软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2865−2876 [doi: 10.13328/j.cnki.jos.004726]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于聚类的直推式学习的性能分析*

张 新, 何 苯, 罗铁坚, 李东星

(中国科学院大学 计算机与控制学院,北京 101408)

通讯作者: 张新, E-mail: zhangxin510@mails.ucas.ac.cn

摘 要: 近年来,Twitter 搜索在社交网络领域引起越来越多学者的关注.尽管排序学习可以融合 Twitter 中丰富的特征,但是训练数据的匮乏,会降低排序学习的性能.直推式学习作为一种常用的半监督学习方法,在解决训练数据的稀少性中发挥着重要的作用.由于在直推式学习的迭代过程中会生成噪音,基于聚类的直推式学习方法被提出.在基于聚类的直推式学习方法中有两个重要的参数,分别为聚类的阈值以及聚类文档的数量.在原有工作的基础上,提出使用另外一种不同的聚类算法.大量在标准 TREC 数据集 Tweets11 上的实验表明,聚类的阈值以及聚类过程中文档数量的选择都会对模型的检索性能产生影响.另外,也分析了基于聚类的直推式学习模型的鲁棒性在不同查询集上的表现.最后,引入名为簇凝聚度的质量控制因子,提出了一种基于聚类的自适应的直推式方法来实现 Twitter 检索.实验结果表明,基于聚类的自适应学习算法具有更好的鲁棒性.

关键词: 聚类;直推学习;Twitter 检索;自适应;性能

中图法分类号: TP181

中文引用格式: 张新,何苯,罗铁坚,李东星.基于聚类的直推式学习的性能分析.软件学报,2014,25(12):2865−2876. http://www.jos.org.cn/1000-9825/4726.htm

英文引用格式: Zhang X, He B, Luo TJ, Li DX. Performance analysis of clustering-based transductive learning. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2865−2876 (in Chinese). http://www.jos.org.cn/1000-9825/4726.htm

Performance Analysis of Clustering-Based Transductive Learning

ZHANG Xin, HE Ben, LUO Tie-Jian, LI Dong-Xing

(School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 101408, China)

Corresponding author: ZHANG Xin, E-mail: zhangxin510@mails.ucas.ac.cn

Abstract: Recently, Twitter search has drawn much attention of researchers in social networks. Although rich features of Twitter can be incorporated into rank learning, the retrieval effectiveness can be hurt by the lack of training data. Transductive learning, as a common semi-supervised learning method, has been playing an import role in dealing with the lacking of training data. Due to the fact that noise is generated during the iterative process of transductive learning, a clustering-based transductive method is proposed. There exist two important parameters in the clustering-based transductive approach, namely the threshold of clustering and the number of the documents that will be clustered. This paper extends the method by utilizing a different clustering algorithm. As shown by extensive experiments on the standard TREC Tweets11 collection, both of the two parameters have an effect on the retrieval effectiveness. Furthermore, the robustness of the clustering-based transduction approach on different query sets is also studied. Finally, the paper proposes an adaptive clustering-based approach by introducing a so called cluster coherence as quality controller. The experimental results show that the robustness of the proposed method is better.

Key words: clustering; transductive learning; Twitter search; adaptive; performance

随着在线社交网络的飞速发展,微博作为一种非常受欢迎的社交网络服务,正吸引着越来越多的互联网用

* 基金项目: 国家自然科学基金(61103131, 61472391); 教育部留学回国人员科研启动基金; 北京市自然科学基金(4142050)
收稿时间: 2014-05-05; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

户[1].在微博上,每天甚至每一分钟都有大量的更新,从而导致用户搜寻相关信息的难度增加.对给定的查询,如何准确找到与该查询相关且具有时效性的信息,成为微博检索中广为关注的问题.

有学者通过将时效性因子融合到传统的语言模型[2]中,或者将时效性因子引入到伪相关反馈[3]中来实现微博检索.这些研究工作虽然在一定程度上提高了微博检索的性能,但却忽略了微博固有的丰富的特征[1],如作者权威度、@、转发、话题标记等.

由于排序学习可以融合大量的微博特征,因此近年来,排序学习在微博检索中引起广泛的关注.排序学习[4]是指利用训练数据自动学习排序模型的一类机器学习方法.目前,主流的排序学习方法主要为监督式学习.但是在实际的应用中通常会缺少足够的训练数据,从而降低排序学习的性能影响排序学习的应用范围.直推学习[5,6]作为一种常见的半监督学习方法被用来解决排序学习中训练数据稀疏的问题.在信息检索中,直推学习通常遵循下列步骤[7−9]:

(1) 假定初始检索结果中排名靠前的文档是和查询相关的,排名靠后的文档是和查询无关的;
(2) 利用这些文档作为初始训练数据,训练一个排序模型 M;
(3) 利用 M 对初始检索结果中的其他未标记数据进行预测,将可信度高的数据分别作为正负样本加入到原始的标记数据中;
(4) 若满足终止条件则停止;否则,继续下面的步骤;
(5) 重复步骤(2)~步骤(4).

上述直推学习虽然缓解了训练数据不足的情况,但是由于忽略了查询本身的特性,因此,迭代过程中会有大量的噪音产生.最直观的例子是,在 Tweets11 上的两个查询 MB001 和 MB049:MB001 在使用 KLIM 模型[10]得到的初始检索结果中的前 100 个文档,有超过半数是和该查询相关的;而对于 MB049 来讲,只有一个文档是和该查询相关的、更值得引起注意的是:在整个语料集中,也只有两个文档被官方判定为和 MB049 相关[11].很显然,如果将 MB049 作为训练查询集的元素之一,势必会在迭代过程中引入大量的噪音,从而影响了排序学习的性能[12].

为缓解噪音数据对排序学习检索性能的影响,基于聚类的直推式学习方法被提出[13].该方法分为两步:第 1 步通过聚类的方法选择出高质量的查询;第 2 步通过再次聚类挑选出初始的训练数据.基于聚类的直推学习方法的优势在于在整个过程中不需要人工干预.因此,该方法适合无标记或者只有少量标记数据的应用场景.

本文在原有工作的基础上提出使用基于模型的聚类方法来检验基于聚类的直推式学习方法的性能.另外,在基于聚类的直推式学习中有两个重要的参数,即聚类阈值和聚类文档的数量.在本文中,我们在没有训练数据的情况下,重点研究了这两个参数在不同聚类算法、不同查询集合的应用场景中对模型鲁棒性的影响.最后,我们对原始模型进行了改进,提出了基于聚类的自适应学习方法.

本文第 1 节表述半监督学习在信息检索中的相关研究.第 2 节对基于聚类的直推式学习方法进行简要概述.第 3 节提出一种使用簇凝聚度作为质量控制因子的自适应聚类学习方法.第 4 节对实验数据集、评测指标、实验方法等进行表述.第 5 节分别展示并分析各种方法的实验结果.最后是全文总结以及对未来工作的展望.

1 相关工作

在微博检索中,时效性是非常重要的.如果在检索中忽略了查询的时效,用户体验就会受到影响[14].因此,有研究工作将时间因子融合到传统的语言模型中来实现检索[2].此方法考虑了微博的发布时间,并将文档的先验概率设定为条件概率.Efron 和 Golovchinsky 提出,通过集成排名靠前的文档的发布时间在查询似然模型中,并最大化后验概率[15].Efron 还提出,将时间因子融合到伪相关反馈中[3].因此,文档生成概率就会向该文档的时间戳倾斜,从而能挑选出具有时间特性的查询扩展词.在微博中存在大量体现微博特性的特征[1,16,17],例如:作者权威度体现了微博发布者对其他微博用户的影响力;时效性特征表明该条微博是否是陈旧的.此外,微博还有一些特有的特征,如转发、@、话题标记等.相比以前的研究,排序学习在融合了微博大量丰富的特征下,在微博检索中具有更大的优势.实际上,已经有很多将排序学习应用在微博检索中[17−19]的研究工作.

然而当训练数据匮乏时,排序学习的检索性能势必会受到影响.例如在 TREC 2011 年的微博检索中,由于缺少训练数据,在所有的参赛者中,使用 KLIM[10]模型的检索性能超过了排序学习的性能[11].为了缓解训练数据稀缺的问题,有些研究开始尝试半监督学习方法.半监督学习方法的基本思想是:基于一定的策略,给未标记数据赋予相关性标记,从而增加训练数据的数量[20].有大量的研究将半监督学习应用到分类问题中.Dempster 等人利用 EM 算法估计生成模型的参数和未标记数据的相关性[21].Blum 和 Chawla 提出基于相似性定义的图来决定未标记数据的相关性[22].Blum 和 Mitchell 尝试从多维度、利用不同的学习算法来建立模型[23].Sellamanickam 等人提出了一种基于对的排序模型,该学习过程同时从正文本和未标记数据中学习.他们建立的模型使得正样例的得分高于未标记样例的得分,最后,通过设定阈值来达到分类的目的[24].Huang 等人提出使用 co-training 的方法来对传统的伪相关反馈进行改进,即对候选查询扩展词进行分类,然后选择最有可能的查询词加入到训练样本中[8].Huang 等人继续提出了一种自适应的 co-training 方法来选择好的反馈文档,该方法引入一种阈机制来自动监控新增加的训练数据的质量[7].近几年,也有尝试采用半监督学习方法解决排序问题的研究.Duh 和 Krichhoff 提出了一种直推式学习的框架,该框架从测试数据中生成更好的特征,并将这些特征集成到测试数据中,使之能够更好地适应测试查询集合[25].Li 等人提出使用传统的检索模型和监督式学习方法来共同决定未标记数据的相关性标签[6].直推式学习作为一种常用的半监督学习方法,在缓解微博检索训练数据匮乏的问题中发挥着巨大的作用.Zhang 等人在无标记数据或只有少量标记数据的情况下检验了直推式学习在微博检索中的可行性[9].Zhang 等人提出了利用直推式学习建立查询偏重模型的方法,该查询偏重模型考虑了不同查询的特有的特性[26].具体地,文章以 KL-Divergence 权重作为度量,针对每一查询,抽取能表征该查询的特有的特征词,然后使用半监督方法迭代生成伪标记数据,建立查询偏重模型,以捕捉查询特有的特性;最后,此查询偏重模型和基于通用特征集建立的通用模型加权组合,对初始检索结果列表重排.大量实验表明:该查询偏重模型不仅能有效捕捉每个给定查询特有的特性,从而提高了微博实时检索的正确率;而且该模型非常适合建立在线实时系统.

以上提到的研究试图利用半监督学习对未标记数据打标签,从而达到增加训练数据数量的目的,因此在迭代增加训练数据的过程中会不可避免地引入噪音.基于聚类的直推式学习方法被提出,就是为了减轻迭代过程中的噪音.

2 基于聚类的直推式方法

2.1 选择高质量的查询集

分类学习中存在大量针对噪声数据的研究,此类研究可以分为 3 类:特征选择方法、数据选择方法以及建立容忍噪声模型的方法[12].在这里,我们想通过数据选择的方法来减少迭代过程中噪音数据的引入.在本文中,我们希望通过聚类的方法筛选出高质量的查询,在接下来的直推式学习中,只利用这些高质量的查询来迭代生成伪标记数据,以期减少噪音数据的引入,这也是本文的直接动机.

在传统的直推式学习中,通常假定初始检索结果中排名靠前的文档为正样本,排名靠后的文档为负样本.在无训练数据的情况下,这些正负样本将被作为初始训练集,然后通过一个迭代过程生成伪标记数据.但是这种假设由于忽略了查询的质量,导致在使用半监督学习迭代生成伪标记数据的过程中会因为某些查询的相关文档稀疏且分散而引入低质量甚至错误的伪标记数据,因此并不总是有效的.比如:对于某些查询,语料集中会有大量的相关文档,而且这些相关文档会集中在排名靠前的初始检索结果中;而对于另外一些查询,它们的相关文档数目不仅比较少,还可能会分布得很分散.我们将前者称为容易的查询,后者称为困难查询.如果在使用半监督学习算法迭代生成训练数据的过程中引入这些困难的查询,则势必会有大量的噪音数据生成,从而影响排序学习的检索性能.

很显然,传统的直推学习忽略了不同查询的特性.如果能够通过某种方法将容易的查询挑选出来,将困难的查询移除,则直推式学习迭代过程中生成的训练数据质量会有所提高,从而有助于提高排序学习的检索性能.基于聚类的直推学习的第 1 步就是通过使用聚类方法来选择容易的查询,即高质量的查询.接下来,我们只将半监督算法应用在高质量的查询中,低质量的查询只作为测试集的子集用来检验伪标记数据的有效性.算法的具体

流程如图 1 所示.

输入:
N:每一查询中要聚类的文档数目;
F:用来表示每个文档的特征集;
b:聚类阈值;
输出:
SQ:“容易”的查询集.
算法:
(1) 对属于 *N* 的每个文档用特征集 *F* 表示;
(2) 应用聚类算法对 *N* 个文档聚类为两簇,分别标记为 *G*1,*G*2;
(3) 分别计算 *G*1,*G*2 内排名靠前的文档所占的百分比,分别标记为 *B*1,*B*2;
(4) 选择出满足|*B*1−*B*2|>=*b* 的查询集,标记为 *SQ*

Fig.1 Algorithm of selecting easy queries

图 1 选择容易查询集的算法

2.2 生成初始标记数据

为了进一步减少噪音,基于聚类的直推式学习对传统的生成初始标记数据的方法进行了改进.在传统的直推式学习中,直接假设初始检索结果中排名靠前和靠后的文档分别为相关和不相关.与传统直推学习方法不同,基于聚类的直推学习对 *G*1,*G*2 分别进行二次聚类,并采用投票策略来决定初始标记数据.

经过两步聚类,直推式学习分别筛选出容易的查询集和这些查询集的初始标记数据.直推式学习利用这些初始的标记数据,通过一定次数的迭代最终生成训练数据,从而缓解了排序学习中无训练数据的情况.

2.3 聚类算法

目前存在多种聚类算法,如层次聚类、划分聚类、基于密度的聚类、基于模型的聚类等[11].在本文中,除了利用先前工作中使用的 *K*-Means 算法之外,还采用了基于模型的 EM 算法[21,27].EM 算法又称为最大期望算法,在统计计算中,EM 算法旨在在概率模型中寻找参数最大似然估计或者最大后验估计,其中,在概率模型中通常有无法观测到的隐含变量存在.EM 算法是包含两个步骤的交替迭代算法,其中,第 1 步是 E 步骤,即利用对隐藏变量的现有估计值,计算其最大似然估计值;第 2 步是 M 步骤,即最大化在 E 步骤上求得的最大似然值来计算参数的值.

假设给定的训练样本是$\{x^{(1)},x^{(2)},\ldots,x^{(m)}\}$,EM 算法的目的是找到每个样例隐含的类别 *z*,能使得 $p(x,z)$最大.$p(x,z)$的最大似然估计如下:

$$L(\theta;X)=p(X\mid\theta)=\sum_{Z}p(X,Z\mid\theta).$$

EM 算法的第 1 步是 E 步:计算 log 似然函数关于 $P(Z|X,\theta^{(t)})$的期望值:

$$Q(\theta\mid\theta^{(t)})=E_{Z|X,\theta^{(T)}}[\log L(\theta;X,Z)].$$

第 2 步是 M 步:找到$\theta^{(t+1)}$,使得 $Q(\theta|\theta^{(t)})$最大:

$$\theta^{(t+1)}=\arg\max_{\theta}Q(\theta\mid\theta^{(t)}).$$

通过此迭代过程,EM 算法可以达到局部最优化.

2.4 排序学习算法

目前的排序学习算法可以分为 3 大类,分别为 pointwise,pairwise 和 listwise[4].其中,pairwise 和 listwise 的性能明显好于 pointwise,因此得到了广泛应用[4].在本文中,鉴于查询集的元素比较少,因此我们采用 pairwise 方法.Pairwise 方法从原始训练集中根据标注的相关性的某种大小关系构建偏序文档对集合,从而将排序学习问题转化为二元分类问题[28].Rank SVM[29]作为一种 pairwise 的学习排序算法在信息检索、微博相关性检索[17]等得到了广泛的引用,在本文中,我们使用 Rank SVM 作为排序学习算法.

3 改进的自适应方法

基于聚类的直推式学习针对每一查询中排名靠前的 N 个文档分别聚类,这种聚类方法将聚类文档的数量作为参数.我们在此基础上提出了一种自适应的基于聚类的直推式学习方法,此方法通过引入质量控制因子,对聚类文档的质量进行自动监控.质量控制因子即簇凝聚度,是指每一簇中文档和簇心的余弦平均相似度.改进的自适应方法的算法描述如图 2 所示.在自适应方法中,初始聚类针对排名靠前的 20 个文档,我们分别计算结果簇中的簇凝聚度,若低于簇凝聚度阈值,则将此查询抛弃;否则继续增加聚类文档的数量.当簇凝聚度再次低于阈值时,停止聚类.最终,我们选取整个聚类过程中具有最大簇凝聚度的聚类文档数目作为最优的聚类文档数量.基于聚类的自适应直推式方法的最大优点是:可以自动监控聚类过程中聚类文档的质量,既可以保证每一簇中的文档的相似性,又可以有效减少非相似性文档的引入.

输入:

N:每一查询中要聚类的文档数目集合,*N* 分别取值为 20,40,60,100,200,300;

F:用来表示每个文档的特征集;

C:簇凝聚度阈值;

输出:

SQ:容易的查询集;

ON:簇凝聚度最大的聚类文档数量 *N*1,*N*2.

算法:

(1) *SQ* 初始化为所有的查询集中的元素;

(2) 对属于 *N* 的每个文档用特征集 *F* 表示;

(3) 当 *N*=20 时,应用聚类算法对 *N* 个文档聚类为两簇,分别标记为 *G*1,*G*2;

(4) 分别计算 *G*1,*G*2 的簇凝聚度,并标记为 *C*1,*C*2;

(5) 若 *C*1 或 *C*2 低于阈值 *C*,则停止继续聚类,并将此查询移除 *SQ*,聚类结束;

(6) 否则,增加聚类文档的数量,应用聚类算法对文档重新聚类为两簇,分别标记为 *G*1,*G*2;

(7) 分别计算 *G*1,*G*2 的簇凝聚度,并标记为 *C*1,*C*2;

(8) 若 *C*1 或 *C*2 低于阈值 *C*,则停止继续聚类;否则,重复步骤(6)~步骤(8);

(9) 聚类结束,针对每一查询分别选取 *G*1 和 *G*2 中簇凝聚度最大的聚类文档数量 *N*1,*N*2.

Fig.2 Algorithm of adaptive clustering-based transduction learning

图 2 基于聚类的自适应直推式学习算法

算法中的输入参数簇凝聚度阈值,是通过交叉验证获得.

4 微博检索实验设计

4.1 实验数据集

采用 Tweets11[11]作为实验数据集,Tweets11 是 2011 年 TREC 发布的用来评价实时微博检索的标准数据集.在 TREC 2011 中,我们共成功爬取了 13 401 964 条数据.但是由于这些微博中的数据具有动态性(这些数据可能会被删除或者被设置访问权限),在 TREC 2012 中,我们使用官方发布的 id-status.01-May-2012.gz 对 13 401 964 条数据进行了过滤.在 TREC 2011 中包含 49 个查询,分别标记为 MB001 至 MB049;在 TREC2012 中包含 59 个查询,分别标记为 MB051~MB110(鉴于 MB076 无相关文档,因此在最终的评价中被移除).之所以选择 Tweets11,是在此数据集上可以利用各种特征,这些特征既包括文档内容相关的,又包括@,RT 等 Tweet 本身的特征.在实验中,我们采用了文献[10]中用到的特征集.

建立索引和检索过程都采用了标准的去除停用词和 Porter 词干还原,索引的建立和检索是基于 Terrier 3.0[30]的二次内部扩展实现的.

4.2 评测标准

我们采用 TREC 微博检索中的官方评价标准,即 Precision at 30(P30)和 Mean Average precision(MAP).

4.3 测试方法

直推式学习算法在检索任务中的一种直接应用,就是将所有可得的查询都作为训练查询集来迭代生成伪标记数据,我们把这种方法标记为 TL.为检验基于聚类算法的直推式学习(CTL)的有效性,我们采用 3-折交叉验证做了一系列的实验.在实验中,我们将 TL 作为基准实验,比较当 P30 和 MAP 作为评测标准时,CTL 和 TL 的实验对比结果,而 TL 也是相关研究通常选用的一个基准[7,26].3 折交叉验证方法,即根据查询编号将查询随机分为 3 折:一折用于训练,一折用于测试,另外一折用于验证.其中,训练数据是通过基于聚类的直推式学习方法得到的.我们采用 *K*-means 和 EM 两种聚类算法对每个查询的初始检索结果进行聚类,分别标记为 KM_CTL 和 EM_CTL.*K*-means 是一种基于划分的聚类算法,之所以选择 *K*-means 对初始检索结果聚类,是因为该算法不仅框架简单,而且具有很好的伸缩性和高效性,时间复杂度仅为 $O(KNt)$,其中,K 为簇的数目,N 为聚类对象的数目,t 为迭代次数.更重要的:在我们的实验中,对于高质量的查询和低质量的查询,它们的文档分布呈现明显的区别.即在高质量的查询中,排名靠前和靠后的文档具有更相对清晰的分界线;而在低质量的查询中,排名靠前和靠后的文档混杂在一起,二者的分界线不明显.*K*-means 在处理簇之间具有显著差异性的情境中聚类的效果很好,因此非常适合我们的实验.EM 是一种基于模型的聚类算法,之所以选择 EM 作为聚类算法,是因为该模型相比 *K*-means 更具有一般性,因为它可以使用各种类型的分布;而且 EM 可以发现任何不同大小或者椭球形状的簇.在实验中,我们设置不同的聚类阈值以及聚类的文档数量,通过一系列的实验分析了基于聚类的直推式学习算法的鲁棒性.其中,聚类阈值和聚类文档的数量的参数设置详见表 1.

Table 1 Values of clustering threshold and the number of clustering documents

表 1 聚类阈值和聚类文档数目

参数	值
聚类阈值	0.4, 0.6
文档数量	20, 40, 60, 100, 200, 300

5 实验结果与性能分析

5.1 基于聚类的直推式学习方法的实验结果及分析

由于 TREC 2011 和 TREC 2012 上的查询集表现出不同的特性(例如,TREC 2011 上较高的 MAP 预示着在 TREC 2011 上的查询相对容易,然而 TREC 2012 上的查询相对困难),为了检验在混合查询集上基于聚类的直推学习方法的检索性能,我们将 TREC 2011 和 TREC 2012 合并作为一个新的混合数据集.

以下的表格和图中 TREC 2011 代表从 MB001~MB049 的查询;TREC 2012 代表从 MB051~MB110 的查询(不包含 MB076);TREC 2011 & TREC 2012 代表从 MB001~MB110 的查询(不包括 MB050 和 MB076).显著性测试是在 0.05 级别上的对随机测试[31]上进行的.

5.1.1 聚类阈值对排序模型的影响

聚类的阈值决定了训练查询集的数量,同时也影响了训练查询集的质量.在文档数目固定的情况下,我们分别将聚类阈值设置为 0.4 和 0.6,以此检验聚类阈值对排序模型的性能造成的影响.

表 2 和表 3 分别展示了不同数值的聚类阈值 b 对实验结果的影响,其中,黑色加粗的结果表示相对于基准实验 TL 具有显著性提高.当使用 EM 对初始检索结果进行聚类时,在 TREC 2011 上,EM_CTL 相比 TL 并无显著性提高,甚至当以 MAP 为评测准则时,EM_CTL 的检索性能还稍微下降;在 TREC 2011 & 2012 上,不论以 P30 还是 MAP 为评测标准,EM_CTL 的实验结果相比 TL 都有显著性提高.在 TREC 2011 & 2012 上,以 P30 为评测标准时,EM_CTL 相比 TL 具有显著新提高.当使用 *K*-means 作为聚类算法时,实验结果与 EM_CTL 类似.总的来讲:CTL 在 TREC 2011 上的实验结果相比 TL 无显著性提高;在 TREC 2012 以及 TREC 2011 & 2012 上,CTL 相比 TL 都具有统计上地显著性提高.我们认为,这是因为在 TREC 2011 上的查询集本身的质量就很高(相比 TREC2012,在 TREC 2011 上 TL 的 P30 和 MAP 就比较高),这预示着在 TREC 2011 上可能没有必要进行聚类筛选高质量的查询,相反,聚类反而只是筛选出了部分高质量的查询遗漏掉了另一部分高质量的查询,从而导致查

询质量并无显著性提高.而在 TREC2012 上,整体的查询集质量偏低,使用基本模型得到的初始检索结果携带的相关信息相对偏少,因此,通过使用 EM 或者 *K*-means 对每一个查询的初始检索结果进行聚类,以期得到高质量的查询.同时我们还观测到,*b* 的最优取值和聚类算法的选择有密切关系.当 *b*=0.4 时,KM_CTL 的性能优于 EM_CTL,尤其是在 TREC 2012 上;而当 *b*=0.6 时,EM_CTL 的性能优于 KM_CTL.实验结果也表明了基于聚类的直推式学习方法在 TREC 2012 以及 TREC 2011&2012 上的有效性.

Table 2 Experimental results obtained by applying the clustering algorithm (*b*=0.4)

表 2 当 *b*=0.4 时,基于聚类算法的实验结果

	TL	KM_CTL	EM_CTL
TREC 2011			
P30	0.4	0.408 8	0.406 8
MAP	0.342 5	0.339 9	0.338 5
TREC 2012			
P30	0.337 3	**0.367 2**	**0.363 3**
MAP	0.212 1	**0.233 6**	**0.226 3**
TREC 2011&2012			
P30	0.365 7	**0.386 1**	**0.380 6**
MAP	0.271 3	0.278 9	0.275 4

Table 3 Experimental results obtained by applying the clustering algorithm (*b*=0.6)

表 3 当 *b*=0.6 时,基于聚类算法的实验结果

	TL	KM_CTL	EM_CTL
TREC 2011			
P30	0.4	0.406 8	0.406 8
MAP	0.342 5	0.333 5	0.341 9
TREC 2012			
P30	0.337 3	**0.362 1**	**0.365 0**
MAP	0.212 1	**0.230 6**	**0.237 8**
TREC 2011&2012			
P30	0.365 7	**0.381 8**	**0.384 0**
MAP	0.271 3	**0.280 3**	**0.283 9**

我们可以看出:无论采用哪种聚类方法,聚类阈值 *b* 都会对排序学习的性能造成一定的影响,而且这种影响通常具有统计显著性.总体来说:在 TREC 2011 上,*b* 的影响会小于在其他查询集上的影响.我们认为:这是由于在 TREC 2011 中的查询更容易,因此导致查询的质量较高,因而这些查询的初始检索结果中,排名靠前的文档携带了更多和给定查询相关的信息.因此,无论采用哪种聚类方法选择的初始训练数据质量都比较高,这在一定程度上保证学习性能的稳定性.对于 TREC 2012 中的查询,初始检索结果列表携带的与主题相关的信息相对较少,这意味着 TREC 2012 上的查询更困难些.虽然采用两步聚类的方法提取高质量的查询和初始伪标记数据之后,最终的排序学习性能有显著性提高,但是由于聚类阈值 *b* 控制着聚类的聚合度,当聚类阈值较小时,表示每一簇的聚合度越小,这也表示同一簇中的文档的相关性越小,簇之间的差异性越小,因而会有越多的噪音数据引入,从而对模型的检索性能造成了一定影响.因此,基于两步聚类的直推式学习方法在 TREC 2012 上的鲁棒性较差.

5.1.2 聚类文档数目对排序模型的影响

聚类文档数量作为另一个基于聚类的直推式学习方法的参数,势必会对排序学习的性能造成影响.在本实验中,我们分别对初始检索结果中排名靠前的 20,40,60,100,200,300 个文档聚类,通过一系列的实验,研究聚类文档数量对排序学习性能的影响.

表 4 和表 5 向我们展示了以 P30 和 MAP 为评价指标,当聚类阈值 *b* 固定的情况下,聚类文档数量对排序学习模型的检索性能造成的影响.不难看出:无论使用 EM 还是 KM 聚类方法,文档数量对 TREC 2011 的影响都明显小于 TREC 2012 以及 TREC 2011 & 2012,无论聚类阈值 *b* 怎么选择,聚类文档的数量对 TREC 2012,TREC 2011 & 2012 的影响都明显大于 TREC 2011.虽然在 3 个查询集上取得最优值的文档数量不同,但是不可否认的是:在 *b* 确定的情况下,只要聚类文档数量得到了优化,最终的 P30 就会达到最优.同时我们还注意到:无论采用哪

一种聚类算法,最终得到的最优化的 P30 无显著性差异.但是聚类文档的数量仍然在不同查询集上对检索性能具有不同的影响,相对而言,在 TREC 2011 上,聚类文档数量对 MAP 的影响则相对较少.我们坚持认为:这是由于 TREC 2011 中排名靠前的文档的高聚合度造成的,即在 TREC 2011 中所有查询的 MAP 明显高于 TREC 2012. 这意味着相比 TREC 2012,TREC 2011 中存在更多的查询,它们的初始检索结果中排名靠前的文档与给定查询的相关性更高.

Table 4 Experimental results obtained by applying CTL with different numbers of clustering documents (b=0.4)

表 4 当 b=0.4 时,不同数目的聚类文档设置下 CTL 的实验结果

		20	40	60	100	200	300
TREC 2011							
TL	P30	0.400 0					
	MAP	0.342 5					
KM_CTL	P30	0.400 0	0.398 0	**0.408 8**	0.393 2	0.404 8	0.355 1
	MAP	**0.339 9**	0.333 4	0.337 9	0.320 6	0.330 0	0.298 5
EM_CTL	P30	0.395 2	0.403 4	0.403 4	**0.406 8**	0.401 4	0.393 9
	MAP	0.329 7	0.334 6	**0.338 5**	0.335 7	0.327 4	0.323 6
TREC 2012							
TL	P30	0.337 3					
	MAP	0.212 1					
KM_CTL	P30	0.342 9	0.362 1	**0.367 2**	0.357 1	0.358 8	0.349 2
	MAP	0.220 3	**0.233 6**	0.224 0	0.211 6	0.230 8	0.210 6
EM_CTL	P30	0.349 2	0.353 7	0.355 9	0.355 9	**0.363 3**	0.345 2
	MAP	0.224 0	**0.226 3**	0.222 2	0.210 0	0.219 5	0.206 9
TREC 2011&2012							
TL	P30	0.365 7					
	MAP	0.271 3					
KM_CTL	P30	0.368 8	0.378 4	**0.386 1**	0.373 5	0.379 6	0.351 9
	MAP	0.274 6	**0.278 9**	0.275 6	0.261 0	0.275 8	0.250 5
EM_CTL	P30	0.370 1	0.376 2	0.377 5	0.379 0	**0.380 6**	0.367 3
	MAP	0.272 0	**0.275 4**	0.274 9	0.267 0	0.268 4	0.259 9

Table 5 Experimental results obtained by applying CTL with different numbers of clustering documents (b=0.6)

表 5 当 b=0.6 时,不同数目的聚类文档设置下 CTL 的实验结果

		20	40	60	100	200	300
TREC 2011							
TL	P30	0.400 0					
	MAP	0.342 5					
KM_CTL	P30	0.406 1	0.400 7	0.399 3	0.390 5	0.389 8	**0.406 8**
	MAP	0.333 5	0.323 5	0.325 8	0.320 9	0.331 9	**0.340 1**
EM_CTL	P30	0.404 1	0.406 1	0.397 3	0.391 2	**0.406 8**	**0.406 8**
	MAP	**0.341 9**	0.332 2	0.320 0	0.329 7	0.339 4	0.332 5
TREC 2012							
TL	P30	0.337 3					
	MAP	0.212 1					
KM_CTL	P30	0.361 0	0.358 2	**0.362 1**	0.348 6	0.357 1	0.361 0
	MAP	0.222 1	0.219 0	0.222 9	0.217 2	0.228 1	**0.230 6**
EM_CTL	P30	0.355 4	0.348 0	0.358 2	0.350 8	**0.365 0**	0.346 3
	MAP	0.215 9	0.219 3	0.219 4	0.216 9	**0.237 8**	0.206 2
TREC 2011&2012							
TL	P30	0.365 7					
	MAP	0.271 3					
KM_CTL	P30	0.381 5	0.377 5	0.379 0	0.367 6	0.371 9	**0.381 8**
	MAP	0.272 6	0.266 4	0.269 5	0.264 2	0.275 2	**0.280 3**
EM_CTL	P30	0.377 5	0.381 8	0.375 9	0.369 1	**0.384 0**	0.373 8
	MAP	0.273 1	0.270 5	0.265 1	0.268 1	**0.283 9**	0.263 5

我们可以总结得出:在 TREC 2011 上,基于聚类的直推式学习方法对聚类文档的数量具有更好的鲁棒性;而

在 TREC 2012 上,模型的鲁棒性较差.归根结底,这是因为在 TREC 2011 中,初始检索结果列表中排名靠前的文档提供了与相应主题更切合的信息,因而导致这些文档之间的差异性较小,相似性更大,因此,即使聚类文档数量变化,也不会引入大量的噪音文档,因此模型表现了较强的鲁棒性;而在 TREC 2012 中,由于基本模型返回的初始检索结果列表携带的相关信息较少,导致随着聚类文档数量的增加,原本不相关的文档会被错认为相关,导致有越来越多的噪音文档引入,从而降低了检索结果的性能,也导致模型表现出较差的鲁棒性.

5.2 基于聚类的自适应直推式学习方法的实验结果及分析

实验中仍然使用 TL 作为基准实验结果.表 6 中黑色加粗的部分表明,我们提出的自适应的直推式学习方法(Adapt_CTL)相比 TL 具有统计上的显著性提高.例如:当使用 EM 算法对初始检索结果聚类时,自适应学习方法 Adapt_CTL 在 TREC 2012 上将 P30 提高了 5.01%;当使用 *K*-means 聚类时,不管以 P30 还是 MAP 为评测标准,在 TREC 2012 上,Adapt_CTL 相比 TL 都具有统计上的显著性提高(在 P30 上提高了 6.70%,在 MAP 上提高了 6.98%).下面着重对 Adapt_CTL 的鲁棒性进行分析.

Table 6 Experimental results obtained by applying the adaptive transduction method

表 6 自适应直推式学习的实验结果

	TREC11		TREC12		TREC11&12	
	TL	Adapt_CTL	TL	Adapt_CTL	TL	Adapt_CTL
	EM					
P30	0.4	0.403 4	0.337 3	**0.354 2**	0.365 7	**0.376 5**
MAP	0.342 5	0.340 1	0.212 1	0.213 7	0.271 3	0.271 1
	K-means					
P30	0.4	0.406 8	0.337 3	**0.359 9**	0.365 7	**0.376 9**
MAP	0.342 5	0.341 9	0.212 1	**0.226 9**	0.271 3	0.272 9

在以下的图中:Iter_MAP 和 Iter_P30 分别代表不同迭代次数设置的情况下,使用自适应基于聚类的直推式学习方法的得到的 MAP 和 P30;MAP 和 P30 则分别代表参数优化后的自适应聚类方法得到的检索结果.图中横坐标代表迭代次数,纵坐标代表 MAP 或 P30.

图 3 展示了使用 *K*-means 作为聚类算法时,不同迭代次数设置情况下和参数优化后基于聚类的自适应算法在 TREC 2011,TREC 2012,TREC 2011 & 2012 检索结果的比较.

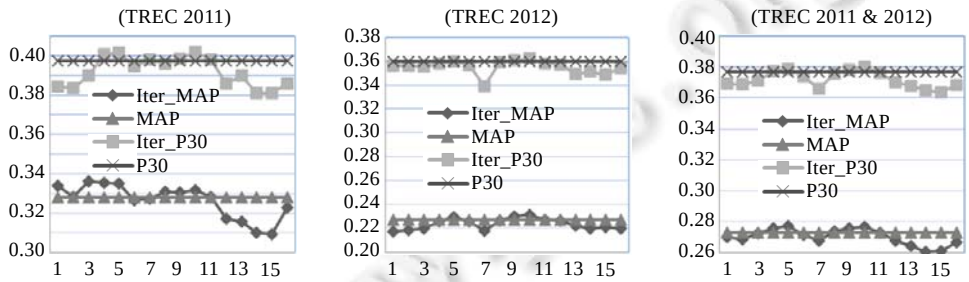

Fig.3 Taking KM as the clustering algorithm, the retrieval effectiveness of the proposed adaptive clustering approach on TREC 2011, 2012, 2011 & 2012 with optimal parameters and different settings of the iterative count

图 3 使用 KM 聚类,参数优化的自适应聚类算法和不同迭代次数设置下,在 TREC 2011,2012,2011 & 2012 上的检索结果

可以看出:相比在第 5.1 节中检索结果的大幅度变化,自适应的聚类学习算法具有更好的鲁棒性.具体地,虽然随着迭代次数的增加(尤其是当迭代次数大于 12 时),自适应算法在 TREC 2011 上的检索结果有微小波动,但是当迭代次数控制在 11 之内时,自适应学习算法的鲁棒性能表现良好;在 TREC 2012 以及 TREC 2011 & 2012 上,自适应学习算法的性能随着迭代次数的变化基本稳定,表现了更强的鲁棒性.我们猜测:在引入了簇凝聚度作为质量控制因子的情景下,基于聚类的直推式学习方法能够自动监控不同数目的聚类文档的聚合度,从而增

强了模型的鲁棒性.至于在 TREC 2011 上,当迭代次数大于 12 时,自适应学习算法的检索性能波动幅度较大,这可能是因为在 TREC 2011 上,排名靠前的初始检索结果携带了更多与查询主题相关的信息,而排名越靠后的文档携带的相关信息更少,因此与查询主题不符;当迭代次数大于 12 时,有更多的排名靠后的文档加入到伪标记数据中,因此降低了伪标记数据的质量,从而也导致了排序学习性能的略微降低.

图 4 展示了使用 EM 作为聚类方法时,基于聚类的自适应学习算法在不同迭代次数下检索性能的比较.相比 KM 聚类得到的检索结果,基于 EM 的自适应学习算法的鲁棒性能稍差一些.这主要表现在 TREC 2011 上当迭代次数控制在 2 以内时,自适应学习算法的检索性能变化幅度较大.但是总体来讲,基于聚类的自适应学习算法的鲁棒性相比在无质量控制因子控制下的鲁棒性能更好.前两个子图是自适应学习算法在 TREC 2012, TREC 2011 & 2012 上随着迭代次数的增加检索性能的变化.可以很明显地看出:不管是以 P30 还是 MAP 作为评价指标,尽管迭代次数从 1 变化到 16,基于聚类的自适应学习算法的检索性能基本保持稳定.这从另一侧面反映了在引入簇凝聚度作为质量控制因子后,基于聚类的直推式学习算法的鲁棒性有了较大提高.

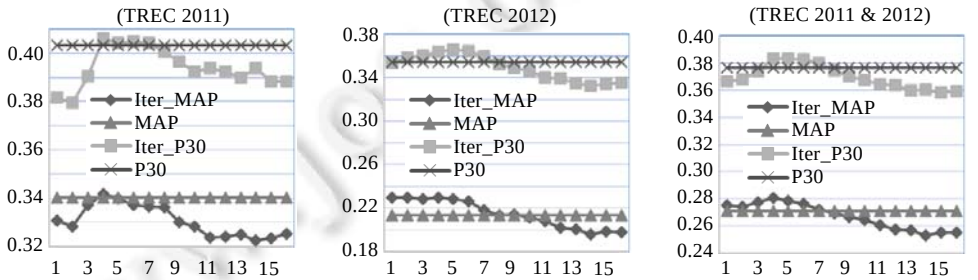

Fig.4 Taking EM as the clustering algorithm, the retrieval effectiveness of the proposed adaptive clustering approach on TREC 2011, 2012, 2011 & 2012 with optimal parameters and different settings of the iterative count

图 4 使用 EM 聚类,参数优化的自适应聚类算法和不同迭代次数设置下,
在 TREC 2011,2012,2011 & 2012 上的检索结果

综上所述,基于聚类的自适应直推式学习(Adapt_CTL)相比 CTL 对迭代次数的敏感性更低,这也意味着 Adapt_CTL 模型具有更好的鲁棒性.我们认为:在引入了凝聚度作为质量控制因子来自动监控聚类文档的相似度之后,Adapt_CTL 能够实时对每一簇中的聚类文档的聚合度进行检测,从而能够保证在半监督学习的迭代过程中将高度相似的文档加入到相关簇中,相似度低的文档舍弃.通过这种方法,Adapt_CTL 有效减少了迭代过程中噪音数据的引入,因此模型也具有更好的鲁棒性.

6 结论与展望

训练集的质量对排序学习算法的性能有直接的影响.基于聚类的直推式学习方法虽然可以应用在无训练数据的情况下,但是在聚类过程中的两个参数聚类阈值以及聚类文档数量对模型的性能都会有敏感影响.在本文中,我们通过一系列的实验,重点分析了两个参数对模型性能的影响.此外,我们还提出了利用基于聚类的 EM 算法来检验直推式学习的性能,并通过大量的实验对比分析了 EM 和 KM 两种聚类算法对排序学习模型性能的影响.实验结果表明:最终的排序学习的性能对聚类阈值和聚类文档数量都是敏感的;相比 TREC 2012,TREC 2011 & 2012,对于 TREC 2011 上的查询集排序学习的性能随着聚类阈值和文档数量的变化更具有相对的稳定性;无论采用 EM 还是 KM 聚类算法,只要聚类阈值和文档数量得到了优化,最终的排序学习的性能就能得到优化,尤其是采用 P30 作为评价指标时.在对基于聚类的直推式学习方法进行了详细的鲁棒性分析之后,我们将簇凝聚度作为质量控制因子引入到基于聚类的直推式学习方法中,提出了基于聚类的自适应学习算法.实验结果表明,基于聚类的自适应学习算法在不同查询集上表现了更强的鲁棒性.

下一步研究的重点将是检验在有少量标记数据的情况下,聚类阈值和聚类文档数量对基于聚类的直推式

学习方法的影响.另外,我们也计划会在LETOR数据集上对基于聚类的直推式学习方法的鲁棒性进行进一步的分析.

References:

[1] Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media? In: Proc. of the 19th Int'l World Wide Web (WWW) Conf. New York: ACM Press, 2010. 591−600.

[2] Li XY, Croft WB. Time-Based language models. In: Proc. of the twelfth Int'l Conf. on Information and Knowledge Management (CIKM 2003). New York: ACM Press, 2003. 469−475.

[3] Efron M, Golovchinsky G. Estimation methods for ranking recent information. In: Proc. of the 34th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2011). New York: ACM Press, 2011. 495−504.

[4] Liu TY. Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 2009,3(3):225−331.

[5] El-Yaniv R, Pechyony D. Stable transductive learning. In: Proc. of the 19th Annual Conf. on Learning Theory (COLT 2006). 2006. 35−49.

[6] Li M, Li H, Zhou ZH. Semi-Supervised document retrieval. Information Processing and Management, 2009,45:341−355.

[7] Huang JX, Miao J, He B. High performance query expansion using adaptive co-training. Information Processing and Management, 2013,49(2):441−453.

[8] Huang X, Huang YH, Wen M, An A, Liu Y, Poon J. Applying data mining to pseudo-relevance feedback for high performance text retrieval. In: Proc. of the IEEE Int'l Conf. on Data Mining Series (ICDM). IEEE, 2006. 295−306.

[9] Zhang X, He B, Luo TJ. Transductive learning for real-time Twitter search. In: Proc. of the Int'l Conf. on Weblogs and Social Media (ICWSM). AAAI, 2012. 611−614.

[10] Amati G, Amodeo G, Bianchi M, Celi A, Nicola CD, Flammini M, Gaibisso C, Gambosi G, Marcone G. Fub, IASI-CNR, UNIVAQ at trec 2011. Technical Report, Gaithersburg: TREC, 2011.

[11] Ounis I, Macdonald C, Lin J, Soboroff I. Overview of the TREC 2011 microblog track. Technical Report, Gaithersburg: TREC, 2011.

[12] Geng XB, Qin T, Liu TY, Cheng XQ, Li H. Selecting optimal training data for learning to rank. Information Processing and Management, 2011,47(5):730−741.

[13] Zhang X, He B, Luo TJ, Li DX, Xu JG. Clustering-Based transduction for learning a ranking model with limited human labels. In: Proc. of the 22nd ACM Int'l Conf. on Information & Knowledge Management (CIKM 2013). New York: ACM Press, 2013, 1777−1782.

[14] Dong A, Chang Y, Zheng ZH, Mishne G, Bai J, Zhang RQ, Buchner K, Liao C, Diaz F. Towards recency ranking in Web search. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining (WSDM 2010). New York: ACM Press, 2010. 11−20.

[15] Efron M, Golovchinksy G. Estimation methods for ranking recent information. In: Proc. of the 34th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2011). New York: ACM Press, 2011. 495−504.

[16] Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring user influence in twitter: The million follower fallacy. In: Proc. of the Int'l Aaai Conf. on Weblogs and Social Media (ICWSM). AAAI, 2010.

[17] Duan YJ, Jiang L, Qin T, Zhou M, Shum HY. An empirical study on learning to rank of tweets. In: Proc. of the 23rd Int'l Conf. on Computational Linguistics (COLING 2010). Stroudsburg: Association for Computational Linguistics, 2010. 295−303.

[18] Metzler D, Cai C. USC/ISI at TREC 2011: Microblog track. Technical Report, Gaithersburg: TREC, 2011.

[19] Miyanishi T, Okamura N, Liu XX, Seki K, Uehara K. TREC 2011 microblog track experiments at KOBE University. Technical Report, Gaithersburg: TREC, 2011.

[20] Vapnik VN. Statistical Learning Theory. New York: Wiley, 1998.

[21] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1977,39(1):1−38.

[22] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: Brodley CE, Danyluk AP, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). San Francisco: Morgan Kaufmann Publishers, 2001. 19−26.

[23] Blum A, Mitchell TM. Combining labeled and unlabeled data with co-training. In: Proc. of the 11th Annual Conf. on Computational Learning Theory (COLT'98). New York: ACM Press, 1998. 92−100.

[24] Sellamanickam S, Garg P, Selvaraj SK. A pairwise ranking based approach to learning with positive and unlabeled examples. In: Berendt B, de Vries A, Fan WF, Macdonald C, Ounis I, Ruthven I, eds. Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management (CIKM 2011). New York: AC Press, 2011. 663−672.

[25] Duh K, Kirchhoff K. Learning to rank with partially-labeled data. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2008). New York: ACM Press, 2008. 251−258.

[26] Zhang X, He B, Luo TJ, Li B. Query-Biased learning to rank for real-time Twitter search. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management (CIKM 2012). New York: ACM Press, 2012. 1915−1919.

[27] Alldrin N, Smith A, Turnbull D. Clustering with EM and *k*-means. Technical Report, California: University of San Diego, 2003.

[28] Niu SZ, Cheng XQ, Guo JF. Noise sensitivity in learning to rank. Journal of Chinese Information Process, 2012,26(5) (in Chinese with English abstract).

[29] Joachims T. Optimizing search engines using clickthrough data. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2002). 2002. 133−142.

[30] Ounis I, Amati G, Plachouras V, He B, Macdonald C, Lioma C. Terrier: A high performance and scalable information retrieval platform. In: Proc. of the ACM Workshop on Open Source Information Retrieval (SIGIR 2006). New York: ACM Press, 2006.

[31] Smucker MD, Allan J, Carterette B. A comparison of statistical significance tests for information retrieval evaluation. In: Proc. of the 16th ACM Conf. on Information and Knowledge Management (CIKM 2007). New York: ACM Press, 2007. 623−632.

附中文参考文献:

[28] 牛树梓,程学期,郭嘉丰.排序学习中数据噪音敏感度分析.中文信息学报,2012,26(5).

张新(1988－),女,山东泰安人,博士生,主要研究领域为社会计算,半监督学习.
E-mail: zhangxin510@mails.ucas.ac.cn

罗铁坚(1962－),男,博士,教授,博士生导师,主要研究领域为信息检索与推荐系统,大规模系统性能优化,大数据分析与知识发现,自适应学习云平台.
E-mail: tjluo@ucas.ac.cn

何苯(1979－),男,博士,副教授,主要研究领域为社会计算,信息检索.
E-mail: benhe@ucas.ac.cn

李东星(1988－),男,硕士生,主要研究领域为信息检索.
E-mail: lidongxing12345@163.com