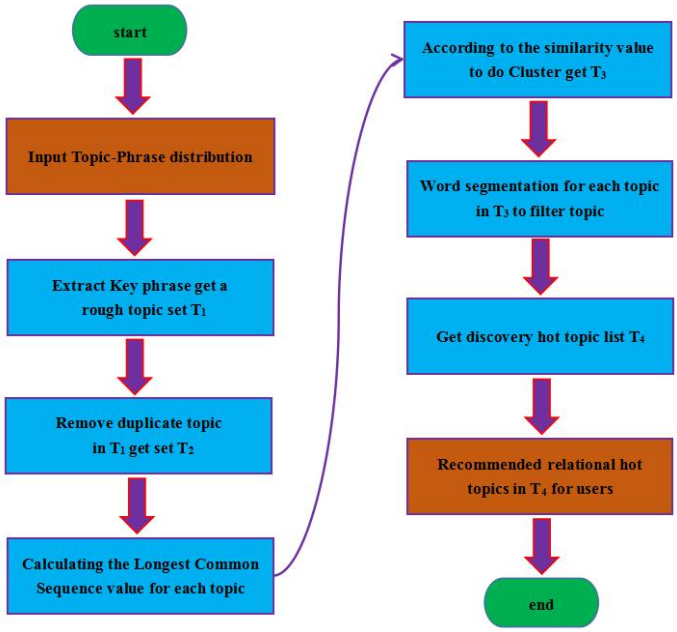

图 3：热点话题发现及推荐方法流程图

5 实验设置

5. 1 数据集介绍

为了充分验证 KPLDA 模型在军事舆情热点话题发现上的性能。 选取了三个与军事舆情领域相关的真实文本数据集来评估模型性能。数据集共包含了 1500 篇与军事舆情领域相关的文本文档，主要涉及台军、朝核试验、航母这 3 个军事舆情子领域。 表 2 列出了每个子领域中所包含的文档数量。

表 2：实验数据集分布情况

专题名称	台军	朝核试验	航母	总计
文档数量	500	500	500	1500

如上表所示，实验中所使用的是一个平衡数据集。

5.2 实验分析

为客观地验证 KPLDA 模型的性能，实验中采用了三个核心指标来评估模型的性能：（1）训练模型的耗时量；（2）困惑度；（3）发现热点话题的质量。

通过训练不同的模型并记录相应的时间消耗，然后经过对比分析，就可以有效的评估模型性能。 在上述实验中选择的话题数量在 5 和 100 之间。图 4 展示了 KPLDA 模型和 LDA 模型在不同主题数与训练模型耗时量之间的关系。

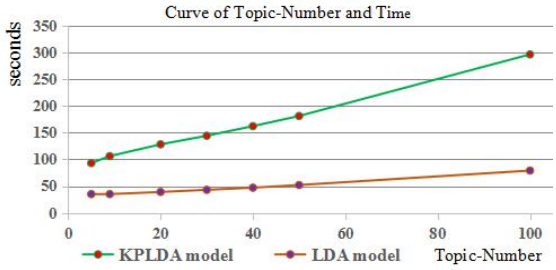

图 4：话题编号和时间消耗的曲线

根据图 4 的结果，LDA 模型训练过程相对于 KPLDA 模型具有很大的优势。 随着话题数量的增加，LDA 模型的时间消耗量变化较小，但 KPLDA 模型的训练耗时量会随着话题数量的增长而线性增长。 当话题数量达到 100 时，耗时量大约比 LDA 模型高 3.7 倍。

困惑度是评估 LDA 模型性能的常用指标。 在 LDA 模型的评估中，它最初由 Blei 引入[4]，困惑度越小，说明模型的性能越好。 为了合理的评估 KPLDA 模型在话题发现上的性能，我们将发现话题的数量设置为 5 到 100 之间，并且使用困惑度作为评估标准。困惑度计算公式如下：

$$Perplexity = \exp\left\{-\frac{\sum_{d=1}^{M}\ln(P(w_d))}{\sum_{d=1}^{M}N_d}\right\} \tag{8}$$

其中 P（w_d）是文档 d 中短语特征的生成概率。 图 5 展示了在测试数据集中 LDA 和 KPLDA 模型在不同主题数目下的平均困惑度变化情况。

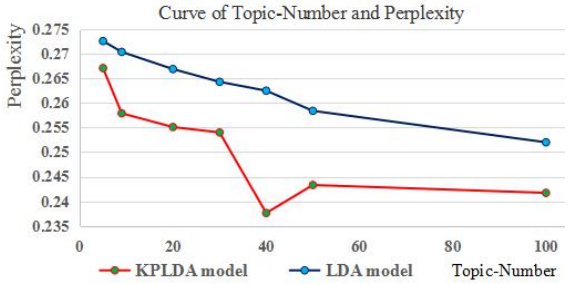

图 5：主题数量和困惑度的变化情况

实验结果表明，困惑度会随着话题数量的增加而减少，LDA 模型的困惑度减少趋势相对 KPLDA 模型而言更显平缓。实验中，当话题数量设为 100 时，困惑度会达到最小。这意味着在实验数据集上，当话题数设为 100 时，LDA 模型的性能达到最优。然而，KPLDA 模型的困惑度变化更为明显，当话题数设为 40 时，困惑度最小。意味着 KPLDA 模型在实验数据集上的最佳话题数量为 40。 KPLDA 模型的困惑度整体上均小于 LDA 模型，这充分说明 KPLDA 在热点话题发现任务中要比 LDA 模型表现更好。

为进一步验证 KPLDA 模型在热点发题发现上的性能，在三个与军事舆情领域相关的数据集上做了实验验证。在三个数据集上所发现的热点话题结果分别如图 6 至图 8 的词云所示：

图 6：KPLDA 模型在台军数据集上发现的热点话题

图 7：KPLDA 模型在朝核试验数据集上发现的热点话题

图 8：KPLDA 模型在航母数据集上发现的热点话题

从图 6 至图 8 所展示的词云效果中可以看出，KPLDA 模型在军事舆情领域的热点话题发现上表现出了较好的效果。最后，将 KPLDA 模型所发现的热点话题与用户所设定的个人偏好信息相关联，从而将发现的热点话题推荐给相关用户。

6 结论

热点话题的发现和推荐是一项有价值的研究工作。目前可用的技术还不成熟。为了改善这种现状，经过大量的实验探索，提出了基于关键短语和 LDA 模型的军事舆情热点话题发现及推荐方法(KPLDA)。通过实验比较，发现 KPLDA 模型在军事舆情领域的热点话题发现效果比传统的 LDA 模型要更优，但训练模型的耗时也会相应更长，因此在实际使用中应合理设置热点话题的数量。

参考文献

[1] Salton, G.: A vector space model for automatic indexing. *Communications of the Acm, 18*(11), 613-620 (1975)

[2] Deerwester, S.: Indexing by latent semantic analysis. Journal of the Association for Information Science & Technology, 41(6), 391-407 (1990)

[3] Chen, Y., Yin, X., Li, Z., Hu, X., & Huang, J. X.: Promoting Ranking Diversity for Biomedical Information Retrieval Based on LDA. IEEE International Conference on Bioinformatics and Biomedicine, pp. 456-461. IEEE (2012)

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. J Machine Learning Research Archive, 3, 993-1022 (2003)

[5] Minka, T.: Expectation-propagation for the generative aspect model. Proc. Conference on Uncertainty in Artificial Intelligence(UAI) (2002)

[6] Zhao, Z., Xu, W., & Chen, D.: EM-LDA model of user behavior detection for energy efficiency. IEEE International Conference on System Science and Engineering ,pp.295-300.

(2014)
[7] Zou, J., Ye, Q., Cui, Y., Wan, F., Fu, K., & Jiao, J.: Collective motion pattern inference via locally consistent latent dirichlet allocation. Neurocomputing, 184, 221-231 (2016)
[8] Adams, P. H., & Martell, C. H.: Topic Detection and Extraction in Chat. *IEEE International Conference on Semantic Computing* ,pp.581-588. IEEE Computer Society (2008)
[9] Lane, I., Kawahara, T., Matsui, T., and Nakamura, S.: Out-of-domain utterance detection using classification confidences of multiple topics, IEEE Transactions on Audio, Speech and Language Processing, vol. 15, No. 1, pp.150-161 (2007)
[10] Blei, D. M., Lafferty, J., D.: Dynamic topic models, in:Proceedings of the 23rd International Conference on Machine Learning, pp.113-120 (2006)
[11] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., & Yan, H., et al.: Comparing twitter and traditional media using topic models. European Conference on Advances in Information Retrieval(Vol.6611/2011, pp.338-349). Springer-Verlag (2011)
[12] Lu, F., Shen, B., Lin, J., & Zhang, H.: A Method of SNS Topic Models Extraction Based on Self-Adaptively LDA Modeling. Third International Conference on Intelligent System Design and Engineering Applications (Vol.37, pp.112-115). IEEE Computer Society (2013)
[13] Huang, C. M., & Wu, C. Y.: Effects of Word Assignment in LDA for News Topic Discovery. IEEE International Congress on Big Data,pp.374-380. IEEE Computer Society (2015)
[14] Zhai, C., & Lafferty, J.: A study of smoothing methods for language models applied to Ad Hoc information retrieval. International ACM SIGIR Conference on Research and Development in Information Retrieval(Vol.22, pp.334-342).ACM. (2001)
[15] Wang , Z. , Feng, Y. , & Li, F.: The improvements of text rank for domain-specific key phrase extraction. International Journal of Simulation Systems, Science & Technology,17(20), 111–115 (2016)
[16] Yeh, J. F., Tan, Y. S., & Lee, C. H.: Topic detection and tracking for conversational content by using conceptual dynamic latent dirichlet allocation. Neurocomputing, 216, 310-318.(2016)
[17] Apostolico, A., & Guerra, C.: The longest common subsequence problem revisited. Algorithmica,
2(1-4), 315-336.(1987)