文章编号：1003-0077(2011)01-0041-07

基于 LDA 模型的博客垃圾评论发现

刁宇峰，杨 亮，林鸿飞

（大连理工大学 信息检索研究室，辽宁 大连 116024）

摘 要：Blog(博客)作为一种新兴的网络媒体，在很大程度上增强了互联网的开放性，Blog 已经成为互联网上的主要信息源之一，这也使得 Blog 空间中的垃圾评论成倍增长，因此如何识别垃圾评论成为面临的重要问题。该文首先借鉴处理垃圾邮件的方法，针对 Blog 本身的特点，使用规则初步过滤垃圾评论，然后对剩余评论，利用 Latent Dirichlet Allocation(LDA) 这种能够提取文本隐含主题的产生式模型，对博客中的博文进行主题提取，并结合主题信息进行判断，从而识别 Blog 空间的垃圾评论。通过实验验证，该方法可以发现大多数垃圾评论，实验取得了较好的结果，使 Blog 信息更加准确、有效的为用户使用。

关键词：Blog；博文；LDA；主题；垃圾评论

中图分类号：TP391 **文献标识码**：A

LDA-Based Opinion Spam Discovering

DIAO Yufeng, YANG Liang, LIN Hongfei

(Information Retrieval Laboratory, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: As well-known, Blog has become one of the main information sources on the Internet, and the opinion spam also grows fantastically in Blog. The paper focuses on identifying the opinion spam. Firstly, it adopts the method of email spam identification. Considering the characteristics of Blog, it establishes the rules of comments to filter the opinion spam, and then it utilizes the Latent Dirichlet Allocation Model (LDA) to extract the topics information from text content in Blog. Finally, with the topics information integrated, it judges the opinion whether spam or not. Experiments prove it can identify most of the spam opinions, effectively bringing more accurate and efficient Blog information for users.

Key words: Blog; Blog content; LDA; topic; opinion spam

1 引言

Blog 的全名是 Weblog，意思是“网络日志”，是继 E-mail、BBS、IM 之后出现的第四种网络交流方式，是网络时代的个人读者文摘，是一种表达个人思想、网络链接、内容的日志，按照时间顺序排列，并且不断更新的出版方式。简言之，Blog 就是以网络作为载体，简易、迅速、便捷地发布自己的心得，及时、有效、轻松地与他人进行交流，再集丰富多彩的个性化展示于一体的综合性平台。因此，与传统的论坛(BBS)相比，Blog 更能展示个性，更有针对性；与普通网页相比，Blog 又拥有更强大的互动功能。Blog 信息的形式包括文本、图片、音频、视频等多种媒体格式，Blog 由博文和评论两部分构成。

随着整个互联网业的迅速发展，作为一种新兴事物，Blog 正处于高速的发展时期，互联网上的 Blog 数量一直在急剧的增长，Blog 已经成为互联网

收稿日期：2010-08-30 **定稿日期**：2010-10-25

基金项目：国家自然科学基金资助项目(60673039,60973068)；国家社科基金资助项目(08BTQ025)；国家 863 高科技计划资助项目(2006AA01Z151)；教育部留学回国人员科研启动基金和高等学校博士学科点专项科研基金资助项目(20090041110002)

作者简介：刁宇峰(1987—)，女，硕士生，研究方向为情感分析和意见挖掘；杨亮(1986—)，男，博士生，研究方向为情感分析和意见挖掘；林鸿飞(1962—)，男，博士，教授，博导，研究方向为搜索引擎、文本挖掘、情感计算和自然语言理解。

上一种重要的信息源。Blog 赋予了数以百万计人自由地发表言论的权利，因此，Blog 评论信息数量庞大，并且具有多样性，同时也会含有大量的垃圾评论(Opinion spam)，会严重地干扰用户的阅读和使用。近年来，人们越来越热衷于分析在线的评论信息及其极性，更希望在相关评论上搜集有用信息，因此，垃圾评论的发现也是愈加重要。

本文在这里主要讨论在 Blog 领域中垃圾评论的信息。在现今，垃圾信息的研究已经成为一个重要的研究领域，比如说研究热点垃圾网页(Web spam)。在当今 Blog 大量信息下，由于经济或宣传等效应，Web spam 是普遍存在的。Web spam[1]的目标是为了吸引人们浏览这些页面，而采取若干手段使该网页在搜索引擎中享有较高的排名。Web spam 有诸多相关的因素，主要的有：垃圾内容(Content spam)和垃圾链接(Link spam)[2]。Link spam 是指在链接上存在垃圾信息，此链接并没有指向真正的评论信息。Content spam 试图在目标网页中添加与内容无关的词，用以欺骗搜索引擎排名。而 Opinion spam 同 Content spam 类似，均是在内容上对信息进行分析处理，但是在 Blog 领域内，在评论中除了相关评论，还会存在与博文的主题毫不相关的其他评论信息，即为本文研究的 Opinion spam。

Opinion spam 的初步处理同垃圾邮件(E-mail spam)类似，对大多数用户，Email spam 大部分都是没有主动订阅的广告、电子期刊等宣传品，其基本特征是“不请自来”、带有商业目的或者政治目的，实际上，垃圾邮件的大部分都是采用基于规则的方式进行处理[3]。在初步处理中，Opinion spam 也可以采用该方法，但是对于剩余 Opinion spam，需要寻找更有效的方法进行检测发现。

在国内外，无论是工业界还是学术界，越来越多的研究者关注产品中的评论信息[4-5]。Liu 等人首次调研产品评论中的垃圾信息，并提出行之有效的解决方法[6-7]，要将 Opinion spam 分为三类：(1)非可信评论，(2)品牌效应评论，(3)无内容评论。对于第二、三类，主要是看作二值分类问题进行处理。他们通过研究 Opinion spam 信息固有的特点，发现充分能够代表 Opinion spam 的特征，并用这些特征建立分类模型，识别垃圾评论。在这里，本文主要处理的是 Blog 领域的 Opinion spam 问题，与上述方法不同的是会考虑到 Blog 的博文信息。

本文对 Blog 中评论信息进行分析，发现主要有两类 Opinion spam：显式垃圾评论和隐式垃圾评论。针对上述两种类型，本文分析 Opinion spam 的特点，在新浪博客语料集上，对于显式垃圾评论，采取类似于处理垃圾邮件(Email spam)的处理方式[3]，使用基于规则的方法进行识别。在此基础上，对于隐式垃圾评论，本文采用 LDA 这种主题模型来发现主题信息，通过基于主题的特征选取和基于主题的检索模型两种方法，进而发现隐式垃圾评论，最终过滤 Opinion spam，这样能够帮助人们将大量的 Blog 评论信息按话题相关程度进行组织，并将过滤后的评论呈现给用户。

本文主要解决垃圾评论问题，具体方法在下面详细介绍：第 2 节主要介绍一些相关工作；第 3 节主要介绍本文的核心算法—基于 LDA 的垃圾评论发现，主要有基于主题的特征选取和基于主题的检索模型两种方法；第 4 节是实验流程以及结果分析；最后，在第 5 节中总结工作并计划下一步工作。

2 相关工作

2.1 情感词汇本体

本文使用的外部资源是大连理工大学信息检索实验室的情感词汇本体[8](以下简称情感本体)，该情感本体将情感分为 7 大类 20 小类。情感词汇本体通过一个三元组来描述：

$$\text{Lexicon}=(B, R, E)$$

其中 B：表示词汇的基本信息，主要包括编号、词条，对应英文、词性、录入者和版本信息。R 代表词汇之间的同义关系，即表示该词汇与哪些词汇有同义的关系。E 代表词汇的情感信息，包括情感类别、情感强度、情感极性，是情感词汇描述框架中比较重要的一部分。

情感本体的基本知识主要来源于现有的一些词典和语义网络。其中词典包括《现代汉语分类词典》、《汉语褒贬义词语用法词典》、《汉语形容词用法词典》、《中华成语大词典》、《汉语熟语词典》、《新世纪汉语新词词典》。语义知识网络有《知网》和 WordNet。另外还加入了《汉语情感系统中情感划分的研究》中的部分词汇。因此覆盖面是比较全面的。

目前，该情感词汇本体收录情感词汇共 17 156 个，为句子级、段落级和篇章级的情感计算提供了词汇基础和分析依据。

2.2 LDA 模型基本思想

Latent Dirichlet Allocation(LDA)模型是 Blei 等在 2003 年提出的[9]，属于主题模型(Topic Models，是当前文本表示研究的主要范式)的一种。作为一种产生式模型，LDA 模型已经成功的应用到文本分类，信息检索等诸多文本相关的领域[1,3,9-13]。

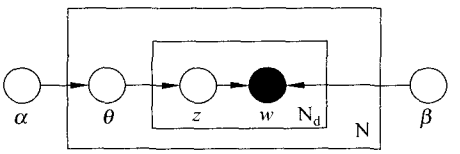


图 1　LDA 的图模型表示形式

LDA 是一个多层的产生式全概率生成模型，是典型的有向概率图模型，是一种对文本数据的主题信息进行建模的方法[15]，如图 1 所示，包含词、主题和文档三层结构。给定一个文档集合，LDA 将每个文档表示为一个主题集合，每个主题是一个多项式分布，用来捕获词之间的相关信息。在 LDA 中，这些主题被所有文档所共享；每个文档有一个特定的主题比例。LDA 由文档层的参数(α,β)确定，α 反映了文档集中隐含主题间的相对强弱，β 代表了所有隐含主题自身的概率分布。θ 代表文档中各隐含主题的比重，z 表示文档分配在每个词上的隐含主题比重，w 是文档的词向量表式。N 为文档集中文档个数，N_d 表示该文档的词总数。

LDA 模型较之 LSI/PLSI 等模型有着突出的优点[14]：首先 LDA 模型是全概率生成模型，因此具有清晰的内在结构，并且可以利用高效的概率推理算法进行计算；再者，LDA 模型是通过无监督方法进行训练的，与训练样本数量无关，因此更适合处理大规模文本语料。近几年，LDA 模型、LDA 的扩展模型以及它们在自然语言和智能信息处理中的应用得到充分的重视和深入的研究[9-13]，但还没有人基于 LDA 发现垃圾评论，由于 LDA 可以挖掘隐含主题这种特性，本文将 LDA 模型应用到垃圾评论发现领域上，用来发现 Blog 中博文的主题信息，进而发现 Blog 中的垃圾评论。

3 垃圾评论的识别

本文在 Blog 领域内进行研究，主要考虑的是评论的文本信息，不涉及图像的识别。针对 Blog 的评论和博文进行分析，并结合 Blog 固有的特点，本文将垃圾评论的类型总结为两类。

第一类：显式垃圾评论。经分析，主要有三种类型，(1)广告及链接等，(2)与评论无关的信息如大量随机字符等，(3)重复评论。这类垃圾评论的分析与处理垃圾邮件类似，主要通过基于规则的方式进行发现。

第二类：隐式垃圾评论，主要是与 Blog 中博文内容不相关的评论。这类垃圾评论不能依靠基于规则的方法发现，需要引入 LDA 模型，结合挖掘出的隐含主题信息进行隐式的分析，主要是通过基于主题的特征选取和基于主题的检索模型两种方法来发现。具体流程如下：

(1) 对所有评论，通过基于规则的方法，初步过滤显式垃圾评论。

(2) 对剩余评论，需要发现隐式垃圾评论。这里主要采用基于主题的方法，引入 LDA 模型，对 Blog 的博文信息进行隐式分析，进而挖掘隐式主题信息，最后通过基于主题的特征选取和基于主题的检索模型两种方法发现隐式垃圾评论。

3.1 基于规则的垃圾评论识别

设置一些规则，只要符合这些规则的一条或几条，则认为是显式垃圾评论。这些规则通常有：

(1) 垃圾关键词精确匹配：

在本文，通过分析语料，定义一些反映垃圾评论特征的关键词或关键短语，如："欢迎到我的博客来"、"保证有你想要的"、"交流群"、"百度贴吧欢迎你"、"联系电话"、"欢迎咨询"、"订购热线"、"24 小时人工服务"、"欢迎到我的博客一游"等明显的垃圾词语，当在评论中发现若干条关键词或短语，则判断为显式垃圾评论。

(2) 重复评论发现：

本文发现，在同一博文或者不同博文之间，均会含有大量的同一评论者或者不同评论者发表的相似或者完全相同的评论，称之为重复评论，也属于显式垃圾评论。对于重复评论的判断，在这里，本文采用 Jaccard Distance[16] 的方法检测重复评论。首先，对所有评论建立 2-gram 语言模型，然后对两个评论 A,B，计算相似值 $J(A,B)$，具体公式如下：

$$J(A,B)=\frac{A\cap B}{A\cup B} \tag{1}$$

此相似值越大，证明两评论为重复评论的可能性越大。经过分析和计算，本文发现重复评论主要有四种类型，(1)不同的评论者对同一博文发表相同

的评论，(2)不同评论者对不同博文的相同评论，(3)同一评论者对不同博文的相同评论以及(4)同一评论者对同一博文的相同评论。对上述四类相同评论，本文分析得到：第一、四类评论是重复评论需要过滤掉，第二类是相关评论，而第三类评论可能是隐式垃圾评论，因此在下一步要着重检测。

(3) 其他垃圾特征发现：

例如，评论中的文字较少，但却含有大量的超级链接；评论中包含大量的随机字符或者特殊字符等；非垃圾评论的字数量虽多但字体很小，而垃圾字体设置为正常字体等，这样不但可以保证垃圾评论的视觉效果，又因为含有大量的相关评论，欺骗搜索引擎的搜索，并造成用户的诸多不便，这些均判断为显式垃圾评论。

3.2 基于主题的垃圾评论识别

传统的垃圾评论发现只是在基于规则的基础上发现显式垃圾评论，对于剩余评论，再进行简单的特征选择最后进行分类，从而发现隐式垃圾评论，是一个简单的二值分类问题。与此不同，本文充分考虑到Blog博文的主题信息，提出基于主题的特征选择和基于主题的检索模型两种方法。该算法基于以下的基本假设：(1)每类博文讨论若干个主题，类间主题的相关程度低于类内主题；(2)一个具体博文讨论的主题是该博文所属的主题集合的子集。

3.2.1 基于主题的特征选择方法

对于该方法的与评论相关的特征集合的构造，这里主要有四类信息：(1)评论的内容，(2)评论者，(3)博文作者以及(4)博文的内容。因此，特征集合主要由以上4类信息组成，具体如表1所示。

表1 特征集合

		特征集	特征集主要元素
主题特征集	简单特征集	评论	情感词典，评论长度，评论时间
		评论者	评论者等级，访问量
		作者	作者相关信息
		博文	隐含主题信息

对于评论本身的内容，我们主要考虑评论中包含情感倾向的词和评论的长短。本文从情感本体中提取表达较强的情感强度的情感词，如“支持”、“同意”、“赞同”等，并从语料中提取Blog特有的表达情感的词和短语，如“路过”、“留名”、“杯具”、“稀饭”、“板凳”、“顶一下”、“sofa”、“马扎”等词，共同建成本文所需的情感词典。同时，评论的长度、时间都会对识别垃圾评论造成一定的影响，长评论的关注度相对较高，而早期发表的评论的关注相对也会较多，这些都是识别垃圾评论的标注。

对于评论者，在这里，主要考虑评论者本身的等级和评论者的访问量，用以参考评论者的信誉度。

分析Blog的评论，本文发现评论中很多是对博文作者的评价，因此，博文作者以及和作者紧密相关的信息均作为特征。

Blog中的评论不仅可以针对博文作者，也可以是针对博文的内容。但是，由于博文的内容过于庞大而且繁杂，不能全部作为特征。在本文，对Blog的博文集合，使用LDA模型建模，抽取隐含的主题集合，将这个主题集合作为特征使用。

本文采用SVM方法进行分类[17]。

3.2.2 基于主题的检索模型方法

由于采用基于主题的特征选取方法来发现垃圾评论需要标注和训练，为了节省这种大规模训练，本文受到文献[15]中的方法启发，采用概率检索模型来发现垃圾评论，这种基于统计的无监督方法，无需训练集，不用着重筛选特征集合。在这里，将评论和博文的问题看作是检索问题。评论C假设为查询串，博文B当作文档，博文集合是文档集合，在未引入主题信息前，计算博文产生该评论的概率，建立简单的概率检索模型[15]。公式如下：

$$P(C \mid B) = \prod_{w \in C} P(w \mid B) \qquad (2)$$

其中，B 为博文，C 为 B 的评论集合中的一条，w 是 C 中的一个词，假定 B 中词与词之间相互独立，$P(C|B)$ 为 B 产生 C 的概率，$P(w|B)$ 为 w 在 B 中出现的概率。

在该模型中，由于未考虑到博文的稀疏性以及词之间隐含的主题信息，也没有对 $P(w|B)$ 这项进行平滑，该公式还有待改进，需要在该概率检索模型中加入隐含的主题信息[15]。本文在上述概率检索模型的基础上，加入引入LDA模型后发现的隐含主题集合，用于进行平滑 $P(w|B)$，即主题检索模型。与其他平滑模型不同，LDA模型建立一种全新的博文模型。与其他聚类模型不同，LDA模型将博文看作是包含多个主题的集合，而不只是单一主题集合。看作单一主题集合这种假设对于大规模博文语料来讲过于局限，相反，LDA模型将博文看作多个主题集合并以不同比例进行区分，增强了灵活性。本文结合隐含的主题信息，共同建立主题检索模型。

具体公式如下：

$$P(C \mid B) = \lambda \prod_{w \in C} p(w \mid B) + (1-\lambda) \prod_{w \in C} \sum_{t \in t_B} \prod_{w \in t} p(w \mid t) \times p(t \mid B) \quad (3)$$

其中，t_B 为博文 B 的主题集合，t 为 t_B 中的一个主题，λ 为参数，$p(w|t)$ 为词 w 在主题 t 中出现的概率，$p(t|B)$ 为主题 t 在博文 B 中出现的概率。

对 Blog 博文中的所有评论，均要计算 $P(C|B)$ 建立主题检索模型，然后若此概率值小于某一阈值，则判定为隐式垃圾评论，反之，为相关评论。

4　实验流程及结果分析

4.1　语料来源及实验流程

实验的语料来自新浪博客下载的博文以及评论，作者为博客总人气排行榜前 10 名，选取其中这些作者的部分博文共 100 篇博文（每人 10 篇），并从中选取评论共有 5 980 条，经标注，共发现垃圾评论 1 282 条。本文使用中科院的分词工具 ICTCLAS [18]。具体相关数据见表 2。

表 2　博客统计表

作者	总点击率	总评论数	抽取评论数	垃圾评论	垃圾比例
徐静蕾	1 135 029	11 037	1 092	256	0.234 4
韩寒	1 098 137	10 892	972	217	0.223 3
极地阳光	989 316	8 921	803	213	0.265 3
李承鹏	860 093	7 285	526	94	0.178 7
当前明月	806 037	6 097	695	88	0.126 6
郭敬明	744 517	4 893	457	106	0.231 9
三峡在线	485 835	4 207	413	103	0.249 4
马未都	323 426	3 225	572	47	0.082 2
洪晃在 Ilook	253 135	3 187	450	158	0.351 1
董路	223 117	0	0	0	0

通过观察语料，本文发现在新浪博客人气排行榜上，对每个博文作者，都有数量不少的垃圾评论的存在。在人气排名靠前的作者博文中，评论者和垃圾评论者均对其作者博文的关注度相对较高，反之，在人气排名靠后的作者博文中，评论者和垃圾评论者均对其作者博文的关注度相对较低。经分析，本文得出以下 Blog 领域垃圾评论的分布规律：在高点击率的作者博文中，评论垃圾比例相对也较高；在低点击率的作者博文中，评论垃圾比例相对较低。本文可以在该数据集上进行合理的验证上述寻找垃圾评论的方法。

对于显式垃圾评论，本文主要采用基于规则的方法进行判断。在此基础上，对于隐式垃圾评论，本文采取基于主题的特征选取和基于主题的检索模型两种方法进行实验。

在基于主题的特征选取的方法中，本文主要是针对 Blog 博文和评论二者进行分析，筛选特征集合。具体的特征集合见表 1。本文选取其中 4 000 条评论为训练集，1 980 条评论作为测试集。实验具体流程如下：

（1）从网上下载 Blog 语料，并人工鉴定每条评论是垃圾评论还是相关评论。

（2）对所有的评论进行基于规则的初步过滤，判断评论是显式垃圾评论还是相关评论，并记录可以得到结果的评论。

（3）将搜集到的 Blog 特定用语导入分词的扩展词典，利用中国科学院的分词完成博文和剩余的评论切分等预处理工作。

（4）对于评论长度少于 5 个字符的评论，若包含本文建立的情感词典中的词或短语，认为是相关评论，否则，为隐式垃圾评论。

（5）构建简单特征集合和主题特征集合，进行对比实验。

（6）对于剩余评论，利用 SVM 在分类特征集上进行隐式垃圾、非垃圾分类。

（7）评估结果的正确率和召回率。

在基于主题的检索模型的方法中，要对于每条评论的词或短语，需要统计该词在博文中出现的概率，该词在主题中出现的概率以及主题在博文中出现的概率，然后建立主题检索模型。再通过公式（3），判断评论是隐式垃圾评论还是相关评论。

建立 LDA 模型时，经分析，发现抽取的主题数量对结果的影响很大，而 LDA 模型使用交叉熵作为评价概率模型的性能指标之一，当此熵值越小时，LDA 模型的性能越佳。如图 2 所示，可以得到，当主题数目 T 等于 110 的时候，此时交叉熵最小，建立的 LDA 模型性能最佳。在 LDA 模型中，需要给出 Dirichlet 先验 (α, β)，在本文，令 $\alpha = 50/T$，$\beta = 0.01$，T 为主题数目（此为经验值，多次实验表明，这种取值在本实验的语料集上有较好表现）。

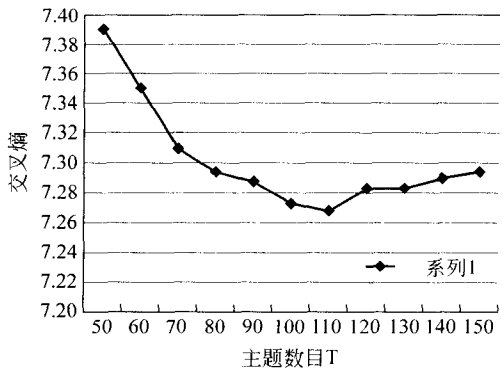


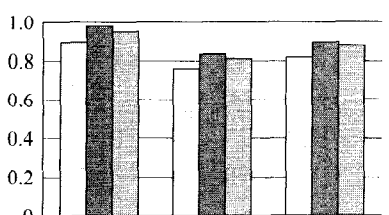
图2 LDA模型主题数目

4.2 实验结果与分析

在单独使用基于规则的方法时，共发现显式垃圾评论872条，充分说明Blog评论中含有大量的广告链接、重复评论等垃圾信息。

在此基础上，对于隐式垃圾评论，本文使用基于主题的特征选取方法（特征集分别为简单特征集和主题特征集两种）和基于主题的检索模型方法（主题检索模型）进行实验。在基于主题的特征选取方法中，本文与Liu等人在处理产品评论[6]中采用的处理垃圾评论的方法进行对比，Liu等将垃圾评论发现看作一个分类问题，特征集合主要由表1提到的情感词典、评论长度以及评论者等级等组成，即为简单特征集。而本文在此基础上，引入了使用LDA模型后对博文进行抽取出的主题信息，即为主题特征集，具体见表1。在基于主题的检索模型方法，本文主要按3.2.2介绍的主题检索模型进行建模并判断。

本文使用SVM-light分类器，3倍交叉验证，结果具体见图3。



	准确率	召回率	F值
□简单特征集	0.898 8	0.762 1	0.824 8
▩主题特征集	0.983 4	0.833 9	0.902 5
▨主题检索模型	0.956 2	0.817 5	0.881 4

图3 垃圾评论的实验结果

基于主题特征集的方法与基于简单特征集方法对比，由图3可以知道，各项指标都有所提升。这是因为主题特征集中加入了与博文相关的主题信息。例如：该博文的作者是徐静蕾，博文主要讲述头晕健康问题，有一条评论为“注意身体，千万别晕”，使用LDA模型后，可以发现该评论是相关评论，而有一条评论为“我喜欢看《杜拉拉》”，此评论的主题为电影问题，与博文的主题头晕健康问题无关，则本文判断为垃圾评论，而用基于简单特征集方法则不能判断。

使用主题检索模型方法的各项指标均高于基于简单特征集的方法，主要原因在于判断该评论的词语在博文、主题，以及该主题出现的概率，而去掉重复评论之后，垃圾评论的词语相对为低频词，使得垃圾评论的概率相对低于相关评论。例如：该博文的作者是洪晃在Ilook，该博文主要介绍新西兰之行，有一条评论为“新西兰的云层也很厚哇，人类活动剧烈，天空污染越来越大”，经本文的主题检索模型方法判断，确实和博文同属一个主题，判断为相关评论，而又有一条评论为“人生可以没有辉煌、没有精彩，但不能没有感恩的心！…”，此评论的主题是感恩，与博文主题不符，用该主题检索模型方法判断为垃圾评论。而简单特征集中没有这些主题信息，故不能正确判断该评论。

基于主题的特征选取方法的各项指标均高于主题检索模型方法，主要原因在于前者是有监督方法，有了人工标注的训练集，相对准确，后者是无监督方法，是基于概率统计的方法，准确率等指标不如前者。由图3可知，在没有标注语料的前提下，本文使用基于主题检索模型的方法也达到了较好的效果。

在隐式垃圾评论发现中，尚存在着一些暂时无法解决的难度较大的隐式情感评论问题无法识别。如“众里寻他千百度，蓦然回首，那人却在灯火阑珊处”、“衣带渐宽终不悔，为伊消得人憔悴”、“有一种花开叫绽放，有一种鼓励叫赞扬，有一种激情叫释放，有一种美丽叫善良”等诸如此句，使用本文的方法均不能判断评论，需要加入更深层次的语法分析、语义消歧等方法进行情感分析，从而判断评论是垃圾评论还是相关评论。

5 结束语与下一步工作

本文通过分析Blog垃圾评论的特点，主要将垃圾评论分为两大类。对于第一类显式垃圾评论，主要是用基于规则的方法进行识别。而对于第二类隐式垃圾评论，本文采用LDA模型来对博文抽取隐

含主题信息，然后通过这些主题信息，使用基于主题的特征选取和基于主题的检索模型两种方法，进而发现垃圾评论，经实验验证，该方法是行之有效的。

在 Blog 这个开放性平台，评论者可以自由地发表言论，很多评论都是由诗歌、散文等隐式表达情感构成的，需要通过更深层次的方法来挖掘其隐式含义，进而判断与博文是否相关，最终判断该评论是垃圾评论还是相关评论。这也是本文下一步需要解决的工作。目前，博客中的博文和评论在研究方面的语料还不够丰富，因此，本文的语料是作者手工收集和整理，语料的丰富和校验工作还需进一步进行，同时情感词典也需要进一步完善，并且未进行情感词强度的考虑。以上情况都有待作进一步细致的研究。

参考文献

[1] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna. A Reference Collection for Web Spam[C]//ACM SIGIR Forum, 2006, 40(2): 11-24.

[2] Dennis Fetterly, Mark Manasse, Marc Najork, Spam, Damn Spam, and Statistics Using Statistical Analysis to Locate Spam Web Pages[C]//Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004, Paris, France, 1-6.

[3] 潘文锋. 基于内容的垃圾邮件过滤研究[D]. 北京：中科院计算技术研究所，2004.

[4] M. Hu and B. Liu. Mining and Summarizing Customer Reviews[C]//Proceedings of the tenth International Conference on Knowledge Discovery and Data Mining (KDD2004), Seattle, WA, USA, 2004: 167-177.

[5] N. Jindal and B. Liu. Product Review Analysis [M]. Technical Report, The University of Illinois at Chicago, 2007.

[6] Nitin Jindal and Bing Liu, Opinion Spam and Analysis [C]//Proceedings of the International Conference on Web Search and Data Mining (WSDM2009), Palo Alto, California, USA, 2009: 219-230.

[7] N. Jindal and B. Liu. Analyzing and Detecting Review Spam[C]//Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), Omaha, Nebraska, USA, 2007: 547-552.

[8] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008，27(2)：180-185.

[9] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[10] 李文波，孙乐，黄瑞红，冯远勇，张大鲲. 基于 Labeled LDA 模型的文本分类新算法[C]//第三届全国信息检索与内容安全学术会议，苏州，2007.

[11] D. Blei and J. Lafferty, Correlated topic models [C]//Advances in Neural Information Processing Systems 18, MIT Press, Cambridge, MA. 2006.

[12] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai, Topic Sentiment Mixture: Modeling Facets and Opinions in Web logs [C]//Proceedings of the 16th international conference on World Wide Web (WWW 2007), Banff, Alberta, Canada: 171-180.

[13] Yue Lu, Chengxiang Zhai. Opinion Integration Through Semi-supervised Topic Modeling[C]//Proceedings of the 17th International Conference on World Wide Web (WWW 2008), Beijing, China: 121-130.

[14] 曹娟，张勇东，李锦涛，唐胜. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报，2008，31(10)：1780-1787.

[15] Xing Wei, W. B. Croft, LDA-based Document Models for Ad-hoc Retrieval[C]//Proceedings of the 29th SIGIR Conference, Seattle, Washington, USA, 2006: 178-185.

[16] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data [M]. Springer, 2007.

[17] Vapnik V., The Nature of Statistical Learning Theory [M]. New York: Springer, 1995.

[18] 中科院分词系统：http://ictclas.org[DB/OL].