网络大数据的文本内容分析

程学旗，兰艳艳
中国科学院计算技术研究所 北京 100019

摘要
文本内容分析是实现大数据的理解与价值发现的有效手段。尝试从短文本主题建模、单词表达学习和网页排序学习3个子方向，探讨网络大数据文本内容分析的挑战和研究成果，最后指出未来大数据文本内容分析的一些研究方向和问题。

关键词
文本内容分析；短文本主题建模；单词表达；排序学习

doi: 10.11959/j.issn.2096-0271.2015029

Text Content Analysis for Web Big Data

Cheng Xueqi, Lan Yanyan
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100019, China

Abstract
Text content analysis is an effective way to understand and acquire the “value” of big fata. The challenges and research results were investigated in the three hot topics: topic modeling for short texts, word embedding and learning to rank for web pages. In the end, some remaining problems in this area were proposed.

Key words
text content analysis, topic modeling for short texts, word embedding, learning to rank

1 引言

伴随着互联网技术的迅猛发展和普及以及用户规模的爆发式增长，互联网已经步入了“大数据”时代。网络大数据的“大”，不仅仅体现在其体量巨大（大数据的起始计量单位至少是Petabyte[1]、Exabyte[2]或Zettabyte[3]），而且还体现在其增长异常迅猛（通常是指数级的速率），数据类型多样（包括了文本、图像、声音、视频等），数据质量良莠不齐并且关联关系复杂。同时，网络大数据另外一个突出的特点就是其价值密度低，大数据中包含了大量重复、噪声和垃圾数据，存在大量共现但又毫无意义的关联模式，如果缺乏有效的信息处理手段提取网络大数据中潜在的价值，网络大数据不仅不能成为一个价值“宝藏”，反倒会成为一个数据的“坟墓”。

文本内容分析是网络信息处理的关键技术。网络大数据对于文本内容分析而言是一把双刃剑：一方面，网络大数据提供了需要文本分析丰富的数据源，大规模的样本资源可以更好地支持文本分析关键技术的发展；另一方面，网络大数据复杂的内在特征对传统文本分析技术提出了严峻的挑战。例如，网络大数据越来多地存在于电商、问答等私有化网络或者深网中，包括了结构化数据、半结构化数据和非结构化数据，数据的获取和存储更加困难；数据庞大的规模、复杂的关联关系，使得传统的文本分析和挖掘技术在计算的时空复杂度上激增；另外，迅猛的数据增长速率、巨大的数据体量也使得传统的全量计算模式（依赖于全体样本的计算模式）不再适用。本文从短文本主题建模、单词表达学习和网页排序学习3个子方向探讨网络大数据文本内容分析的挑战和研究成果。

2 文本内容分析关键技术

2.1 短文本主题建模

随着Web2.0、社交媒体和移动互联网等技术的发展，每个网民都成为了互联网上信息的创造者与传播者，促使网上文本信息爆炸式增长。与此同时，互联网上的文本内容形式也在不断变化。从博客到轻博客和微博、从邮件到论坛和即时通信、从新闻到评论等，一个显著的特点就是这些文本信息的长度越来越短。这是因为短文本信息编写简单随意，发布更为便捷。同时，短文本信息比长文本更简约、紧凑，能节省其他用户阅读消息的时间和精力。短文本信息比传统文本信息来源要广得多，更新频率也快得多，大大加速了互联网上信息产生与传播的速度。

海量的短文本数据中蕴藏着大量有价值的信息，但也给现有文本语义分析技术带来了新的挑战。与长文本相比，短文本信息内部上下文信息缺乏。此外，普通用户常常用语不规范，新词、多义词等比较普遍。因此，对一条短文本信息的理解要比长文本要困难得多。在以往的长文本语义分析领域，一种普遍的方法就是利用概率话题模型（如LDA[1]和PLSA[2]）对文档内部的话题结构进行建模，然后利用统计推断手段学习出文档集合中潜在的话题以及话题结构。这些模型的一个基本假设是文档是话题的一个混合分布，其中每个词来源于一个话题。当文档长度较长时，可以较准确地根据文档中的词推断出文档的话题属性。然而，当文档很短（只有几个或十几个词，甚至少于话题的个数）时，由于数据不足，难以准确推断出文档中话题混合

1 1 Petabyte=1×10^{15} byte

2 1 Exabyte=1×10^{18} byte

3 1 Zettabyte=1×10^{21} byte

分布的参数以及每个词的话题属性，从而影响话题学习的效果。

为克服短文本信息的数据稀疏性，一种简单做法是利用外部数据（如Wikipedia、搜索结果）扩充文档的表示，再利用现有的长文本语义建模工具处理。但这种方式的效果严重依赖于原短文本文档与扩充的外部数据的相关程度。对于一些实时性强的数据（如微博），要找到合适的外部数据源是很困难的。为此，很多人尝试利用内部数据扩充文档的表示，如伪相关反馈、加入短语特征[3]、相关消息聚合[4]等。无论是利用外部数据扩充，还是利用内部数据扩充，都面临着扩充数据选择不当带来额外噪音的风险。另外，这两种方法并没有从模型上带来任何改变，只是治标不治本。另外，一些研究者[5,6]则提出一条短文本消息只包含一个话题，将短文本消息用单词混合模型（mixture of unigrams）建模。该方式虽然可缓解参数估计时的数据稀疏性问题，但对短文本消息的建模过于简化。现实情况下，一条短文本消息仍然可能包含多个话题，尤其是在话题粒度较细的时候。此时，单词混合模型无法区分。

由于短文本消息和长文本文档显著不同，传统面向长文本的话题建模方法并不能简单地套用到短文本文档上。为了更好地对短文本进行语义建模，提出了一种新的话题建模方法——双词话题模型（biterm topic model，BTM）[7]。BTM和传统基于文档产生式建模的话题模型的最大区别是，它通过建模文档集合中双词的产生来学习话题。这里，双词指的是在同一个上下文中共现的词对。由于一条短文本消息很短，可以简单地认为每条消息是一条上下文[4]。比如在“短文本语义建模”中，可以抽取出3个双词：（“短文本”，“语义”）、（“短文本”，“建模”）、（“语义”，“建模”）。其直接体现了词的共现关系，因此采用双词作为建模单元。直观地讲，两个词共现次数越多，其语义越相关，也就越可能属于同一话题。根据这一认识，假设每个双词由同一个话题产生，而话题从一个定义在整个语料集合上的话题混合分布产生。与LDA相比，BTM通过直接建模双词（即词共现模式）来学习话题，从而避免短文本文档过短导致的文档建模困难问题。二者的图模型表示如图1所示。实验结果表明，BTM在短文本上的效果相比LDA等传统方法有明显提升，而且在长文本上的效果也不输于LDA。

除了长度短之外，互联网上的短文本大数据还具有规模大、更新快的特点。为此，提出了BTM的两种在线学习算法：在线BTM（oBTM）和增量BTM（iBTM）[8]。

4
对于较长的文本，可认为在一个固定长度的窗口内的文本片段为一个上下文。

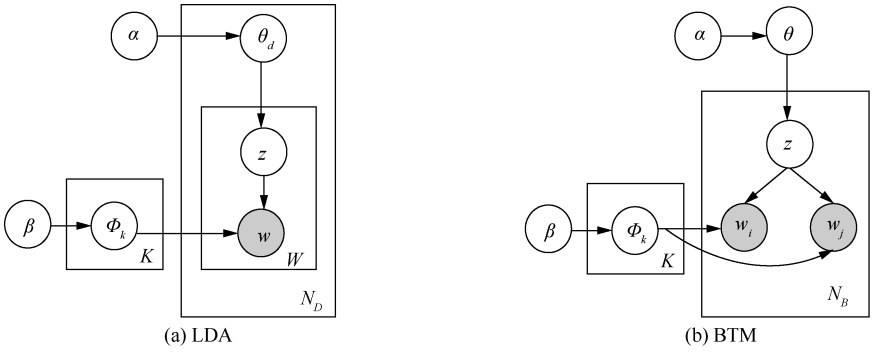


图1 LDA和BTM的图模型表示

这两种算法的主要思想是用最近时间段内接收到的数据来持续更新模型，而不必反复计算较久远的历史数据。这两种算法不仅可以用来处理大规模流式短文本数据，其学到的模型还可以即时反映话题的动态变化，比较适合用于大规模流式短文本语义建模。在微博等互联网应用中，短文本信息还具备很强的时效性，因此其潜在的话题结构也会剧烈变化。尤其受一些突发和热点事件、活动的影响，每天都可能涌现出大量的突发话题。为了对微博中突发话题建模，在BTM的基础上提出了一种突发双词话题模型（BBTM或Bursty BTM）[9]。BBTM的做法是利用双词的突发性来指导突发话题的建模。原因是双词的突发性可以根据其时序频率估算出来，突发性越强、共现次数越多的双词，越可能来源于某个突发话题。基于这一思想，BBTM首先将文档集合中的话题分为突发和非突发两类，然后将双词的突发性作为一个双词话题类别判定的先验。在实验验证过程中，发现BBTM识别出来的突发话题的准确性和可读性都显著优于现有的启发式方法。

尽管在短文本语义建模方向取得了一些进展，但和人对短文本信息的认知能力相比，目前的研究仍然非常初步。在该方向上，笔者认为目前值得深入探索的一些研究点如下。

（1）多源异质数据下的短文本语义建模

在大数据时代的背景下，如何广泛利用其他数据源中的知识（如Wikipedia、开放网页、知识库等），进一步提高计算机短文本的理解与处理能力，是进一步提升短文本语义建模的必经之路。

（2）复杂结构语义建模

目前研究的话题模型结构都比较简单，只有一层潜在语义结构，话题的数目也很有限。这种简单结构的话题模型只能大概反映文本中的语义，难以准确、全面地描述文本内容。真实文本数据中的语义结构很复杂，可以考虑采用层次、图状等结构提升模型的表达能力。

（3）融合上下文特征的话题建模

目前的语义建模方法大多仍然局限在利用词共现信息上。在真实的应用环境中，短文本消息还包含大量的上下文信息（如词性、词序等内源特征）以及作者、地点、人物关系、时间等外源特征。丰富的上下文特征对解决短文本内容稀疏性会有很大帮助。

（4）与应用结合

对短文本语义建模能力的提升最终还需要落地于具体应用中。要最大限度地提升应用效果，需要研究与具体应用相关的语义建模方法。

2.2 单词表示学习

单词表示一直是人工智能、自然语言处理、信息检索等领域的一个基本核心问题。

自然语言处理等相关领域最常用的单词表示方法是one-hot表达，将单词表示为一个高维向量，这个向量的维度是词表大小，其中绝大多数元素为0，只有一个维度的值为1，这个维度就代表了当前的词。这种one-hot表达如果采用稀疏方式存储，非常简洁、高效，配合上最大熵、SVM、CRF等算法，已经很好地完成了NLP（natural language processing，自然语言处理）领域的各种主流任务。

但是，这种表达有着根本性的缺陷，其假设所有词都是独立无关的，抛弃了单词之间的相关性。这意味着，即使是两个同义词，在此表达下，它们的相似度也是0，这显然是不合理的。同时，也因为每个单词都是一个孤立的离散单元，在许多实际问题中会遇到维度灾难问题。例如语言模型中，对于一个词汇集合为100 000的模型，即使只使用二元组，其可能的组合情况便

可以达到100亿种。这样就面临严重的稀疏问题，传统的语言模型必须使用各种复杂的平滑技术来估计那些没有在语料中出现的组合的概率。

为解决语言模型中的维度灾难和稀疏问题，Bengio等人提出了神经网络语言模型[10]。此模型将每个单词表示为一个连续的低维稠密实数向量，这样得到的语言模型自带平滑，无须传统n-gram模型中那些复杂的平滑算法。这样的连续低维稠密实数向量叫做分布式表达（distributed representation），最早由Hinton提出[7]，有别于传统语义网络中用一个独立节点表示一个概念的局部表达（local representation）的方式。而其真正意义上流行开来，始于Bengio在语言模型上取得的成功。现在，单词的分布式表达已经广泛应用于自然语言处理的各个方面，如机器翻译、情感分析和词性标注等。

使用语言模型来学习单词表达涉及在给定前文下预测下一个单词出现的概率，因此需要在整个词汇集合中进行归一化操作，这是非常耗时的一个操作。而当年Bengio的神经网络语言模型在今天看来并不很大的语料上训练了4个月之久，即使后来的C&W的工作，也花了两个月时间才得到一份单词的表达。这在更大的数据上几乎是不可忍受的。早期的单词分布式表达工作主要集中在如何加速训练上面。

代表性工作有Bengio等人2005年提出的层次神经网络模型，输出端不再是一个平坦的softmax层，而是一个树状输出，利用WordNet将一个多项分布分解为一串伯努利分布来优化[11]。AndriyMnih和 Geoffrey Hinton提出Log-Bilinear模型，去掉了网络中隐层的非线性，在此基础上又提出hierarchical log-bilinear模型，同样也是将模型的输出改为层级输出，从而加速模型的训练，并且效果也有一定的提升[12,13]。此后，Mnih将噪声对比估计（noise contrastive estimation，NCE）用于近似优化神经网络语言模型中的sofmax目标函数[14]。而在这方面走得最远的当属目前最受关注的Mikolov等人的工作——Word2Vec。Mikolov在循环神经网络语言模型的工作中发现，将单词的表达学习与语言模型的学习分离开来，可以获得很好的结果。于是提出了continuous bag-of-words（CBOW）和skip-gram（SG）两种单词表达学习模型[15]。这两种模型的目标不再是学习语言模型，而是直接利用自然语言处理中的分布式假设（distributional hypothesis）来学习单词表达。这个假设认为一个单词的语义由其周围的上下文决定，因此出现在相似上下文中的词，其含义也相似。CBOW模型利用上下文单词的表达，直接预测当前词的表达；而SG模型则是使用当前词预测上下文中的每一个词。这两种模型都可以使用哈夫曼树或者negative sampling加速优化。

单词表达学习算法大体都是基于一个同样的假设——分布式假设。其假设一个单词的语义由其周围的上下文决定。由于单词之间存在横向（syntagmatic）和纵向（paradigmatic）两种关系，如图2所示。其中，横向关系主要关注的是词与词之间在上下文中的共现关系，是一种组合性关系；而纵向关系则关注的是词与词之间具有相似上下文的关系，是一种替代性关系。根据所使用的分布信息的不同，单词表达学习方法就可以分为两大类：基于横向关系和基于纵向关系。现有模型都只单独考虑了一种关系。如隐式语义索引（latent semantic indexing，LSI），假设在文档中共现的单词具有相似的语义，其利用了横向关系；而Word2Vec这类方法认为，如果两个单词其周围上下文相似，则其语义相似，其利用了纵向关系。

如图2所示，如果仅仅使用横向关系，不能得到wolf和tiger相似，这并不合理；另一方面，如果只是用纵向关系，则wolf和fierce也不相似。可见，单独使用任一关系，都不能很好地捕捉单词间的关联。在ACL2015的工作[16]提出了两种新的单词表达学习模型（如图3所示），有别于现有模型只建模单词间的横向关系或纵向关系，以并列（PDC模型）或层次（HDC模型）的方式同时建模这两种关系，以得到更好的单词表达。PDC模型和HDC模型对应地扩展了Word2Vec中CBOW和HDC模型，在其基础上，利用文档表达来预测文档中出现的单词，以捕捉单词间的横向关系。

在单词的类似与相似度任务上，这两个模型均取得了state-of-the-art结果。

分布式表达的假设自身也有不足之处，比如不能很好地处理反义词情形。因为互为反义词的两个词，经常出现在同样的上下文中，所以往往反义词之间的相似度反而高于其他近义词。针对此问题，主流思路都是利用外部的知识库来辅助单词的表达学习。这类工作的思路大体类似，都是利用外部知识库如Wikipedia、WordNet约束单词表达的学习过程，比如让更新同义词表达、限制反义词表达等。

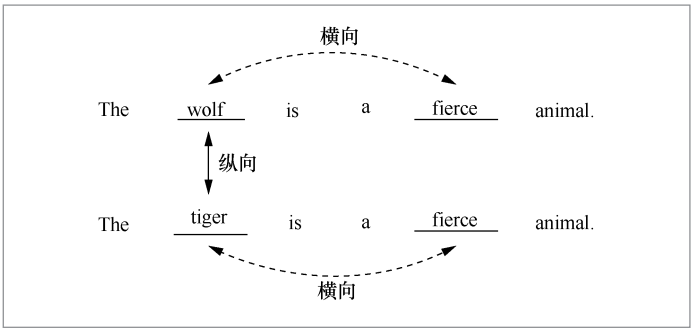


图 2 纵向与横向关系示例

此外，分布式表达的假设也不能很好地处理那些出现次数很少的单词。因为这些单词的上下文信息太少，不足以学到一个很好的表达。比较直接的办法就是直接利用语素信息来改善单词的表达，如果两个单词具有相同的词根，则其语义相似。

另外，目前单词的表达学习主要还是无监督的学习。因此，评价更多地集中在对单词本身的语义表达性能，如各种word similarity和word analogy任务。然而，这些任务并不能反映单词表达在真实的自然语言处理或者信息检索任务中的性能，所以更应该使用真实的任务作为实验。但这样带来的一个问题就是前端表达学习作为无监督学习，与后端的具体任务是脱节的。这也导致许多研究反映，虽然不同的

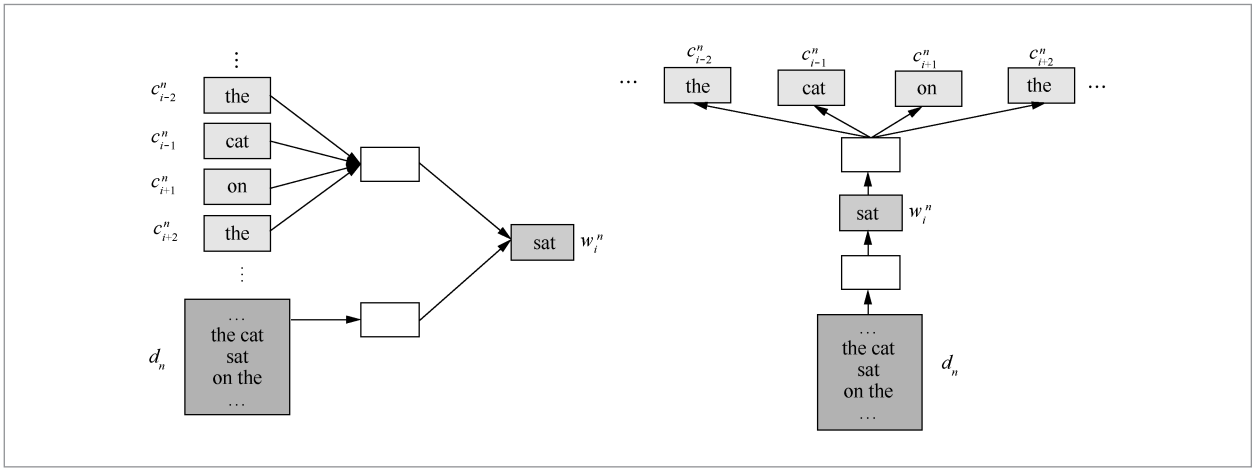


图 3 PDC 模型和 HDC 模型

单词学习模型在单词表达任务上可能性能差异很大，但是在具体实际任务中并没有显著差异。但如果直接根据任务设计有监督的单词学习模型，又会面临可用的标注数据太少的问题。一种可行的方案可能是先利用大规模数据进行无监督学习，得到初始的单词表达，然后根据具体的监督任务调整单词表达，以期望实现更好的应用性能。

2.3 网页排序学习

网络搜索引擎已经成为人们日常生活中的重要工具，在搜索引擎的各个组成部分中，排序模型直接决定了人们看到的搜索结果，因此这种模型对于搜索引擎的性能起着至关重要的作用。

在信息检索发展的历史上，人们提出了很多排序模型，如进行相关性排序的BM25模型[17]和语言模型[18]以及进行搜索结果多样化的MMR[19]模型等。这些模型对推动搜索技术发展起到了一定作用，但是也存在一些问题：有的模型建立在人们对搜索问题的主观理解之上，需要根据经验人为设定模型参数；还有一些模型虽然可以从大量网页中学习，不断调整参数，但无法利用用户的反馈信息对模型参数进行优化。由于用户提交不同的查询词或者不同用户提交相同的查询词都有可能代表不同的信息需求，因此仅从研究者的主观理解或者仅从网页数据中学习排序模型，都无法很好地解决复杂的网络搜索问题。在这样的背景下，近年来研究人员开始尝试使用有监督的机器学习方法，即从用户标注或者反馈中学习最优的相关性排序模型，称为排序学习（learning to rank）[20]。

为了学习最优的相关性排序模型，需要一个训练数据集。该集合包含随机抽取的查询词、与查询词有关的网页以及这些网页的标注。这些标注可能是由用户显式提供的绝对信息，如一个网页和查询词是非常相关、相关还是不相关等；也可能是从搜索引擎的用户行为中挖掘出来的相对信息，如某个网页是否比另外一个网页更加相关。为了从这些标注数据中学到最优的排序模型，通常需要定义3个部分：一是表征网页信息的特征向量（如词频、页面分级（PageRank）等）和网页间关系的特征向量（如网页相似度等）；二是模型的基本形式（如线性、非线性等）；三是用来控制学习过程的损失函数（它衡量了当前模型的排序结果和标注信息之间的差别）。极小化损失函数可以得到与标注数据最吻合的模型参数。经过优化的模型将用于回答新的查询词。给定新的查询词，首先通过倒排表找到包含该查询词的网页，然后为每个网页提取特征向量，并将排序模型应用到这些特征向量上，从而给每个网页输出一个分数，最后将网页按照分数的降序进行排列并返回给用户。

目前针对相关性的排序学习算法效果已经做得很好，部分算法甚至还应用到了搜索引擎的部分模块中。然而一个好的排序不仅依赖于相关性，多样化也是一个重要考虑。其目标在于在排序结果的顶部尽量多地展现不同子话题的网页，因此在排序的同时需要考虑网页间的相似度。然而，这种解决方案的难点在于传统的排序算法都以查询和单个文档作为输入，计算查询—文档相关性很难将文档间的关系融入排序模型内。

为了解决上述问题，有的研究者们直接利用结构化支持向量机直接优化多样化排序评价准则[21]，乐（Yue）等[22]也利用结构化支持向量机寻找最佳文档子集。然而，由于上述算法没有对排序模型进行本质上的改变，模型仍然难以胜任多样化排序任务。

朱(Zhu)等人[23]提出了关系排序学习模型R-LTR，其基本思想是：利用传统的搜索结果多样化模型MMR的思想，使用序列文档选择的方式构造文档排序，即从序列的顶部开始，依次选择排在每一个位置上的文档。在每一次进行文档选择时，考虑查询—文档的相关性和当前文档与已选择文档间的相似性，如图4所示。

因此，R-LTR模型的参数分成两个部分：第一部分为相关性参数，其对应的特征描述了与查询—文档之间匹配的情况和文档的重要性等；第二部分为文档关系参数，其对应的特征描述了文档—文档之间的关系，包括文档在话题、词等级别的相似性等。在训练过程中，R-LTR通过最大化似然的方式进行参数估计。在TREC标注数据集合上的测试表明，在搜索结果多样化的任务上，R-LTR能够超过传统的排序学习模型，取得了显著的效果提升。

夏(Xia)等人[24]针对R-LTR算法只利用了“正例”排序(如α-NDCG=1的最佳排序)进行训练的问题，提出了PAMM算法，其主要思想是：同时利用“正例”排序和“负例”排序进行训练；在排序过程中直接优化多样化排序评价准则。实验结果表明，上述改进方法进一步改善了搜索结果多样化的排序效果，且使得算法具有优化制定的评价准则的能力。

尽管上述各项工作取得了一定的成功，但是由于搜索结果多样化任务本身的复杂性，且评价准则本身不连续、不可导，使得直接对其进行优化仍然存在很多困难。相关的学习算法可能无法收敛或者很容易陷入局部极值点。总体上讲，这个方向还面临很多挑战，需要不断探索。另外，是否能够利用深度学习的方法自动学习多样性排序的特征和样本之间的依赖关系也是一个非常有前景的方向。

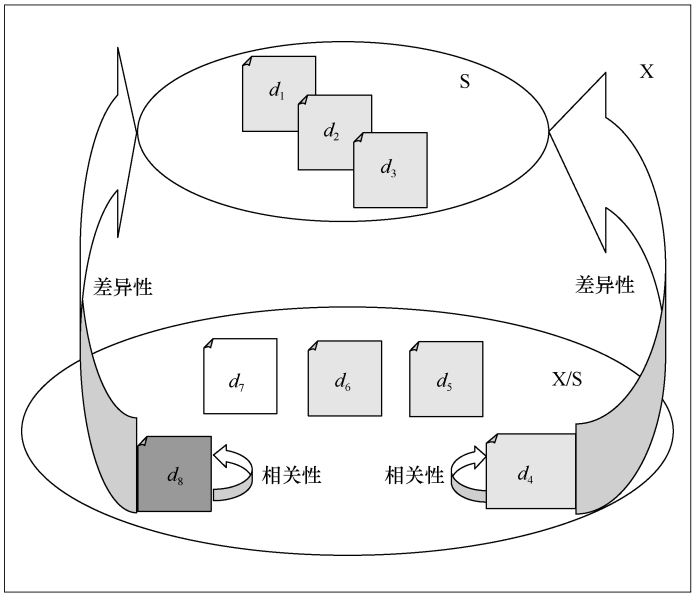


图 4 顺序文档选择过程

3 结束语

综上所述，内容分析成为理解网络大数据的重要手段。其中，短文本主题建模、单词表达学习和多样性排序学习是网络大数据内容分析的热点问题。随着网络大数据的发展，这些方向还存在很多值得探讨的科学问题，例如多源异构数据的主题建模、如何有效利用监督信息得到特定主题的单词表达以及如何使用深度学习的方法来自动学习多样性的特征等。这些问题的解决有助于更好地理解和挖掘网络大数据，从而达到内容分析的目的，为精准检索、推荐等应用提供支持。

参考文献

[1] Hofmann T. Probabilistic latent semantic analysis. Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 1999

[2] Blei D M, Ng A Y, Jordan M I. Latent

dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993~1022

[3] Metzler D, Dumais S, Meek C. Similarity measures for short segments of text. Proceedings of the 29th European Conference on IR Research, Rome, Italy, 2007: 16~27

[4] Hong L, Davison B. Empirical study of topic modeling in Twitter. Proceedings of the 1st Workshop on Social Media Analytics, Washington DC, USA, 2010: 80~88

[5] Zhao W, Jiang J, Weng J, *et al*. Comparing Twitter and traditional media using topic models, Proceedings of the 33rd European Conference on IR Research, Dublin, Ireland, 2011: 338~349

[6] Lakkaraju H, Bhattacharya I, Bhattacharyya C. Dynamic multi-relational Chinese restaurant process for analyzing influences on users in social media. Proceedings of the 12th IEEE International Conference on Data Mining, Brussels, Belgium, 2012

[7] Yan X H, Guo J F, Lan Y 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

[8] Cheng X Q, Yan X H, Lan Y Y, *et al*. BTM: topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2928~2941

[9] Yan X H, Guo J F, Lan Y Y, *et al*. A probabilistic model for bursty topic discovery in microblogs. Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin Texas, USA, 2015

[10] Bengio Y, Ducharme R, Vincent P, *et al*. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137~1155

[11] Morin F, Bengio Y. Hierarchical probabilistic neural network language model. Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, Barbados, 2005

[12] Mnih A, Hinton G. Three new graphical models for statistical language modelling. Proceedings of the 24th International Conference on Machine Learning, New York, USA, 2007: 641~648

[13] Mnih A, Hinton G E. A scalable hierarchical distributed language model. Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 2009

[14] Mnih A, Kavukcuoglu K. Learning word embeddings efficiently with noise-contrastive estimation. Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, USA, 2013

[15] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of Workshop of ICLR, Arizona, USA, 2013

[16] Sun F, Guo J F, Lan Y Y, *et al*. Learning word representation by jointly modeling syntagmatic and paradigmatic relations. Proceedings of the 53rd Annual Metting of the Association for Computational Linguistics, Beijing, China, 2015

[17] Robertson S E. Overview of the okapi projects. Journal of Documentation, 1997, 53(1): 3~7

[18] Zhai C, Lafferty J. A study of smoothing methods for language models applied to Ad Hoc information retrieval. Proceedings of the 24th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, New Orleans, USA, 2001: 334~342

[19] Carbonell J, Goldstein J. The use of mmr, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st Annual International ACM SIGIR Conference on Research & Development on Information

Retrieval, Melbourne, Australia, 1998: 335~336

[20] Liu T Y. Learning to Rank for Information Retrieval. New York: Springer-Verlag New York Inc, 2011

[21] Liang S S, Ren Z C, Maarten D R. Personalized search result diversification via structured learning. In Proceedings of the 20th ACM SIGKDD, New York, USA, 2014: 751~760

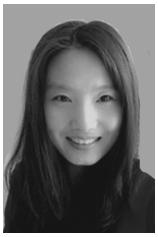
[22] Yue Y, Joachims T. Predicting diverse subsets using structural svms. Proceedings of the 25th ICML, Helsinki, Finland, 2008:1224~1231

[23] Zhu Y, Lan Y, Guo J, *et al*. Learning for search result diversification. Proceedings of the 37th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, Gold Coast, QLD, Australia, 2014: 293~302

[24] Xia L, Xu J, Lan Y Y, *et al*. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 2015

作者简介

程学旗，男，中国科学院计算技术研究所研究员、博士生导师，中国科学院“网络数据科学与技术”重点实验室主任，目前主要从事网络数据科学和社会计算等研究领域的工作，主持和参与多项国家“973”计划、“863”计划、国家自然科学基金项目和国家自然科学基金委杰出青年基金项目，并多次荣获国家科技进步奖一等奖。近年来，在IEEE TKDE、ACM SIGIR、WWW等本领域顶级期刊与国际会议发表论文40余篇，并荣获CIKM最佳论文奖和SIGIR最佳学生论文奖。



兰艳艳，女，中国科学院计算技术研究所副研究员、硕士生导师，目前主要从事机器学习与数据挖掘领域的研究工作，在ACM SIGIR、NIPS、ICML等本领域顶级会议发表论文20余篇，并荣获SIGIR最佳学生论文奖。

收稿日期： 2015-08-16

基金项目： 国家重点基础研究发展计划（“973”计划）基金资助项目（No.2014CB340402），国家自然科学基金资助项目（No.61170008，No.61272055）

Foundation Items: The National Basic Research Program of China(973 Program)(No.2014CB340402), The National Natural Science Foundation of China (No.61170008, No.61272055)

论文引用格式： 程学旗，兰艳艳. 网络大数据的文本内容分析. 大数据, 2015029

Cheng X Q, Lan Y Y. Text content analysis for web big data. Big Data Research, 2015029