基于实体关系网络的微博文本摘要

薛竹君 杨树强 束阳雪
(国防科学技术大学计算机学院 长沙 410073)

摘 要 在解析微博文本语法的基础上,结合实体关系的定义和形式化表示,提出了采用关系网络有向图模型的方法来反映文本之间的结构关系,较好地表达了文本的语义信息,弥补了词频特征刻画的不足之处。利用改进后的TPR(Topic-PAGERANK)测算各节点对应的度来表现关系元组的重要程度,按序输出关系元组对应的原博文语义字段作为摘要。最后,通过实验证明了基于关系网络的文本自动文摘方法抽取出的摘要涵盖信息更全面,冗余更少。

关键词 实体关系,短文本,文本表示,语法分析,Topic-PAGERANK
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.014

Microblog Text Summarization Based on Entity Relation Network

XUE Zhu-jun YANG Shu-qiang SHU Yang-xue
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract On the basis of syntax parsing, combining the definition of entity relationship and formalized representation, this paper put forward a method based on directed graph model to reflect the structured relationship between texts, expressing text semantic information, making up for the shortcomings of word frequency characteristics. After that, the corresponding value of each node is measured with improved TPR (Topic-PAGERANK) to represent the importance of the relationship group. Then the corresponding original microblog text of relational tuples is sequentially outputed. Finally, it is proved by experiments that the text summarization extracted by automatic text summarization method based on relational tuple is more comprehensive and less redundant.

Keywords Entity relationship, Short text, Text expression, Syntax parsing, Topic-PAGERANK

1 引言

微博是当前流行的社交网络平台,在为用户提供丰富多样的信息的同时,也使得用户面临海量数据的选择问题。如何在规模庞大且噪声较多的微博数据集中提炼出有价值的重要信息,是微博文本摘要研究需要解决的重要问题。

近年来,随着面向微博等社会化短文本的自动文摘研究的深入,研究者们不断提出各种方法。Harabagiu等[1]通过抽取事件的结构特征和分析用户行为模型两种方法对来自多个微博平台中关于相同事件的讨论进行相关性评分,在不超过摘要长度上限的条件下,选择得分最高的几条消息作为摘要。Long等[2]针对微博数据特征,根据文档频率和消息标签选择话题词,并根据词的共现情况对话题词聚类,基于二部图获得描述事件的话题词,最后选择能代表各事件内容、信息含量大且考虑了事件时间跨度的消息作为摘要。

以上方法仅仅针对微博消息的结构特征采用了不同的处理,缺乏对微博文本词语之间语义关联的考虑。由于交互式的微博文本提供了丰富的上下文,博文经常省略前文提到的内容,或者采取指代方式。Zhao等[3]将消息集中的词表达成无向图,基于上下文感知的话题PageRank(context-sensitive Topical PageRank,cTPR)及信息检索中的概率模型从Twitter中抽取与话题相关的关键短语作为摘要。基于图模型的方法之所以有效,是因为它通过迭代的计算能有效地获取图的全局信息即文本的全局信息,从而在判断句子重要程度上更为准确。相比采用一系列特征为句子打分的方法,这种方法有更好的鲁棒性,同时它通常不需要预先设定任何参数。但是Zhao提出的方法只考虑了短语之间机械的上下文关系,并没有考虑到句子与句子间的相似度,这将低估相似度高的短语的重要程度。

Zhong等[4]提出基于事件关系图的文本事件重要度排序方法,其通过构造事件影响因子矩阵来描述文本集合中事件之间的关联强弱,阐述基于事件影响关系识别文本集合中重要事件的方法。以上以事件为单元勾画文档的内容的方法的不足之处在于需要对文本进行大规模事件标注之后才能进行文本摘要。

一些语言学家从语言学的角度给出了事件的定义和事件结构。Davidson提出语义逻辑事件,认为事件不仅包括了表示动作行为的动词,还包括与之相关的名词和修饰成分。郭喜跃等[5]提出将语言学中的SVO(Subject-Verb-Object)结构与实体关系结构对应,主语、宾语、核心谓词组成了实体关系。

到稿日期:2015-07-01 返修日期:2015-08-22 本文受国家863高技术研究发展计划(2012AA01A401)资助。
薛竹君(1991-),女,硕士生,主要研究方向为海量数据挖掘、自然语言处理,E-mail:xzjyouruo@126.com;杨树强(1969-),男,博士生导师,主要研究方向为海量数据挖掘、分布式计算;束阳雪(1991-),女,硕士生,主要研究方向为社交网络分析。

因此，实体关系结构能够反映事件发展的客观知识，用其代替词语、句子、段落等语义单元，应用到文摘技术中能够使文摘更加简洁全面。

本文首先定义了实体关系及关系网络的相关概念；然后根据关系网络的定义给出了微博关系网络构建的方法和过程；最后基于该关系网络采用 Topical-PAGERANK(TPR)算法计算节点的重要程度，形成微博摘要。实验阶段在对多类微博话题语料进行分析后，选取代表性话题博文作为实验语料，将基于关系网络生成微博摘要的方法与已有微博文摘方法作对比，实验结果证明了本文方法的有效性。

2 实体关系及关系网络定义

定义 1(实体关系) 指在某个特定的时间和环境下发生的、由若干角色参与、描述参与者之间动作特征的一种状态。形式上，实体关系可表示为 E，定义为一个四元组：

$E=(O,T,P,C)$

其中，关系元组中的元素称为关系要素，分别表示实体(Object)、时间(Time)、环境(Place)、约束(Constraint)。

O(实体)：发生关系的施动者(主体)和受动者(客体)。句子中核心谓词是支配其它成分的中心成分，而其本身却不受其他任何成分的支配，由于句子中的实体必定会作为一个短语结构出现在依存结构中，因此这种依存关系也必然会反映出相应实体之间的关系特征。例如“许多民众自发悼念事故遇难者。”的句法分析结果如图 1 所示，“民众”作为“悼念”的名词主语(nominal subject)，“遇难者”作为“悼念”的直接宾语(direct object)，“许多”和“民众”是数值修饰(numeric modifier)，“事故”和“遇难者”是名词组合形式(noun compound modifier)。因此关系词“悼念”的实体为：〈许多民众，事故遇难者〉。

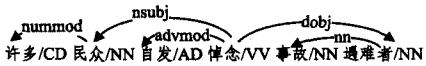


图 1 语法分析示例

T(时间)：关系发生的时间段分为绝对时间段和相对时间段。绝对时间段是指句子中显式包含的时间数据；相对时间段是指句子中隐含了表示时间的词语，需要经过推理才能得到关系发生的绝对时间。

P(环境)：关系发生的场所及其特征等。

C(约束)：约束由关系发生的前置条件、中间约束以及后置条件构成。前置条件指为形成该关系，各要素应当或可能满足的约束条件，它们可以是关系发生的触发条件；中间约束指关系发生时的约束限定；关系发生后，关系各要素将引起变化或者各要素状态的变迁，这些变化和变迁后的结果将成为关系的后置条件。在上述例子中“自发”和“悼念”之间是状语修饰(adverbial modifier)的关系，因此“自发”是“悼念”的中间约束条件。

图 2 给出了一个简单的关系元组的例子。

关系名： 击落
实体： MH17 马航 导弹
时间： 2014/07/17
环境： 乌 俄 边境
中间约束： 被

图 2 关系元组示例

定义 2(微博实体关系间的相互依赖)

(1)因果相关：在每条微博文本中，关系 e_1 的发生导致了关系 e_2 的发生，表示为 $RCE(e_1,e_2)$。

(2)并列相关：在每条微博文本中，关系 e_1 和关系 e_2 同时发生，表示为 $RU(e_1,e_2)$。

(3)时序相关：在微博文本转发树中，关系 e_2 跟随关系 e_1 之后发生，表示为 $RF(e_1,e_2)$。

图 3 给出了微博文本中实体关系依赖的例子。

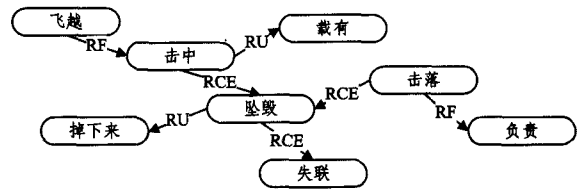


图 3 实体关系依赖模型

定义 3(实体关系相似度) 指关系的相似程度，通常用区间[0,1]之间的值来表示。设关系集合存在任意两个关系 e_i 和 e_j，根据关系要素对应的相似度计算关系元组的相似度，则有：

$$S(e_i,e_j)=\sum_{k\in(o,t,p,c)}\omega_k s(e_{ik},e_{jk})$$

其中，$S(e_i,e_j)$是指 e_i,e_j 之间的相似度，e_{ik} 表示关系 e_i 的第 k 个要素，e_{jk} 表示关系 e_j 的第 k 个要素，ω_k 表示关系各要素在计算关系相似度时的权重，关系相似度的权重介于[0,1]，即 $\sum\omega_k=1$。

根据关系要素在文本的地位，将 ω_k 设定为 $\omega_o=0.5,\omega_t=\omega_p=0.2,\omega_c=0.1$。通过实验观察，当关系相似度 $S(e_i,e_j)\geqslant 0.7$ 时，可以认为(e_i,e_j)是相似的。

定义 4(关系网络) 指一组包含一系列关系节点及相连边的有向图的集合，节点表示实体关系元组，边表示关系元组之间的相互依赖。形式化表示为：

$RNT=[\ \text{RNode},\text{Links}\ ]$

$RNode:\{e_1,e_2,e_3,\cdots,e_n\}$

$Links:\{L(e_1,e_2),L(e_1,e_3),\cdots,L(e_i,e_j),\cdots\}$

其中，RNT 是关系网络。$RNode$ 表示关系节点集合，节点集合 $V=\{e_1,e_2,\cdots,e_n\}$中的每个节点 e_i 代表一个实体关系元组，n 为整个图结构的节点个数。$Links$ 表示关系节点间的相互联系，在有向边的集合 $E=\{\cdots,l_{ij},\cdots\}$中，每条有向边 l_{ij} $(i,j=1,2,\cdots,n,\text{且 } i\neq j)$代表两个邻接节点 e_i 和 e_j 间的依赖性。

3 微博关系网络的构建

为构建微博关系网络，首先在博文集合 D 中抽取出关系词以及对应的关系要素，分析关系词之间的相互依赖性，得到微博文本 D 的关系节点集合 $RNode=\{e_1,e_2,\cdots,e_i,e_j,\cdots,e_n\}$，并在此基础上构建关系网络，构建步骤如下。

1)初始化节点集合 $V=\{\ \}$、有向边集合 $E=\{\ \}$；

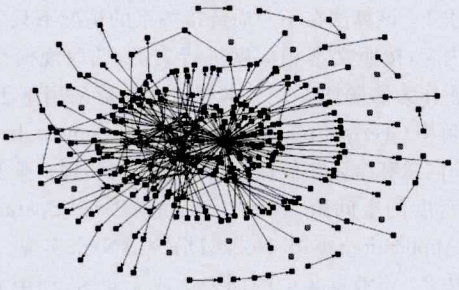
2)依次将文本 D 的事件集合 $RNode=\{e_1,e_2,\cdots,e_i,e_j,\cdots,e_n\}$中的单位事件映射至事件网络图结构中的节点，得到节点集合 $V=\{v_1,v_2,\cdots,v_i,v_j,\cdots,v_k\}$；

3)在节点集合 V 中取节点 v_i，并在节点集合 V 中依次查找与 v_i 相关联的节点 v_j，根据节点 v_i 和 v_j 间的依赖关系添加有向边→；

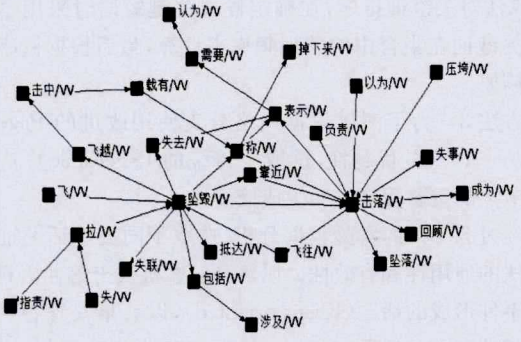
4)在节点集合 V 中任取节点 v_i，依次遍历集合 V 中其它节点 v_j，计算它们对应的关系节点 e_i 和 e_j 的相似度，如果相似度大于或等于阈值(这里阈值设定为 0.7)，则在 v_i 和 v_j 之间添加两条有向边→和←；

5)根据步骤3)和步骤4)可以得到有向图集合 $E(D)$,从而得到文本 D 的关系网络有向图。

下面以实例说明微博文本实体关系网络的构建,文本来自新浪微博话题#马航客机被击落#,选取400条话题相关的博文数据,使用Stanford parser语法分析工具[8]逐条进行语法分析,并按上述方法得到文本 D 的关系网络有向图,该图包含815个节点和399条有向边。通过NetDraw工具将文本可视化,每个节点使用实体关系词进行标识,通过节点间带箭头的线表示关系间的相互依赖,如图4所示。



(a)全局图



(b)局部图

图4 微博话题网络有向图实例

4 基于关系网络形成摘要

为了形成摘要,在对微博文本集中关系词的重要度进行排序后需将关系词所代表的句子全部串联起来。图5为抽象化的实体关系网络表示模型。

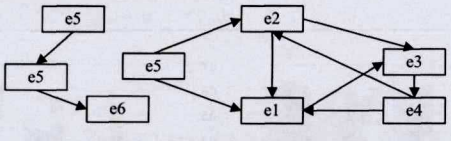


图5 实体关系网络表示模型

从图5可以看出,关系节点间存在一条或多条有向边相互连接,若某节点存在 K 条边链出,N 条边链入,则表示当前关系的发生影响 K 个其它关系,同时受 N 个其它关系的影响。由于话题中存在不同主题,网络中可能存在多个网络分支,经过实验发现不同的连通分支在主题聚类划中分有很好的表现,已完成大量的实验工作。本文将不同的连通分支看作不同的主题,在一个连通分支中节点间的链入与链出边通过矩阵的方式来表示。

$$W=\begin{bmatrix} l(e_1,e_1) & l(e_1,e_2) & \cdots & l(e_1,e_n) \\ l(e_2,e_1) & \ddots & & l(e_2,e_n) \\ \vdots & & l(e_i,e_j) & \vdots \\ l(e_n,e_1) & l(e_n,e_2) & \cdots & l(e_n,e_n) \end{bmatrix}$$

其中,$l(e_i,e_j)$ 表示节点 e_i 指向 e_j 方向的一条边。如果该边存在,则 $\sum_i^n l(e_i,e_j)=1$,否则 $l(e_i,e_j)=0$。

关系节点的重要程度采用基于主题的PageRank(Topical-PageRank,TPR)算法进行排序,图中每个节点的度计算公式为:

$$R_t(e_i)=d\sum_{j:e_j\to e_i} R_t(e_j)/L(e_j)+(1-d)/n_t, d\in[0,1]$$

其中,e_i,e_j 是关系图中的任意节点,$R_t(e_i)$ 表示在主题 t 中 e_i 的重要度。$R_t(e_j)$ 表示在主题 t 中 e_j 的重要度。$L(e_j)$ 表示 e_j 连接线指向别的节点 $e_k(k\in n_t)$ 的个数。n_t 是主题 t 包含的节点数;d 是参数,为一个衰减因子,又称阻尼系数[0,1],通常取 $d=0.85$。

由于微博是交互性文本,除了具有转发结构外,还有人们对博文的赞同程度,表现为"点赞数",因此在此基础上,加入博文赞数作为影响因子 δ,$\delta\in(0,1)$。改进后的公式为:

$$R_t(e_i)=d\sum_{j:e_j\to e_i} R_t(e_j)/L(e_j)+(1-d)*\delta/n_t, d\in[0,1]$$

文摘句的排序应按照关系的重要程度及发展过程进行。首先将关系元组按重要程度进行排序;其次,比较关系元组的时间要素,按照关系发生的先后顺序排序;最后删除语句中对信息无贡献的子句,即删除不包含任何关系元组及要素的子句,以达到消除冗余的目的。

5 实验结果与分析

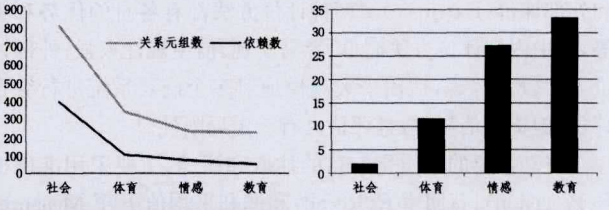
5.1 实验数据及评价性能分析

由于目前在公开的数据集中未有针对微博的标准数据集,因此本文采用的数据集的原始数据来源于新浪微博。该数据集收集了从2014年6月10日至8月10日发布的3164478条微博。

结合Zhao等人[6]的分析,我们认为微博话题分为突发事件形成的话题和长期存在的话题两种不同类型。因此本文从这两类话题中抽取了4个话题关键词在24小时间的博文数据,共4587条,社会、体育类代表突发事件形成的话题,情感、教育作为长期存在的话题。经过去广告、去过短文本等处理后,每类话题的博文数目及各自摘取的5项代表性关系词如表1所列。各类语料的关系元组数目以及依赖数目的统计分析如图6所示。

表1 语料数据统计表

语料关键词	语料类型	清洗后的博文数	代表性关系词
马航客机被击落	社会	1169	失联 击落 遇难 压垮 告慰
巴西对战克罗地亚	体育	936	怒骂 上演 预言 取胜 直播
早安青春	情感	407	荒废 遗憾 放弃 成长 希望
四级必过	教育	652	加油 考 祈求 及格 学习

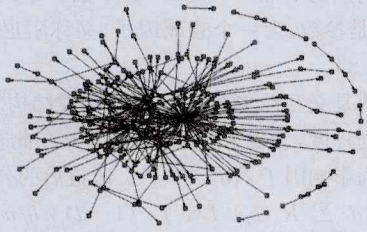


(a)关系元组与依赖数对比图 (b)孤立关系元组(节点)比例

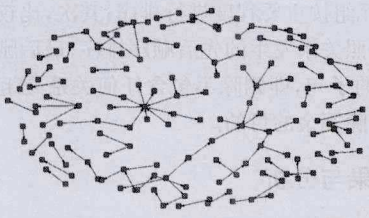
图6 关系元组及依赖数目统计图

图6(a)展示了相同博文条数下不同类型的话题包含的关系元组与依赖数目的对比,由此可知,社会类事件发展变化

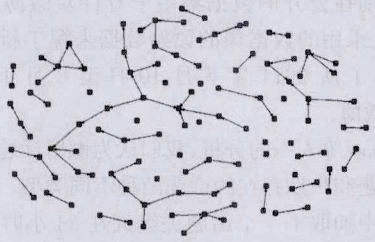
迅速，具有时效性，因此表示事件发展的关系元组和关系依赖数目比较多。从图 6(b)中可以看出，社会类话题存在的孤立节点的比例非常少，体育类次之，这是由于社会类话题讨论的主题往往比较集中，关系之间彼此依赖，因此孤立元组非常少。图 7 是 4 种不同类型话题抽取 400 条微博形成的实体关系网络有向图，可以看出图 7(a)所示的社会类事件有向图的连通性强，而图 7 中由(a)、(b)与(c)、(d)的对比发现，社会类与体育类话题具有比较强的连通性，连通分支较少，而情感类和教育类话题连通分支较多、较离散。



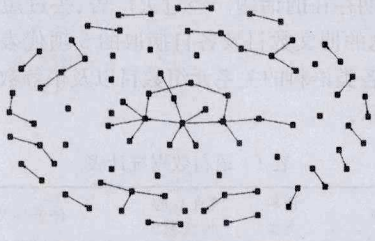
(a)社会类话题实体关系网络有向图



(b)体育类话题实体关系网络有向图



(c)情感类话题实体关系网络有向图



(d)教育类话题实体关系网络有向图

图 7　微博话题的实体关系网络有向图

目前，自动摘要的评价方法通常采用内部评价(Intrinsic)和外部评价(Extrinsic)。两种评价方法都有各自的优势和劣势，其中内部评价方法简单、容易实现，但主观性太强；外部评价方法较为客观，适用于大规模地对多个摘要系统进行综合评价，但资源消耗多，且评价具有一定局限性。

自动摘要的本质是信息的抽取和压缩，主要采用准确率 P(Precision)、召回率 R(Recall)和调和平均值 F(F-Measure) 3 个指标对自动摘要系统进行内部评价。摘要召回率反映摘要对原文主题信息的覆盖程度，是对评价摘要质量的一个重要标准。摘要准确率反映摘要表现原文主题信息的准确程度。摘要准确率 P 定义为：

$$P=|x\cap y|/|y|*100\%$$

摘要召回率 R 定义为：

$$R=|x\cap y|/|x|*100\%$$

F 值即为准确率和召回率的调和平均值：

$$F=P*R*2/(P+R)$$

5.2　实验过程及结果

为了验证本文自动摘要方法的有效性，本文选取了 2 个具有代表性的基准测试算法和 1 个传统的重要度计算方法与之比较。

方法 1　该算法在不考虑链接关系的情况下只关注消息文本的内容，按照文本相似度进行聚类，选取规模最大的 K 个聚类簇作为话题摘要。文本相似度计算采用基于词频-逆向文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)向量和余弦相似性度量的方法来实现。聚类算法采用基于密度的空间聚类算法(Density-Based Spatial Clustering of Applications with Noise，DB-SCAN)来实现。

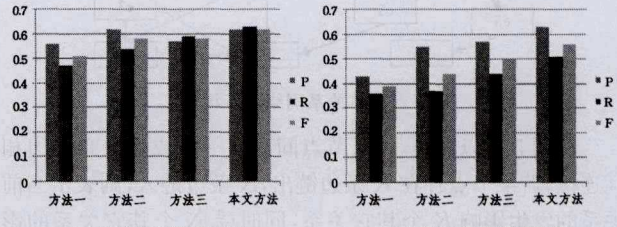
方法 2　将消息集中的词表达成无向图，利用上下文感知的话题 PageRank(context-sensitive Topical PageRank，cTPR)算法对关键词排序，在确定候选关键短语时采用不同主题下关键词在集合中的共现频率来计算，最后根据概率模型输出摘要。

方法 3　为了测试根据网络分支使用改进的 PageRank 算法——TPR 的优越性，选取了传统的 PageRank 算法计算关系节点的重要程度进而生成摘要。

经过 5.1 节的实验数据分析，选取不同类型话题证明本文方法的通用性和有效性。以＃马航客机被击落＃语料代表突发事件形成的话题(Event-oriented)，以＃早安青春＃语料代表长期存在的话题(Long-standing)作为实验语料，实验结果如表 2 所列。

表 2　实验结果分析

实验方法	突发事件形成的话题 (Event-oriented)			长期存在的话题 (Long-standing)		
	P	R	F	P	R	F
方法 1[8]	0.56	0.47	0.51	0.43	0.36	0.39
方法 2[9]	0.62	0.54	0.58	0.55	0.37	0.44
方法 3	0.57	0.59	0.58	0.57	0.44	0.50
本文方法	0.62	0.63	0.62	0.63	0.51	0.56



(a)突发事件话题语料　(b)长期存在的话题语料

图 8　实验结果分析图

由表 2 中不同方法下不同语料生成的摘要对比(见图 8)可看出，在突发事件话题语料中，方法 2 的准确率与本文方法相当，但本文提出的方法在召回率上表现突出，高于其他方法，且本文方法对长期存在的话题做出的摘要更为有效，在准确率、召回率和调和平均值上提高较大。

结束语　对于热点事件形成的话题和长期存在的微博话题，本文根据实体关系间的相互依赖构建有向图网络，带有向

图的关系网络不仅能帮助我们清晰地理解事件发展趋势，还能根据关系网络有向图的连通性勾画出话题的主题性，运用基于主题的 PAGERANK 计算各节点的相对重要度，从而得到各子主题在整个话题中的重要程度，并在关系元组的重要程度排序的基础上按照时间要素先后进行排序。最后逐步删除排序了的语句集合中对信息贡献最小的子句，直到剩余的句子长度之和达到目标文摘长度。实验证明这种方法能更全面、更简洁地概括出文本的主要内容。

参 考 文 献

[1] Harabagiu S, Hickl A. Relevance modeling for microblog summarization [C]// Fifth International AAAI Conference on Weblogs and Social Media. 2011

[2] Long R, Wang H, Chen Y, et al. Towards effective event detection, tracking and summarization on microblog data[M]// Web-Age Information Management. Springer Berlin Heidelberg, 2011: 652-663

[3] Zhao W X, Jiang J, He J, et al. Topical keyphrase extraction from twitter[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 379-388

[4] Zhong Z, Liu Z. Ranking events based on event relation graph for a single document [J]. Information Technology Journal, 2010, 9(1): 174-178

[5] Guo Xi-yue, He Ting-ting, Hu Xiao-hua, et al. Chinese Named Entity Relation Extraction Based on Syntactic and Semantic Features[J]. Jounal of Chinese Informatio Processing, 2014, 28(6): 183-189(in Chinese)
郭喜跃，何婷婷，胡小华，等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报，2014，28(6)：183-189

[6] Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2011: 338-349

[7] Chen Dan-qi, Manning C D. A Fast and Accurate Dependency Parser using Neural Networks[C]// Proceedings of EMNLP 2014. 2014

[8] Nenkova A, Maskey S, Liu Y. Automatic summarization[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011. Association for Computational Linguistics, 2011

[9] Zhao W X, Jiang J, He J, et al. Topical keyphrase extraction from twitter[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 379-388

(上接第 60 页)

准确率的影响是显而易见的，数据样本越多，则预测结果的准确率越高。

从与 SVM 算法的比较结果来看，NLDA 的预测准确率基本上维持在 90%左右。最终在 300000 个测试数据上进行测试，得到 91.69%的准确率，比 SVM 的 89.64%有明显提升，结果如图 7 所示。

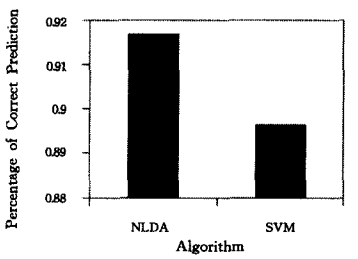


图 7 NLDA 和 SVM 算法之间的比较

结束语 本文提出了一个用于学习和识别网络异常行为分类的主题模型。该模型是对网络安全的一个应用并且在 KDDCUP'99 测试集上的效果显著。本文主要的创新点在于使用主题模型对网络行为进行分类并在参数分析的基础上提高了识别的准确率。

在接下来的工作中，将在单机和分布式环境下做进一步的实验。考虑使用 Spark 和 Petuum 平台作为分布式环境。之后，将对网络流量如何转换为网络特征做进一步研究。

参 考 文 献

[1] Garcia-Teodoro P, Diaz-Verdejo J, Macia-Fernandez G, et al. Anomaly-based network intrusion detection: Techniques, systems and challenges[J]. Computers & Security, 2009, 28(1/2): 18-28

[2] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. J. Mach. Learn. Res., 2003, 3: 993-1022

[3] Blei D M. Probabilistic topic models[J]. Commun. ACM, 2012, 55(4): 77-84

[4] Fei-Fei L, Perona P. A bayesian hierarchical model for learning natural scene categories[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005(CVPR 2005). IEEE, 2005, 2: 524-531

[5] Cramer, Christopher, Carin L. Bayesian topic models for describing computer network behaviors[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011: 1888-1891

[6] Newton B D. Anomaly Detection in Network Traffic Traces Using Latent Dirichlet Allocation[OL]. http://www.cs.unc.edu/~bn/BenNetwonFinalProjectReport.pdf

[7] Huang J, Kalbarczyk Z, Nicol D M. Knowledge Discovery from Big Data for Intrusion Detection Using LDA[C]// 2014 IEEE International Congress on Big Data (BigData Congress). IEEE, 2014: 760-761

[8] Kasliwal B, Bhatia S, Saini S, et al. A hybrid anomaly detection model using G-LDA[C]//2014 IEEE International Advance Computing Conference (IACC). IEEE, 2014: 288-293

[9] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html