面向搜索的微博短文本语义建模方法

寇菲菲 杜军平 石岩松 杨从先 崔婉秋 梁美玉 石 磊

(北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876)

摘 要 微博中包含大量具有时间、用户等信息的短文本数据,通过挖掘其语义信息来实现精准搜索已受到广泛关注.将传统的主题模型应用于微博短文本语义建模时通常会存在以下问题.一方面,微博的短文本会引起语义稀疏性;另一方面,由于传统的主题模型仅建模文档之间的信息,不能充分挖掘文档内部的上下文信息,因此其仅能捕获全局语义.针对以上问题,文中提出了面向搜索的微博短文本语义建模方法,该方法包含三部分:基于词向量的短文本扩展算法、基于扩展的微博主题模型和微博搜索.首先,所提扩展算法以具有局部语义的词向量为基础,通过计算单词间相似度对微博短文本进行扩展,以此缓解短文本的语义稀疏性并实现局部语义与全局语义的相互补充.其次,将扩展后的长文本作为所提主题模型的输入所提主题模型,以扩展后的长文本作为输入,通过建模双词进一步克服语义稀疏性,并同时利用微博多种特征(文本、时间、用户信息)来约束主题的生成过程从而提高短文本语义表示的质量.最后,基于生成的统一语义表示,可以计算短文本间相似度从而实现微博搜索.本文在真实的新浪微博数据集上进行了多组实验,对所提的微博短文本语义建模方法语义建模方法得到的语义表示进行了分析与评价并将其应用于微博搜索,实验结果验证了所提方法的有效性.

关键词 社交网络;微博;短文本;语义建模;搜索

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2020.00781

Microblog Short Text Semantic Modeling Method for Search

KOU Fei-Fei DU Jun-Ping SHI Yan-Song YANG Cong-Xian

CUI Wan-Qiu Liang Mei-Yu SHI Lei

(*Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract Microblogs contain lots of short text data with time and user information. It has received widespread attention to achieve accurate search by mining the semantics of Microblogs. When applying the traditional topic models to Microblog short text semantic modeling task, they usually will face the following issues. First, traditional topic modeling methods cannot deal with the problem of semantic sparsity that caused by the shortness of Microblogs. Second, since topic models only acquire semantic at document-level, they cannot mine the local semantic existing in contexts. Therefore, rough semantic representation will result in inaccurate search results. In order to obtain high-quality semantic representation and realize precise search, we propose a Microblog short text semantic modeling method for search (MSSMS), which contains three components: a short text expansion algorithm based on embedding vector, a microblog topic model based on expansion and Microblog search. The short text expansion algorithm aims to expand short text into long text. To realize this purpose, it utilizes the embedding vectors to construct

收稿日期:2018-09-04;在线出版日期:2019-07-08.本课题得到国家重点研发计划(2018YFB1402600)、中国博士后科学基金资助项目(2019M660564)、国家自然科学基金项目(61772083,61532006,61877006,61802028)和广西科技重大专项(桂科 AA18118054)资助.

寇菲菲,博士,研究方向为社交网络分析、数据挖掘、机器学习. E-mail: koufeifei000@126.com. **杜军平**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能、机器学习、模式识别. E-mail: junpingdu@126.com. **石岩松**,硕士研究生,主要研究方向为机器学习、信息检索. **杨从先**,硕士研究生,研究方向为机器学习和语义建模. **崔婉秋**,博士研究生,主要研究方向为数据挖掘、机器学习. **梁美玉**,博士,副教授,主要研究方向为信息搜索、数据挖掘. **石 磊**,博士研究生,主要研究方向为信息搜索、数据挖掘.

similar-word sets for each word in the short text. As the embedding vectors contain local semantics, by using the expanded long text as the input of the topic model, the local semantics contained in embedding vectors and the global semantics acquired through the topic model can be combined. Besides, as short texts have turned into long texts, the semantic sparsity of short text can be weakened. In the proposed Microblog topic model, to further alleviate the semantic sparsity of short text, we introduce the bi-term pattern, which assigns word-word pairs to share the same topic. In addition, the proposed Microblog topic model also models multiple characteristics (text, time, user information) simultaneously. This operation could further improve the quality of the generated semantic representation, for the reason that multiple characteristics can constrain the generation of topics. After that, we can acquire the document-topic distributions, topic-word distributions, topic-time Beta distributions, and topic-user distributions, as these multiple characteristics (text, time, user information) are all mapped into the topic semantic space, they can be seemed as the unified semantic representation. Based on the generated unified semantic representation, we can calculate the similarities between short texts. Through sorting these similarities, we can realize the precise Microblog search. Finally, to verify the effectiveness of the proposed MSSMS, we conduct extensive experiments on real-world datasets of Sina Weibo, and these experiments are divided into two categories. One is to evaluate the semantic modeling ability of the MSSMS, and the other is to apply the MSSMS into Microblog search. In order to comprehensively evaluate the semantic modeling ability of the MSSMS, we not only use objective evaluation metric to measure the topic coherence but also use subjective evaluation methods to access the quality of the generated semantic representation. The experimental results show that compared with the comparison algorithm, the semantic representation generated by the proposed MSSMS method has the highest quality, and the MSSMS method has the best semantic modeling ability. In addition, the microblog search experiment results also verify that the proposed MSSMS method can achieve accurate microblog search.

Keywords social network; Microblogs; short text; semantic modeling; search

1 引 言

近年来,微博已成为一种非常流行的社交网络平台,人们可以方便地在微博上发表意见,分享他们的所见所闻以及所想所做.许多事件通过用户的发布、转发与评论等操作而发展为热门话题[1].因此,通过微博搜索已成为一种趋势[2].为了实现精准搜索,需要准确理解搜索与待搜索数据的语义[3].主题模型是一种常用的语义建模方法[4-5],然而由于微博短文本通常不超过 140 个字符[6],将其用于微博语义建模时将会面临语义稀疏性的问题.因此,对微博短文本进行建模,并通过其生成的语义表示实现精准搜索是一个极具挑战性的任务.

为了增强语义建模效果,许多研究者对传统的主题模型进行了改进.然而现有的主题模型方法通常假设每个文档属于同一主题分布,在主题生成过程中没有考虑词序关系,不能充分挖掘上下文信息,因此通过传统主题模型仅能得到全局语义而不能得到局部语义[7].除主题模型以外,以 Word2vec[8] 为代表的词嵌入模型是另一类典型的语义建模方法,该方法在训练过程中通过挖掘上下文信息,可以得到数据的局部语义信息.现有研究成果表明将局部语义与全局语义进行结合,可以有效地提高短文本语义表示的质量[9-10].尽管如此,由于微博除包含文本信息外,还包含时间和用户等属性信息[11-12],因此,仅仅将局部语义与全局语义相结合是不够的.如何利用微博多种属性,得到高质量的语义表示,从而为实现精准搜索提供基础,仍是一项亟待解决的挑战.

针对短文本语义稀疏性,常用的解决方法可以分为两类,一类是短文本扩展方法,即通过将短文本扩展为长文本,来丰富短文本的语义.作者信息[12]、地理位置信息[13]、外部知识库[14]、词共现频率[15]以及单词语义相似度[16]等均可用来对短文本进行聚

合和扩展.然而,以属性信息(作者、地理位置)进行聚合的方式通常具有数据敏感性;以外部知识聚合的方式,引入的知识通常太笼统,不能精确地表示语义;以词共现频率或者单词语义相似性对语义进行扩展会引入无关词,虽然可以在一定程度上降低语义稀疏性,但是该类方法获得的语义表示的质量仍有待提高.

另一类方法是利用微博数据自身的属性信息来应对语义稀疏性.如 Cheng 等人[17]通过建模词对之间共享一个话题来增强语义学习效果.Wang 等人[11]通过将时间建模为 beta 分布,以时间来约束话题的产生.Cai 等人[13]建模了主题与时间、主题与标签的关系,通过建模多种数据本身的信息对话题的产生形成一种约束,进而提高语义学习效果.这类方法对于语义质量的提升可以起到一定的作用,然而并不能在本质上解决短文本的语义稀疏性.

采用语义模型得到的语义表示进行搜索已被广泛采用[18],然而将现有的语义模型应用于社交网络短文本数据时,其得到的语义表示由于质量较低而不足以实现精准搜索.针对以上挑战,本文提出了一种面向搜索的微博短文本语义建模方法(Microblog Short Text Semantic Modeling Method for Search,MSSMS).该方法主要包含三部分:基于词向量的短文本扩展方法、基于扩展的微博主题模型和微博搜索.基于词向量的短文本扩展方法以具有局部语义信息的词向量为基础,结合单词的语义信息与位置信息将原始社交网络短文本扩展为长文本.基于扩展的微博主题模型,以扩展后的长文本为输入,引入了双词模式[17],通过结合短文本扩展和双词模式来解决语义稀疏性.另外,所提基于扩展的微博主题模型在建模双词主题的同时建模了时间特征与用户特征,以多种特征约束主题模型的生成过程,从而得到高质量的语义表示.为实现微博搜索,首先利用所提主题模型生成的语义表示来计算查询项与待搜索数据之间的相似性,通过对相似性进行排序,最终实现精准的微博搜索.

本文的主要贡献如下:

(1)提出了一种面向搜索的微博短文本语义建模方法 MSSMS,可以有效地建模微博短文本语义,实现微博的精准搜索.

(2)提出了一种基于词向量的短文本扩展算法,可以实现短文本的精确扩展,通过将其与主题模型结合可以缓解语义稀疏性并能够在全局语义中引入局部语义,从而实现局部语义与全局语义的相互补充.

(3)设计了一种基于扩展的微博主题模型,与传统的 LDA(Latent Dirichlet Allocation)主题模型不同,该模型以扩展后的长文档作为输入,引入了双词模式,同时建模了微博多种特征(文本、时间和用户),通过多种因素共同作用获取了高质量的语义表示.

2　相关工作

许多研究人员致力于短文本语义建模,从而为实现微博精准搜索提供基础.这些工作主要分为三类:(1)对短文本进行扩展,将短文本扩展为长文本,以解决语义稀疏性;(2)二是局部语义与全局语义相结合的语义表示方法,通过将局部语义与全局语义相结合提高语义表示质量;(3)是充分利用微博数据的内部信息(如文本、时间、用户),来约束话题的生成过程.在本节中,将对微博短文本语义建模方法进行回顾.

2.1　短文本扩展方法

解决短文本语义稀疏性的一个重要方法是对短文本进行扩展,通过将短文本扩展为长文本,克服其语义稀疏性.短文本扩展的方法主要可以分为两类方法.一类是通过公共属性对短文本进行聚合.如利用作者[11]或地理位置[13]等信息将具有公共属性的短文本聚合为长文本.这类方法通常具有一定的数据敏感性,如果所用数据集中的文本内容与作者或地理位置等公共属性信息关联不大,则效果不明显.另一类方法则是通过借用外部知识库对短文本进行扩展.例如,Chen 等人[14]使用 WordNet 的外部知识库和词汇的词义关系,引入与短文本相关的外部知识,从而对短文本进行扩展.然而由于微博文本通常非常口语化,而采用这类方法引入的外部知识通常太过笼统,所以并不能精确的表示语义信息[19].除此之外,还有学者将具有单词共现关系[15]或者语义相似性[16]的单词聚合在一起,实现短文本向长文本的转变.这种方法对数据具有普遍适用性,在一定程度上可以解决语义稀疏性,然而仅仅使用词频或者语义相似性对短文本进行扩展,会引入一些无关或者关联性不太大的单词,因此这类方法的语义建模能力仍有待提高.

2.2　局部语义与全局语义相结合的语义表示方法

除主题模型外,以 Word2vec[8]为代表的深度学习方法同样广泛用于语义表示.该类方法在训练过程中充分利用了单词间的关系,因此通过其得到的

词向量可以看作是一种局部语义[20]. 近年来,研究发现将局部语义与全局语义结合,可以显著提高语义表示质量. 如 TWE(Topical Word Embeddings)算法[7]设计了三种不同的主题嵌入方式,通过在词向量的训练过程中引入主题词,来实现主题全局语义与词向量局部语义的结合. GPU-DMM(GPU-Dirichlet Multinomial Mixture)算法[9]则通过将 Word2vec 产生的词向量引入到主题的生成过程中,实现局部语义与全局语义的相互补充. DREx(Distributed Representation-Based Expansion)算法[16]利用单词的向量表示计算单词间的语义相似性,通过语义相似词对短文本进行扩展,并将扩展后的文本作为主题模型的输入. 由于主题模型设定整个文档共享同一主题分布,因此获得的语义是一种全局语义. 该方法在主题的生成过程中引入了具有局部语义的相似词,同样简单有效地将局部语义引入到了全局语义中. 将主题模型与词嵌入方法相结合,可以实现局部语义与全局语义的相互补充,进而获取高质量的语义表示. 然而此类方法没有利用社交网络的多种属性,仅仅将局部语义与全局语义相结合得到的社交网络短文本语义表示仍不能实现精准搜索.

2.3 微博信息建模

微博数据包含文本、时间、用户、标签等多种信息. 很多学者通过对社交网络数据的单种或者多种信息建模来约束话题的生成过程,从而提高所生成话题的质量. 典型的方法如 Yan 等人[21]提出的双词话题模型,假设一个滑动窗口内出现的两个单词共享一个话题,以此对话题的生成过程进行约束,可以有效地缓解语义稀疏性,提高话题的语义表示质量. Wang 等人[11]在建模主题单词分布的同时建模了主题时间分布,实验结果表明加入时间信息可以提高模型的语义建模能力. Zhou 等人[22]同时建模文本、时间、地理位置信息,提出了一种 LTT(Location-Time Constrained Topic)模型. Cai 等人[13]则在建模主题单词分布的同时,建模了时间、标签、图像信息. 通过建模多种数据信息提高话题质量,然而此类方法仅约束话题的生成过程,并不能在本质上解决语义稀疏性.

3 面向搜索的微博短文本语义建模方法的提出

3.1 问题描述与定义

为了获取高质量的微博短文本语义表示并实现微博精准搜索,本文提出了面向搜索的微博短文本语义建模方法 MSSMS,其框架如图 1 所示. 所提 MSSMS 主要包含三部分:基于词向量的短文本扩展、基于扩展的微博主题模型和微博搜索.

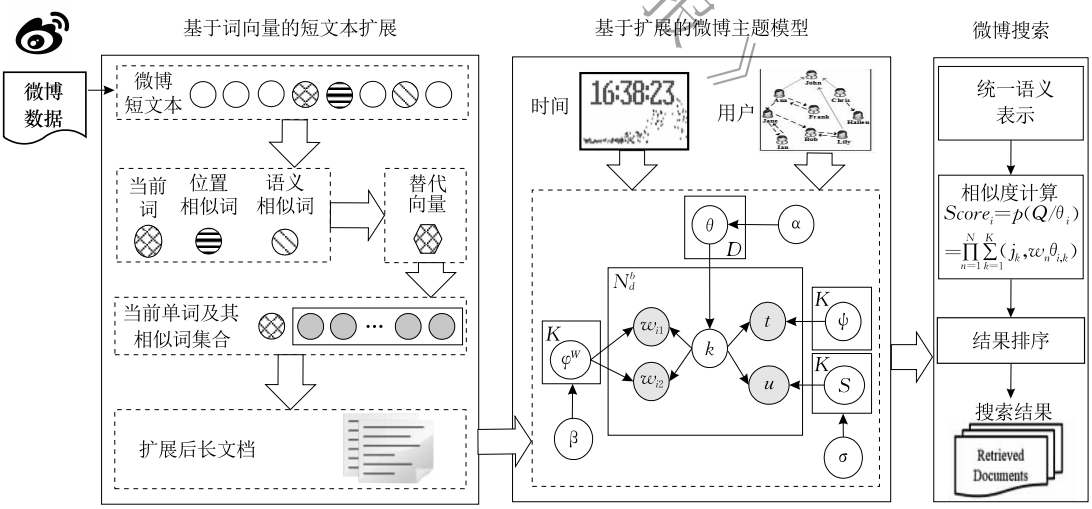


图 1 面向搜索的微博短文本语义建模方法 MSSMS 框架

首先获取微博短文本中各单词的词向量,采用基于词向量的短文本扩展方法对微博短文本进行扩展. 利用上下文中单词间语义相似性和位置相似性获取单词的相似词集合. 由于语义相似性通过具有局部语义的词向量计算得到,且位置相似性是通过单词的位置关系得到的,因此,采用该相似词集合对微博短文本进行扩展时可以引入文本的上下文局部语义信息.

然后,将扩展后的长文档、归一化的时间信息和用户信息作为基于扩展的微博主题模型的输入数据,通过建模微博多种特征(双词模式、时间特征、用户特征)获取语义表示. 所提主题模型采用扩展后的

长文档作为输入，可以有效地削减语义稀疏性，并实现局部语义与全局语义的结合. 此外，由于引入了双词模型，且同时建模多种特征(时间特征和用户特征)，可以进一步克服语义稀疏性并生成高质量的语义表示.

根据所提主题模型生成的语义表示，可以计算查询项与每条待搜索文档之间的相似度，根据相似度进行排序，最终获得搜索结果. 为了便于描述所提MSSMS，首先给出以下定义.

定义 1.　微博短消息 m. 每条微博短消息 m 可记作一个三元组 (p_m, t_m, u_m)，表示微博中的用户 u_m 在时刻 t_m 发布了具有文本 p_m 的短消息 m. p_m 表示文本，t_m 表示时间戳，u_m 表示用户.

定义 2.　单词的替代向量 $\boldsymbol{r}_i$. 在对预处理后的短文本进行扩展时，采用 Word2vec 对语料预训练，得到每条文本 p_m 中的每个单词 $w_i, i\in[1,\cdots,N_m]$ 的向量表示 $\boldsymbol{v}_i$. 需要查找 p_m 中与当前词 w_i 最具语义相似性和位置相似性的单词. 在本文中将 w_i 的后续邻接词作为其位置相似性单词，以 $\boldsymbol{s}_i$ 和 $\boldsymbol{c}_i$ 分别表示语义相似词和位置相似词的词向量，对 $\boldsymbol{v}_i, \boldsymbol{s}_i, \boldsymbol{c}_i$ 加权求和，得到替代向量 $\boldsymbol{r}_i$. 以 $\boldsymbol{r}_i$ 代替 $\boldsymbol{v}_i$ 的好处在于通过语义相似性和位置相似性对单词的语义进行补充与约束，从而可以查找到更贴切的相似词集合.

定义 3.　扩展后长文档 d_m. 对每条微博短消息 m 中的文本 p_m 进行扩展，用 d_m 表示经扩展后长文档.

定义 4.　统一语义表示. 主题模型经过多次迭代达到收敛，从而可以输出多项主题分布：文档主题分布、主题单词分布、主题时间分布和主题用户分布. 由于微博的多种特征(文本、时间和用户)均通过主题模型映射到了公共主题语义空间中，因此我们将这些分布称为统一语义表示.

3.2　基于词向量的短文本扩展算法的提出

针对如定义 1 所示的微博短消息中的短文本 p_m，为缓解其语义稀疏性并在主题模型中引入局部语义，提出了一种基于词向量的短文本扩展算法. 该方法旨在为微博短文本中的单词构造相似词集合，通过加入相似词将原始的微博短文本扩展为长文档.

在构造相似词集合的过程中，为了避免无关词的引入，我们同时考虑了单词的语义相似性和位置相似性. 即首先在当前单词所在文档中找到与其具有语义相似性和位置相似性的单词，并将这些单词与当前单词组合形成如定义 2 所描述的替代向量，然后通过该替代向量查找当前单词的相似词集合.

为了获取单词间的语义相似性，需要获取每个单词的向量表示，通过计算单词向量表示之间的余弦距离得到单词之间的语义相似度，最终筛选出当前词的语义相似词. 在以卷积神经网络(Convolution Neural Network，CNN)、循环神经网络(Recurrent Neural Network，RNN)、Word2vec 等为代表的深度学习的向量表示方法中，Word2vec 是使用尤其广泛的词向量表示方法，可以实现大规模语料的向量化表示[16]. 并且该方法在训练过程中充分考虑了单词间的上下文信息，得到的词向量具有局部语义信息[20]. 因此在本文中采用 Word2vec 对文本语料进行预训练，获取单词的向量表示.

然而由于单词的词向量不能随着语境的变化而变化，因此单纯采用语义相似性选择相似词，容易引入大量无关词. 为减少无关词的引入，在为每个单词查找相似词集合时，同时考虑了单词间的位置相似性. 在同一个文档中相邻出现的单词往往具有相同的主题，本文将这种单词间的相邻关系定义为位置相似性. 如定义 2 所描述，在本文中将每个单词的后续邻接词作为其位置相似性单词. 通过将当前词、其语义相似词和其位置相似词的词向量加权求和得到其替代向量，以替代向量为每个单词查找相似词集合. 最终通过逐个为原始文本中每个单词查找相似词，以所有单词的相似词集合对原始短文本进行扩展实现短文本向长文本的转变，形成如定义 3 所示的扩展后长文档. 我们在算法 1 中描述了基于词向量的短文本扩展算法.

算法 1.　基于词向量的短文本扩展算法.

输入：经过数据预处理后的微博短文本 p_m

输出：扩展后的长文档 d_m

1. FOR 每条短文本 $p_m, m\in[1,\cdots,M]$ DO
2. IF $N_m=1$，返回前 L 个相似词
3. ELSE
4. FOR 每个单词 $w_i, i\in[1,\cdots,N_m]$ DO
5. 获取其词向量 $\boldsymbol{v}_i$
6. 采用余弦距离计算在本条短文本中与当前词语义最相似词，将其词向量记作 $\boldsymbol{s}_i$.
7. 获取当前词 w_i 的后续邻接词的词向量 $\boldsymbol{c}_i$
8. 计算当前词的替代词向量 $\boldsymbol{r}_i$：$\boldsymbol{r}_i=(1-\lambda_1-\lambda_2)v_i+\lambda_1 s_i+\lambda_2 c_i$
9. 采用余弦距离计算得到与 $\boldsymbol{r}_i$ 距离最小的前 L 个相似词.
10. END FOR
11. END IF
12. 返回经过相似词集合扩展的长文档 d_m
13. END FOR

计算机学报

3.3 基于扩展的微博主题模型的提出

如定义 1 所示，微博信息中主要包含文本、时间、用户信息. 基于扩展后的长文档和微博的多种特征，提出了一种基于扩展的微博主题模型. 基于扩展的微博主题模型如图 2 所示，其中的符号及描述如表 1 所示.

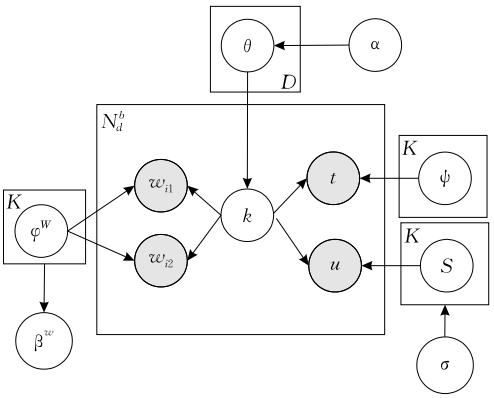


图 2 基于扩展的微博主题模型

表 1 符号及其描述

符号	描述
k,w,b	话题、单词、双词
b_i,w_{i1},w_{i2}	第 i 个双词和其中的两个单词
t_i,u_i	第 i 个双词的时间信息、用户信息
D	文档总数
N_d^b	第 d 个文档中的双词总数
K,W,B	话题总数、单词字典大小、双词字典大小
α,β,σ	狄利克雷参数
φ^W	主题-单词分布
ψ	时间贝塔分布
S	主题-用户分布

基于扩展的微博主题模型通过以扩展后的长文本作为输入来结合局部语义，并在一定程度上缓解了短文本的语义稀疏性. 由于双词共享同一主题模式已经被广泛证明可以削弱语义稀疏性[23]，因此在模型的生成过程中通过假设双词共享同一主题引入了双词模式. 此外，为进一步提高语义表示质量，所提主题模型充分挖掘微博多种特征信息，同时建模了时间信息与用户信息，因此，该模型可以将局部语义和全局语义相结合，并在最大程度上克服语义稀疏性，从而得到高质量的语义表示.

3.3.1 所提主题模型的生成过程

假设共有 K 个主题，每条扩展后的长文档共享一个主题分布 θ_d. 提取同一个上下文窗口内的双词 b_i，并假设其共享同一个主题 k. 在微博多特征主题模型中，每个主题 k 同时融合了主题-单词分布 φ、主题-时间分布 ψ、主题-用户分布 S 三种主题分布. 通过对多种特征的共同建模，可以有效地提高生成话题的质量. 给定一条微博短消息，以其扩展后的长文档作为文本输入，同时输入其时间信息、用户信息，所提基于扩展的微博主题模型的生成过程如下：

对每个聚合后文档 $d=1,\cdots,D$，根据参数 α，抽取文档的主题分布 $\theta_d\sim Dirichlet(\alpha)$.

对每个话题 $k=1,\cdots,K$，抽取主题-时间分布 $Beta(\psi_k)$，分别根据参数 β,σ 抽取主题-单词分布 $\varphi_k\sim Dirchlet(\beta)$ 和主题-用户分布 $s_k\sim Dirchlet(\sigma)$.

对文档中的每个双词 b_i：

(1) 根据以 θ_d 为参数的多项式分布，抽取双词的主题：$k\sim Multi(\theta_d)$；

(2) 根据以 φ_k 为参数的多项式分布，抽取每个单词：$w_{i1},w_{i2}\sim Multi(\varphi_k)$；

(3) 根据以 ψ_k 为参数的贝塔分布，抽取双词的时间戳；

(4) 根据以 s_k 为参数的多项式分布，抽取双词的用户信息.

3.3.2 所提主题模型的推理

在本文提出的基于扩展的微博主题模型中，需要对四个参数 θ、φ、ψ、S 和一个潜在变量 k 进行推断. 鉴于 Collapsed Gibbs 采样方法的简洁有效性，因此使用其来估计参数. 所提主题模型的 Gibbs 采样算法如算法 2 所示.

算法 2. Gibbs 采样算法.

输入：短文本 p_m、归一化的时间戳、用户 ID、主题数 K、先验参数 α、β、σ、迭代次数 N_{iter}

输出：文档-主题分布、主题-单词分布、主题-时间分布、主题-用户分布

```
对短文本随机初始化主题设置
FOR 迭代次数=1 TO N_iter DO
  FOR d=1 TO D DO
    FOR 每个双词 b_i, i∈[1,…,N_d^b] DO
      根据式(3)抽取主题
      更新参数 n_d^k, n_k^u, n_k^{w_i1}, n_k^{w_i2}
    END FOR
  END FOR
  FOR 主题=1 TO K DO
    根据式(5)更新 ψ_k
  END FOR
END FOR
根据式(4)计算各项分布的后验估计.
```

在基于扩展的微博主题模型中需要为每个双词的话题进行采样，采样公式如下：

$$p(k_i|K_{\neg i},T,B,U)\propto\frac{p(K,T,B,U|\Theta)}{p(K_{\neg i},T_{\neg i},B_{\neg i},U_{\neg i},|\Theta)} \tag{1}$$

其中，Θ 表示所有的参数集合，$\neg i$ 表示除 i 之外. 为了推导上述条件概率公式，计算联合概率分布：

$$\begin{aligned} p(K,B,U,T|\Theta) &= p(K|\alpha)p(B|K,\beta)\times \\ &\quad p(U|K,\sigma)p(T|K,\psi) \\ &= \prod_{d=1}^{D}\frac{\Delta(\boldsymbol{n}_d^K+\alpha)}{\Delta(\alpha)}\prod_{k=1}^{K}\frac{\Delta(\boldsymbol{n}_k^B+\beta)}{\Delta(\beta)}\times \\ &\quad \prod_{k=1}^{K}\frac{\Delta(\boldsymbol{n}_k^U+\sigma)}{\Delta(\sigma)}\prod_{i=1}^{N^b}p(t_i|\psi_{k_i}) \end{aligned} \tag{2}$$

其中，$\boldsymbol{n}_d^K$ 是一个 K 维向量，每一维元素表示每个主题出现的次数. $\boldsymbol{n}_k^B$ 是一个 B 维向量，每一维元素表示每个双词属于主题 k 的次数. $\boldsymbol{n}_k^U$ 是一个 U 维向量，每一维元素表示每个用户属于主题 k 的次数.

结合式(1)和式(2)，可以推导得到采样式(3)：

$$\begin{aligned} p(k_i=k|K_{\neg i},T,B,U)\propto & \\ \frac{n_{k,\neg i}+\alpha}{\sum_{k=1}^{K}n_{k,\neg i}+K\alpha}\times\frac{(1-t_i)^{\psi_{k1}-1}t_i^{\psi_{k2}-1}}{B(\psi_{k1},\psi_{k2})}\times & \\ \frac{(n_{k,\neg i}^{w_{i1}}+\beta)(n_{k,\neg i}^{w_{i2}}+\beta)}{(\sum_{w=1}^{W}n_{k,\neg i}^{w}+W\beta)\cdot(\sum_{w=1}^{W}n_{k,\neg i}^{w}+1+W\beta)}\times & \\ \frac{(n_{k,\neg i}^{u_i}+\sigma)}{(\sum_{u=1}^{U}n_{k,\neg i}^{u}+U\sigma)} & \end{aligned} \tag{3}$$

迭代地执行上述采样规则直到稳定状态，可以获得潜在变量的值. 通过式(4)和式(5)进行参数估计，得到如定义 4 所述的统一语义表示：

$$\theta_{d,k}=\frac{n_d^k+\alpha}{\sum_{k=1}^{K}n_d^k+K\alpha},\varphi_{k,w}=\frac{n_k^w+\beta}{\sum_{w=1}^{W}n_k^w+W\beta},s_{k,u}=\frac{n_k^u+\sigma}{\sum_{u=1}^{U}n_k^u+U\sigma} \tag{4}$$

$$\psi_{k1}=\bar{t}_k\left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2}-1\right),\psi_{k2}=(1-\bar{t}_k)\left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2}-1\right) \tag{5}$$

采用矩量法对话题时间 Beta 分布的两个参数 ψ_{k1} 和 ψ_{k2} 进行估计，$\bar{t}_k$ 表示均值，s_k^2 表示方差.

3.4　微博搜索

根据待搜索文档的文档主题分布以及主题单词分布可以计算得到每个待搜索文档生成输入查询的概率，概率越大则表示与查询输入的相似度越高. 以 $Q=\{w_1,w_2,\cdots,w_n\}$ 表示输入查询，θ,φ 分别表示统一语义表示中的文档主题分布和主题单词分布. 每条待检索文档与查询之间的概率生成相似度可通过式(6)计算得到.

$$Score_i=p(Q/\theta_i)=\prod_{n=1}^{N}p(w_n/\theta_i)=\prod_{n=1}^{N}\sum_{k=1}^{K}(\varphi_{k,w_n}\theta_{i,k}) \tag{6}$$

采用如算法 3 所示的基于概率生成相似度的微博搜索算法，可以得到最终的微博搜索结果.

算法 3.　基于概率生成相似度的微博搜索算法.

输入：搜索项 Q、待搜索文档-主题分布 θ，主题-单词分布 φ

输出：搜索结果

1. 将搜索项 Q 中的每个单词 w_n 转换为模型中的字典编码
2. FOR 每条待搜索文档 DO
3.　　找到其文档主题分布 θ_i
4.　　根据式(6)计算与搜索项的相似度
5.　　根据相似度大小进行排序
6.　　返回搜索结果
7. END FOR

3.5　MSSMS 复杂度分析

本部分从时间复杂度和空间复杂度两个方面出发对 MSSMS 的语义建模复杂度进行分析，并将其与典型语义建模方法 LDA 的复杂度进行比较. 首先定义如下变量：D 表示文档的数量，M 表示文档的平均长度，V 表示字典大小，K 表示话题数量，N_{iter} 表示话题聚类时的迭代次数，U 表示用户数，N_{vd} 表示词向量的维度. 因此，根据文献[23]，LDA 的时间复杂度可以表示为：$O(N_{iter}DMK)$，其空间复杂度可以用需要存储的变量数进行表示并记作：$DK+VK+DM$.

由于所提方法 MSSMS 的语义建模是由基于词向量的短文本扩展算法和基于扩展的微博主题模型两部分完成的，因此，需要对这两部分的复杂度进行分析.

基于词向量的短文本扩展部分的时间复杂度分析如下. 对于一个单词，获取其替代向量的时间复杂度为 $O(M)$，在字典中寻找替代向量的相似词集合时，需要对字典进行遍历，时间复杂度为 $O(V)$，因此，要对 D 条文档中的 M 个单词进行词扩展时，其时间复杂度为 $O(DM^2V)$.

基于扩展的微博主题模型部分，使用了大小为 l 的滑动窗口来生成双词[21]，因此双词的总数量可以表示为 $Dl(l-1)/2$，遍历所有双词的时间复杂度为 $O(Dl^2)$. 在模型的每次迭代过程中，需要对每个话题进行采样，采样每个话题时，需要对所有双词进行扫描，因此一个迭代的时间复杂度为 $O(Dl^2K)$. 当对基于扩展的微博主题模型进行 N_{iter} 次迭代时，其时间复杂度为 $O(N_{iter}Dl^2K)$. 因此 MSSMS 的时间复杂度为上述两部分的总和 $O(N_{iter}Dl^2K+DM^2V)$.

MSSMS 的空间复杂度分析如下. 在基于词向

《计算机学报》

量的短文本扩展部分，需要存储每个单词的词向量，对应的空间复杂度为 VN. 在基于扩展的主题模型部分需要存储的参数有双词信息、文档主题矩阵、主题单词矩阵、主题用户矩阵以及主题时间分布的参数，分别对应的空间复杂度为 $Dl(l-1)/2$、DK、KV、KU、K. 因此 MSSMS 的空间复杂度为 $VN_{vd}+Dl(l-1)/2+DK+KV+KU+K$.

将 MSSMS 的复杂度与 LDA 的复杂度进行对比可以发现，MSSMS 在时间复杂度和空间复杂度上均比 LDA 高. 这是因为 MSSMS 以增加时间和内存消耗为代价，通过对短文本进行扩展并建模时间特征与用户特征，来减少语义稀疏性，从而提高语义表示质量.

4 实验结果与分析

4.1 实验设置

为验证面向搜索的微博短文本语义建模方法 MSSMS 的有效性，本文设置了两类实验. 一类是语义质量表示评价，分为主观评价和客观评价. 一类是实际的搜索应用，即将语义表示应用于实际的微博搜索，以搜索性能衡量其语义建模能力. 由于在所提算法中包含了四种关键因素（短文本扩展、双词、时间、用户）. 因此，除了现有的短文本语义建模方法之外，实验中还使用了所提方法的简化版方法作为对比方法对所提方法中不同因素的作用进行验证.

4.1.1 数据集

采用新浪微博数据作为实验数据集. 数据集信息如表 2 所示. 数据集中文档长度的分布如图 3 所示，通过图 3 可以看出，新浪微博文本长度主要分布在 11～15 个字之间，单条文本的长度较短，具有很强的语义稀疏性.

表 2 数据集描述

数据集	微博数	时间周期	用户数
数据集 1	72053	2014.01.01～2014.01.31	6445
数据集 2	21437	2012.03.01～2012.10.31	6897

数据集 1. 以随机用户 ID 为依据，采集用户在一段时间内发布的微博.

数据集 2. 以关键字“暴雨”为输入，采集相关微博.

通过数据集 1 与数据集 2，分别验证用户特征和时间特征在 MSSMS 中作用. 选择“暴雨”作为关键词是因为暴雨灾害类事件通常发生在雨季，是一种时间敏感的事件类别.

为了更有效地进行实验，对采集的数据进行以下预处理操作：去除转发微博，保留原创微博；去除少于 10 个字以及多余 100 字的微博；过滤广告；去除不活跃用户（时间周期内所发微博数少于 5 条的用户）；对特别活跃用户的微博进行随机采样，保留其 20 条微博；对微博文本进行分词，除去停用词和低频词.

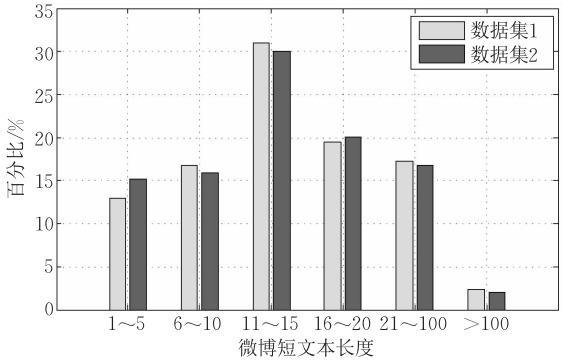


图 3 新浪微博文本长度（字数）分布图

4.1.2 对比算法

UCT[11]：通过利用社交网络用户特征，理解短文本的语义.

PTM[24]：通过设定伪文档数，将短文本自动聚合为长伪文档，从而削弱语义稀疏性.

BTM[17]：设定整个语料库共享一个主题分布，建模了双词共享同一个话题模式.

WNTM[15]：将根据词频将短文本扩展成长文本进行语义建模.

GPU-DMM[9]：在该主题模型的生成过程中，融合了词向量，从而使的局部语义与全局语义相互补充.

为了进一步验证所提 MSSMS 的有效性，我们设计了 MSSMS 的简化版本，MSSMS 的简化版本如下：(1) MSSMS^{-E} 表示不对短文本进行扩展并以原始短文本作为输入的方法，将经过所提短文本扩展算法得到的长文本作为所提主题模型的输入，可以将局部语义引入到全局语义中，实现局部语义与全局语义的相互补充. 因此，通过 MSSMS 与 MSSMS^{-E} 之间的对比，可以分析将局部语义与全局语义相互补充相比仅使用全局语义的优越性，从而验证基于词向量的短文本扩展算法的有效性；(2) MSSMS^{-B} 表示不建模双词共享同一主题的简化版算法，即去掉双词模式的的简化算法，因此，通过 MSSMS 与 MSSMS^{-B} 之间的对比，可以分析引入双词模式的优越性；(3) MSSMS^{-U} 表示不建模用户特征的情形，将 MSSMS 与 MSSMS^{-U} 对比，可以验证建模用户特征的有效性；(4) MSSMS^{-T} 表示不建模时间特征的情形，将 MSSMS 与 MSSMS^{-T} 对比，可以验证建

模时间特征的有效性.此外,由于双词模式、用户特征和时间特征是所提基于扩展的微博主题模型的组成因素,因此通过将 MSSMS 与 MSSMS^{-B}、MSSMS^{-U}、MSSMS^{-T}对比,可以有效验证所提主题提模型的有效性.

4.1.3　评价指标

(1)语义表示质量

针对所提方法产生的语义表示质量,本文分别从主观与客观两种角度进行了评价.对于主观评价,我们挑选了几个主题,并列出了所提算法与对比算法中每个主题下的前 10 个概率最大的单词,单词与主题相关度越高且排名越靠前则说明语义表示质量越好.

对于客观评价,由于点对互信息(Pointwise Mutual Information,PMI)[17]能够反映主题的语义一致性与可理解性,已被广泛用作主题模型的话题质量客观评价指标评价.因此,本文选择 PMI 指标对语义表示质量进行客观评价,对每个主题计算其前 N 个单词的 PMI 值,N 取 5、10、20.

当给出一个主题 k 和这个主题上前 N 个可能的单词时,以 $p(w_i)$ 表示单词出现的概率,以 $p(w_i,w_j)$ 表示两个词共同出现的概率,主题 k 的 PMI 值可以通过式(7)计算得到:

$$PMI(k)=\frac{2}{N(N-1)}\sum_{1\le i\le j\le N}\log\frac{p(w_i,w_j)}{p(w_i)p(w_j)} \tag{7}$$

(2)搜索任务

在搜索任务中,采用如文献[25]的方法,邀请了 10 名实验室人员作为志愿者对搜索结果与检索项之间的相关度进行评价.通过略掉算法名称,将不同算法得到的搜索结果以匿名的方式递交给志愿者,从而降低主观意愿的影响.对所有志愿者的评价结果进行平均,从而进一步避免人为因素的影响.采用两种搜索的常用评价指标:平均准确率均值(Mean Average Precision,MAP)[26]和归一化折损累积增益(Normalized Discounted Cumulative Gain,NDCG)[27]对搜索性能进行评价,指标取值越大表示搜索性能越好.MAP 中的平均精度(AP)可以通过式(8)计算得到:

$$AP=\frac{1}{R'}\sum_{r=1}^{R}prec(r)\delta(r) \tag{8}$$

在式(8)中,R 表示所有返回的搜索结果的数量,R'表示与查询相关的返回的搜索结果的数量,$prec(r)$表示第 r 个返回的搜索结果的精度,$\delta(r)$是指标函数,其中 1 表示结果具有相关性,0 表示相关性无相关性.NDCG 定义如下:

$$N(n)=Z_n\sum_{j=1}^{n}(2^{r(j)}-1)/\log(1+j) \tag{9}$$

其中 Z_n是归一化因子,$r(j)$是第 j 个搜索结果的相关性评分.

4.1.4　参数设置

MSSMS 主要的参数有:超参数 α 和 β、主题数 K、迭代次数 N_{iter}.仿照文献[15],分别将超参数 α 和 β 设置为 $50/K$ 和 0.01,主题数 K 依次设置为 10、30、50,迭代次数设置为 2000 次.在短文本扩展过程中,为了生成合适的替代向量,对语义相似词与位置相似词设置相同的权重系数,并将当前词本身的权重系数设置为略高于两者的权重系数,因此根据经验将 λ_1 和 λ_2 均设定为 0.3.为了获取适中长度的扩展后文档,根据本文所用数据集的文档平均长度,将相似词集合长度 L 设置为 3.

4.2　微博语义表示质量分析

4.2.1　MSSMS 与对比算法的 PMI 值比较

采用 PMI 衡量微博语义表示质量,为 MSSMS 与对比算法设置相同主题数,并应用在数据集 1 和数据集 2 上.通过实验可以得到每个主题下的单词集合,分别对每个主题中的前 5、前 10 和前 20 个单词的平均 PMI 值进行计算,从而对不同算法的语义表示质量进行客观的衡量和比较.数据集 1 和数据集 2 上的实验结果如表 3 所示.另外,为分析算法在不同数据集上的稳定性,对数据集 1 与数据集 2 的 PMI 值求平均,并将其以柱状图的形式进行展示,实验结果如图 4 所示.

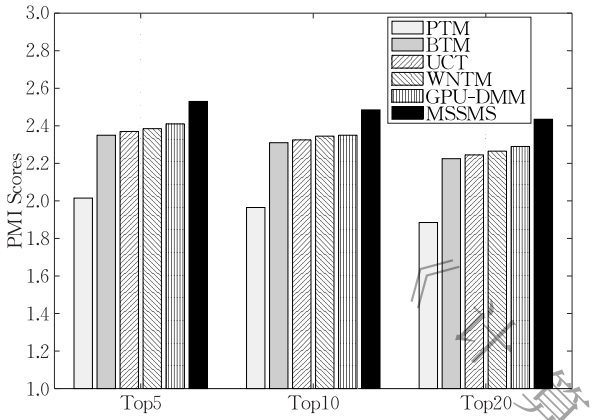
通过表 3 和图 4 可以分析得到不同算法在主题数下的鲁棒性.当主题数较小时,MSSMS 与对比算法均表现不佳,这是因为主题数较小时不能充分表达语义信息.当主题数增加到 30 时,MSSMS 与对比算法均表现良好.当主题数增加到 50 时,UCT、WNTM 与 GPU-DMM 的话题质量相比主题数为 30 时分别平均下降 5.16%、4.62%和 4.46%,这是因为增加主题数会使得语义稀疏性更强.

当主题数增加到 50 时,PTM 与 BTM 的话题质量相比主题数为 30 时略有提升,分别平均提升 0.23%和 2.07%,这是因为 PTM 通过建立伪文档削弱了语义稀疏性,BTM 通过设定整个语料共享同一个主题分布,并且建模双词共享同一主题模式,增强了语义空间密度.当主题数增加到 50 时,MSSMS 话题质量相比主题数为 30 时有较大程度的提升,平均提升 4.36%,这是因为 MSSMS 在削弱短文本语义稀疏性的基础上,建模了双词共享同一主题的模式,通过该模式可以有效地增强语义空间密度,从而可以生成高质量的话题.

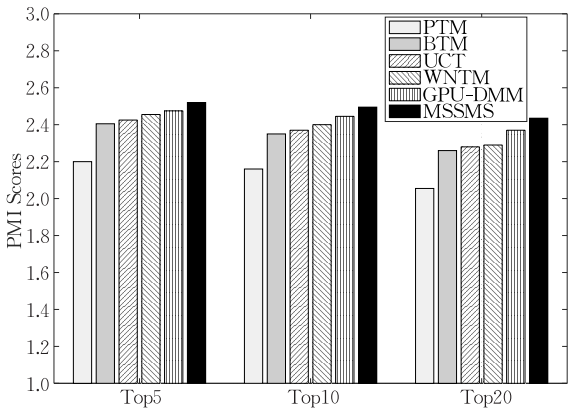
《计算机学报》

表 3 MSSMS 与对比算法的 PMI 值比较

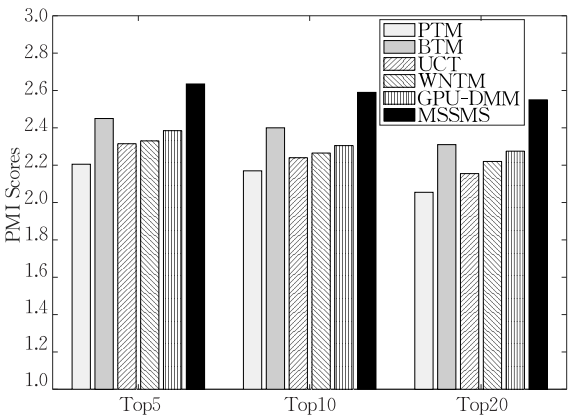
数据集	主题数 算法	10			30			50		
		Top5	Top10	Top20	Top5	Top10	Top20	Top5	Top10	Top20
数据集 1	PTM	1.91	1.87	1.81	2.13	2.09	1.97	2.14	2.10	1.97
	BTM	2.31	2.28	2.22	2.39	2.32	2.26	2.44	2.37	2.31
	UCT	2.33	2.29	2.24	2.41	2.33	2.28	2.31	2.24	2.19
	WNTM	2.35	2.31	2.25	2.45	2.37	2.27	2.32	2.26	2.21
	GPU-DMM	2.39	2.31	2.29	2.47	2.44	2.39	2.40	2.32	2.30
	MSSMS	2.51	2.46	2.41	2.51	2.48	2.42	2.64	2.57	2.52
数据集 2	PTM	2.12	2.06	1.96	2.27	2.23	2.14	2.27	2.24	2.14
	BTM	2.39	2.34	2.23	2.42	2.38	2.26	2.46	2.43	2.31
	UCT	2.41	2.36	2.25	2.44	2.41	2.28	2.32	2.24	2.12
	WNTM	2.42	2.38	2.28	2.46	2.43	2.31	2.34	2.27	2.23
	GPU-DMM	2.43	2.39	2.29	2.48	2.45	2.35	2.37	2.29	2.25
	MSSMS	2.55	2.51	2.46	2.53	2.51	2.45	2.63	2.61	2.58



(a) 主题数为10时的PMI值比较



(b) 主题数为30时的PMI值比较



(c) 主题数为50时的PMI值比较

图4 MSSMS 与对比算法的在两个数据集上的平均 PMI 值比较

本文提出的 MSSMS 话题质量表现最好，且随着主题数的变化表现一直稳定，这是因为 MSSMS 通过短文本扩展克服了语义稀疏性，引入了双词共享同一主题模式，并且建模了多种特征，最大程度地提高了语义质量. PTM 由于不能根据语料自动确定最佳伪文档数，因此其 PMI 值较低，语义质量并不是很好，相比 MSSMS 的话题质量平均低 17.48%. BTM、UCT、WNTM 和 GPU-DMM 比我们提出的 MSSMS 在话题质量上表现差，分别平均低 7.12%、8.59%、7.58% 和 6.04%，是因为 BTM 仅建模了双词共享同一主题模式，UCT 仅利用了用户特征，WNTM 仅对短文本进行了扩展，GPU-DMM 仅将局部语义与全局语义相结合，而 MSSMS 则同时融合了多种因素（短文本扩展、双词模式、用户特征与用户特征）.

4.2.2 MSSMS 与其简化版本的 PMI 值比较

本实验通过比较 MSSMS 与其简化版本的语义建模能力，来验证 MSSMS 中不同因素（短文本扩展、双词模式、用户特征与用户特征）对于其语义建模能力提升的有效性. 将主题数设置为 30，分别计算不同算法生成主题的 Top5，Top10 与 Top20 的 PMI 值，通过 PMI 值对 MSSMS 及其简化版本的语义建模能力进行客观分析与评价，实验结果如表 4 所示.

表 4 不同因素下 MSSMS 的 PMI 值比较

	算法	Top5	Top10	Top20
数据集 1	$MSSMS^{-E}$	2.44	2.37	2.32
	$MSSMS^{-B}$	2.51	2.49	2.41
	$MSSMS^{-U}$	2.47	2.41	2.37
	$MSSMS^{-T}$	2.55	2.51	2.44
	MSSMS	2.59	2.55	2.49
数据集 2	$MSSMS^{-E}$	2.45	2.44	2.27
	$MSSMS^{-B}$	2.49	2.48	2.33
	$MSSMS^{-U}$	2.53	2.51	2.47
	$MSSMS^{-T}$	2.46	2.45	2.28
	MSSMS	2.58	2.56	2.52

通过表 4 可以得到以下观察：(1) 在数据集 1 和数据集 2 中表现最差的方法均为 $MSSMS^{-E}$，平均

比 MSSMS 降低 6.54%，可见在四种因素中短文本扩展对于语义质量的提升最为关键. 这是因为通过短文本扩展可以使得局部语义与全局语义相互补充，相比仅仅使用全局语义更具优越性. 此外，由于文档由短变长，可以在一定程度上克服短文本的语义稀疏性；(2) $MSSMS^{-B}$ 相比 MSSMS 平均降低 3.80%，由此可以验证引入双词结构可以有效提高模型的语义建模能力；(3) $MSSMS^{-U}$ 与 $MSSMS^{-T}$ 相比 MSSMS 分别平均降低了 3.47% 和 3.92%，并且在数据集 1 中表现较差的方法为 $MSSMS^{-U}$，在数据集 2 中表现较差的方法为 $MSSMS^{-T}$，这说明建模用户特征在与用户具有高度相关性的数据中作用显著，建模时间特征在与时间具有高度相关性的数据中作用显著，因此可以根据数据集的特点而灵活选用最具影响力的特征；(4) 由于双词模式、用户特征和时间特征是所提基于扩展的微博主题模型的组成因素，因此通过上述三种特征的有效性，可以有效验证所提主题提模型的有效性.

4.2.3　MSSMS 与对比算法的主观评价结果与分析

为了对 MSSMS 语义表示质量进行评价，我们分别选取了“暴雨”、“抢险”、“受灾”三个主题，通过列出主题下 Top10 个最可能出现的单词对不同算法的语义表示质量进行主观评价. 实验结果如表 5 所示.

为了更直观地对实验结果进行展示，我们依据单词与主题的相关度将单词分为三类：与主题强相关单词、与主题弱相关单词、与主题不相关单词. 在表中分别以粗体、斜体和下划线对以上三类相关度的单词进行标记. 对每个主题而言，相关度越高的单词越多且排序越靠前则说明主题聚类效果好，语义表示质量越高，反之，语义表示质量则越低.

从表 5 中可以观察得到，对比算法中均包含了一定数量排序靠前的与主题弱相关词和与主题不相关词. 如在“暴雨”主题下，对比算法中出现了“毫米”、“河流”、“气象台”、“天气”“水位”等弱相关词，和“工作”、“水库”、“阵风”这种不相关词. 在“抢险”主题下，对比算法中出现了“水”、“排水沟”、“抽水泵”这类弱相关词和“转移”、“提醒”、“车辆”等不相关词. 在“受灾”主题下，对比算法中出现了“转移”、“安置”、“倒塌”这些弱相关词，和“路面”、“房屋”这样的无关词. 虽然在所提算法 MSSMS 生成的主题中也包含了个别弱相关词，但是整体可以看出，所提算法在不同主题下生成的强相关单词最多，且排序最靠前. 因此，与对比算法相比，其语义表示质量最高.

表 5　MSSMS 与对比算法语义表示质量主观评价

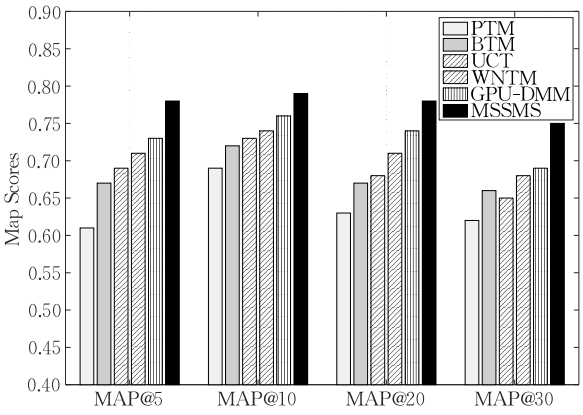
算法	“暴雨”主题	“抢险”主题	“受灾”主题
PTM	**暴雨**、**雨量**、*毫米*、*天气*、工作、水库、**下雨**、**积水**、*雷电*、*气象台*	**抢险**、**沙袋**、**战士**、坑、**清淤**、车辆、行车、**防汛**、*水*、*排水沟*	**受灾**、**灾害**、**损失**、**倒塌**、**死亡**、**灾情**、路面、转移、**被淹**、房屋
BTM	**暴雨**、**阵雨**、阵风、**积水**、工作、毫米、**降水**、河流、*气象台*、*天气*、	**抢险**、**营救**、**救灾**、坑、**清淤**、转移、抽水泵、**排涝**、*水*、*排水沟*	**受灾**、**灾害**、**围困**、**灾情**、倒塌、**死亡**、转移、**被淹**、路面、房屋
UCT	**暴雨**、**积水**、**阵雨**、工作、水库、**降水**、*毫米*、*天气*、**下雨**、河流	**抢险**、**营救**、**防汛**、坑、**清淤**、车辆、**救灾**、行车、*水*、*排水沟*	**受灾**、**灾害**、**灾情**、倒塌、**死亡**、转移、**被淹**、**损失**、房屋、安置
WNTM	**暴雨**、**雨**、**阵雨**、水库、**积水**、阵风、毫米、**降水**、河流、**下雨**	**抢险**、**沙袋**、**防汛**、**疏导**、**清淤**、**战士**、转移、**疏散**、*水*、*排水沟*	**受灾**、**灾害**、**失踪**、遭遇、**围困**、转移、**被淹**、**损失**、路面、**灾情**
GPU-DMM	**暴雨**、**雨**、**阵雨**、**雨量**、**洪水**、河流、天气、阵风、**下雨**、**积水**	**抢险**、**沙袋**、**营救**、**疏散**、提醒、**清淤**、抽水泵、**排涝**、**救灾**、**疏导**	**受灾**、**灾害**、**死亡**、**失踪**、**围困**、*安置*、*转移*、**被淹**、**损失**、**受损**
MSSMS	**暴雨**、**雨**、**下雨**、**强降雨**、**洪水**、**积水**、**阵雨**、**雨量**、**毫米**、**水位**	**抢险**、**沙袋**、**战士**、**防汛**、**疏导**、**清淤**、**抽水泵**、**排涝**、**救灾**、**营救**	**受灾**、**灾害**、**死亡**、**失踪**、**围困**、**被淹**、**损失**、**受损**、**遭遇**、**灾情**

4.3　微博搜索结果与分析

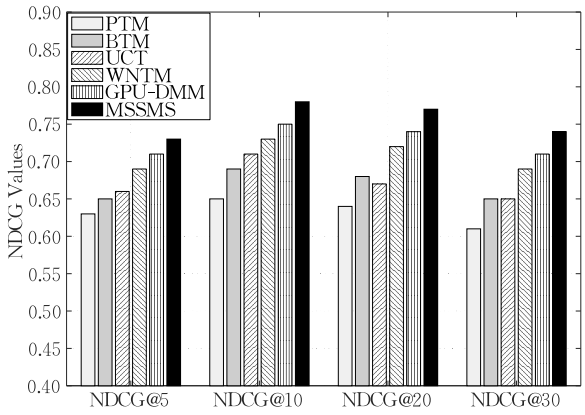
4.3.1　MSSMS 与对比算法的搜索性能比较

主题模型获得的语义表示可用于相似文档的搜索，语义表示越贴切则搜索性能越好. 通过设置搜索实验，验证 MSSMS 与对比算法的语义表示能力及其搜索性能. 将主题数设置为 30. 与其它搜索任务相似，采用 MAP 值和 NDCG 值来衡量搜索的性能. 图 5 和图 6 分别展示了在不同数据集中 MSSMS 与对比方法的 MAP 值和 NDCG 值.

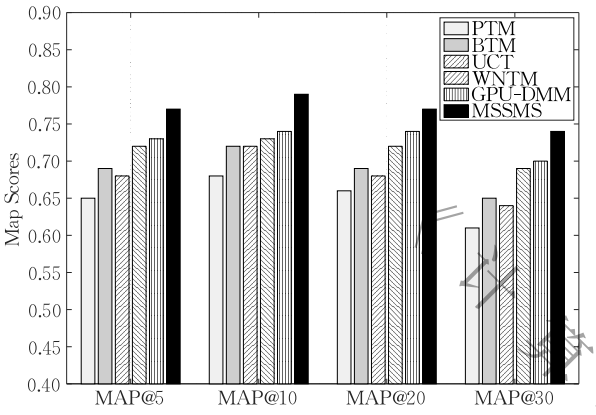
通过图 5 和图 6，可以看到 MSSMS 的 MAP 值和 NDCG 值均比对比方法高，这是由于它融合了多种因素，具有较高的语义表示能力. 相比 GPU-DMM，MSSMS 不仅引入了局部语义，而且采用双词、时间和用户特征可多方面地削弱语义稀疏性，提高语义表示能力；相比基于伪文档数对短文本进行扩展的 PTM 和基于词频对短文本扩展的 WNTM，MSSMS 除利用语义相似性与位置相似性对短文本扩展之外，还引入了双词共享同一主题模式，并且同时建模了时间、文本、用户特征；相比以用户对短文本进行聚合并且建模双词的 UCT 方法，MSSMS 不



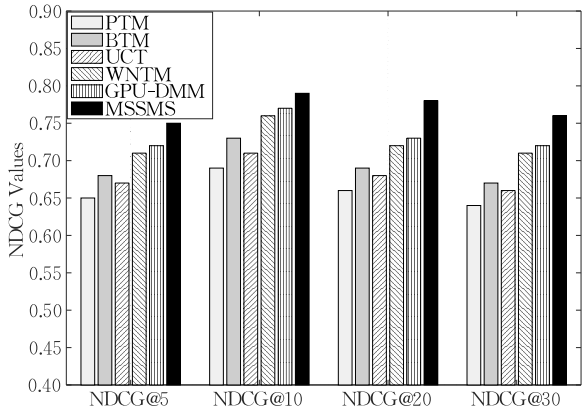
(a) 数据集1上的MAP值



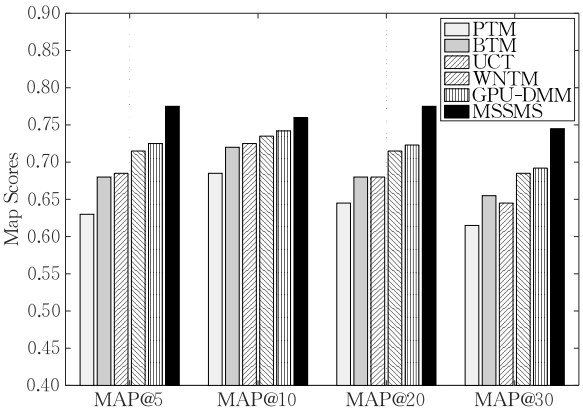
(a) 数据集1上的NDCG值



(b) 数据集2上的MAP值

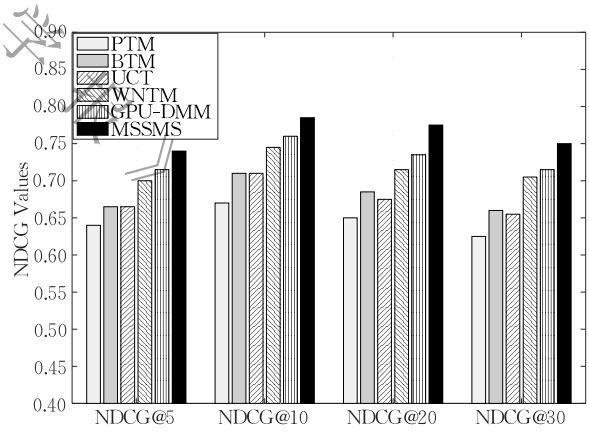


(b) 数据集2上的NDCG值



(c) 两个数据集上的平均MAP值

图 5 MSSMS 与对比算法在不同数据集上的 MAP 值



(c) 两个数据集上的平均NDCG值

图 6 MSSMS 与对比算法在不同数据集上的 NDCG 值

仅同时利用双词与用户,并且基于语义相似性与位置相似性对短文本进行了扩展,此外还同时建模了时间特征;相比 BTM 简单的设定整个语料共享同一个话题分布,MSSMS 设定每个聚合后的长文档共享同一个主题分布,而且同时建模多种特征,从而提高了语义表示能力.

4.3.2 MSSMS 与其简化版本的微博搜索性能比较

本实验通过比较 MSSMS 与其简化版本的搜索性能,来验证 MSSMS 中各因素的有效性.将所有算法的主题数均设置为 30,并将其生成的语义表示用于搜索,采用 MAP 和 NDCG 对搜索性能进行了比较,实验结果如表 6 所示.

从数据集 1 和数据集 2 上的搜索实验中可以看出:MSSMS 的 MAP 值和 NDCG 值总是最高,搜索性能最好,这说明同时融合多种因素(短文本扩展、双词特征、用户特征、时间特征)有利于提高微博搜索性能;$MSSMS^{-E}$ 的 MAP 值和 NDCG 值总是最低,搜索性能最差,这是因为短文本扩展一方面可

以使得局部语义和全局语义相结合，另一方面可以削弱语义稀疏性，因此所提基于词向量的短文本语义扩展算法对于 MSSMS 的搜索性能的提升最为关键. $MSSMS^{-B}$、$MSSMS^{-U}$与 $MSSMS^{-T}$相比 MSSMS 的搜索性能，均有一定程度的降低. 由于双词特征、用户特征和时间特征是所提基于扩展的微博主题模型的重要组成特征，因此该实验结果有效地验证了所提基于扩展的微博主题模型的搜索性能. $MSSMS^{-U}$和 $MSSMS^{-T}$分别在数据集 1 和数据集 2 中的搜索性能略高于 $MSSMS^{-E}$，这说明对于 MSSMS 的微博搜索性能而言，在数据集 1 中用户特征是仅次于短文本扩展因素的第二关键因素，在数据集 2 中时间特征是仅次于短文本扩展因素的第二关键因素，这是由于数据集 1 和数据集 2 分别是用户相关的数据集和时间相关的数据集，虽然用户特征与时间特征会由于数据集的不同而表现不同的重要程度，但是整体来讲多种特征在不同的数据集上都起到了积极的作用.

表 6　MSSMS 与其简版版本的搜索性能比较

	数据集 1 上 MAP 值				数据集 2 上 MAP 值				数据集 1 上 NDCG 值				数据集 2 上 NDCG 值			
	@5	@10	@20	@30	@5	@10	@20	@30	@5	@10	@20	@30	@5	@10	@20	@30
$MSSMS^{-E}$	0.62	0.70	0.62	0.61	0.66	0.69	0.67	0.62	0.64	0.66	0.63	0.59	0.68	0.71	0.67	0.65
$MSSMS^{-B}$	0.76	0.77	0.76	0.73	0.75	0.77	0.75	0.73	0.76	0.79	0.77	0.71	0.79	0.8	0.76	0.74
$MSSMS^{-U}$	0.73	0.75	0.72	0.69	0.74	0.75	0.73	0.72	0.70	0.74	0.71	0.69	0.72	0.75	0.73	0.71
$MSSMS^{-T}$	0.74	0.76	0.74	0.71	0.73	0.74	0.72	0.71	0.72	0.76	0.74	0.7	0.70	0.73	0.71	0.69
MSSMS	0.78	0.79	0.78	0.75	0.77	0.79	0.77	0.74	0.78	0.81	0.79	0.72	0.81	0.83	0.78	0.76

5　总　结

本文提出了一种面向搜索的微博语义建模方法 MSSMS. 该方法有效地融合了局部语义与全局语义并最大程度地缓解了语义稀疏性. 通过建模微博的多种特征(文本、时间和用户)，获取了高质量的语义表示，并基于该语义表示实现了精准的微博搜索. MSSMS 在词向量的基础上同时利用语义相似性和位置相似性对短文本进行扩展，因此可以获取具有语义一致性的相似词集合. 此外，MSSMS 以扩展后的长文档为输入，引入了双词模式并同时建模了微博多种特征，因此，可以生成微博短文本的高质量统一语义表示. 在真实的新浪微博数据中的实验表明，所提算法在微博话题质量和搜索性能上的表现均优于对比算法.

由于 MSSMS 将微博的多种特征映射到了公共的主题语义空间中，得到了多特征的统一语义表示. 因此，在以后的工作中我们可以进行多种形式的搜索，如搜索相似主题的用户、搜索某一话题在特定时间段内的热度等，从而实现微博的智慧精准搜索.

参　考　文　献

[1] Zhu C, Du J. Background feature clustering and its application to social text. Information Processing Letters, 2018, 136: 44-48

[2] Kou F, Du J, He Y, et al. Social network search based on semantic analysis and learning. CAAI Transactions on Intelligence Technology, 2016, 1(4): 293-302

[3] Wang Xiao-Yang, Zheng Xiao-Qing, Xiao Yang-Hua. Entity-relation modeling and discovery for smart search. Journal on Communications, 2015, 36(12): 17-27(in Chinese)
(王晓阳，郑骁庆，肖仰华. 智慧搜索中的实体与关联关系建模与挖掘. 通信学报，2015，36(12)：17-27)

[4] Zhou D, Wu X, Zhao W, et al. Query expansion with enriched user profiles for personalized search utilizing folksonomy data. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(7): 1536-1548

[5] Chen T, Salaheldeen H M, He X, et al. VELDA: Relating an image tweet's text and images//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 30-36

[6] Zha H, Zha H, Zha H, et al. Cost-effective online trending topic detection and popularity prediction in microblogging. ACM Transactions on Information Systems, 2016, 35(3): 18

[7] Liu Y, Liu Z, Chua T S, et al. Topical word embeddings// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2418-2424

[8] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality// Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 3111-3119

[9] Li C, Wang H, Zhang Z, et al. Topic modeling for short texts with auxiliary word embeddings//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 165-174

[10] Xun G, Li Y, Zhao W X, et al. A correlated topic model

《计算机学报》

using word embeddings//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2017: 4207-4213

[11] Wang X, Mccallum A. Topics over time: A non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 424-433

[12] Zhao Y, Liang S, Ren Z, et al. Explainable user clustering in short text streams//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 155-164

[13] Cai H, Yang Y, Li X, et al. What are popular: Exploring twitter features for event detection, tracking and visualization //Proceedings of the 23rd ACM International Conference on Multimedia. Brisbane, Australia, 2015: 89-98

[14] Chen Z, Mukherjee A, Liu B, et al. Discovering coherent topics using general knowledge//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, USA, 2013: 209-218

[15] Zuo Y, Zhao J, Xu K. Word network topic model: A simple but general solution for short and imbalanced texts. Knowledge and Information Systems, 2016, 48(2): 379-398

[16] Bicalho P, Pita M, Pedrosa G, et al. A general framework to expand short text for topic modeling. Information Sciences, 2017, 393: 66-81

[17] Cheng X, Yan X, Lan Y, et al. BTM: Topic modeling over short texts. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(12): 2928-2941

[18] Wang Y, Li D, Liu H, et al. Generalized ensemble model for document ranking. IEEE Transactions on Knowledge & Data Engineering, 2017, 41(2): 367-395

[19] Tang Y K, Mao X L, Huang H, et al. Conceptualization topic modeling. Multimedia Tools & Applications, 2018, 77(3): 3455-3471

[20] Xun G, Li Y, Gao J, et al. Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 535-543

[21] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 1445-1456

[22] Zhou X, Chen L. Event detection over twitter social media streams. The VLDB Journal, 2014, 23(3): 381-400

[23] Kou F, Du J, Lin Z, et al. A semantic modeling method for social network short text based on spatial and temporal characteristics. Journal of Computational Science, 2017, 28: 281-293

[24] Zuo Y, Wu J, Zhang H, et al. Topic modeling of short texts: A pseudo-document view//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 2105-2114

[25] Cui W, Du J, Wang D, et al. Extended search method based on a semantic hashtag graph combining social and conceptual information. World Wide Web-Internet & Web Information Systems, 2018: 1-22

[26] Wang Y, Lin X, Wu L, et al. Effective multi-query expansions: Collaborative deep networks based feature learning for robust landmark retrieval. IEEE Transactions on Image Processing, 2017, 26(3): 1393-1404

[27] Jiang Z, Dou Z, Wen J R. Generating query facets using knowledge bases. IEEE Transactions on Knowledge & Data Engineering, 2016, 29(2): 315-329

KOU Fei-Fei, Ph. D. Her major research interests include social network analysis, data mining and machine learning.

DU Jun-Ping, Ph. D., professor. Her major research interests include artificial intelligence, machine learning, and pattern recognition.

SHI Yan-Song, M. S. candidate. His major research interests include machine learning and information retrieval.

YANG Cong-Xian, M. S. candidate. His major research interests include machine learning and semantic modeling.

CUI Wan-Qiu, Ph. D. candidate. Her major research interests include data mining and machine learning.

LIANG Mei-Yu, Ph. D., associate professor. Her major research interests include information search and data mining.

SHI Lei, Ph. D. candidate. Her major research interests include information search and data mining.

Background

This paper focus on the semantic analysis and precise search of short text in social network. In recent years, Microblog has become a prevalent social networking platform, which is full of hot topics, therefore, searching on it has

become a trend. However, as texts in the social network are so short and sparse that will bring obstacles for precise text search. To overcome this problem, many methods have been proposed and they can be classified into two categories: short text expansion methods and newly designed topic models. However, the existing methods still cannot overcome the semantic sparseness of microblog short texts well. For the reason that the short text expansion methods usually will introduce too many unrelated words and bring in noise, and the newly designed topic models often cannot mine the context and attribute information inside the short texts and only capture the global semantics between texts.

Consequently, we propose a microblog short text semantic modeling method for search (MSSMS), which is implemented by combining a word vector based short text extension method, an expansion based microblog topic model and Microblog search. The proposed short text expansion method can effectively expand the short text by fusing the local semantic information and the position information. Then, by using the expanded short text as the input of the proposed topic model, the sparsity of short text can be alleviated well, and the local semantic and global semantic can complement each other. Moreover, in our proposed topic model, we also simultaneously model multi-information of the short text. In particular, we first construct a bi-term pattern by mining the semantic of the word-word pairs, and then, we introduce the time feature and the user feature in the process of semantic modeling. Experimental results show that MSSMS achieves better search results compared with the baseline methods.

This work is supported by the National Key R&D Program of China (2018YFB14026000), the project funded by China Postdoctoral Science Foundation (2019M660564), the Natural Science Foundation of China (61772083, 61532006, 61877006, 61802028) and the Science and Technology Major Project of Guangxi (GuikeAA18118054).

《计算机学报》