

# 基于点击日志的搜索引擎用户满意度评价研究

邓晓妹, 武 刚

DENG Xiaomei, WU Gang

北京林业大学 信息学院, 北京 100083

School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China

DENG Xiaomei, WU Gang. Evaluating user satisfaction of search engine using click log. Computer Engineering and Applications, 2015, 51(8):245-249.

**Abstract:** To solve the problem that traditional search engine evaluation method is of large cost and low efficiency, this paper presents a method to evaluate the user satisfaction of search engine based on user click log in this paper. Based on the analysis of search engine user click log, search result ranking, click rate and average browsing time are selected as indicators to estimate user satisfaction. Then it builds search engine evaluation models using linear regression analysis, logistic regression analysis and BP neural network method respectively. Experimental results verify the validity of the models by comparing the performance of different models.

**Key words:** search engine evaluation; user satisfaction; user click log; Back Propagation (BP) neural network

**摘 要:** 针对传统的搜索引擎人工评价方法效率低、成本大的问题, 提出一种利用用户点击日志来评价搜索引擎用户满意度的方法。通过分析搜索引擎的用户点击日志, 选择网页搜索结果排名、网页点击率、网页平均浏览时间作为用户满意度特征, 分别运用多元线性回归分析、多元对数回归分析和BP神经网络方法, 建立了基于用户点击日志的搜索引擎用户满意度评价模型。结合具体的实验数据集, 通过实验对线性回归模型、对数回归模型和BP神经网络模型的结果进行了比较与分析, 验证了模型的有效性。

**关键词:** 搜索引擎评价; 用户满意度; 用户点击日志; 反向传播(BP)神经网络

**文献标志码:** A **中图分类号:** TP183 doi:10.3778/j.issn.1002-8331.1305-0330

## 1 引言

搜索引擎评价问题是搜索引擎研究领域的重要课题之一, 是搜索引擎系统优化、算法改进及日常维护的重要保障<sup>[1]</sup>。目前, 传统的搜索引擎评价主要采用人工评价的方式<sup>[2-3]</sup>, 需要巨大的时间和人力成本。随着搜索引擎数据处理规模的日益庞大, 传统的搜索引擎评价方法已无法满足搜索引擎全面快速的评价要求<sup>[4]</sup>。搜索引擎的用户点击日志所记录的用户行为信息, 隐含了用户对搜索引擎结果的判断, 因此, 利用用户点击日志来评价搜索引擎成为搜索引擎研究领域的一个新方向。

使用用户点击数据进行搜索引擎评价的思路是由Joachims<sup>[5]</sup>首次提出的, 他通过收集用户在搜索引擎中

的查询信息和点击信息, 根据用户对不同网页的点击倾向性对搜索结果进行评价。基于这种思路, 何靖等<sup>[6]</sup>提出一种用户浏览时间模型, 通过用户浏览时间来推测搜索结果网页的相关性程度, 取得了一定的效果, 但是未考虑其他指标的影响。刘建等<sup>[7]</sup>提出一种基于用户行为分析的网页用户兴趣度估计方法, 通过分析用户行为特征与兴趣度之间的关系, 选择兴趣度估计参数, 建立了用户兴趣度估计模型。本文在前人研究的基础上, 提出一种利用用户点击日志来评价搜索引擎的方法, 通过分析搜索引擎的用户点击日志, 从中提取出能够反映搜索引擎用户满意度的特征作为评价指标, 运用数学方法建立搜索引擎用户满意度评价模型。

**基金项目:** 中央高校基本科研业务费专项资金(No.BLYX200928)。

**作者简介:** 邓晓妹(1988—), 女, 硕士研究生, 研究方向为信息管理与信息系统; 武刚(1962—), 男, 通讯作者, 博士, 教授, 研究方向为林业信息化。E-mail: wugang@bjfu.edu.cn

**收稿日期:** 2013-05-24 **修回日期:** 2013-08-07 **文章编号:** 1002-8331(2015)08-0245-05

**CNKI网络优先出版:** 2013-09-17, <http://www.cnki.net/kcms/detail/11.2127.TP.20130917.1056.001.html>

2 搜索引擎用户行为分析

2.1 用户点击日志分析

用户点击日志是记录用户在使用搜索引擎时提交查询词、点击搜索结果网页等用户行为信息的日志,在搜索引擎中有着广泛的应用。搜索引擎用户点击日志记录的信息包括用户提交的查询词、在搜索结果页点击的网页、该网页在搜索结果中的排名、用户点击该网页时的序号、用户ID、点击时间等。其中用户ID是搜索引擎在记录用户行为信息时,为区分不同用户而分配给用户的唯一标识符。搜索引擎用户点击日志的详细格式见表1。

表1 搜索引擎用户点击日志格式

字段	字段含义	字段描述
Query	查询词	用户输入的查询词
URL	URL	用户点击网页的URL
Rank	网页排名	用户点击的网页在查询结果中的位置排名
Order	点击序号	用户对查询结果的第几次点击
UserID	用户ID	搜索引擎分配给用户的唯一标识符
ClickTime	点击时间	用户点击时间

表2是来自于搜索引擎用户点击日志的一个实例,描述了ID为6478452674857780的用户在使用搜索引擎时真实的查询点击过程。用户输入查询词“清华大学”,得到若干条搜索结果。首先,用户在17点33分06秒时点击了排在搜索结果第1位的清华大学首页,然后在17点33分52秒时点击了搜索引擎的第2条结果清华大学图书馆首页,然后在17点34分16秒时重复点击了第1条结果,之后分别点击了搜索引擎的第3、4条结果后,结束了查询过程。

实例中用户2次点击了清华大学首页,而且浏览时间较长,说明相对于其他网页,清华大学首页能更好地满足用户需求,用户对该网页的满意度较高。用户点击日志所记录的用户与搜索引擎交互过程中的行为信息,隐含了用户对搜索引擎结果的判断。因此可以根据用户点击日志设计用户满意度评价指标来评价用户对搜索引擎结果的满意度。

2.2 用户满意度评价指标

用户满意度指用户在搜索引擎中检索查询词时,对搜索引擎结果中的单个网页的满意程度。基于对搜索引擎的用户行为和用户点击日志的分析,本文选择网页搜索结果排名、网页点击率、网页平均浏览时间3个评

价指标来衡量用户满意度。表3列出了搜索引擎用户满意度评价及其描述。

表3 用户满意度评价指标

变量名	变量描述
Rank	网页搜索结果排名
Click	网页点击率
Time	网页平均浏览时间

各评价指标变量的定义及计算方法如下:

(1)网页搜索结果排名Rank:用户检索查询词时,网页在搜索引擎查询结果中的排名。搜索引擎的查询结果是按照网页与查询词之间的相关性排序的,因此可以用网页排名代表网页与查询词之间的相关程度。网页搜索结果排名越小,网页在搜索引擎查询结果中的位置越靠前,网页与查询词之间的相关性越高,用户满意度越高。

(2)网页点击率Click:在单个用户会话过程内,网页被用户点击的次数与用户的总点击次数之比,为该会话内网页的点击率。对所有会话内网页的点击率取平均值,即网页的点击率。用户对网页的点击次数越多,点击率越高,说明用户对网页的满意度越高。

(3)网页平均浏览时间Time:用户点击该网页与点击下一网页的时间差,即用户浏览该页面的时间。浏览时间越长,说明用户对该网页的满意度越高。对所有用户对该网页的浏览时间取平均值,即该网页的平均浏览时间。为保证数据的有效性,去除超时离开的无效记录,需设置浏览时间阈值。研究表明,时间阈值为25.5 min时比较接近真实会话<sup>[8]</sup>,因此,将网页浏览时间的阈值设为25.5 min,即1 530 s。

3 搜索引擎用户满意度评价模型

3.1 数据来源

实验数据来源于国内一家搜索引擎的用户点击日志。根据相关文献统计<sup>[9]</sup>,用户提交的查询词比较集中,少量的高频查询词代表了大多数用户的查询请求,而大量的低频查询词只被少量用户提交。因此本文对查询词进行查询频率统计,从查询次数大于100的查询词中随机抽取若干个查询词,每个查询词对应若干个搜索结果网页,计算每个网页的各个用户满意度特征值,并请标注人员对网页进行满意度评价人工打分。剔除死链网页等不合格网页,最后得到100个有效网页作为实验

表2 用户查询点击过程实例

用户ID	点击序号	查询词	URL	点击时间	网页排名
6478452674857780	1	清华大学	www.tsinghua.edu.cn/	17:33:06	1
6478452674857780	2	清华大学	www.lib.tsinghua.edu.cn/	17:33:52	2
6478452674857780	3	清华大学	www.tsinghua.edu.cn/	17:34:16	1
6478452674857780	4	清华大学	app.edu.qq.com/cgi-bin/CollegeInfo/collegeindex?CollegeId=3	17:35:33	3
6478452674857780	5	清华大学	baike.baidu.com/view/1563.htm	17:35:49	4

样本数据。

数据集中的一个样本记作  $(q, p, Click, Time, Rank, y)$ , 其中  $q$  代表查询词,  $p$  代表网页,  $y$  表示用户对该查询词下的该网页的满意度。  $Click, Time, Rank$  三个指标值根据用户点击日志中的所有用户信息进行统计计算得出, 用户满意度  $y$  由标注人员打分。标注人员根据经验, 对每一个  $(q, p)$ , 给出相应的用户满意度打分  $y$ 。打分采用 0~4 分的评价等级<sup>[10]</sup>, 0 分和 4 分分别表示非常不满意和非常满意。为了尽量减轻打分的主观性, 请两名具有网页用户满意度打分经验的标注人员独立打分, 最终结果根据两人打分共同决定。为了保证人工标注结果的可靠性, 使用公式  $K = \frac{P(A) - P(E)}{1 - P(E)}$  计算 Kappa 统计值来衡量两名标注人员标注的一致性程度<sup>[11]</sup>。其中  $P(A)$  是两名标注人员标注结果一致的比例,  $P(E)$  是标注结果不一致的比例。如果最终得到的  $K > 0.8$ , 认为标注过程非常可靠。两名标注人员的结果并不完全一致, 但比较接近, Kappa 统计值为  $0.86 > 0.8$ , 人工标注结果可靠。

实验数据的部分样本如表 4。

表 4 实验数据样本				
序号	Click	Time	Rank	用户满意度
1	168	1 225	1	4
2	133	1 092	1	4
3	60	695	3	3
4	56	597	2	3
5	11	151	5	2
6	11	179	9	2
7	4	80	8	1
8	2	28	10	1
9	1	10	18	0
10	2	13	18	0

3.2 多元回归模型

假设多元线性回归模型的函数公式为  $y = a_0 + a_1 \times Rank + a_2 \times Click + a_3 \times Time$ 。其中,  $a_0, a_1, a_2, a_3$  为回归系数,  $Rank$  为网页搜索结果排名,  $Click$  为网页点击率,  $Time$  为网页平均浏览时间。利用 MATLAB 求解回归模型, 得到的多元线性回归模型如下:

$$y = -0.523\,8 \times Rank + 0.408\,1 \times Click + 0.323\,5 \times Time + 0.190\,4$$

(1)

将多元线性回归模型加以改进, 建立多元对数回归模型。经 MATLAB 计算, 得到的多元对数回归模型如下:

$$y = -0.179\,9 \times \ln Rank + 0.248\,2 \times \ln Click + 0.282\,5 \times \ln Time + 1.780\,7$$

(2)

为了比较多元线性回归模型和多元对数回归模型性能的优劣, 将 2 个模型的决定系数  $R^2$ ,  $F$  统计量值, 以及与  $F$  值对应的概率值  $p$  等统计量进行比较, 结果如表 5。

表 5 两个模型统计量比较			
模型方法	$R^2$	$F$	$p$
多元线性回归模型	0.718 6	60.441 2	0
多元对数回归模型	0.832 6	117.753 0	0

表 5 中, 多元线性回归模型的  $R^2 = 0.718\,6$  说明用户满意度  $y$  的 71.6% 可由多元线性回归模型确定, 多元对数回归模型的  $R^2 = 0.832\,6$  说明用户满意度  $y$  的 83.26% 可由多元对数回归模型确定, 两个模型的  $F$  值都大于  $F$  检验的临界值,  $p$  都小于置信水平  $\alpha$  (缺省值 0.05), 说明两个模型都具有显著的统计学意义, 因而模型有效。

与多元线性回归模型相比, 多元对数回归模型的决定系数  $R^2$  值和  $F$  统计量值都有大幅度提高, 证明多元对数回归的模型优于多元线性回归模型。这说明用户满意度与各指标之间并不是简单的线性关系, 可能存在复杂的非线性关系。鉴于 BP 神经网络在处理非线性问题方面的优势, 本文把 BP 神经网络的理论应用于搜索引擎评价中, 运用 BP 神经网络建立用户满意度评价模型, 以量化估计用户满意度与各指标之间的非线性变化规律。

3.3 BP 神经网络模型

BP 算法是一种用于前向多层神经网络的反向传播学习算法, 由鲁梅尔哈特 (D.Rumelhart) 等于 1986 年提出<sup>[12]</sup>。采用 BP 算法的神经网络模型一般称为 BP 神经网络, 它由输入层、隐含层、输出层组成。

BP 神经网络的输入神经元为 3 个用户满意度评价指标  $Rank, Click, Time$ , 输出神经元为用户满意度打分  $y$ 。

关于隐含层数的选择, 根据 kolmogorov 定理<sup>[13]</sup>, 仅含一个隐含层的神经网络可以逼近任意连续函数, 而网络精度的提高可以通过增加隐含层神经元的个数来实现。因此采用只含有一个隐含层的三层 BP 神经网络模型。

隐含层节点数的选择参考公式<sup>[14]</sup>:

$$n = \sqrt{n_i + n_o} + a$$

(3)

式中  $n$  为隐含层节点数;  $n_i$  为输入层节点数;  $n_o$  为输出层节点数;  $a$  为常数, 取值在 1~10 之间。

对于本文研究的用户满意度模型, 输入层共设置 3 个神经元, 即  $n_i = 3$ , 输出层为单神经元,  $n_o = 1$ , 经实验选取隐含层神经元个数为 7。

本文采用累计误差的批处理方式对 BP 算法进行训练, 可以有效减小全局误差; 隐含层和输出层的传递函数都选择 S 型 (Sigmoid) 函数。假设  $X$  为输入层输入向量,  $O$  为隐含层输出向量,  $Y$  为输出层输出向量,  $D$  为期望输出向量,  $M$  为训练样本总数,  $V$  为输入层与隐含层之间的权值矩阵,  $W$  为隐含层与输出层之间的权值矩阵,  $H$  为隐含层神经元阈值向量,  $G$  为输出层神经元阈值向量,  $m$  为样本计数器,  $q$  为训练次数计数器,  $\eta$  为



学习率,  $E$  为误差,  $E_{\min}$  为目标精度。BP 神经网络的具体步骤如下:

步骤1 对权重矩阵  $V$ ,  $W$  和阈值矩阵  $H$ ,  $G$  随机赋值, 令  $m=1$ ,  $q=1$ ,  $E=0$ ,  $\eta=0.05$ ,  $E_{\min}=0.05$ 。

步骤2 计算隐含层输出向量  $O$  和输出层输出向量  $Y$  的值:

$$O=f(VX-H)=\frac{1}{1+e^{H-VX}} \quad (4)$$

$$Y=f(WO-G)=\frac{1}{1+e^{G-WO}} \quad (5)$$

步骤3 计算BP神经网络的误差:

$$E=\sqrt{\frac{1}{2}\sum_{m=1}^M(D_m-Y_m)^2} \quad (6)$$

步骤4 若  $m < M$ , 则  $m=m+1$ , 重复步骤2。

步骤5 计算各层误差信号值:

$$\delta^Y=(D-Y)(1-Y)Y \quad (7)$$

$$\delta_j^O=\delta^Y W_j(1-O_j)O_j \quad (8)$$

其中  $j=1, 2, \dots, 7$  为隐含层神经元序号。

步骤6 调整权重矩阵  $W$ ,  $V$  和阈值矩阵  $G$ ,  $H$  的值。

$$W_j^{q+1}=W_j^q+\eta\delta^Y Y_j \quad (9)$$

$$V_{ij}^{q+1}=V_{ij}^q+\eta\delta_j^O X_i \quad (10)$$

$$G_j^{q+1}=G_j^q+\eta\delta^Y \quad (11)$$

$$H_{ij}^{q+1}=H_{ij}^q+\eta\delta_j^O \quad (12)$$

其中  $i=1, 2, 3$  为输入层神经元序号;  $j=1, 2, \dots, 7$  为隐含层神经元序号。

步骤7  $q=q+1$ 。

步骤8 若  $E < E_{\min}$ , 算法结束; 否则  $E=0$ ,  $m=1$ , 重复步骤2。

#### 4 实验结果

由于样本数据量较小, 为了提高模型的有效性和泛化性, 客观地评估模型预测准确率, 采用5折交叉验证的方法进行实验<sup>[15]</sup>。首先采用随机抽样的方式把样本数据集平均分为5个子集, 每个子集中含有20条样本数据。每次以其中的4个子集作为训练样本训练模型, 剩下的1个子集作为测试集, 重复5次, 使得每个子集都得到训练和测试。

本文通过以下2个指标来评判各个模型的评价效果:

(1) 均方误差  $MSE$ 。计算公式为:

$$MSE=\frac{1}{N}\sum_{i=1}^N(y_i-\hat{y}_i)^2 \quad (13)$$

(2) 准确率  $P$ 。计算公式为:

$$P=\frac{1}{N}\text{Count}(\{y_i=\hat{y}_i\}) \quad (14)$$

其中,  $y_i$  为专家标注的用户满意度打分,  $\hat{y}_i$  为模型计算出的用户满意度预测值,  $N$  为测试集大小。

将交叉验证每次实验的测试集的预测结果与人工标注的满意度分值进行比较, 用5次实验结果的均方误差和准确率的平均值来对模型性能进行比较。各个模型预测结果的均方误差和准确率见表6。

表6 模型性能比较

模型	均方误差	准确率
多元线性回归模型	0.58	0.66
多元对数回归模型	0.46	0.72
BP神经网络模型	0.36	0.76

从结果中可以看出, BP神经网络模型的均方误差最小而且准确率最高, 因此整体看来, 在该实验数据集上, BP神经网络模型的性能优于线性回归模型和对数回归模型。

表7列出了各个模型的指标权重的比较结果。结果显示, 用户满意度与三个评价指标之间都具有一定的相关性。其中, 用户满意度与网页搜索结果排名成负相关关系, 网页搜索结果排名越大, 即网页在搜索结果中的位置越靠后, 用户满意度越小; 用户满意度与网页点击率和网页平均浏览时间成正相关关系, 即网页点击率越高, 网页平均浏览时间越大, 用户满意度越大。此外, 各个模型权重最大的指标各不相同。对于多元线性回归模型, 网页搜索结果排名的权重最大; 对于多元对数回归模型, 网页平均浏览时间的权重最大; 对于BP神经网络模型, 网页点击率的权重最大。这说明对于不同的模型, 各个指标的影响作用不尽相同。

表7 模型指标权重比较

模型	Rank	Click	Time
多元线性回归模型	-0.523 8	0.408 1	0.323 5
多元对数回归模型	-0.179 9	0.248 2	0.282 5
BP神经网络模型	-3.849 5	5.926 9	4.689 7

为了进一步评估各模型对不同的分值的网页的评价性能, 将各个模型对整个数据集中0~4各个分值的网页的预测评价效果进行比较, 结果如图1。

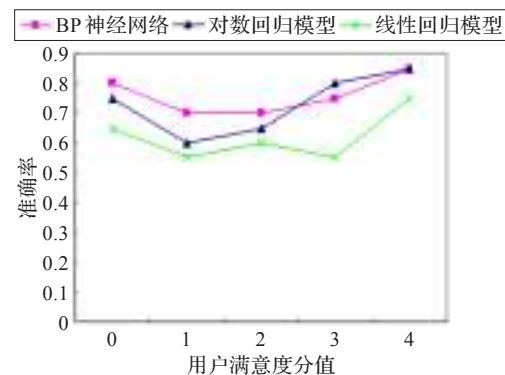


图1 各模型在不同分值下的准确率比较

从图1可以看出, 总体上BP神经网络模型的评价结果的准确度比线性回归模型和对数回归模型都要好。但是, 对数回归模型对用户满意度分值为3的网页

的评价结果好于BP神经网络模型。BP神经网络模型和对数回归模型对用户满意度分值为1和2的网页的评价结果较差,对用户满意度分值为4的网页的评价结果最好。线性回归模型对用户满意度分值为1,2,3的网页的评价结果相差不大,说明线性回归模型无法准确描述用户兴趣度与各指标之间的关系。这说明各指标与用户满意度之间可能存在复杂的非线性关系,因此,相对于多元回归模型,运用BP神经网络处理这种非线性问题能够取得更好的效果。总之,各个模型能较为准确地预测估计用户满意度;同时也说明网页搜索结果排名、网页点击率和网页平均浏览时间等特征能较为明显地反映用户对搜索引擎结果的满意度的差别,可以有效评价搜索引擎结果。

## 5 结束语

本文提出一种利用用户点击日志来评价搜索引擎的方法,通过分析搜索引擎用户点击行为和用户点击日志,选择用户满意度评价指标,分别运用多元线性回归分析、多元对数回归分析和BP神经网络方法,建立了基于用户点击日志的搜索引擎结果评价模型。实验证明,总体上BP神经网络模型的性能优于多元线性回归模型和多元对数回归模型,这说明各指标与用户满意度之间可能存在复杂的非线性关系,因此,相对于多元回归模型,运用BP神经网络处理这种非线性问题能够取得更好的效果。

## 参考文献:

- [1] 岑荣伟.基于用户行为分析的搜索引擎评价研究[D].北京:清华大学,2010.
- [2] 苏君华.搜索引擎评价研究综述[J].情报杂志,2011,30(4):

28-33.

- [3] 魏超,陈飞,许丹青,等.网页质量评价体系的研究[J].中文信息学报,2011,25(5):3-8.
- [4] 刘奕群,岑荣伟,张敏,等.基于用户行为分析的搜索引擎自动性能评价[J].软件学报,2008,19(11):3023-3032.
- [5] Joachims T.Optimizing search engines using click through data[C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2002:133-142.
- [6] 何靖.基于点击日志分析的搜索引擎质量评价方法[D].北京:北京大学,2011.
- [7] 刘建,孙鹏,倪宏.基于神经网络的用户兴趣度估计[J].计算机工程,2011,37(7):187-189.
- [8] Catledge L, Pitkow J.Characterizing browsing strategies in the world wide web[J].Computer Networks and ISDN Systems,1995,27(6):1065-1073.
- [9] Silverstein C, Marais H, Henzinger M, et al.Analysis of a very large web search engine query log[C]//ACM SIGIR Forum,1999,33(1):6-12.
- [10] 费巍.搜索引擎检索功能的性能评价研究[D].武汉:武汉大学,2010.
- [11] Carletta J.Assessing agreement on classification tasks:the kappa statistic[J].Computational linguistics,1996,22(2):249-254.
- [12] 胡守仁.神经网络导论[M].长沙:国防科技大学出版社,1993:23-24.
- [13] 肖魏娜,张为群,王玲玲.一种基于BP神经网络的软件需求分析风险评估模型的研究[J].计算机科学,2011,38(4):199-202.
- [14] 宫运启.基于神经网络的机械加工工序能耗预测[J].计算机工程与应用,2012,48(21):235-239.
- [15] 邱龙金,贺昌政.神经网络稳定性的交叉验证模型[J].计算机工程与应用,2010,46(34):43-45.

(上接239页)

- [6] Wei Z, Zhang S, Qi N.Reliable facility systems based on the p-median problem subject to edge failures[J].Information Technology for Manufacturing Systems,2012,3:671-675.
- [7] Wei Z.Reliability analysis and design of facility systems based on edge failures:a review[J].Advanced Materials Research,2013,715:2721-2725.
- [8] 王艳敏.基于可靠性的供应链设施选址问题的优化模型[J].科学技术与工程,2012,12(11):2517-2520.
- [9] 李东,晏湘涛,匡兴华.考虑设施失效的军事物流配送中心选址模型[J].计算机工程与应用,2010,46(11):3-6.
- [10] 李东,晏湘涛,匡兴华.考虑路段失效的军事物流配送中

心选址模型[J].武汉理工大学学报:交通科学与工程版,2011,35(6):1143-1146.

- [11] 沙丹,许建修.基于最小费用/可靠性比值路的物流配送中心选址问题[J].上海师范大学学报:自然科学版,2010,39(1):35-42.
- [12] 姜阳光,庞大钧.基于集合覆盖模型的城市ULS物流节点选址分析[J].物流科技,2009(10):54-55.
- [13] 魏宗田.网络抗毁性和设施系统可靠性研究[D].西安:西北工业大学,2013.
- [14] Daskin M S.What you should know about location modeling[J].Naval Research Logistics,2008,55:283-294.
- [15] Church R L, ReVelle C.The maximal covering location problem[J].Papers Reg Sci Assoc,1974,32:101-118.