计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20180363
55(8):1641-1652,2018

面向微博短文本的社交与概念化语义扩展搜索方法

崔婉秋[1] 杜军平[1] 寇菲菲[1] 李志坚[1] Lee JangMyung[2]
[1](智能通信软件与多媒体北京市重点实验室(北京邮电大学) 北京 100876)
[2](釜山国立大学电子工程系 韩国釜山 46241)
(wanqiucui@foxmail.com)

The Social and Conceptual Semantic Extended Search Method for Microblog Short Text

Cui Wanqiu[1], Du Junping[1], Kou Feifei[1], Li Zhijian[1], and Lee JangMyung[2]
[1](*Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia* (*Beijing University of Posts and Telecommunications*), *Beijing* 100876)
[2](*Department of Electronics Engineering*, *Pusan National University*, *Busan*, *Korea* 46241)

Abstract Mining the semantics of the microblog texts to realize accurate search is an essential task in microblog search. Because the content of the short texts in microblog has the characteristics of sparsity and semantic limitation, the traditional search methods which only analyze the semantics of literal text for short texts understanding and similarity matching have certain restriction. Therefore, we propose an extended search algorithm based on social and conceptual semantics. By exploiting the unique social attributes such as the #hashtag#, the mention "@" and the link information *URL* in the social network, we further extend the short texts in microblog through the social semantics. The method combines the conceptual words obtained from literal analysis of short texts with the potential associated hashtags information in a graph structure formed by social relationships. It performs the feature representation of short texts in two semantic extensions and achieves the precise search based on full mining of short texts meaning. Finally, we conduct experimental comparisons with traditionally extended search algorithms in the microblog datasets. The results show that the proposed algorithm can capture more semantics and has semantic enhancement function in the search for short texts of microblog. Moreover, the search performance has been significantly improved in the short texts of microblog.

Key words short text in microblog; social and conceptual semantics; extended search; conceptual words; associated hashtags

摘 要 充分挖掘微博短文本的语义以实现精准搜索是一项重要任务.由于微博文本内容具有稀疏性和语义局限性的特点,使得仅通过分析字面语义来进行短文本理解和相似性匹配的传统搜索方法受到了一定的限制.因此提出了一种社交与概念化语义结合的扩展搜索方法,通过挖掘社交网络独特的社交属性如#标签#、"@"和链接信息*URL*,对微博短文本实现进一步的社交语义扩展.该方法将文本字面分析获取的概念词语和社交关系中潜在的关联标签信息相结合,对短文本进行2种角度下的语义特征

收稿日期:2018-05-18;**修回日期**:2018-06-08
基金项目:国家自然科学基金项目(61532006,61320106006,61772083)
This work was supported by the National Natural Science Foundation of China (61532006, 61320106006, 61772083).
通信作者:杜军平(junpingdu@126.com)

表示,实现了基于微博短文本语义充分理解的精准搜索.在微博数据集上的对比实验表明,与已有的扩展搜索方法相比所提方法能捕捉更多的语义特征,微博搜索的性能也得到了显著的提升.

关键词 微博短文本;社交与概念化语义;扩展搜索;概念词语;关联标签

中图法分类号 TP391

近年来,社交网络信息在微博等社交平台上得以广泛的传播和共享.由于社交媒体上用户发布信息的字数受平台限制,并且描述具有随意性,缺乏语义的短文本大量涌现.为了用户在社交媒体平台上更加有效、及时的交流和信息获取,针对短文本的语义分析和搜索成为了目前的研究热点[1].传统的文本语义分析技术中,通常仅考虑文本字面语义[2],对于短文本的查询则采用基于语义的扩展方法[3-4].然而,微博短文本的稀疏性和语义局限性使单纯的文本语义分析方法不能很好地实现对文本语义的有效挖掘.本文利用微博短文本特有的社交属性,包括引导性辅助信息标签、"@提及"和链接信息*URL*.这些辅助信息大多被作为外部元数据来描述文本内容,对其充分挖掘会使微博文本在搜索中体现更多的社交语义信息.本文在微博搜索中结合概念化语义信息,融合标签引导的话题趋向以及微博互动文本内部的关联,解决微博短文本的语义稀疏性问题,使搜索更具话题指导性.

社交网络搜索是目前研究的热点问题.贾焰等人[5]对社交网络智慧搜索技术进行了全面地分析与总结.传统的扩展搜索方法如用搜索结果扩展的伪相关反馈方法[6],或通过外部知识库的概念化扩展方法[7-8]都是单纯地从文本语义出发,没有考虑文本之间的联系.此外,在已有利用标签[9-10]和微博结构信息[11]进行辅助搜索的研究中,没有分析文本的语义.本文融合文本语义和标签等社交结构信息,用社交语义对文本做进一步扩充,使用户在标签指引的主题一致性和社交紧密度等特征的辅助下,挖掘微博文本之间更多的潜在语义关系,提高微博短文本搜索的精确性.

在相关短文本研究的基础上,本文基于维基百科的显性语义分析算法(explicit semantic analysis, ESA),获取短文本的概念化语义信息.利用标签和社交关系构建语义标签图模型,生成关联标签特征,从而表示短文本的社交语义信息.提出基于社交与概念化语义的扩展搜索(expanded search based on social and conceptual semantic, SCS-ES)算法.通过实验验证,SCS-ES 算法能够有效地提高微博短文本搜索的性能.

本文的主要贡献有 4 方面:

1) 提出一种基于社交与概念化语义的扩展搜索方法 SCS-ES,挖掘社交网络中特有的社交属性,扩充短文本语义,从而提高搜索准确性.

2) 提出利用 Wikipedia 生成的最具语义表示的概念词作为纯文本的标签.解决微博数据集中标签使用少的问题,并实现微博结构的统一表示,为抽取标签构建图模型奠定基础.

3) 在概念空间内利用社交关系,构建社交语义标签图模型.挖掘出社交关联下语义相似的文本,丰富短文本的社交语义.

4) 在不同微博数据集上的实验结果表明,本文提出的 SCS-ES 算法能够有效地增强短文本的语义,改进微博短文本的搜索准确性.

1 相关工作

随着社交网络的发展短文本形式不断涌现,针对短文本的研究变得非常重要.由于其自身短小的特点,在搜索中文本语义分析和扩展方法具有重要的研究价值和意义.本节对短文本在搜索中的相关技术进行概述.

1.1 文本语义分析

近年来,在文本搜索领域存在大量的研究成果,包括大量地从纯文本中分析语义提高搜索有效性的方法.经典的 LSA 和 LDA 等语义分析方法,基于统计思想来构建语义学习模型,忽略了词语间的语义关系,因而容易造成文本语义信息的缺失.

此外,由于概念模型在短文本语义学习中的有效性,被认为是一种有效的语义挖掘方法,受到了广泛的关注.如基于 Wikipedia 知识库[12-13]的 ESA[14]方法,以及基于 Probase 的 LexSA[15]等计算语义的方法[16-17],它们提供了跨领域背景知识的相关概念,构建了概念袋向量,用于表示短文本的深层语义结构,扩展了短文本的含义.以上方法优于词袋和主题模型等仅关注词及统计思想的方法,能够提供更多可参考的文本语义信息.

1.2 短文本搜索

由于短文本自身简短、描述随意且缺乏语义信息的特点,使得在搜索中单纯的字面语义分析面临一定的困难.王仲远等人[1]针对短文本的理解方法分析了多种处理技术,在搜索中提供给用户可理解的语义关系解释.针对短文本的搜索方法主要着重于扩展的形式,可以增强文本特征词的语义描述能力,特征词数量的增加也可以在一定程度上解决短文本特征向量稀疏性问题.如基于时空特性[18]的扩展、概念袋[7-8]及两者结合的反馈概念模型[19],这些方法采用相关文本或概念携带的信息来扩充文本的表示.

基于概念语义的搜索利用外部知识源提供在文档集合和查询中没有显性表现出来的额外背景知识和上下文.与主题模型等方法相比,基于外部知识库的方法具有更好的词汇覆盖率,计算模型可以有效地应用于不断更新的语料库.但单纯的概念袋及概念反馈扩展模型,仅从文本表面进行分析,虽然在语义理解上形成了统一的知识模式,但是在微博短文本中,短小和歧义性会导致概念化的方法受到一定的局限.因此,充分挖掘微博的社交属性,利用社交结构和标签等辅助信息进行语义挖掘是面向微博搜索的发展趋势[9-11].

2 基于社交与概念化语义扩展的搜索算法

2.1 问题描述

在微博搜索中,将微博短文本扩展为一种包含3个域的虚拟文本结构:原始文本域 ST(short text)、概念化语义特征域 CS(conceptual semantics)和标签社交特征域 HS(hashtag semantics).该扩展过程在离线阶段完成,数据集中短文本形成的语义特征结构 ST' 表示为

$$ST'=\{ST+CS+HS\}, \tag{1}$$

其中,将微博中的标签和纯文本部分统一称为短文本,即原始文本域 $ST=\{ST_1,ST_2,\cdots,ST_n\}$,$n$ 为微博文本处理后的短文本总数,包括文本和标签2部分,短文本由一组词或短语组成,表示为 $ST_i=\{t_1,t_2,\cdots,t_k\}$;$CS$ 为借助 Wikipedia 外部知识库对 ST 进行概念化后生成的概念化语义特征域;在标签短文本被概念化的基础上,将其根据微博中的社交属性构建语义标签图模型,并通过模型生成短文本的社交特征域 HS.

给定查询 Q,将其类似于短文本的扩展表示为3部分语义结构 $Q'=\{Q+CS+HS\}$.其中,CS 的获取过程与微博短文本生成过程相同,社交特征则通过与微博中标签(短文本)生成的 HS 进行语义相似性计算,获取最相近的一组标签作为 Q 的标签社交特征.在2.3节中将给出其详细的生成过程.针对微博短文本的搜索结果由 Q' 与 ST' 的 top-K 相关性排序分数给出,计算为

$$Rank(Q',ST')=\sum_{t_i\in Q'} idf(t_i)\times\frac{weight(t_i,ST')}{k_1+weight(t_i,ST')}, \tag{2}$$

其中,$idf(t_i)$ 为典型的逆文档频率,$weight$ 为词在扩展短文本所有域中的累积权重,本文2.4节将对式(2)进行详细说明.

SCS-ES 算法的实现由3部分组成,分别为对微博短文本的概念语义特征扩展、社交语义特征扩展以及基于语义扩展的微博短文本搜索,如图1所示.下面将分别对每部分进行说明.

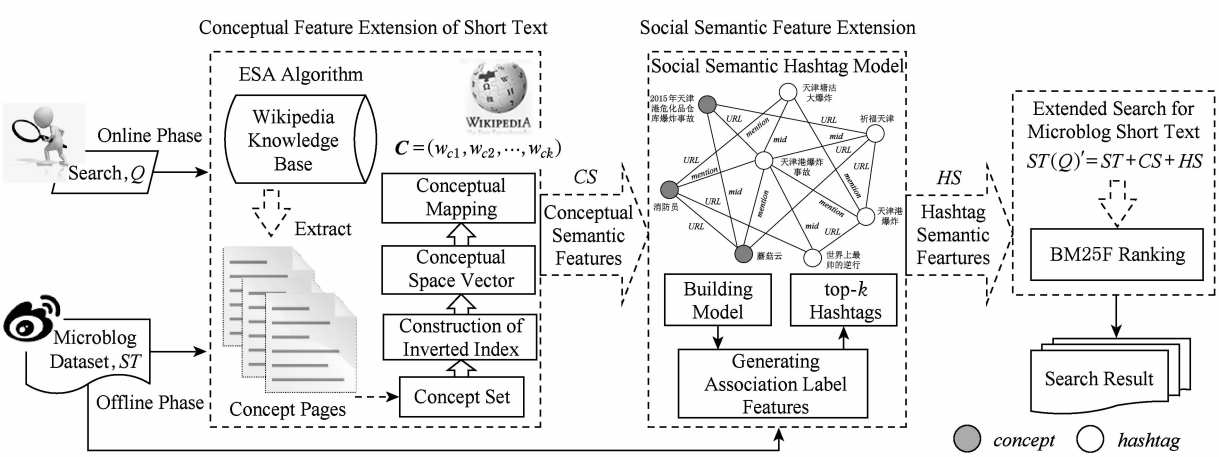


Fig. 1 Short text extended search algorithm based on social and conceptual semantics

图1 基于社交与概念化语义的短文本扩展搜索算法

2.2 短文本概念特征扩展

由于 Wikipedia 是目前最大的知识密集型网络仓库,它是动态更新、快速增长的事件资源,具有一定的新闻价值和事件覆盖性,因此它非常适合作为在社交网络信息快速传播的环境下进行搜索的外部参考资源. 将 Wikipedia 作为微博短文本概念化语义生成的知识库,将它的文章标题作为概念,页面内容作为相应概念的描述,从而构建一个概念语义空间模型. 由于概念空间的概念词语是从 Wikipedia 抽取出来的,因此短文本通过概念化表示后语义属性具有良好的可读性.

2.2.1 概念化

采用显性语义分析算法 ESA[14] 进行短文本的概念语义分析. 对 Wikipedia 页面描述中显性存在的词语进行分析,将每个概念都表示为相应页面词语的属性向量,即转换为概念空间的向量模型来表示,并在概念空间中进行概念提取和估计. 利用倒排索引技术将短文本中每个词项映射为与其相关的 Wikipedia 概念的加权序列,在该方式下原始文本被表示为概念化空间下的权重向量 $\boldsymbol{C}$. 其中,短文本中的词语 t 在相应概念 c 下的映射概率为 $P(c|t)$,以概念的代表性分数表示为

$$P(c\mid t)=\frac{count(t,c)}{\sum_{t\in c_i} count(t,c_i)}, \tag{3}$$

其中,$count(\cdot)$为词语 t 与概念 c 的共现次数;M 是在 Wikipedia 中抽取出来的概念页面的集合,$c_i\in M$.

短文本进行概念映射后,从词向量空间转换为概念空间,被向量化为 $\boldsymbol{C}=(w_{c1},w_{c2},\cdots,w_{ck})$. 每一项为短文本 ST_i 在概念 c_i 下的对应权重,表示概念与短文本的关联强度,计算为

$$w_{c_i}=\log\sum_{t_i\in ST_i,T_i\in T_{c_i}} T_i\times idf_c(t_i)\times icf(c_i), \tag{4}$$

其中,T_i 表示词语 t_i 的概念映射概率,每个词语在概念匹配时对应一组概念,即词语由一组概念关联性概率表示,$T_{c_i}=\{T_i|P(c_i|t_i)\}$. 为了使词语 t_i 和概念 c_i 具有很好的区分能力,引入逆文档频率 idf_c 和逆概念频率 icf,分别表示词语在概念页面和概念在整个概念集合中的代表性. 计算如下:

$$idf_c(t_i)=\log\frac{|M|}{N(t_i)+1}, \tag{5}$$

$$icf(c_i)=\log\frac{|M|}{N(c_i)+1}, \tag{6}$$

其中,$|M|$ 为 Wikipedia 的概念页面总数;$N(t_i)$ 为词语 t_i 在概念页面 c_i 内出现的次数;$N(c_i)$是概念 c_i 在所有词语映射中出现的次数.

2.2.2 概念特征扩展

在离线和在线阶段分别对微博数据集和查询(统称为短文本)进行概念化,使其服从概念空间内的概念关联权重分布. 将 ST 基于 Wikipedia 的概念知识映射到统一的概念空间下,生成一列相关概念集合,由概念与文本的相应概率表示. 取排序 top-k 的最具代表性的概念词语组成 CS. 原始文本被统一扩展为 $ST+CS$.

例 1. 微博短文本"天津滨海新区爆炸现场发巨响腾起蘑菇云",对其分词和去停用词等预处理后,概念化特征扩展生成的概念特征为 $CS=${天津市,爆炸,蘑菇云,2015 天津港危化品仓库爆炸事故,开发区}.

2.3 社交语义特征扩展

从短文本中提取概念后,整个建模过程转化到概念空间内. 结合概念空间与社交属性,对标签短文本进行建模,生成社交语义标签图. 同时对短文本实现社交语义(关联标签特征)的扩展,作为 SCS-ES 算法的主要部分.

2.3.1 社交语义标签图模型构建

微博文本中的辅助信息"@提及"指向一个活跃的微博用户,该用户讨论了此话题或者在该问题上具有一定的权威性. 通过"@"作为微博实体的链接,用户可以将他们共同感兴趣的话题联系起来. 同样地,涉及到相同 URL 链接的用户,他们讨论和关注了同一个链接内容. 因此通过上述信息可以将整个微博网络形成一种潜在的关联,以图的形式来表征和计算社交网络中的自然语义,充分考虑了社交特征的关联,并在概念空间下将标签之间的相关性组织起来.

在对短文本进行语义概念化转换后,它们被映射到统一的概念空间中. 不同概念之间是相互独立的,我们通过在概念空间的基础上构建社交语义标签图,将概念之间通过社交关系、标签共现等辅助属性进行连接,形成具有话题一致性的图结构. 由于微博文本信息中包含标签的仅占少部分,因此我们对微博中纯文本内容进行标签信息的补充. 由于概念化词语是对短文本内容的语义总结,与标签微博中的话题总结作用类似,所以将纯文本的 top-n 相关的概念词作为标签,将微博数据空间统一格式为:#标签#和纯文本. 将整个微博数据之间建立社交关联,从而生成社交语义信息.

构建完整紧凑的社交概念化语义标签图 $G=\langle V,E,W\rangle$.其中,V 为标签表示的节点;E 表示标签之间的关联关系;W 为连接边的权重,反映了标签之间相关性的紧密程度,包括每条边上各种社交关系的平均累加权重以及标签概念化词语的重叠度,在关联标签的生成过程中,连接标签之间重叠的概念化词语越多,标签表达的语义越相关.将社交概念化语义标签图导入适用于社交网络图形结构的 Neoj4 图数据库[20],使节点关系更直观,关联操作更灵活.为了使社交语义标签图结合微博信息传播中的社交关系,定义标签之间的连接规则如下:

规则 1. 标签共现在同一短文本中,则标签之间存在连接(表示 2 个标签具有同一话题倾向性);

规则 2. 标签出现的文本中具有相同的 *URL*,则标签之间形成连接边(包含相同链接信息的文本内容,文本可能讨论同一事件或主题);

规则 3. 标签出现的文本包含@同一个人或组织,标签之间建立连接(2 个标签的内容与同一个人或组织相关时,它们在语义上相似.例如,2 个微博内容中都@天津防火,则其均涉及"天津爆炸"的话题).

图 2 为社交语义标签图模型中同一事件部分节点的放大实例表示,节点分为标签和概念(由网格节点表示)2 类.节点属性包括 *mid*,*mention*,*URL*,*concept*.其中 *mid* 为微博编号,*mention* 为微博内容中带有@字段的信息,*URL* 为链接信息,*concept* 为节点信息(标签或文本)概念化映射生成的一组概念.由于每个节点可能出现在同一话题下的多条微博信息中,因此节点所在微博内的 *mid*,*mention* 和 *URL* 会对应多个不同的取值,将其作为节点相应属

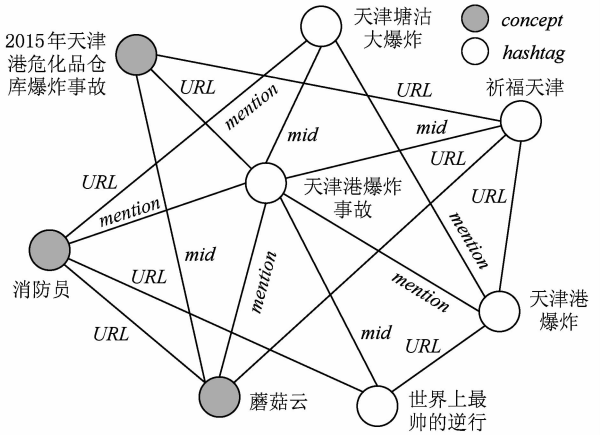


Fig. 2 Social semantic hashtag graph model with an instance of node enlargement for a part of the same event

图 2 社交语义标签图模型同一事件部分节点放大实例

性的值.其次 *concept* 属性能够使标签图充分地融合文本语义信息.如标签节点#天津港爆炸事故#,其 4 个属性内容如下:

mid={CvA0nw9EH,CvA33xInC,CvA3ggmJV,…};

mention={天津消防,人民日报,法制晚报,…};

URL={http://t.cn/RL3J6I4,http://t.cn/RL39DD,…};

concept={天津市,爆炸,蘑菇云,2015 天津港危化品仓库爆炸事故,消防员,…}.

由于标签节点〈#天津塘沽大爆炸#,#天津港爆炸#,#祈福天津#,#世界上最帅的逆行#〉、概念节点〈蘑菇云,消防员,2015 年天津港危化品仓库爆炸事故〉与节点#天津港爆炸事故#分别具有 *mid*,*mention*,*URL* 不同的连接关系,所有这些节点间存在连接边.

2.3.2 社交语义特征的生成

在关联标签生成的过程中,利用共享概念及社交规则实现节点之间连接紧密度的测量,进而生成语义一致性的标签集合,通过关联标签对短文本进行社交语义的扩展.

针对微博短文本关联标签特征的扩展,在概念空间内进行关联标签的度量.定义语义一致性分数来评估标签 n 与短文本 st 中的目标标签 v 的相关概率.在社交语义标签图中,针对标签对应的概念集,生成一组与给定短文本包含的标签语义一致性的标签集.原始标签节点 v 与其邻居标签 n 的语义一致性分数计算为

$$S(n_i|v,st_i)=e(n_i,v)+S_{\text{sim}}(n_i|v,st_i), \quad (7)$$

其中,$e(n_i,v)$为 n_i 与 v 节点之间边的权重,即两节点之间满足连接规则的情况下,3 种规则中权重的平均值.对于每一个规则,边的权重是 2 个标签节点共现的属性数与目标节点中包含该属性的数量的比值.例如,2 个标签节点共现 *mid* 的个数与目标节点包含的 *mid* 个数的比值.利用每个标签的概念词语重叠的数量衡量标签之间的语义相似性,计算为

$$S_{\text{sim}}(n_i|v,st_i)=\frac{\sum_{c_i\in C_{n_i}\cap C_v} p(c_i|n_i)\times p(c_i|v)}{|C_{n_i}\cap C_v|}, \quad (8)$$

其中 $C_{n_i}\cap C_v$ 为标签节点 n_i 与 v 对应的公共概念集,两节点重叠的概念词语越多,说明语义越相近.式(8)中的分子为标签 n_i 和 v 在公共概念集中的概念代表性分数的乘积和,其中 $p(\cdot)$为标签 n_i 和 v 对应 c_i 的概率 w_{ci}(计算如式(4)所示).选择语义

一致性分数高的 top-k 个标签作为社交语义特征 HS. 表 1 为标签或纯文本在社交语义标签图模型下生成的社交语义特征扩展实例,其中纯文本包含 2 个概念词语.

Table 1 Extended Instance of Hashtag/Text Generated Associated Hashtags Feature

表 1 标签/纯文本生成的关联标签特征扩展实例

Hashtag/Plaint Text	Association Hashtags
#天津港爆炸事故#	#天津塘沽大爆炸#,#天津滨海新区码头爆炸#,#祈福天津#,#突发#,#今夜与滨海同在#,etc.
暴雨	Concepts"雨","强降雨","泥石流",#暴雨#,#武汉暴雨#,#暴雨直播#,#暴雨成灾#,etc.

由于查询短文本的社交语义扩展是在线阶段实现的,所以更新标签图模型存在响应时间的问题,将查询短文本概念化后生成的一组概念语义特征,表示为 CS^Q. 离线阶段生成微博短文本(除去标签部分)的概念化特征表示为 CS^{st}. 将两者进行相似性匹配,选择 CS^{st} 中与 CS^Q 相关性 top-k 的概念集所对应的标签,作为 Q 的关联标签扩展部分 HS. 其中,根据 CS^{st} 和 CS^Q 中的公共概念来表示语义相关性,如式(9)所示:

$$Sim(CS^Q,CS^{st})=\frac{\sum\limits_{c_i\in CS^Q\cap CS^{st}} p(c_i|CS^Q)\times p(c_i|CS^{st})}{|CS^Q\cap CS^{st}|},\tag{9}$$

其中,$|CS^Q\cap C^{st}|$ 为 2 个概念集公共概念的个数;$p(c_i|CS^Q)$ 与 $p(c_i|C^{st})$ 表示公共概念集内 c_i 在对应集合内的概念代表性分数.

对微博短文本和查询分别进行社交语义特征的生成计算,如对例 1 中的微博短文本进一步获得社交语义特征为 HS={天津塘沽大爆炸,天津滨海新区码头爆炸,祈福天津,突发,今夜与滨海同在}. 最后所有短文本扩展为 2 部分语义特征表示的形式 $ST'=ST+CS+HS$. 即例 1 中短文本扩展表示为 ST'={天津,滨海新区,爆炸现场,蘑菇云;天津市,爆炸,蘑菇云,2015 天津港危化品仓库爆炸事故,开发区;天津塘沽大爆炸,天津滨海新区码头爆炸,祈福天津,突发,今夜与滨海同在}.

2.4 微博短文本扩展搜索

对原始短文本的 2 部分扩展操作使文本具有了结构化特征,每个域代表短文本不同的语义信息. 选择适用于该类型文本数据的搜索技术 BM25F[21] 排序算法,在搜索中体现 3 个域对文本解释的不同重要程度,在扩展的紧密语义空间内得出相似性最高的搜索结果.

由于标签在微博事件发展的一致性中起到聚集和指导作用,因此标签在搜索中十分重要,将 HS 域在搜索中定义为最大的权值,概念化语义在语义层面上对搜索起到理解作用,因此将 CS 域赋予高于原始文本 ST 的权重. 词在所有域中的累积权重可计算为

$$weight(t_i,ST)=\sum_{c\in ST}\frac{occurs_{t_i,c}^{ST}\times boost_c}{((1-b_c)+b_c\times\frac{l_c}{avl_c})},\tag{10}$$

其中,l_c 为域的长度,avl_c 为域 c 的平均长度,$occurs_{t_i,c}^{ST}$ 为在 ST 短文本的域 c 中 t_i 出现的次数. 根据文献[21],我们设置以下参数的取值:b_c 是与域长度相关的调节因子,搜索中设置 $b_{HS}=0.4$,$b_{CS}=0.4$,$b_{ST}=0.3$;$boost_c$ 是用在域 c 上的增强因子,定义 $HS=5$,$CS=3$,$ST=2$. 根据经验参数设置非线性饱和函数 $weight/(weight+k_1)$ 中的 $k_1=1.7$,得到的排序函数如式(2)所示,其中 $idf(t_i)$ 可计算为

$$idf(t_i)=\log\frac{N-df(t_i)+0.5}{df(t_i)+0.5},\tag{11}$$

其中,N 是在数据集中的文档数,df 是出现词 t 的文档数.

3 微博短文本搜索实验

采用本文提出的 SCS-ES 算法对微博短文本进行搜索测试,分别对数据集及预处理、实验中参数的设置、对比算法和评价指标进行介绍,最后展示搜索实验的结果.

3.1 数据集及预处理

采用维基百科离线数据集作为外部知识库,用于短文本概念语义的映射. 在新浪微博 3 个事件组成的数据集上评估 SCS-ES 算法的有效性.

3.1.1 维基百科外部知识库

从中文维基百科网站① 下载数据 zhwiki-20170701-pages-articles.xml.bz2,分别以数据库的形式

① https://dumps.wikimedia.org/zhwiki

存储,如 page.sql,interlinks.sql 等.为了减少对概念化操作产生噪声和干扰,对概念进行筛选.删除数据中对语义理解和扩展没有作用的页面,包括非概念页面(如 Talk)、字数少于 200 的短页面以及在关系结构中链接少于 3 的概念页面.将候选概念页面通过 wikiExtractor 解析为文本格式,由 opencc 进行繁化简操作.下载的数据大小为 1.4 GB,解析后包含 948 835 篇文章,筛选后剩余 167 328 候选概念页面.

将抽取出来的信息利用 Apache Lucene① 构建倒排索引,完成概念化过程中词与概念的映射和匹配操作,同时加速短文本的概念映射.在映射中只对描述事件的关键词语进行概念化,避免生成噪声概念词对搜索产生影响.

3.1.2 微博数据集及预处理

对新浪微博中与国民安全相关的热点突发事件的数据通过相应的关键词进行爬取,采用关键词及组合形式进行搜索,得到包含精确事件及少量噪声的数据集.爬取了 2 个主要事件的 168 199 条数据,将其处理形成以下 4 个数据集.

1) 数据集 1.它是由"天津、塘沽、爆炸、仓库、滨海新区"及各自组合获得的单一事件数据.

2) 数据集 2.它是由关键词"暴雨,湖北,武汉,防汛、灾害"及各自组合获得的单一事件数据.

3) 数据集 3.它是由数据集 1 和数据集 2 合并组成的混合事件数据.

4) 数据集 4.它是参照文献[22],选取实验室 10 名成员对数据集 1 进行分类标注获得带有分类标签的数据集 4.为了减少标注过程的误差,给定类别描述的关键词及对应的类别标签:爆炸现场(0)、消防员救援(1)、医疗伤亡(2)、祈祷(3)、爆炸无关信息(4).每个成员分别对每一条微博文本进行标注,在 10 个类别标签中选取数量最多的作为该文本的类别标签.

将每一个数据集中的内容进行抽取,获得纯文本、#标签#、"@"提及和链接信息 *URL* 字段(可以为空).将纯文本和标签内容进行繁化简、分词和去噪声处理,以便概念化.通过对社交辅助信息之间的关联关系进行计算,构建社交语义标签图模型,并将其导入图数据库 Neoj4 中,形成标签节点和概念节点的连接图结构.抽取后微博内容中包含的辅助信息数量及分别占数据集的比例统计如表 2 所示:

Table 2 Auxiliary Information of Microblog Datasets

表 2 微博数据集辅助信息数量

Dataset	# Microblog	*hashtag*		*mention*		*URL*	
		Number	Ratio/%	Number	Ratio/%	Number	Ratio/%
Dataset 1 Dataset 4	86 929	19 124	22	8 997	10.3	13 039	15
Dataset 2	81 270	13 156	16	7 398	0.91	11 031	13.5
Dataset 3	168 199	32 280	19	16 395	0.97	24 070	14

3.2 评价指标

信息检索的评价指标主要包括搜索到相关文档的能力和对相关文本正确排序的能力.采用准确率 $P@K$、平均准确率 *MAP* 和归一化折扣累积增益 *NDCG* 指标来评估搜索算法的性能(K 为搜索返回结果的阈值).相应的评价指标和计算公式如下:

1) 准确率

$$P@K=\frac{tr}{tr+fr}, \tag{12}$$

其中,tr 是返回结果中正确的文档数;fr 为结果中错误文档的数量.两者之和为结果列表中文档的总数 K.

2) *MAP* 值在准确率的基础上考虑了返回文档在列表中的位置信息.其公式为

$$MAP=\frac{1}{Q_n}\sum_{q=1}^{Q_n}\left(\frac{1}{R}\times\sum_{r=1}^{R}\frac{r}{position(r)}\right), \tag{13}$$

其中,Q_n 为查询次数,$position(r)$为第 r 个相关文档在返回列表中的位置,R 表示相关文档的个数.

3) *NDCG* 是衡量排序质量的指标.它为连续值的索引,基于返回的前 K 个搜索结果进行计算.

$$NDCG(Q,k)=\frac{1}{|Q|}\sum_{j=1}^{|Q|}Z_{j,k}\sum_{m=1}^{k}\frac{2^{R(j,d)}-1}{\log(1+m)}, \tag{14}$$

其中,$R(j,d)$是文档相关性等级,m 是文档返回的位置.

① http://lucene.apache.org

3.3 实验设置

3.3.1 对比算法

我们参照文献[10]的方法,邀请10名经验丰富的微博用户参与搜索任务,对搜索结果进行评价.其中,每人任意给出10条事件相关的搜索作为搜索集Q.为了评估SCS-ES的有效性,分别选取了概念化和主题扩展的4种对比搜索算法进行实验.每种算法分别返回top-K个搜索结果组成结果集合.为了避免由多个搜索之间的学习效果引起的认知偏差,对每个搜索算法的搜索结果集进行匿名和随机组合分配给实验参与者,并让参与者在其中指定K个与搜索最相关的结果.计算被用户标记为相关的结果与每个算法返回的K个结果之间的相关匹配程度.此外,为了验证人为因素对算法有效性的影响,我们采用文献[3]客观评价的方法,在分类数据集中将搜索的返回结果是否与查询属于同一类别作为搜索的正确性依据.

对比算法介绍如下.

1) ESA-ES[14] (explicit semantic analysis-extended search).基于ESA的概念搜索扩展方法,通过显性语义分析对文本进行概念扩展.

2) Topic-ES[23] (topic-extended search).利用主题词代替概念,利用LDA生成短文本的主题分布,并作为短文本的扩充部分.

3) SEMD[10] (semantically enriched microblog document).语义化扩充微博文档算法.对微博短文本标签和文本信息分别进行概念化扩展,文档结构包含5个域:纯文本、标签、分词后的标签、维基百科链接实体扩展的分词标签和纯文本.

4) ESAC[24] (explicit semantic analysis confidence).通过结合ESA和概念-词语之间的关联规则置信度对查询进行概念化扩展,进而实现短文本扩展搜索.

3.3.2 参数设置及讨论

文献[25]表明在伪相关反馈中,作为扩展词进行扩展搜索的词数设定为20时,对于扩展搜索的效果最佳,因此设定ESA-ES的概念扩展词个数$k=20$.Topic-ES算法中选取主题模型LDA.根据文献[23],设定模型参数$\alpha=50/l$(设主题数$l=10$),$\beta=0.01$,吉布斯采样迭代次数为1 000,每个主题返回20个主题词进行搜索扩展.此外,在SCS-ES算法中由于概念和标签长度不固定,分词后扩展的词语数量不能保持一致,因此根据经验将概念特征扩展和关联标签特征扩展的词数均设置为$k=5$.

利用top-n纯文本的概念化词语作为文本标签进行标签图的构建,参数n的取值会影响算法构建社交语义标签图的结构紧凑性,从而影响关联标签特征的生成,因此SCS-ES算法的搜索有效性会随着n值的变化而改变.为了验证n值何时能够使SCS-ES算法达到最佳状态,在3个数据集上进行了实验比较,图3为SCS-ES算法在参数n的不同取值下搜索$MAP@10$指标的敏感性结果.

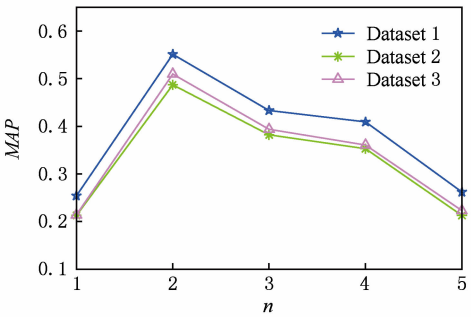


Fig. 3 The influence of parameter n on the search accuracy of SCS-ES algorithm

图3 参数n对SCS-ES算法搜索准确率的影响

如图3所示,当$n=2$时,在3个数据集上搜索的平均准确率均相对较高,说明取2个概念词语作为标签能够拟合微博环境下标签存在的实际数量,使社交网络结构紧凑且标签不冗余,保证了社交语义标签图生成社交特征的有效性.此外,数据集1表现的搜索效果最佳,数据集3中的搜索效果优于数据集2.结合表3内统计的各数据集中微博辅助信息的数量可知,SCS-ES算法的搜索性能与数据集内微博文本包含的辅助信息的比例成正相关,说明SCS-ES算法中标签数量对扩展搜索性能起到了一定的作用,所占比例越大,指导搜索的效果越好.

3.4 SCS-ES与对比算法搜索性能的比较

为了分析SCS-ES算法在微博短文本数据集上的搜索性能,在$P@K$,MAP和$NDCG$三个指标上验证其搜索的有效性,并与对比算法进行了比较和分析.

3.4.1 $P@K$指标上的对比结果与分析

为了充分展示SCS-ES算法的搜索效果,以下实验中均设置作为纯文本标签的概念化词语的个数top-n中$n=2$.此外,由于数据集1包含的标签等辅助信息数量最大,使SCS-ES算法能够充分发挥优势,因此为了排除数据集对SCS-ES算法的影响,并充分验证SCS-ES在搜索上与对比算法性能的优势,本节在数据集1中分别选取搜索结果返回值K

为 10,20,30,40,50 时,计算 SCS-ES 和对比算法搜索准确率 $P@K$ 的结果,如表 3 所示.随着 K 值的增大,SCS-ES 与所有对比算法的 $P@K$ 值均呈逐渐下降的趋势.

Table 3 Comparison of $P@K$ on Dataset 1

表 3 在数据集 1 中 $P@K$ 指标对比

Algorithm	$P@10$	$P@20$	$P@30$	$P@40$	$P@50$
ESA-ES	0.518	0.490	0.430	0.400	0.380
Topic-ES	0.579	0.527	0.450	0.434	0.417
SEMD	0.663	0.535	0.459	0.441	0.420
SCS-ES	0.734	0.601	0.524	0.462	0.447

从表 3 结果中进一步分析可知,SCS-ES 算法的总体性能最佳,相比 SEMD 算法 $P@K$ 值提升了 10%左右,SEMD 与 Topic-ES 算法在搜索准确率上差距不大,均优于 ESA-ES 算法,这表明单纯的概念化方法并不理想,利用标签的 SCS-ES 和 SEMD 方法具有较好的准确率.当 $K=10$ 时各算法准确率达到峰值,说明 $K=10$ 时每个搜索算法的性能均最显著,SCS-ES 算法与对比算法相比能够实现最好的搜索准确率.

为了排除人为因素的影响,选择数据集 1 和数据集 4(相同的数据,区别为有无分类信息)进行搜索实验.在数据集 4 的每一类数据中选取 10 条作为 2 个数据集的搜索集合,并根据人为判断和客观类别区分的方法计算 $P@10$ 的结果,对比结果如图 4 所示.从图 4 中可以看出,对于同一算法在 2 个数据集上的 $P@10$ 结果非常接近,并且客观评价的数据集 4 上的效果优于人为评价的数据集 1,说明通过人为评价的方式更加严格,在返回结果相关性的判断上,会更加的趋于用户搜索意图.验证了本文实验设置中无分类数据集中所用的评价方式具有一定的可信度.

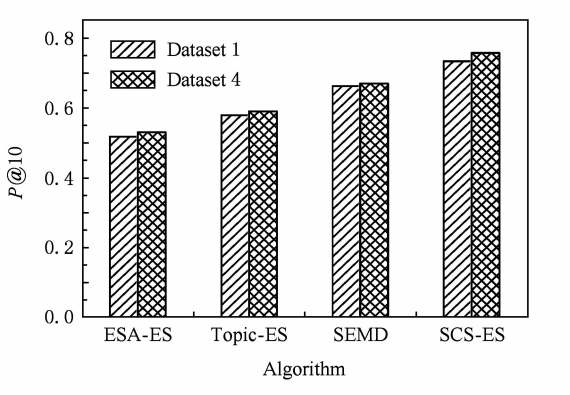


Fig. 4 Comparison of $P@10$ of algorithms under two datasets

图 4 2 个数据集下算法 $P@10$ 指标对比

3.4.2 *NDCG* 指标上的对比结果与分析

为了研究不同数据集对实验中算法性能的影响,选取单一事件数据集 1 和数据集 2 以 *NDCG* 指标进行实验,得出数据集上 SCS-ES 与对比算法的搜索性能.

图 5 为在数据集 1 和数据集 2 中 SCS-ES 及对比算法 *NDCG* 指标的对比.实验结果表明,当 $K=30$ 时数据集 1 和数据集 2 中 SCS-ES 及所有对比算法均达到最好的搜索效果.从图 5 中可以看出,SCS-ES 和 SEDM 算法在数据集 1 中的搜索效果优于数据集 2 的对应实验结果,因此数据集的变化对 SCS-ES 和 SEMD 算法的影响最明显,说明在利用了标签信息的算法中,标签所占数据集的比例会影响算法的整体性能,而其他对比算法在 2 个数据集上的 *NDCG* 值并没有明显变化,说明利用微博辅助信息进行扩展搜索的性能会受到微博内容中标签等辅助信息数量的影响,且包含的辅助信息越多,搜索效果越好.

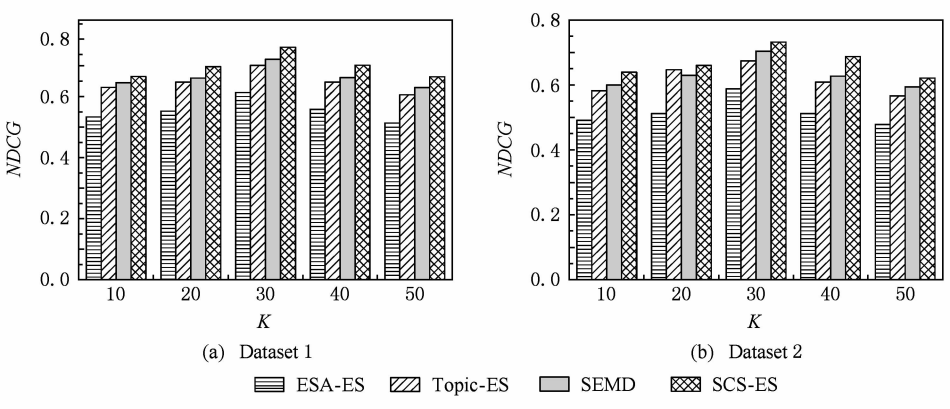


Fig. 5 Comparison of *NDCG* under two datasets

图 5 2 个数据集下 *NDCG* 指标对比

图 5(b)展示为数据集 2 中 SCS-ES 和对比算法的实验结果. 虽然数据集 2 内包含的辅助信息数量并没有使算法 SCS-ES 达到最佳搜索效果,但是其搜索结果仍优于其他对比算法. 对于概念化方法 ESA-ES 只分析了文本的字面语义,通过外部知识库进行概念转换时,短文本和描述随意的特点使概念表示存在噪声,因此搜索结果最差. 主题词扩展的 Topic-ES 算法通过分析文本的主题分布获得主题语义,由于其训练语料库存在局限性,搜索的效果并不理想. SEMD 算法利用标签进行了扩展搜索,针对微博短文本的搜索效果比 ESA-ES 和 Topic-ES 算法有一定提升. 但 SEMD 算法仅根据微博中现有的标签进行扩展,微博中存在大量无标签的纯文本,无法获取标签语义,因此搜索效果没有 SCS-ES 算法的 *NDCG* 指标高. 在 SCS-ES 算法中对纯文本进行了标签的补充,同时通过社交关系构建了社交语义标签图模型,使短文本的语义扩展通过文本语义和社交语义 2 部分实现,实验结果表明,SCS-ES 算法能达到最佳的搜索效果.

3.4.3 *MAP* 指标上的对比结果与分析

为了去除单一数据集在搜索中事件一致性的指导效果,对混合事件数据集 3 进行实验. 图 6 为 3 个数据集中 SCS-ES 及对比算法 *MAP* 值的实验结果. 从图 6 中分别可以看出 $K=10$ 时,SCS-ES 及所有对比算法的 *MAP* 值均达到最优,本文算法 SCS-ES 在 3 个数据集中的 *MAP* 值显著优于对比算法. 图 6(a)为数据集 1 中的 *MAP* 结果,SCS-ES 的效果明显优于图 6(b)和图 6(c)中分别表示的数据集 2 和数据集 3 中 SCS-ES 算法的 *MAP* 值. 3 个数据集上 SEMD 算法 *MAP* 值变化幅度大于其他对比算法,受到标签比例的影响. 此外,在单一事件数据集 1 和数据集 2 及混合事件数据集 3 下,SCS-ES 均体现了较好的 *MAP* 值,在数据集 1 上效果最佳. 在数据集 3 上的实验略优于数据集 2 上的实验结果,说明数据集包含的辅助信息(标签、@提及和 URL)的数量对于 SCS-ES 算法具有一定的影响,成正相关. 数据集包含的辅助信息数量越多,算法的性能越好,并且混合事件对 SCS-ES 算法并没有影响和限制. 对上述实验结果进行分析可知,在 $P@K$, *NDCG* 和 *MAP* 值 3 个评价指标中,本文算法 SCS-ES 优于所有对比算法;其次是利用了标签信息分域搜索的 SEMD 算法和主题词扩展的算法 Topic-ES;最差的是概念词扩展算法 ESA-ES.

我们在表 4 中展示了 SCS-ES 及所有对比算法在 3 个数据集上进行的搜索实验中 3 个指标总体性

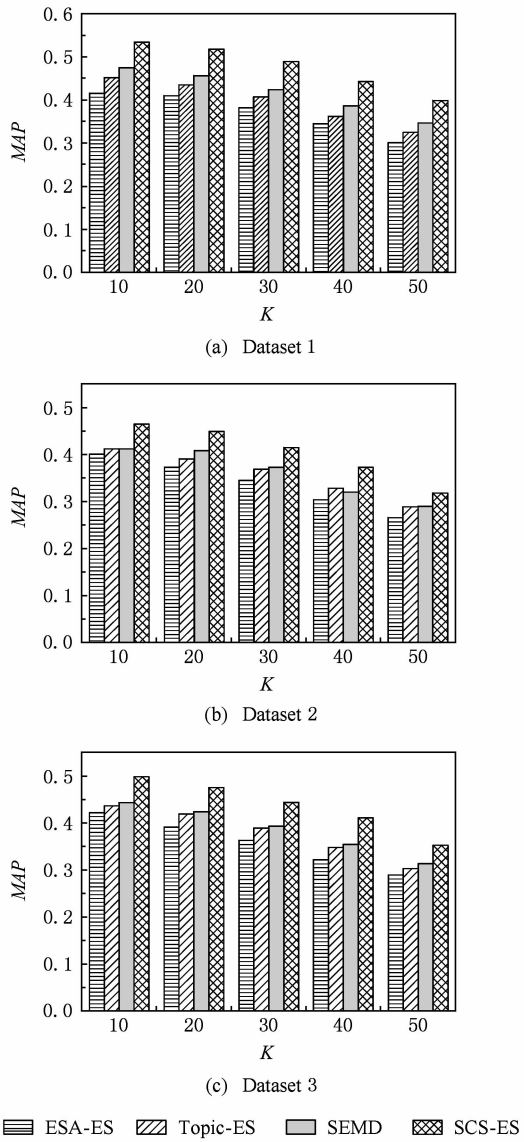


Fig. 6 Comparison of *MAP* under three datasets

图 6 3 个数据集下 *MAP* 指标对比

能的平均值. 从表 4 中可以看出,SCS-ES 算法的搜索性能的平均值最佳,说明该算法具有很好的稳定性. 仅通过概念化(ESA-ES)或主题词进行扩展搜索的算法(Topic-ES)比没有引入标签文本分域的算法(SEMD 和 SCS-ES)的性能好. 由于 SEMD 算法没有考虑各个文本域的语义信息,没有对微博中大量缺乏标签的文本进行处理,也没有融合微博环境下的社交关系等信息,因此搜索效果没有本文提出的 SCS-ES 算法的效果好. SCS-ES 算法在进行微博短文本扩展搜索中,有效地利用社交属性信息扩充了微博短文本的语义,并且对缺乏标签的纯文本进行了补充,在搜索中结合了文本概念化语义和社交语义,满足微博环境下的搜索需求,并且显著提高了微博短文本的搜索效果.

Table 4 Comparison of Average Search Performance on Three Datasets

表 4 在 3 个数据集上的平均搜索性能比较

Algorithm	*P@K*					*NDCG@K*					*MAP@K*				
	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
ESA-ES	0.508	0.481	0.423	0.391	0.373	0.514	0.531	0.601	0.533	0.495	0412	0.391	0.363	0.323	0.285
Topic-ES	0.566	0.515	0.441	0.426	0.407	0.606	0.652	0.688	0.626	0.585	0.433	0.414	0.388	0.345	0.305
SEMD	0.649	0.523	0.448	0.427	0.413	0.645	0.676	0.743	0.675	0.635	0.443	0.429	0.396	0.353	0.316
SCS-ES	0.716	0.583	0.504	0.443	0.421	0.651	0.679	0.749	0.693	0.641	0.498	0.480	0.448	0.408	0.356

3.5 SCS-ES 与对比算法搜索响应时间的比较

在面向搜索的研究中,搜索反馈时间是重要的,虽然本文是针对短文本语义扩展搜索精度的研究,但是为了验证该算法在搜索反馈时间上是否在用户可接受的范围内,我们选取对比算法 SEMD[10]、ESAC[24]与 SCS-ES 算法进行搜索响应时间的对比实验.在搜索数据集大小变化的情况下,验证算法搜索响应时间的有效性,实验结果如图 7 所示:

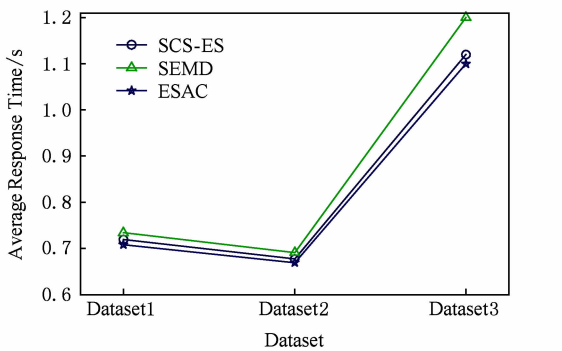


Fig. 7 Comparison of the average search response time

图 7 平均搜索响应时间的对比

图 7 展示了在数据集 1、数据集 2 和数据集 3 上分别进行搜索,返回 *P@*10 结果时算法的平均搜索响应时间.从表 2 数据集信息统计中可知,每个算法的平均搜索响应时间都随着数据量级的增长而增大.SEMD 算法的响应时间略长,这是由于 SEMD 在扩展过程中需要将微博文本分为 5 个域并分别进行语义扩展.ESAC 与 SCS-ES 算法的搜索响应时间非常接近,说明 SCS-ES 在线搜索部分的效率与对比算法是同一量级的,没有激增的现象.因此说明 SCS-ES 算法在注重搜索精度的前提下能够保证搜索效率的稳定.

4 结束语

在微博短文本搜索中,由于文本描述的随意性和不规范性,使得基于文本语义的搜索具有一定的挑战性.本文提出了短文本的社交与概念化扩展搜索方法 SCS-ES,利用概念词语和关联标签丰富短文本语义进行扩展搜索,从而提高搜索质量.实验结果表明本文提出的方法在微博短文本的搜索任务中表现的搜索性能优于其他语义分析及扩展搜索方法,对于 *P@K*,*NDCG* 和 *MAP* 指标有明显的提升.下一步的工作将围绕如何有效地在微博环境下利用标签等社交属性挖掘微博热点话题,以及如何进一步提高微博事件挖掘和搜索的质量.

参 考 文 献

[1] Wang Zhongyuan, Cheng Jianpeng, Wang Haixun, et al. Short text understanding: A survey [J]. Journal of Computer Research and Development, 2016, 53(2): 262-269 (in Chinese)
(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)

[2] Hua Wen, Wang Zhongyuan, Wang Haixun, et al. Understand short texts by harvesting and analyzing semantic knowledge [J]. IEEE Trans on Knowledge & Data Engineering, 2017, 29(3): 499-512

[3] Yu Zheng, Wang Haixun, Lin Xuemin, et al. Understanding short texts through semantic enrichment and hashing [J]. IEEE Trans on Knowledge & Data Engineering, 2016, 28(2): 566-579

[4] Wang Yashen, Huang Heyan, Feng Chong. Query expansion based on a feedback concept model for microblog retrieval [C] //Proc of the 26th Int Conf on World Wide Web. New York: ACM, 2017: 559-568

[5] Jia Yan, Gan Liang, Li Aiping, et al. Research progress and development trend of online social network smart search [J]. Journal on Communications, 2015, 36(12): 9-16 (in Chinese)
(贾焰,甘亮,李爱平,等. 社交网络智慧搜索研究进展与发展趋势[J]. 通信学报, 2015, 36(12): 9-16)

[6] Albishre K, Li Y, Xu Y. Effective pseudo-relevance for microblog retrieval [C] //Proc of the Australasian Computer Science Week Multiconference. New York: ACM, 2017: 51-57

[7] Kotov A, Zhai C X. Tapping into knowledge base for concept feedback: Leveraging conceptnet to improve search results for difficult queries [C] //Proc of the 5th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2012: 403-412

[8] Wang Zhongyuan, Zhao Kejun, Wang Haixun, et al. Query understanding through knowledge-based conceptualization [C] //Proc of the 24th Int Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 3264-3270

[9] Wang Yuan, Liu Jie, Huang Yalou, et al. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs [J]. IEEE Trans on Knowledge & Data Engineering, 2016, 28(7): 1919-1933

[10] Bansal P, Jain S, Varma V. Towards semantic retrieval of hashtags in microblogs [C] //Proc of the 24th Int Conf on World Wide Web. New York: ACM, 2015: 7-8

[11] Luo Zhunchen, Yu Yang, Osborne M, et al. Structuring tweets for improving Twitter search [J]. Journal of the Association for Information Science & Technology, 2015, 66 (12): 2522-2539

[12] Jiang Yuncheng, Bai Wen, Zhang Xiaopei, et al. Wikipedia-based information content and semantic similarity computation [J]. Information Processing & Management, 2017, 53(1): 248-265

[13] Hong K J, Kim H J. A semantic search technique with Wikipedia-based text representation model [C] //Proc of the Int Conf on Big Data and Smart Computing. Piscataway, NJ: IEEE, 2016: 177-182

[14] Egozi O, Markovitch S, Gabrilovich E. Concept-based information retrieval using explicit semantic analysis [J]. ACM Trans on Information Systems, 2011, 29(2): 1-34

[15] Hua Wen, Wang Zhongyuan, Wang Haixun, et al. Short text understanding through lexical-semantic analysis [C] // Proc of the 2015 IEEE 31st Int Conf on Data Engineering. Piscataway NJ: IEEE, 2015: 495-506

[16] Wu Wentao, Li Hongsong, Wang Haixun, et al. Probase: A probabilistic taxonomy for text understanding [C] //Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 481-492

[17] Zhang Libo, Sun Yihan, Luo Tiejian. Calculate semantic similarity based on large scale knowledge repository [J]. Journal of Computer Research and Development, 2017, 54 (11): 2576-2585 (in Chinese)
(张立波, 孙一涵, 罗铁坚. 一种基于大规模知识库的语义相似性计算方法[J]. 计算机研究与发展, 2017, 54(11): 2576-2585)

[18] Kiesel J, Potthast M, Hagen M, et al. Spatio-Temporal analysis of reverted Wikipedia edits [C] //Proc of the 11th Int AAAI Conf on Web and Social Media. Menlo Park, CA: AAAI, 2017: 122-131

[19] Miyanishi T, Seki K, Uehara K. Time-Aware latent concept expansion for microblog search [C] //Proc of the 8th Int AAAI Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2014: 366-375

[20] Cattuto C, Quaggiotto M, Averbuch A. Time-varying social networks in a graph database: A Neo4j use case [C] //Proc of the 1st Int Workshop on Graph Data Management Experiences and Systems. New York: ACM, 2013: 1-6

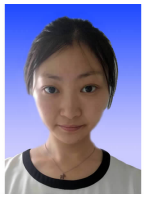
[21] Fresno V, Fresno V, Fresno V, et al. Using BM25F for semantic search [C] //Proce of the 3rd Int Semantic Search Workshop. New York: ACM, 2010: 11-19

[22] Xing Chen, Wu Wei, Wu Yu, et al. Topic aware neural response generation [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 3351-3357

[23] Ye Zheng, Huang Xiangji, Lin Hongfei. Finding a good query-related topic for boosting pseudo-relevance feedback [J]. Journal of the Association for Information Science & Technology, 2014, 62(4): 748-760

[24] Zingla M A, Chiraz L, Slimani Y. Short Query Expansion for Microblog Retrieval [M]. Amsterdam, Netherland: Elsevier Science Publishers, 2016: 225-234

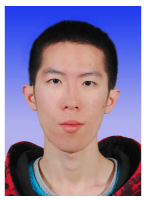
[25] Abderrahim M E A, Benameur S, Abderrahim M A. The number of terms and documents for pseudo-relevant feedback for ad-hoc information retrieval [J]. International Journal of Computer Science Issues, 2013, 10(1): 661-667



Cui Wanqiu, born in 1990. PhD candidate. Her main research interests include social network analysis, machine learning and information retrieval.

Du Junping, born in 1963. Professor, PhD supervisor. Distinguished member of CCF. Her main research interests include artificial intelligence, data mining, social network analysis and search, computer applications.

Kou Feifei, born in 1989. PhD candidate. Her main research interests include semantic learning and multimedia information retrieval and recommendation (koufeifei000@126.com).



Li Zhijian, born in 1994. Master candidate. His current research interests include machine learning and cross-media search (114898070@qq.com).

Lee JangMyung, born in 1957. Research Group Leader for Logistics and IT. His current research interests include intelligent robotic systems, ubiquitous port, and intelligent sensors (jmlee@pusan.ac.kr).