doi:10.3772/j.issn.2095-915x.2018.05.004

基于知识图谱扩展的短文本分类方法

丁连红　孙斌　张宏伟

北京物资学院信息学院　北京　101149

摘 要 概念图谱是微软根据对用户搜索日志的统计分析构建的一个大型知识图谱。为了解决文本分类中短文本的数据稀疏、易受噪声影响和主题不明确等问题，本文提出了一种基于概念图谱的短文本语义扩展表示方法。首先，计算文本特征词与概念图谱中各概念的关联度，选取关联度高的概念构成当前文本的概念词典。然后，将概念词典加入特征词集合得到短文本的语义扩展表示。对来自 Twitter 的短文本进行了扩展前与扩展后的分类实验，实验涉及 5 种分类算法和 6 种关联度计算方法。结果显示，概念化语义扩展表示可以提高短文本的分类效果，且包含可以扩展的特征越多的文本，分类结果提升越显著。

关键词：短文本分类；语义扩展；知识图谱；知识推理

中图分类号：TP391；G35

开放科学（资源服务）标识码（OSID）

Short Text Classification Based on Knowledge Graph Extension

DING Lianhong　SUN Bin　ZHANG Hongwei

School of Information, Beijing Wuzi University, Beijing 101149, China

Abstract The Concept Graph is a large-scale knowledge graph constructed by Microsoft based on statistical analysis of user search logs. In order to solve the problem of sparse data, vulnerability to

基金项目：北京市社会科学基金项目青年项目“社交电商中消费行为演化机制及引导措施研究”（17GLC066），北京物资学院高级别培养项目（GJB20162002）。

作者简介：丁连红（1976–），博士，副教授，研究方向：文本挖掘、社会网络分析，E-mail：lhdingbwu@sina.com；孙斌（1994–），硕士研究生在读，研究方向：自然语言处理、社交网络分析；张宏伟（1992–），硕士研究生在读，研究方向：社交电商。

noise, and unclear topic in short text classification, this paper proposes a short text semantic extension representation method based on the Concept Graph. Firstly, the relevance degree between the feature words and the concepts in the Concept Graph is calculated. Top k concepts with the highest relevance are selected as the concept dictionary of the current text. Then, the concept dictionary is combined with the feature words to obtain the semantic representation of the short text. Dataset from Twitter is adopted to evaluate our method. 5 classification algorithms and 6 correlation calculation methods are involved in the experiments. The experiment results show that the semantic representation through conceptualized extension can enhance the classification of short text. We also find the more the feature words that can be expanded in the short text, the better the classification result is.

Keywords: Short text classification; semantic extension; knowledge graph; knowledge inference

1 引言

随着微博、论坛、微信等应用的普及，互联网中出现了大量需要处理的短文本数据。这类文本主要用于用户的日常网上沟通。分析研究此类文本，有助于话题追踪、舆情检测、情感分析等研究的发展与突破。而短文本的分类与标注又是短文本分析的基础。

文本分类在信息检索领域的应用十分广泛，包括搜索引擎、邮件过滤、资料分检、自动文摘以及信息推送等[1]。在过去几十年里，国内外学者提出并改进了一系列经典的机器学习算法，如 k 近邻分类 (k-Nearest Neighbors，kNN)、朴素贝叶斯分类 (Naive Bayes，NB) 和支持向量机 (Support Vector Machine，SVM) 等[2]，并将其应用于文本分类，取得了较满意的结果。然而相比长文本，短文本具有内容长度短小、噪声大、主题不明确等特点，直接将以上方法应用于短文本分类面临严重的特征稀疏问题，分类效果并不理想[3]。

特征扩展是解决短文本特征维度高且稀疏问题的有效方式。知识图谱旨在通过知识储备，以实现智能检索、智能问答等应用。近年来知识图谱在人工智能领域的应用受到了广泛关注，相关研究工作主要围绕着知识图谱的知识提取、知识融合以及知识推理进行。微软研究院试图通过知识图谱实现实例向概念的转变，即通过实例推理概念信息，并将其应用到智能问答系统。2016 年，微软研究院发布了一个概念图谱（Microsoft Concept Graph）与概念标签模型（Microsoft Concept Tagging Model）①。概念图谱是一个大型的知识图谱系统，通过对来自数以亿计的网页和数年积累的搜索日记的数据进行学习而掌握大量的常识性知识。概念图谱表示形式为实例（instances）、概念（concepts）和关系（relations）的三元组。实例与概念之间为 IsA 关系。如，三元组（苹果，水果，IsA）表示苹果是一种水果。

①https://concept.research.microsoft.com/

doi:10.3772/j.issn.2095-915x.2018.05.004

我们认为话题也是一种概念，而短文本中可能直接包含这个概念词，但更常见的则是包含与这些概念相关的实例信息。如“刚运动完，好舒服！！”，“刚打完篮球，好累。”，这两句话的主题均为运动，第一句直接包含“运动”这个概念词，而第二句则是通过“篮球”这个实例信息来表示“运动”这个概念的。“篮球”是“运动”这个概念的一个实例。通过概念化得到的扩展词在语义上与被扩展词语义相同，且该概念化依赖用户的搜索信息。通常一个人在网络中搜索的内容在很大程度上与其在微博、推特等短文本中发布的内容相近。因此，通过微软知识图谱的概念化对短文本进行语义扩展引入的噪声数据较少，能通过语义的扩展提高短文本分类的准确度。

本文将知识图谱引入短文本的二分类过程中，提出了一种基于微软知识图谱的概念化短文本扩展方法。通过实验分析了不同扩展方法对 5 种分类算法的影响和文本所含可扩展特征的数量与比例对分类结果的影响。首先，通过概念图谱推理获得短文本实例的概念词语，将概念词语用作分类的特征词，实现短文本表示的语义扩展。然后，通过实验分析概念化扩展对短文本分类的影响。我们发现，概念化扩展可以提升短文本的二分类结果，且文本所包含的可以扩展特征越多，分类结果提升越显著。本文第 1 节介绍短文本分类的背景与本文的贡献；第 2 节对短文本特征的扩展方法与应用进行了归纳分析；第 3 节介绍了基于概念图谱的文本概念化扩展方法；第 4 节阐述我们的实验设计、实验结果以及分析；第 5 节对全文进行了总结。

2 相关工作

目前对短文本的扩展方法以基于语料库的特征扩展和基于主题模型的特征扩展为主。基于语料库的特征扩展一般采取同义词词典或互联网知识，提取相似的文本或特征，从而扩充短文本的特征空间：

(1) 引入外部知识库（如维基百科和知网等）对文本进行语义特征扩展以丰富词语间语义关系 [4-5]。赵辉等 [6] 借助维基百科知识库挖掘短文本所蕴含的隐性信息，通过选取在语义层面与特征词高度相关的词，对特征词进行扩展以辅助短文本分类；张素智等 [7] 利用知网对使用基于字的分词策略提取出的文本关键字进行概念映射，扩展其语义表达能力。此类扩展在能一定程度上缓解短文本的数据稀疏问题，但扩展效果依赖于外部知识库的质量，对于未收录的主题无能为力，且计算量大，耗时长。

(2) 借助外部文本，如搜索引擎搜索的结果，进而扩展短文本的特征 [8]。将短文本特征作为关键词，以搜索引擎返回结果对短文本进行扩展。原特征与扩展后的特征之间关联性较高，但实现难度较大，比较耗时，同时在一定程度上依赖于搜索引擎返回结果的好坏。

(3) 基于背景语料库，采用不同算法挖掘背景语料中的潜在特征。Zhang[9] 等人提出了一种基于频繁项集考虑语义关联性的特征扩展方式，通过背景语料库挖掘各类别频繁项集，使用频繁项集对原短文本各类别特征进行扩展。在处理短文本特征稀疏问题的同时考虑特征之间的语义关联，其实验结果显示特征扩展后使用 SVM 分类结果在准确率和召回率方面都有较

大的提升。

主题模型兼具可解释性和高性能降维特点，在文本分类中获得了广泛应用。部分学者通过将原始高维特征词空间映射到低维的潜在语义空间或主题空间，挖掘文本潜在的语义结构。如 LDA 主题模型 (Latent Dirichlet Allocation，LDA) 将文本表示为其隐含主题的概率分布，能改善文本的高维稀疏性，每个主题维度也具有可解释性，因此受到广泛应用。在文本分类中，大多数研究者直接将 LDA 应用到文本分类中。Chen[10] 等人使用 LDA 模型来计算文本之间的相似性。相较于传统 SVM 分类算法，取得了较好的加速效果。姚全珠 [11] 和 Rubin[12] 等人直接在 LDA 主题维上进行文本分类，但由于短文本主题聚焦性差，该方法对于改善短文本的稀疏性效果有限。于是，研究者开始采用主题模型对文本特征进行扩展。胡勇军 [13] 等人基于 LDA 主题模型，通过词与文档的潜在语义关联实现特征扩展，相比于单纯直接应用 LDA 的方法有一定的提升。

此外还有一些其他的扩展方法。胡学钢 [14] 等人提出一种基于自身特征扩展的短文本分类方法。该方法从短文本的自身出发，选择具有高指示性的特征组成特征空间，缩减空间维度。在缩减的特征空间中选择文本相似的特征进行扩展，从而有效减小了短文本的稀疏性。Sriram 等 [15] 将用户信息和推特（Twitter）的特征等作为领域关键词集进行特征扩展，以此实现短文本的分类。

2016 年以来人工智能迎来新一轮发展热潮，以计算机围棋程序 AlphaGo 的连续性突破为代表，人工智能为人类展示了其拥有的巨大潜力 [16]。紧跟世界潮流，我国发布了《新一代人工智能发展规划》，提出“到 2030 年人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工智能创新中心”的宏伟目标 [17]。而知识图谱作为人工智能领域的重要研究内容，也备受重视。2012 年，Google 发布知识图谱项目，并宣布以此为基础构建下一代智能化搜索引擎 [18]。该项目适用于解决与实体相关的智能问答问题。张兆锋 [19] 借助知识图谱，实现了技术功效图自动构建，减少了人工参与。Zhang[20] 将科学文献视为科学研究事件，并提出了一种将事件语义连接成具有多种类型的实体关系网络的方法。以实现从大规模快速增长的文献中获得准确，及时，快速，简短和全面的信息。

3 概念化特征扩展算法

3.1 概念化方法

Wang[21] 等提出了 6 种基于概念图谱进行实例概念化的方法，并在微软知识图谱的官方网站提供了相应 api 函数的调用。分别采用 $P(c|i)$、$P(i|c)$、互信息（Mutual Information，MI）、点互信息（Pointwise Mutual Information，PMI^k）、标准化的点互信息（Normalized Pointwise Mutual Information，NPMI）和概念推理算法（Basic-Level Concepts，BLC）计算实例与概念之间的关联度，进行概念词的筛选。实验表明概念化处理可以帮助搜索引擎更容易理解用户的搜索信息。

$P(c|i)$ 表示对于一个实例 i，它所对应的概念为 c| 的概率。$P(i|c)$ 则表示对于一个概念 c|，

doi:10.3772/j.issn.2095-915x.2018.05.004

它的实例是 i 的概率。二者可以直接用来做推理的评分。

$$P(c|i)=\frac{n(c,i)}{\sum_{i\in c_j}n(c_j,i)} \quad （公式 1）$$

$$P(i|c)=\frac{n(c,i)}{\sum_{i_j\in c}n(c,i_j)} \quad （公式 2）$$

其中 $n(c,i)$ 表示实例 i 和概念 c 同时出现的次数，$\sum_{i\in c_j}n(c_j,i)$ 表示实例 i 和实例 i 属于的所有概念同时出现的次数之和。$\sum_{i_j\in c}n(c,i_j)$ 表示概念 c 和概念 c 包含的所有实例同时出现的次数之和。

互信息（MI）衡量的是两个随机变量之间的相关性，即一个随机变量中包含的关于另一个随机变量的信息量。

$$MI(i,c)=\sum P(i,c)\log\frac{P(i,c)}{P(i)P(c)} \quad （公式 3）$$

其中 $P(i,c)$ 表示实例 i 和概念 c 的联合概率。

点互信息（PMI）可以衡量概念与实例之间的相关性。其公式如下：

$$PMI(i,c)=\log\frac{P(i,c)}{P(i)P(c)}=logP(i|c)-logP(i) \quad （公式 4）$$

其中的 $\log P(i)$ 为常数，以至 PMI 值取决于 $\log P(i|c)$。与概率方法结果近似，所以 Wang 采用标准化的点互信息（NPMI）来计算实例和概念的评分。

$$NPMI(i,c)=\frac{PMI(i,c)}{-logP(i,c)}=\frac{logP(i|c)-logP(i)}{-logP(i,c)} \quad （公式 5）$$

同时，也用 PMI 的 k 次方来进行对比。

$$PMI^k(i,c)=\log\frac{P(i,c)^k}{P(i)P(c)} \quad （公式 6）$$

BLC 算法公式如下：

$$blc(i)=\arg\max_c\max_c Rep(i,c) \quad （公式 7）$$

其中，Rep 的计算公式如下：

$$Rep(i,c)=P(c|i)\cdot P(i|c) \quad （公式 8）$$

3.2 特征概念化扩展算法

我们首先通过以上的概念化方法对短文本的特征进行概念化处理，然后将得到的概念词作为扩展特征加入原始特征集合，得到短文本的语义表示。具体的概念化扩展过程可描述为算法 1。算法的输入包括原始短文本 d，扩展数量 $Topk$，扩展算法 EX 和概念图谱中的实例集合 I，输出为经过知识图谱概念化特征扩展后的文本的语义表示 d^*。其中，$d=\{x_1,x_2\cdots\cdots,x_n,c_j\}$，由特征词 $(x_1,x_2\cdots\cdots,x_n)$ 和主题 c_j 组成；$Topk$ 表示每个特征扩展为 $Topk$ 个概念词；EX 是概念化方法，可以是 3.1 节所介绍的 6 种概念化方法之一；$I=\{i_1,i_2\cdots\cdots,i_n\}$ 为实例的集合。

Algorithm 1 语义表示流程

输入： 原始短文本d,扩展词数量$Topk$,扩展算法EX,概念图谱中实例的集合I

输出： 概念化扩展后的语义表示d^*

1: $Words = splitData(d)$;//对数据集合中的每一条文本，进行分词处理，得到初始特征集合$Words$

2: $Words \leftarrow ReduceStopWords(Words)$;//去掉特征集合中的停用词

3: $I\_Words \leftarrow Words \cap I$;//挑选属于概念图谱中实例集合的特征词，形成实例特征词集合I_Words

4: $Word\_temp \leftarrow AccessAPI(I\_Words, EX, Topk)$;//对I_Words中的每个实例特征，根据不同的EX调用微软提供的相应api-url，该api返回与该特征对应的$Topk$个概念词，构成当前文本的扩展字典$Word_temp$

5: $d^* \leftarrow GetingExtend(Words, Word\_temp)$;//根据$Word_temp$对原始特征集合$Words$进行扩展，得到概念化扩展后的文本语义表示d^*

6: **return** d^*。

4 实验方法

4.1 数据来源与评价指标

实验数据采用已进行分类标注的公开数据集（https://github.com/vinaykola/twitter-topic-classifier），标注类别为 Sports 和 Politics。该数据集的原始数据来源于 Twitter，是典型的短文本数据。

F-measure 可以综合分类结果的召回率和准确率对分类算法进行性能评估。为此，我们采用 *F-measure* 对我们的分类算法进行评价。每个类的 *F-measure* 值计算为：

$$F\_measure = \frac{\left(1+\beta^2\right)\times recall \times precision}{recall + precision \times \beta^2 + 0.01}$$

（公式 9）

其中，*recall*- 召回率是某类的文本被正确地划分到该类的个数与此类文本个数的比；*precision*- 准确率是某类文本被正确地划分到此类的个数与划分到此类文本的总个数的比例；β 用来调整召回率和正确率的比重，通常 β 取值为 1。

4.2 实验过程与结果

首先，我们对数据进行直接扩展，不区分特征词是概念图谱中的概念还是实例，并对扩展后的数据进行分类。我们选择了 5 种分类算法进行分类实验，具体是朴素贝叶斯分类器（Naive Bayes, NB）、*k*- 最近邻分类器（k-Nearest Neighbor, *k*NN）、逻辑回归算法（Logistic Regression, LR）、决策树模型（Decision Trees, DT）和随机森林算法（Random Forest, RF）[22-24]。测试文本的个数设置为 100、200、300、400 和 500 共 5 个等级，从总数据集中随机抽样产生。为避免单次采样对测试结果的影响，对每一数量级，均进行 5 次采样，5 次扩展分类和 5 次的 *F* 值计算。最后计算各组合的 5 次 *F* 值的平均数进行比较。

图 1 是经过扩展的数据分类结果与无扩展数据分类结果的对比。可以发现，5 种分类器算法在文本概念化扩展后 *F* 值均有一定的提高。其中提升最显著的是 *k*NN 算法。

我们对数据集中可扩展的特征的数量和比例满足一定条件的文本数量进行了统计，结果见表 1。表中行是一条文本包含可扩展特征数量的下限，列是一条文本可以扩展的特征词占该文本全部特征的比例的下限。

表中数据是数据集中可扩展特征同时达到数量要求和比例要求的文本数量。如 “99 days to go until the start of #ct13. Did you know Chris Gayle is the only player in the event\u2019s history to be dismissed for 99?” 这个文本经过分词与去除停用词后得到特征词集合：[“99 ”,”days”,”start”,”ct13”,”chris”,”gayle”, “player”,”event”,”u2019s”,”history”,”dismissed”,”99”]。其中，可扩展的特征为 ”99”,”days”,”start”,”gayle”, “player”,”history”,”dismissed” 和 ”99”，数量为 8，占比为 0.667。

对筛选后的数据进行分类实验。从可扩展特征数量和比例满足要求的文本中选择一定数量的文本构成实验数据（select）。从全部数据集中抽取相应数量的文本作为对比数据（no-select）。对两个数据集合进行分类，综合 5 种分类器结果的平均值。图 2(a) 是数量下限为 10，比例下限取不同值时的对比结果。图 3(b) 是比例下限为 0.8，数量下限取值不同时的对比结果。

doi:10.3772/j.issn.2095-915x.2018.05.004

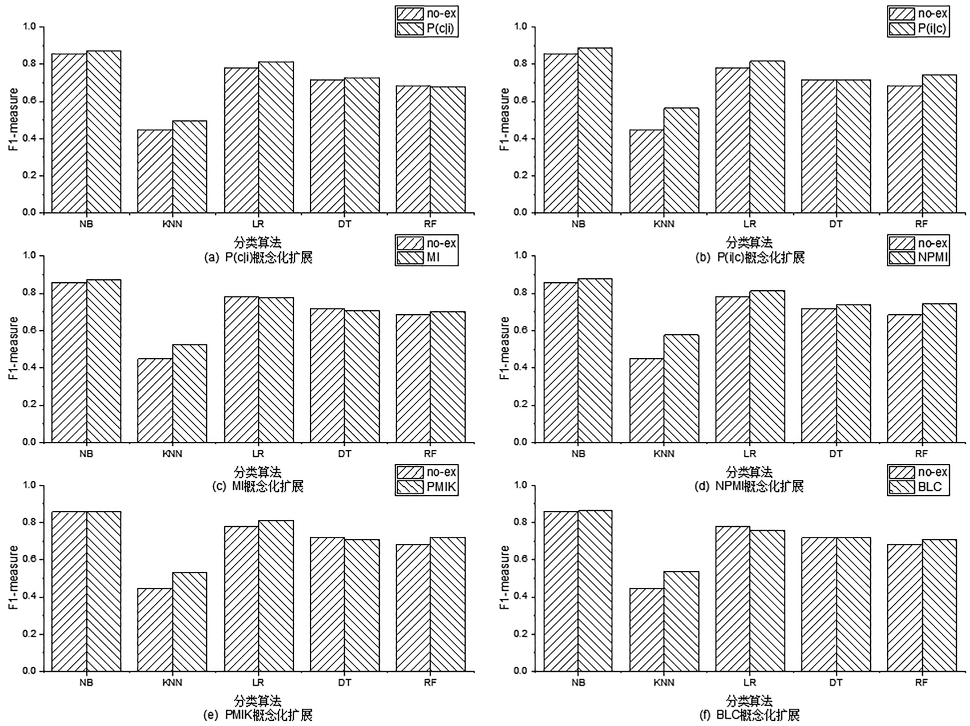

图 1 6 种扩展表示与未扩展表示的分类结果比较

表 1 数据文本中包含可扩展特征的个数及比例

比例 \ 数量	0	3	4	5	6	7	8	9	10
0	6368	5534	5071	4428	3676	2838	1998	1274	727
0.5	5432	5148	4853	4325	3643	2829	1995	1273	727
0.6	4686	4495	4296	3876	3400	2731	1945	1260	725
0.7	3198	3109	3022	2753	2466	2169	1618	1150	678
0.8	1469	1445	1431	1331	1193	1070	938	647	400
0.9	357	333	319	305	273	250	229	212	131

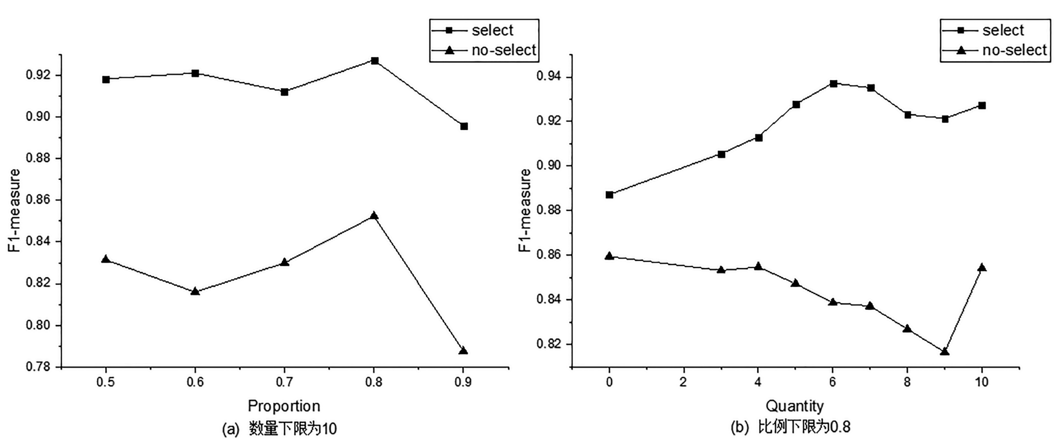

图 2 限定可扩展特征数量和比例时分类结果

由图 2 可知，若待分类文本中可扩展词汇的数量和比例到达一定要求，分类结果提升显著。同时可扩展词汇的数量和比例越高，经概念化扩展后分类结果提升也越显著。

5 总结

本文将知识图谱推理技术引入短文本分类问题中，采用微软的知识图谱，通过大量实验分析基于知识图谱的概念推理在文本分类中的应用。实验表明，经过知识图谱推理产生的扩展信息可以提高短文本的分类结果。另外，当可扩展特征的数量和比例达到一定要求时，分类结果提升显著。后续准备在概念扩展的过程中加入文本的时间信息 [25]，进一步完善语义信息。

参考文献

[1] 孙新，欧阳童，严西敏，等．基于训练集裁剪的加权K近邻文本分类算法[J]．情报工程，2016，2(6):8-16.

[2] 王雯，赵衎衎，李翠平，等．Spark平台下的短文本特征扩展与分类研究[J]．计算机科学与探索，2017(5):732-741.

[3] 王仲远，程健鹏，王海勋，等．短文本理解研究[J]．计算机研究与发展 , 2016, 53(2):262-269.

[4] 朱征宇，孙俊华．改进的基于知网的词汇语义相似度计算[J]．计算机应用，2013，33(8):2276-2279.

[5] 王荣波，谌志群，周建政，等．基于Wikipedia的短文本语义相关度计算方法[J]．计算机应用与软件，2015，32(1):82-85.

[6] 赵辉，刘怀亮．一种基于维基百科的中文短文本分类算法[J]．图书情报工作，2013，57(11):120-124.

[7] 张素智，刘婧姣．基于语义的 kNN 短文本分类算法研究[J]．郑州轻工业学院学报：自然科学版，2012，27(6):1-4.

[8] Wang Z. An improved method of short text feature extraction based on words cooccurrence[C]. Proceedings of the 2nd International Conference on Electric Technology and Civil Engineering, Wuhan, China, May 18-20, 2012. Washington: IEEE Computer Society, 2012:266-269.

[9] Zhang Z F, Miao D Q, Gao C. Short text classification using latent Dirichlet allocation[J]. Journal of Computer Applications, 2013, 33(6):1587-1590.

[10] Chen M G, Jin X M, Shen Dou. Short text classification improved by learning multigranularity topics[C]. Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, Jul 16- 22, 2011. Menlo Park, USA: AAAI, 2011:1776-1781.

[11] 姚全珠，宋志理，彭程．基于 LDA 模型的文本分类研究[J]．计算机工程与应用，2011，47(13):150-153.

[12] Rubin T N, Chambers A, Smyth P, et al. Statistical Topic Models for Multi-label Document Classification[J]. Machine Learning, 2012, 88(1-2):157-208.

[13] 胡勇军，江嘉欣，常会友．基于 LDA 高频词扩展的中文短文本分类[J]．现代图书情报技术，2013(6): 42-48.

[14] 胡学钢，杨超群，张玉红．基于自身特征扩展的短文本分类方法[J]．计算机应用研究，2017，4:1008-1010.

[15] Sriram B, Fuhry D, Demir E, et al. Short text classification in Twitter to improve information filtering [C]. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2010: 841-842.

[16] 戴国强，高芳，徐峰．人工智能在实际应用中的瓶颈及未来研究展望[J]．情报工程，2018，4(1):4-12.

[17] 高芳．全球知名智库对中国《新一代人工智能发展规划》发布与实施情况的评价及启示[J]．情报工程，2018，4(2):26-35.

doi:10.3772/j.issn.2095-915x.2018.05.004

[18] Singhal A, Introducing the Knowledge Graph: things, not strings, May 2012. [EB/OL]. [2018-02-11]. http:// googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.
[19] 张兆锋，张均胜，姚长青. 一种基于知识图谱的技术功效图自动构建方法[J]. 情报理论与实践，2018(3):149-155.
[20] Zhang J, Sun Y, Yao C. Semantically linking events for massive scientific literature research[J]. Electronic Library, 2017, 35(7):724-744.
[21] Wang H, Wang H, Wen J R, et al. An Inference Approach to Basic Level of Categorization[C]. ACM International on Conference on Information and Knowledge Management. ACM, 2015:653-662.
[22] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining. Knowledge and Information Systems[J]. Knowledge and Information Systems 2008, 14(1):1-37.
[23] Cox D R. The regression analysis of binary sequences (with discussion)[J]. Journal of the Royal Statistical Society series B-statistical methodology, 1958(20):215–242.
[24] Breiman L. Random Forests[J]. Machine Learning, 2001(45):5–32.
[25] Zhang J, Yao C, Sun Y, et al. Building text-based temporally linked event network for scientific big data analytics[J]. Personal and Ubiquitous Computing, 2016, 20(5):743-755.