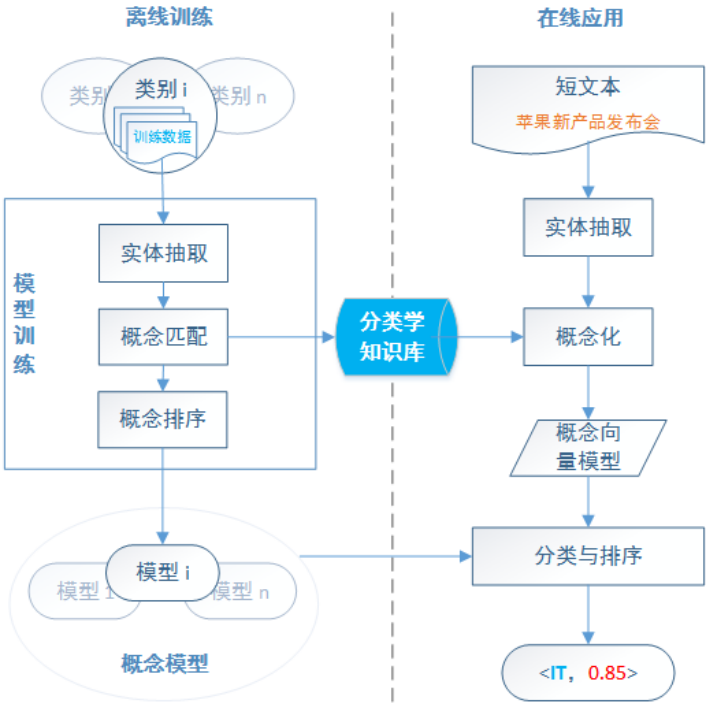
基于概念知识的短文本分类与排序技术

1 技术背景

大数据时代的数据类型多种多样，短文本（Short Text）是其中广泛存在的一种文本类型，常见的短文本包括 Web 查询、微博、在线评论等。这些形式各异的短文本已经成为社会各阶层都广泛接受的信息接收渠道和情感交流手段，深刻改变了亿万民众的沟通方式和生活习惯。如何让计算机自动准确地理解用户短文本语义信息，是大数据时代数据精准分析与处理领域一个极具挑战性的研究热点。这是关乎机器人工智能的基础性研究，对许多实际应用场景都具有重要研究意义。例如，准确理解用户查询短文本所表达的信息需求，将有助于搜索引擎提供更加精准化、个性化和智能化的信息服务。

2 技术方案

短文本分类与排序在搜索引擎中具有重要应用价值，可用于查询分类、查询推荐、广告排序等各个方面。但是，现有的短文本分类方法多利用词袋模型表示文本，常常受字面不匹配的困扰。针对词袋模型在短文本表示中存在的诸多不足，本研究提出了一种基于概念的短文本表示方式，并在此基础上提出了一种新的短文本分类框架，如图 1 所示。首先，利用分类学知识库，为每一个预定义类别学习一个概念模型，用以表示每个类别典型的概念信息；其次，提出了一种改进的短文本概念化（Conceptualization）方法，将给定短文本映射到一组相关概念中；最后，基于相同的概念空间，提出了一种概念相似度计算方法，依此进行短文本分类。



短文本分类与排序框架

3 技术创新点

（1）基于相邻实体改进短文本概念化方法；

（2）基于概念知识的短文本分类方法；

（3）基于概念的短文本多样性排序。

4 应用案例

上述方法实际应用于面向MSN新闻频道的查询推荐。为了便于用户分门别类地浏览新闻内容，MSN和Yahoo!等主流互联网门户网站提供了多种多样的新闻频道。作为一款在线应用，面向MSN新闻频道的查询推荐旨在引导MSN用户的查询需求，当用户正在浏览各个新闻频道时，向其推荐与浏览频道最相关最热门的搜索查询。为实现这一应用，需要解决三个问题：（1）推荐目标（新闻频道）太“短”，新闻热点不断涌现导致频道内容实时更新，缺乏实时的训练数据或用户偏好浏览日志；（2）需要理解搜索查询短文本才能进行推荐；（3）需要分类的同时对结果进行排序，即判断某一搜索查询属于哪一个新闻频道以及对同一频道下的搜索查询进行多样化排序。这些问题也指明了上述方法可能的应用场景。表1为上述方法应用于MSN音乐频道训练得到的概念模型，表2为上述方法应用于搜索查询理解所得到概念化模型。

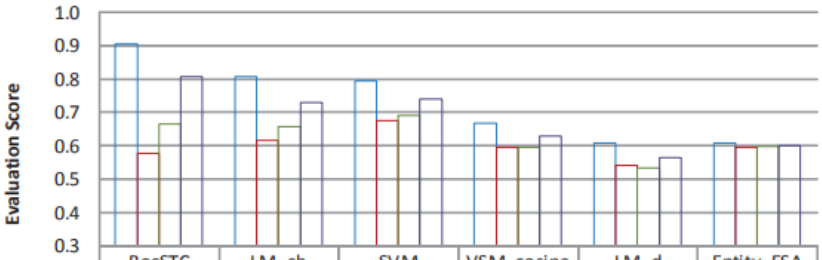
表1 “Channel Music”的概念模型

频道内不同话题下的典型概念		
Singer:	*Song*:	*Instrument*:
performer	good song	musical instrument
pop star	classic song	electronic instrument
pop artist	hip-hop song	string instrument
Music:	*Band*:	*Musician*:
music style	rock band	guitarist
musical genre	metal band	guitar player
musical form	pop band	pianist

表2 查询概念化样例

查询	检测到的实体：实体典型的概念
apple engineer	*apple*: company, corporation, firm *engineer*: professional, expert, occupation
the temptations	*temptation*: artist, popular artist, entertainer
george clooney dated lucy liu	*george clooney*: celebrity, movie star, actor *lucy liu*: celebrity, star, asian actress

采用准确率（Precision）、召回率（Recall）和F值作为分类效果的度量指标。F值是准确率和召回率的调和平均数。本次测评使用了F1和F0.5两种F值，前者表示准确率和召回率具有相同的权重，而后者则更强调准确率的重要性。图2给出了各算法在真实查询数据上的分类表现，其中纵轴表示对应指标在四个频道上的加和平均值。



	BocSTC	LM_ch	SVM	VSM_cosine	LM_d	Entity_ESA
Avg. Precision	0.90	0.81	0.79	0.67	0.61	0.61
Avg. Recall	0.58	0.61	0.67	0.59	0.54	0.59
Avg. F1-Score	0.66	0.66	0.69	0.59	0.53	0.60
Avg. F0.5-Score	0.81	0.73	0.74	0.63	0.56	0.60

5 对接联系

联系人：王芳（信息工程学院博士）

邮　箱：fangwang@bipt.edu.cn