



大数据中基于时态特征和混合式搜索的博客筛选挖掘

张丽娜, 匡泰, 姜迪清

(浙江安防职业技术学院信息工程系, 浙江 温州 325000)

摘要:针对现存很多博客筛选挖掘方法的相关性程度比较松散以及信息检索方法的缺陷,提出一种基于时态特征和混合式搜索的方法。考虑到用户评论是组合证据的重要来源以及时间因素的影响,提出的方法将博客文章的平均评论数量、消息来源的 BM25 的相关性分数、最久博客文章的 BM25 分数和最新相关博文和最旧博文的时间范围作为时态特征集。另外,考虑到线性搜索的局部性优势以及差分进化搜索的全局优势,将两种信息搜索方式组合。实验使用 BlogS06 数据集,由博客主页、XML 源文件和其博客入口页面组成,用于 TREC 2007 和 TREC 2008 的博客筛选挖掘实验。实验结果表明,提出的方法在运行时间和有效性方面获得了满意的效果。

关键词:博客筛选挖掘;时态特征;线性搜索;差分进化;大数据;BM25

中图分类号:TP391

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2017001

Blog screening and mining based on temporal features and hybrid search in big data

ZHANG Lina, KUANG Tai, JIANG Diqing

Department of Information Engineering, Zhejiang College of Security Technology, Wenzhou 325000, China

Abstract: Concerning that the correlation degree of the existing methods of blog screen and mining is loose and the information retrieval of the methods is deficient, a method based on temporal feature and hybrid search method was proposed. Considering the user reviews are important sources of evidence combination, the average number of reviews for blogs, the sources of BM25 relevance scores, the longest blog BM25 scores and time range between the latest related blog paper and the oldest related blog paper are being as the temporal feature sets. In addition, considering local search advantage of linear search(LS) and global search advantage of differential evolution(DE), the two kinds of information search methods were combined. BlogS06 data set was used in the experiment which consists of blog home pages, XML source files and its blog portal pages, it was used for TREC 2007 and TREC 2008 blog mining experiments. Experimental results show that the proposed method can obtain satisfactory results in terms of running time and effectiveness.

Key words: blog screening and mining, temporal feature, linear search, differential evolution, big data, BM25

收稿日期:2016-05-30;修回日期:2016-09-14

基金项目:浙江省 2016 年教育技术研究规划课题支持项目(No.JB139)

Foundation Item: Educational Technology Research Program of Zhejiang Province in 2016(No.JB139)



1 引言

近些年,社交网络及其手机上的应用软件获得了巨大的成功,如微信和 Facebook,其中,微信拥有最多的使用者,其商业价值可想而知。而社交网络的兴起和成功也激发了博客的演变,将其从非正式讨论或非正式站点演变成针对特定领域(如科技、时尚和财经)的专业平台^[1]。创建和管理博客平台促进了互联网上博客社区的大规模发展。这种专业化信息资源使得博客用户的兴趣复杂化。因此,博客信息搜索挖掘^[2]也成了大数据时代的重要研究课题。

博客筛选挖掘一般定义为搜索某一主题博客的过程,可以表述为文本查询^[3]。该任务可以概括为:用某一原理寻找对主题 T 经常感兴趣的博主。对于一个已知主题 T,系统应给出主题 T 的消息来源。

博客筛选挖掘按照提取工作一般可以分为以下两种。

(1)使用博客消息来源^[4]提供的信息,估计博主和某一特定主题的相关性

如参考文献[5]结合消息来源层面(feed)与博客文章层面的相关分数,采用语言模型以检索与一个特定查询相关联的消息来源。参考文献[5]还提出了一个基于维基百科数据提取的查询扩展技术。其中,文本检索会议(text retrieval conference, TREC)的博客筛选提取就采用该扩展技术^[3,4]。参考文献[6]定义了一种投票模型,博客查询与该博客相关的文章,检索到的每篇相关博客文章都被当作对该博客相关性的一次投票。采用不同的文本倾向性处理方法,对于较长篇幅、不同评论和字符权重进行评述,使模型具有倾向识别能力。参考文献[7]遵循资源选择原理,综合利用本体论和 Folksonomy 的优势,进行二级分类。然后在分类算法的基础上结合了用户使用博客兴趣,考虑了不同博客的表征信息。然而,对给定的查询只能给出高排序的博客文章,对高度相关的博客比较有效。参考文献[8]使用线性搜索(linear search, LS)检索博客消息来源,LS 是信息检索中一种简单有效的方法。其参数值被训练样本(针对某一特定评价尺度)优化,训练后的参数值使用一个实验样本进行评价。用一个或两个参数值寻求穷举搜索的最佳参数设定。但优化过程是一个瓶颈,其计算成本会伴随参数数量呈指数增长。另外,多种来源的信息组合问题也可以采用排序学习方法,如粒子群优化^[9](particle swarm optimization, PSO)和差分进化^[10](differential evolution, DE)等。

(2)越过消息来源,尝试检测该主题与个人博客间的关联性^[11]

如博客时间戳、链接分析或外部资源提供的信息,但使用频率相比第一种很小。博客筛选挖掘其实是一个信息融合问题,需要将多种信息组合起来。该过程获得哪些有效特征以及如何使用这些特征是问题的关键。现有的很多方法相关性程度比较松散,为此,本文组合了多个证据源,将关联信号(在博客或文章层面上)与多个异构信号(社会时态特征)相结合,并使用混合方法检索信息。本文的主要工作是:提出了一组基于社会性的时态信息估算博客的相关性;提出了一种混合的博客信息融合方法组合这些信息,以找到潜在的相关博客,整合了全局搜索和局部搜索的优点。

2 博客相关性的估算

2.1 证据组合

一些选定的特征表示内容匹配的证据,与查询主题相关,而基于时态或基于社会特征则独立于查询。本文利用 F 代表消息来源, Q 表示一个查询,相关特征查询用 $qd_feat_k(F, Q)$ 表示。为了结合有效的特征,本文对 $qd_feat_k(F, Q)$ 做如下归一化:

$$qd_feat_{k(norm)}(F, Q) = \frac{qd_feat_k(F, Q)}{\max_{F_i \in C} qd_feat_k(F_i, Q)} \quad (1)$$

其中, C 表示消息来源的集合。而独立查询特征的正则化:

$$q_feat_{k(norm)}(F) = \frac{q_feat_k(F)}{\max_{F_i \in C} q_feat_k(F_i)} \quad (2)$$

其中, F 表示消息来源, $q_feat_k(F)$ 是独立查询特征, C 是消息来源集合。这种类型的正则化被广泛使用,是一种较好的线性检索方案。本文遵循线性结合方法,从独立查询特征和相关查询特征组合中找到证据。

$$Rel(F, Q) = \sum_i \alpha_i \cdot qd_feat_{i(norm)}(F, Q) + \sum_j \beta_j \cdot q_feat_{j(norm)}(F) \quad (3)$$

其中, qd_feat_{norm} 是正则化的独立查询特征, α_i, β_j 是自由参数。通过线性结合获得组合证据是一个简单有效的方法。

2.2 信息检索模型

一般有两种不同模型评估文档间的相关性,即 BM25 和狄氏(Dirichlet)平滑一元语言模型。BM25 基于概率框架^[5],本文使用 Lemur 的 BM25 匹配函数^[12],即:

$$BM25(D, Q) = \sum_{t \in (Q \cap D)} w \cdot \frac{(K_t + 1) f_{t,D}}{K((1-b) + b \times (L_D / L_{ave})) + f_{t,D}} \times \frac{(K_s + 1) f_{t,D}}{K_s + f_{t,D}} \quad (4)$$

其中, $f_{t,D}$ 是文档 D 中 t 的频率, $f_{t,D}$ 是查询 Q 中 t 的频率, L_D 和 L_{ave} 是文档 D 的长度, 即整个采集文档的平均长度。而 w 则是倒排文档频率权值, w 定义如下:

$$w = \lg \left(\frac{N - n + 0.5}{n + 0.5} \right) \quad (5)$$

其中, N 是采集中文档的总数量, n 是包含项 t 的文档数量。狄氏平滑一元语言模型如式(6)所示:

$$Dirichlet(D, Q) = \prod_{i=1}^n \frac{tf(q_i, D) + \mu \cdot P(q_i | C)}{|D| + \mu} \quad (6)$$

其中, D 是一个文档, n 是查询项数量, $tf(q_i, D)$ 是在 D 中 q_i 的原始频率项, $|D|$ 是文档长度, μ 是调整平滑度的参数, $P(q_i | C)$ 是项 q_i 在采集 C 中的出现概率(一般是最大似然估计量), 本文使用默认平化值 $\mu = 1000$ 。

2.3 消息来源层面的相关性特征

消息来源层面上的内容匹配证据已被广泛应用于消息来源的筛选和提取任务中。本文提出的特征集可捕获并查询消息来源的不同子部分(如“标题”“标题+正文”以及“评论”)之间的匹配证据, 并应用不同的搜索和查询扩展模型, 使用下文的特征为消息来源相关性建模。

(1) BM25 消息来源: 标题

通过串联所有标题和消息来源的博客文章得到查询和文档之间的相关性估计。消息来源中的标题序列是博客讨论主题的重要表示方式。

(2) BM25 消息来源: 标题+正文

BM25 和狄氏算法通过串联所有标题和消息来源的正文内容以得到查询和文档之间的相关性估计^[5]。此方法将消息来源全文表现出来, 考虑到每篇博客文章的所有内容。

(3) BM25 消息来源: 评论

BM25 和狄氏算法通过串联所有与任一博客文章相关的评论以得到查询和文档之间的相关性估计。

(4) 扩展消息来源: 标题+正文

应用相关性反馈(pseudo relevance feedback, PRF)和查询扩展(query expansion, QE)是一种高效的方法, 一般情况下, 基于经典概率模型的信息检索和语言建模方法是关联的, 但经典概率模型的主要障碍是需要估计反馈模型, 即求相关类的特征词概率。而相关性模型(RM1 PRF)^[13]只使用查询来估计经典概率模型中的概率, 这解决了同义词和多义词的区别, 可以产生高精度的关联模型。因此本

文选择 RM1 PRF 方法。RM1 模型定义如下:

$$P(w|R) \propto \sum_{d \in C} P(d) P(w|d) \prod_{i=1}^n P(q_i | d) \quad (7)$$

其中, $P(d)$ 为文档的先验知识, $\prod_{i=1}^n P(q_i | d)$ 为文档模型下的查询似然, 通常通过狄氏平滑计算。 $P(w|d)$ 表示文档 d 中的词 w 的重要性。 C 表示采集的文档。在相关性模型下具有最高估计概率项被选择为扩展查询, 本文选择前 100 项作为扩展项, 相关性模型由给定的初始查询中的前 100 个文档中估算得出。

(5) 扩展消息来源: 评论

其相当于扩展消息来源——标题+正文, 但消息来源代表所有评论的串联。

2.4 内消息来源的相关性特征

这些特征并非聚集博客文章的所有内容, 而是通过分析查询和博客文章之间的相关性推断一个博客的相关性。

(1) 查询和博客文章间最大化 BM25 的相关性分数, 由相关度最高的博客文章决定。

(2) 查询和博客文章之间的平均相关性 BM25 分数, 该特征并未代表相关度最高文章的消息来源, 而是选择计算所有博客中的平均估计相关值。以这种方式可以估计消息来源的反复兴趣是否与查询主题重叠。

(3) 查询和博客文章之间的相关性 BM25 得分的方差。为得分趋势建模, 以分辨相关消息来源与非相关消息来源。

(4) 基于索引消息来源的所有博客文章, 对该索引运行查询, 并对属于某一特定消息来源的博客文章进行计数。因此消息来源中的每个检索博客文章为一个查询和消息来源间的相关性进行投票。

以上 4 个特征集有助于捕获消息来源相关性的不同方面。第 1 个和第 4 个特征关注消息来源中最突出的博客文章, 而第 2 个和第 3 个特征则分析相关性分数的整体分布, 以决定该博客在查询主题下是否具有反复兴趣。

3 提出的方法

3.1 提出的社会时态特征

消息来源中的评论流量是衡量博客文章影响度的一个重要指标, 通常流行和权威的博客更能吸引群众的关注, 也收到了更多的用户评论。因此评论成为证据的一个很有意义的来源^[14,15]。另外, 时间也是一个重要因素。例如,



如果相关的博客文章非常过时,那么很可能对用户无用。因此,时态信息在消息来源中很重要,而且其有助于理解不同时间的博客主题。

因此总结如下社会和时态特征集为:博客文章的平均评论数目;消息来源中 BM25 相关性分数,这有助于确定查询主题是否被消息来源的最新内容定址;消息来源中最久文章的 BM25 分数,通过这种方式可以获得查询主题是否在早期被消息来源定址;最新的相关博客文章和最旧的相关博客文章之间的时间范围。本文从所有消息来源中索引所有博客文章,对此索引运行查询,从每个消息来源中提取最新和最旧的博客文章,然后计算这些成对文章的日期(按天计算)。通常一个时间范围狭窄的消息来源只会在一个小的时间窗口内定址查询主题,该消息来源可能比一个时间范围较大的消息来源的相关度更低。

3.2 混合式搜索

本文的信息融合方法是包含参数的方法,信息通过合适加权组合起来,这里的加权值是式(3)中的参数 α_i 和 β_j 。本文选择标准差分进化算法执行证据融合的权重参数优化。此外,还选择了信息检索中常用的局部搜索方法(线性搜索),本文整合了这两种方法的优点,即全局和局部搜索方法,并提出一种混合式方法。

线性搜索从参数空间的一个初始随机点开始,在每个维度中进行搜索,在一个维度中移动每个时间的参数值,同时固定其他维度的参数值。对于每个维度,选择的样本点在轴线中有相同的间距,为了评估每个点的最优性,计算每个样本的适应值,并存储适应性最好的点。通常情况下,LS 沿着高预期的方向。而差分进化^[6]是一个基于群体的搜索方法,根据一个简单的向量交叉和变异计算式,结合现有的候选解,创造出新的候选解。其中心思想是使用差分向量在向量群中产生扰动,解决优化问题。

对于微博的信息融合检索问题,本文将差分进化的全局搜索与线性搜索相结合,以利用这两种方法的优点。全局搜索在搜索空间的不同区域内同时进行探索,而线性搜索则以找到的最佳个体为中心在一个受限的区域内搜索。利用线性搜索为群体中的个体在邻近的区域做导向,对个体进行微调。算法 1 如下所示。

算法 1 混合式算法

- (1) 将群体随机初始化
- (2) repeat
- (3) 对于所有的在查询的博客群体中的个体 x

(4) 取 $x_1, x_2, x_3 \in$ 博客群体,随机得到 $\{x_1, x_2, x_3\}$, 这里的 x 彼此不相同

(5) 使 $R=\{1, \dots, n\}$, 其中, n 是链长;

(6) for $i=1, \dots, n$

(7) r_i 取自均匀分布 $U(0,1)$;

(8) if $(i=R) \vee (r_i < CR)$

(9) $y_i \leftarrow x_{i1} + F(x_{i2} - x_{i3})$;

(10) else

$y_i = x_i$;

(11) $\{y=[y_1, y_2, \dots, y_n]$ 即新产生候选博客个体

(12) if 相关特征查询函数 $f(y) < f(x)$

(13) 用 y 替换个体 x ;

(14) 如果该最佳个体的 $f(x)$ 在三代后没有得到提高,则从群体中选择另一个个体;

(15) 获得一个当前最佳个体 y , 将其作为 LS 算法的初始随机点, 即 LS 算法(算法 2)的原始随机点;

(16) 判断是否达到指定迭代次数或终止条件, 如果是, 则返回满足要求的个体。

算法 2 LS 算法

(1) N =每个维度中样本点的数量, D =维度数量, I =采样间隔;

(2) 选择一个原始随机点;

(3) $\min \leftarrow \max(0, \text{initial_position}[d] - \frac{1}{2})$; (0 是参数下限, d 表示第 d 个维度, $d \leq D$)

(4) $\max \leftarrow \min(1, \text{initial_position}[d] + \frac{1}{2})$; (1 是参数上限)

(5) 增量设置: $\text{increment} \leftarrow \frac{(\max - \min)}{N}$;

(6) 获得最佳原始位置 $\text{best_position}[d] \leftarrow \min$;

// 定义在起点和新计算点之间的一条线。取在式(1)中每个维度里适应值最高的参数值建立该新点。这条线代表高预期方向;

(7) 对于维度中的每个样本点;

(8) 设置步长 $p \leftarrow \min + \text{increment} \times n$;

(9) 获得新位置 $\text{new_position}[d] \leftarrow p$; 如果 $\text{fit}(\text{new_position}[d]) < \text{fit}(\text{best_position}[d])$;

(10) 更新最佳位置 $\text{best_position}[d] \leftarrow \text{new_position}[d]$;

(11) for $d \leftarrow 1:D$ // 对于每个维度;

(12) $\max_dim[d] \leftarrow \max(\text{initial_point}[d], \text{best_position}[d])$;

(13) $\min_dim[d] \leftarrow \min(\text{initial_point}[d], \text{best_position}[d])$;

(14) 增量: $increment[d] \leftarrow \frac{(\max_dim[d]) - \min_dim[d]}{N}$;
for $n \leftarrow 1:N$
for $d \leftarrow 1:D$
(15) 获得新位置 $new_position[d] \leftarrow \min_dim[d] + increment[d] \times n$;
(16) 如果 $fit(new_position) < fit(best_position)$;
(17) 更新最佳位置 $best_position \leftarrow new_position$;
(18) 返回最佳位置 $best_position$ 。

本文仅将线性搜索应用于遗传群体的最佳个体中, 减小了线性搜索所需的额外求值计算量, 利于在进化算法每次迭代末尾执行受限制的最佳个体的搜索。线性搜索的参数 N 被设置为一个很小的值, 实验中 N 取 4。本文设置 DE 算法的群体包含个体 32 个, 迭代次数 50 次。对于 LS 算法, 设置 $N=4$ 意味着每次迭代中需要进行 $32(2 \times 2^4)$ 次适合度值的计算。因此, 线性搜索算法的一次迭代需要的适合度求值次数与进化差分算法的一次迭代求值次数相等, 这有利于提高并行性, 即对群体成员进行求值和 LS 对维度样本求值的并行, 降低了计算成本。另外, 对于每个维度, 使用小的采样间隔, 设其值为 0.5, 在进化算法的每代中的计算量减少了 0.85 倍。最后, 线性搜索的步骤数被设置为 1, 从第一个高预期方向的最终样本点中选择最佳个体。

4 实验评估

为了评估本文模型, 博客筛选提取任务包括在一个给定主题 T 中找到有复发兴趣点的博客。该任务被定义为一个经典的线性搜索问题, 系统必须要检索与一个查询 (BlogS06 数据集) 相关的排名前 100 的博客消息来源。测试平台分别由 45 个和 50 个查询组成。BlogS06 研究集^[17]的具体数据见表 1, 该数据集用于 TREC 2006、TREC 2007 和 TREC 2008 中的博客跟踪。每个 TREC 主题包括 3 个部分 (标题、简介和叙述), 且这些查询对真实用户的网页查询^[3,4]具有很好的代表性。

4.1 适应度函数和评估指标

本文使用平均精度均值 (mean average precision, MAP) 作为线性搜索度量, MAP 是评估给定排序目标有效性的单一度量。对于单一的信息需求, 平均精度是在每次相关目标被检索后, 从排序前 k 位的现有目标中得到的平均精确值, 即:

表 1 BlogS06 集的主要数据

参数	数值
独立博客数目	100、649
RSS	62%
Atom	38%
首次消息来源	2013-12-06
末次消息来源	2014-02-21
提取消息来源数目	753、681
永久链接数目	3、215、171
主页数目	324、880
未压缩总大小/GB	148
消息来源 (未压缩)/GB	38.6
永久链接 (未压缩)/GB	88.8
主页 (未压缩)/GB	2.8

$$MAP = \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} \frac{1}{m_j} \cdot \sum_{k=1}^{m_j} Precision(R_{jk}) \quad (8)$$

其中, 已知查询的相关目标集 $q_i \in Q$, R_{jk} 是检索结果达到排名靠前的目标 o_k 的集合, m_j 是查询 q_j 相关目标的数量。

$$precision(R_{jk}) = \begin{cases} \frac{R_{jk} \text{ 中检索到的相关目标}}{R_{jk}}, & o_k \text{ 相关} \\ 0, & o_k \text{ 不相关} \end{cases} \quad (9)$$

搜索系统的另一个重要评估方面是精度。 $P@10$ 是一种比较流行的精确度量, 表示检索到的相关目标前 10 位的比例。

$$P@10 = \frac{\text{检索到相关目标的前 10 位}}{10} \quad (10)$$

4.2 实验结果讨论

实验运行的平台为小型服务器, 采用四核 AMD 6376 处理器 (64 位)、512 GB 的 DDR3 内存和一个固态 SSD 硬盘。本文将每个查询集 (2007 和 2008) 分为两个同样大小的子集, 采用 10 倍交叉验证, 检查子集的各自表现结果。本文这样做的原因是: 如果不分割查询集, 直接应用 10 倍交叉验证, 最大查询为 5 (测试平台分别由 45 个和 50 个查询组成, 最大查询为 50/10), 不利于统计验证, 会造成 10 倍交叉验证的困难 (查询数少)。而分割成两个子集之后, 查询数大幅增加。很重要的一点是选择怎样的交叉验证, 本文选择 10 倍交叉验证, 因为 10 倍交叉验证是比较公认的误差评估最佳选择。此外, 在信息检索中必须要评估至少 25 个查询, 更多的查询是非常有益的。本文完整的训练和测试配置集合见表 2, 对于每个搜索算法, 训练过程重复 5 次。



表 2 训练和测试配置

查询集	训练次数/次	测试次数/次
2007a	974~995	951~973
2007b	951~973	974~995
2008a	1 076~1 100	1 051~1 075
2008b	1 051~1 075	1 076~1 100

不同方法的比较结果见表 3 和表 4, 评估用的度量是 MAP 和 $P@10$ 。选取的两个基准方法是将消息来源表示为“标题+正文”的串联搜索方式和基于索引消息来源的所有博客文章, 对该索引运行查询的方法, 这两个方法并没有将不同来源的证据结合起来。第一种方法利用 BM25 匹配函数进行检索, 通过将这个模型纳入评估, 可以量化改进的程度; 第二种方法是一个投票模型, 使用在博客发表后的目录中, 每个从消息来源中被检索到的博客文章都视为对查询和消息来源相关性的一张投票。这两种方法没有采用社会时态特征, “线性搜索”“差分进化”“粒子群优化”和提出的方法都采用了社会时态特征, 但搜索的方法不同。前两种方法与其他方法的比较是为了分析社会时态特征的效果, 其他 4 种方法的比较是为了分析搜索方式的影响。

从表 3 可以看出, 所有方法都在基准方法之上。这说明了通过训练组得到的适应值函数的最大值所导出的特征权重在该测试组中表现非常好。

2008a 实验结果见表 4, 其中每栏取 5 次运行的平均

表 3 2007 数据集的测试结果

方法	2007a		2007b	
	MAP	$P@10$	MAP	$P@10$
第 2.3 节(2)所提方法	0.42	0.53	0.46	0.44
第 2.4 节(4)所提方法	0.40	0.51	0.47	0.43
线性搜索	0.52	0.60	0.56	0.52
差分进化	0.53	0.60	0.47	0.51
粒子群优化	0.52	0.58	0.55	0.59
提出的方法	0.59	0.67	0.59	0.69

表 4 2008 数据集的测试结果

方法	2008a		2008b	
	MAP	$P@10$	MAP	$P@10$
第 2.3 节(2)所提方法	0.35	0.35	0.44	0.46
第 2.4 节(4)所提方法	0.32	0.29	0.36	0.38
线性搜索	0.54	0.35	0.55	0.46
差分进化	0.55	0.36	0.56	0.47
粒子群优化	0.52	0.35	0.54	0.46
提出的方法	0.57	0.37	0.58	0.56

值。提出的方法相对于 2.3(2)的 $P@10$ 改进程度非常小, 这可能是因为出现一些“有趣的笑话”, 这些检索在排名最高位置上有较多的垃圾邮件文档(TREC 2008 博客跟踪中包含大约 15%的垃圾邮件文档)。本文不包括垃圾邮件文档的检测。从 2008a 部分中的前 10 排序, 可以观察到投票模型基准方法的表现同样很低, 垃圾邮件的影响比较大。

4.3 各方法的进一步分析

综上所述, 提到的线性搜索、差分进化等方法均在基准方法之上。为了进一步评估, 在数据集 2008b 上的各方法训练集平均精度值与进化点之间的关系如图 1(a)所示。结果显示, 本文提出的混合式方法最优, 差分进化 MAP 位列第二, 线性搜索方法排第三, 其后是粒子群优化。这应该是一个公平的比较, 因为所有的方法都有相同的适应度函数。

从图 1 看, 线性搜索似乎是一个比较好的选择, 在进化过程中, 该算法表现较优。但是线性搜索并没有适当地探索整个搜索空间, 因为其严重依靠原始点。如果最优解的位置在搜索空间的极限处, 那么线性搜索的表现会比全局搜索差得多。图 1(b)给出了计算时间的比较, 从图 1(b)可以发现线性搜索具有较大的劣势, 线性搜索几乎比进化算法的执行时间多了一倍。其他方法用了 0.45 h 完成 50 代, 而线性算法几乎用了 1 h。

尽管本文采用并行代码运行, 但线性搜索有设计上的同步瓶颈。这使得每个进化点的累计时间增长比进化算法快的多, 如图 1(c)所示。本文提出的混合式算法将线性搜索应用于每代中的最佳个体中, 是对有效性和计算量的很好权衡。

本文提出的方法在问题规模变大时是线性的。为评估这一点, 实验使用了在基准中不同数量的查询, 结果如图 2 所示。和预期一样, 在 50 代中计算时间的进化基本上是线性状态。

另外需要说明的是, 许多 TREC 的系统合并来自多个来源的证据(如链接分析和垃圾邮件过滤), 并采用大量的外部数据集(如维基百科)。本文方法无法负担如此大的计算成本, 因为每种类型的特征并不在本文的研究范围之内。

数据集 2008 查询和相关性评价的构建方法可对博客进行非二进制评估, 该评估由人担当, 当一个博客包含足够的主题文章时, 标记为相关消息来源。这种松散的相关性概念可能会导致在主题区域内没有复发兴趣的相关消息来源。而本文方法的一些特征直接估算了查询主题在消息来源中的重复性程度。另外, 本文方法可能并不适用于低相关度消息来源, 但 TREC 系统也没有。

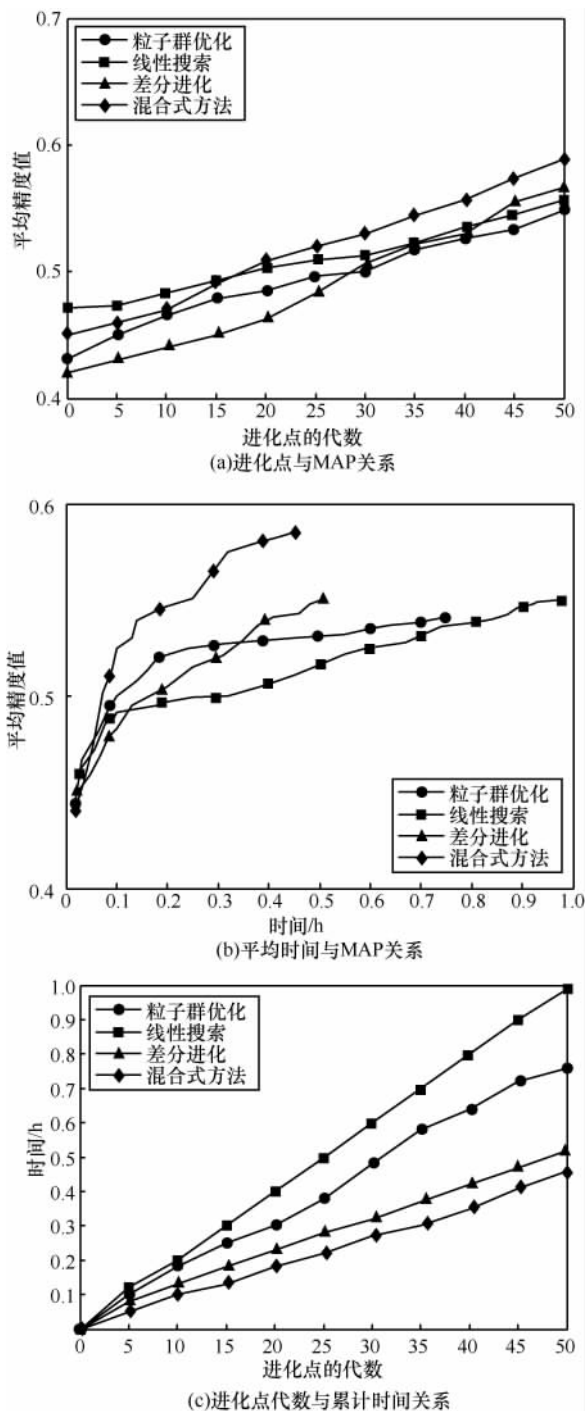


图1 不同方法在数据集 2008b 上的性能比较

5 结束语

本文将博客筛选与提取视为一个信息融合问题。博客圈中大量的消息来源(包括海量的博客文章和评论以及其他估计相关性的信息)将该问题变成一个大数据挑战。本文将不同性质的特征结合以估计消息来源的相关性,使用

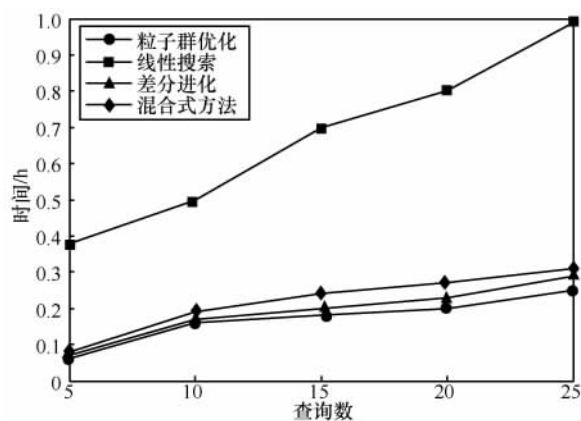


图2 50代的平均执行时间(查询数量为5~25)

群体搜索方法和局部搜索方法结合的混合模式。实验结果表明,提出的博客筛选搜索方法在运行时间和有效性方面获得了满意效果。

未来主要研究特征间的互动和依赖问题,通过引入一个特征选择策略,提取识别度最高的特征。而且 MapReduce、Hadoop 或 Apache Shark 也可以纳入使用。

参考文献:

- [1] RUCCELL M A. 社交网站的数据挖掘与分析[M]. 苏统华, 魏通, 赵逸雪, 等译. 北京: 机械工业出版社, 2015.
RUCCELL M A. Mining the social web[M]. Translated by SU T H, WEI T, ZHAO Y X, et al. Beijing: China Machine Press, 2015.
- [2] 关静怡. 高质量博客检索中核心技术的研究 [D]. 北京: 北京邮电大学, 2011.
GUAN J Y. Research on core technology of high quality blog retrieval [D]. Beijing: Beijing University of Posts and Telecommunications, 2011.
- [3] MACDONALD C, OUNIS I, SOBOROFF I. Overview of the TREC 2007 blog track [C]//16th Text Retrieval Conference, November 6-9, 2007, Gaithersburg, Maryland, USA. New Jersey: IEEE Press, 2007: 1908-1910.
- [4] 翟姗姗, 许鑫, 夏立新. 学术博客中的用户交流与知识传播研究述评[J]. 现代图书情报技术, 2015, 31(Z1): 3-12.
ZHAI S S, XU X, XIA L X. Review of the research on user communication and knowledge dissemination in academic blogs[J]. New Technology of Library and Information Service, 2015, 31(Z1): 3-12.
- [5] ELSAS J L, ARGUELLO J, CALLAN J, et al. Retrieval and feedback models for blog feed search [C]// International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, 2008, Singapore. New York: ACM Press, 2008: 347-354.



- [6] 林旺, 翁彧. 一种面向博客群的主题倾向性分析模型 [J]. 中央民族大学学报(自然科学版), 2014, 23(3): 33-37.
LIN W, WENG Y. A topic opinion analysis model for blogosphere [J]. Journal of Minzu University of China(Natural Science Edition), 2014, 23(3): 33-37.
- [7] 郑美玉. 基于本体的中文博客二级自动分类研究 [J]. 情报科学, 2016, 34(2): 87-90.
ZHENG M Y. Research on two level automatic classification of Chinese blogs based on ontology [J]. Information Science, 2016, 34(2): 87-90.
- [8] 于航. 基于图模型的博客排序系统的研究与实现 [D]. 北京: 北京大学, 2011.
YU H. Research and implementation of blog ranking system based on graph model[D]. Beijing: Beijing University, 2011.
- [9] PARAPAR J, VIDAL M, SANTOS J. Finding the best parameter setting: particle swarm optimization [C]//The 2nd Spanish Conference on Information Retrieval (CERI 2012), June 18-19, 2012, Valencia, Spain. New Jersey: IEEE Press, 2012: 49-60.
- [10] BOLLEGALA D, NOMAN N, IBA H. RankDE: learning a ranking function for information retrieval using differential evolution [C]// Conference on Genetic and Evolutionary Computation, July 12-16, 2011, Dublin, Ireland. New York: ACM Press, 2011: 1771-1778.
- [11] LIN C, LIN C, LIN Z Y, et al. Hybrid pseudo-relevance feedback for microblog retrieval [J]. Journal of Information Science, 2013, 39(6): 773-788.
- [12] 范晨熙, 黄理灿, 李雪利. 基于 Lucene 的 BM25 模型的评分机制的研究[J]. 工业控制计算机, 2013, 26(3): 78-79.
FAN C X, HUANG L C, LI X L. Research on scoring mechanism of BM25 model based on Lucene [J]. Industrial Control Computer, 2013, 26(3): 78-79.
- [13] LAVRENKO V, CROFT W B. Relevance based language models [C]//International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, USA. New York: ACM Press, 2001: 120-127.
- [14] 付仪. 论博客证据[D]. 重庆: 重庆邮电大学, 2013.
FU J. Study on blog evidence [D]. Chongqing: Chongqing University of Posts and Telecommunications, 2013.
- [15] ZHANG S B, ZHANG B, ZHANG Y, et al. A search log sparseness oriented query expansion method [C]// International Conference on Systems and Informatics, November 15-17, 2014, Shanghai, China. New Jersey: IEEE Press, 2014: 1050-1055.
- [16] 许斌, 亓晋, 印溪, 等. 基于多策略离散差分进化的移动互联网个性化服务组合[J]. 电信科学, 2016, 32(2): 1045-1051.
XU B, QI J, YIN X, et al. Personalized service composition based on multi-strategy discrete differential evolution in mobile internet[J]. Telecommunications Science, 2016, 32(2): 1045-1051.
- [17] MACDONALD C, OUNIS I. The TREC blogs06 collection: creating and analysing a blog test collection[EB/OL]. (2016-02-29) [2016-05-27]. https://www.researchgate.net/publication/40704787_The_TREC_Blogs06_Collection_Creating_and_Anal-ysing_a_Blog_Test_Collection.

[作者简介]



张丽娜(1980-),女,浙江安防职业技术学院讲师,主要研究方向为数据挖掘、图形图像、智能算法、云计算。



匡泰(1964-),男,浙江安防职业技术学院信息工程系主任、副教授,主要研究方向为大数据、人工智能。



姜迪清(1965-),男,现就职于浙江安防职业技术学院,主要研究方向为舆情管理、人事管理等。