查询语义图辅助的信息检索性能预测模型

乔亚男,齐 勇
(西安交通大学电子与信息工程学院,陕西西安 710049)

摘 要: 查询性能预测技术试图在进行费时的实际信息检索之前对特定查询的性能进行预测,以便根据预测结果在不影响查询所代表的信息需求的基础上对查询进行调整,提高最终检索结果的精确度.针对传统查询性能预测模型没有考虑查询词间语义关系的问题,本文提出了一种查询语义图辅助的信息检索性能预测模型,该模型将表征查询词间语义关系的查询语义图引入性能预测的过程中,使查询性能预测模型避免了查询词独立性假设.实验结果表明,经过查询语义图加权的性能预测模型的预测精确度明显高于传统的性能预测模型,预测结果与实际检索结果的相关度最高提升了约46.32%.

关键词: 信息检索;查询性能预测;查询语义图;核心词
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2011) 3A-158-05

Predicting Query Performance Using Semantic Chart

QIAO Ya-nan, QI Yong
(*School of Electronic and Information Engineering, Xi' an Jiaotong University, Xi' an, Shaanxi* 710049, *China*)

Abstract: The technologies of predicting query performance try to predict the performances of queries before actual time-consuming information retrievals, and adjust the original queries based on the predicting results to improve the performances without modifying the information needs. For avoiding the terms-independence assumption in traditional query performance predicting models, this paper proposes a query performance predicting model using semantic charts of queries which describe the semantic relations between query terms. The experimental results show that the precisions of semantic charts weighted query performance predicting models are higher than traditional models significantly, and the relativities between predicting performances and actual performances are improved about 46.32% at most.

Key words: information retrieval; query performance predicting; semantic chart; core term

1 引言

信息检索系统(搜索引擎)是当今互联网上用户数量最多且范围最广泛的应用之一,然而大部分用户没有接受过专业训练,构造的低质量查询难以准确反映信息需求,严重影响了检索结果的精确度.如何在低质量查询环境下以尽量少的计算量取得尽量高的信息检索性能,是信息检索领域的一个热点课题.除了优化信息检索算法本身外[1,2],查询性能预测(Query Performance Predicting)[3~6]这一技术亦可很好的解决该问题.查询性能预测技术试图在实施真正的信息检索之前对检索性能进行预先估计,根据预测结果可对用户提供的原始查询进行加工,如冗长查询削减(Verbose Queries Reduction)或查询扩展(Query Expansion)等,最终构造出符合用户的信息需求且性能尽可能高的查询.传统的性能预测模型对查询中每个查询词分别预测"子性能",再使用某种算法(求和、最大值、中位数等)将预测出的"子性能"直接进行合并,即为该查询性能的总预测.传统模型将查询中的各个查询词视为彼此独立的,但任何一个查询的多个查询词之间必然存在着一定的语义关系和依赖性(Terms Dependence),简单地忽略这些相关性虽然可以带来一些模型设计和运算效率上的便利,但势必会影响性能预测模型的精确度.

本文针对传统查询性能预测模型存在的忽略查询词之间语义关系的问题,提出了一种查询语义图辅助的性能预测模型.该模型借助查询语义图这一概念将查询词之间的关系形式化,有效地避免了传统性能预测模型的查询词独立性假设.实验表明查询语义图辅助的性能预测模型与传统性能预测模型相比,在附加了很小的计算代价的前提下,预测精确度有了显著的提高.

收稿日期:2010-10-27;修回日期:2011-01-28
基金项目:国家自然科学基金重点项目(No.60933003)

2 查询性能预测技术

传统的查询性能预测模型可以分为3类:基于查询词清晰度的预测模型、基于反向文档频率的预测模型和基于文档-查询相似度的预测模型.

查询词清晰度(Clarity Score)的概念最早是由 Cronen-Townsend 提出的[7],He[8]在此基础上提出了简化查询词清晰度(Simplified Clarity Score)这一概念.具体定义为:

$$SCS = \sum_{w \in Q} P_{ml}(w \mid Q) \times \log_2 \frac{P_{ml}(w \mid Q)}{P_C(w)} \tag{1}$$

其中 $P_{ml}(w \mid Q)$ 是查询词 w 在查询中出现的概率,$P_C(w)$ 是查询词 w 在整个文档集中出现的概率.

反向文档频率(Inverse Document Frequency)是信息检索领域一个非常著名的指标,最初用于 Salton 提出的向量空间模型[9]之中,用于衡量一个查询词在文档集中的稀有程度,相当于该查询词概率分布的交叉熵.IDF 的具体定义有很多种,但只有一些细微的区别,本文中采用的定义是:

$$IDF_w = \frac{\log_2(N + 0.5)/N_w}{\log_2(N + 1)} \tag{2}$$

其中 IDF_w 为查询词 w 的反向文档频率,N 为文档集的总文档数,N_w 为包含查询词 w 的文档数. IDF_w 可以对只含有一个查询词的查询进行性能预测,对于多查询词的查询,通常会将各查询词 IDF_w 之和作为性能预测值.

文档-查询相似度(Collection-Query Similarity)是 Zhao 在 2008 年提出的[10],Kumaran 进行了一些细节上的修改[11].该指标衡量查询词与整个文档集的相似程度:

$$SCQ_w = (1 + \ln(\frac{f_w}{N})) \times \ln(1 + \frac{N}{N_w}) \tag{3}$$

其中 f_w 为整个文档集中查询词 w 的个数.与 IDF_w 类似,SCQ_w 也可以对只含有一个查询词的查询进行性能预测,对于多查询词的查询,则将各查询词 SCQ_w 之和作为性能预测值.

以上3类模型均认为查询中各个查询词的某种"特征"越明显,则该查询词描述用户信息需求的能力越强,相应的检索性能也就越高,区别只在于这个"特征"如何体现.这些传统的模型使用不同的指标对单个查询词进行评价,然后将这些评价结果进行组合(一般为求和)就成为对于一个查询的总评价.也就是说,任何一个查询词的性能预测值都与其它查询词无关,这显然是不合理的.一个查询中的多个查询词共同体现了用户的一个完整的信息需求,因此查询词之间必然存在着一定的语义关系.如何方便高效地以合适的粒度描述这种关系,是下一节要解决的核心问题.

3 查询语义图辅助的性能预测模型

3.1 查询语义图

查询是用户信息需求的概括,因此一个多查询词查询(Multiple Terms Query)的各查询词之间必然存在着一定的语义关系.本节以 NIST TREC Topic 522 为例来引入查询语义图的概念.

```
<num> Number:522
<title> How is water supplied to the Mojave Desert region?
<desc> Description:
Identify methods being used to provide water to the Mojave Desert region.
```

上述查询的目的是检索出有关 Mojave 沙漠供水方法的文档.将该查询的 Title 和 Description 部分合在一起,经过以下三种处理:保留词干(如将 looks,looking,looked 均简化为 look)、去除停用词(如 a,an,the,that 等没有实际意义的语助词)、去除叙述词(如 identify,contain,describe 等在 TREC Topics 中常用到但和实际信息需求无关的词),可以得到如下查询:

$$Q_{522} = \{\text{water, supply, Mojave, Desert, region, provide, method, use}\} \tag{4}$$

上述查询中的 8 个查询词显然不是相互独立的. supply 的对象是 water,而 water 是要供给 Mojave 沙漠的.如图 1 所示,将这些查询词根据相互间的关系排列为一个树状图,就是该查询的语义图(Semantic Chart).查询中有直接语义关系的词在语义图中是相邻的,没有语义关系的距离就比较远.在本例中 Mojave 是所有语义关系的根节点.

图 1 这种类型的语义图虽然描绘了查询词之间完整的关系,但如果直接用于辅助查询性能预测,还有一定的缺陷.首先,与直接优化信息检索系统本身性能的方法[12,13]相比,查询性能预测技术试图从快速重整用户查询的角度间接提升信息检索系统的性能,为此,查询性能预测模型的运行速度必须要高于实际的信息检索至少一个数量级,即相对于实际进行信息检索的时间,查询性能预测的时间应当是可以忽略不计的.而构造一个完整的查询语义图需要对查询进行详细地分析,计算代价相对较大.其次,式(4)的示例查询基本没有歧义,语义图的画法相对固定;但对于较为复杂的查询,可能有多种语义图的画法,从中如何选择也是一个消耗计算资源的任务.因此,对于图 1 这种原始查询语义图还需一定的简化处理.

本文采用类似图 2 这种紧凑查询语义图(Compact Semantic Chart)来更直接地描绘查询词间的语义关系.图 1 中的 Mojave 是语义图的根节点,所有查询词间语义关系最终都可以归结为和 Mojave 的语义关系.因此

Mojave 是该查询的核心词(Core Term).如图 2 所示,图 1 中所有复杂的语义关系都简化为非核心词与核心词之间的语义关系.也就是说,只要找到一个查询的核心词,该查询的紧凑查询语义图即可唯一确定,构造一个查询语义图的复杂过程就转化为如何确定该查询的核心词了.

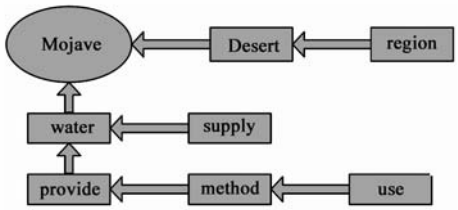

图1 原始查询语义图

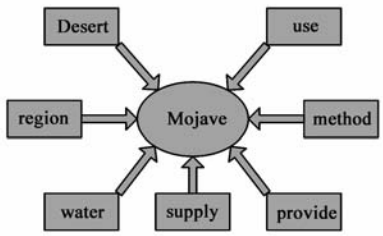

图2 紧凑查询语义图

3.2 查询语义图辅助的性能预测模型

对于查询 $Q=\{w_1,w_2,w_3,\cdots,w_n\}$,设 w_1 为该查询的核心词,使用传统模型对该查询进行性能预测,其一般表达式为:

$$P=\sum_{w_i\in Q}P(w_i) \tag{5}$$

其中 $P(w_i)$为查询词 w_i 的性能预测值,P 为整个查询 Q 的性能预测值.

对于总性能预测值 P 而言,核心词 w_1 的性能预测值 $P(w_1)$无疑是最重要的.但非核心词的性能预测值对整体性能预测的影响则需要进一步讨论.以查询 $Q_1=\{\text{rheum},\text{cold},\text{symptom}\}$为例,核心词为 rheum,cold 和 symptom 同为非核心词,但其性能预测值的重要性却有显著区别.rheum 和 cold 几乎完全是同义词,相似度非常高,rheum 这个清晰度极强的查询词已经能非常准确地反映"感冒"这个概念,因此有很多歧义的查询词 cold 的性能预测值在本查询中其实并不重要.相反地,symptom 和 rheum 的相似度就相对较低,和 cold 相比反映查询 Q_1 信息需求的成分更大,它的性能预测值就相对重要得多.

可见,与核心词相似度高的非核心词,二者之间就有一定的可相互替代性,该非核心词的性能预测值就不太重要,应该加一个较小的权.比如上例 Q_1 中的 cold,从描绘信息需求的角度来说甚至可以说是一个多余的词.与核心词相似度较低的非核心词,二者搭配就意味着更有意义的语义联系,真正体现了信息需求,该非核心词的性能预测值就更重要,应该加一个较大的权.综上所述,查询语义图加权的查询性能预测模型可由下式表示:

$$P=P(w_1)+\sum_{i=2}^{n}C(w_1,w_i)P(w_i) \tag{6}$$

其中 $C(w_1,w_i)$即语义图加权函数,为性能预测值$P(w_i)$的权,由核心词 w_1 和非核心词 w_i 的相似度决定.

可以看出,式(5)和式(6)的主要区别在于对非核心词的处理.查询是用户信息需求的抽象,查询的核心词决定了信息需求的主题,因此在式(6)中核心词性能预测值的处理方式和式(5)是一致的.与此不同的是,查询中的非核心词一般不能独立描绘信息需求(比如查询 Q_1 中的 symptom),必须与核心词搭配,考虑非核心词与核心词之间的语义关系,方能确定非核心词性能预测值的重要性.从这个角度来说,式(6)中的语义图加权函数 $C(w_1,w_i)$的定义也可以是开放的,不必拘泥于核心词和非核心词之间的相似度这一参数,只要能正确地体现核心词与非核心词之间的语义关系对非核心词性能预测值的修正即可.相关扩展研究将在下一步工作中完成.式(6)中的语义图加权函数 $C(w_1,w_i)$在本文中设定为:

$$C(w_1,w_i)=\gamma\times(1-\sin(\frac{\pi}{2}\times\text{sim}(w_1,w_i))) \tag{7}$$

其中 γ 为一个可调节大小的常数,本文实验取 $\gamma=1.45$,$\text{sim}(w_1,w_i)$为核心词 w_1 与非核心词 w_i 的相似度.相似度函数选择基于 Wordnet[14]的 GlossVector[15],该函数将两个词在 Wordnet 中的上下文环境向量(Context Vectors)的夹角余弦作为相似度值.与基于 Wikipedia[16]的相似度函数相比,由于 Wordnet 的体积更小(Wordnet2.1 约 35.4MB,Wikipedia 约 15GB),相似度函数的运算速度也就更快,更适合性能预测模型使用.

另外一个重要问题是查询中的核心词如何确定.核心词识别有很多复杂而精确的方法可以使用,但是过于复杂的方法由于速度缓慢,与查询性能预测的初衷相悖.根据 Bendersky[17]的研究,单纯使用反向文档频率 IDF 来识别核心词(也就是选择 IDF 最大的词)的准确率就可以达到 70%左右,已经基本满足了性能预测的需求.本文扩展了这个思路,在核心词识别上采用自识别的方法,也就是基于 SCS 的模型选用 SCS 最大的查询词作为核心词,基于 IDF 的模型选用 IDF 最大的查询词作为核心词,基于 SCQ 的模型选用 SCQ 最大的查询词作为核心词,这样任何一个模型都不需要额外多余的计算,速度可以得到进一步地提高.

4 实验

4.1 实验设置

为了检验查询性能预测模型的效果,必须在实际的信息检索系统上执行完整的检索,将测试查询的实

际性能与性能预测模型的结果相对照.本文实验选择 Okapi BM25[18]作为实际信息检索系统的检索模型,选择 MAP 即平均精确度(Mean Average Precision)作为实际信息检索的性能评价指标.文档测试集使用 TREC WT10g,包含 1697027 篇文档,总容量约 10.3 GB.TREC topics 451 – 500 对应这个测试集,根据这 50 个 topics 构造了 100 个查询作为测试查询.每个 topic 构造 2 个查询,一个单纯根据 topic 的 title 构造(短查询),另外一个则根据 topic 的 title 和 description 共同构造(长查询).

4.2　实验结果

在实验中使用构造好的 100 个查询通过 BM25 模型对 WT10g 进行完整的信息检索,检索出的文档保留前 1000 位,计算每个查询的平均精确度 MAP.然后再使用 SCS,IDF,SCQ 以及通过语义图加权的 W-SCS,W-IDF 和 W-SCQ 一共 6 个性能预测模型分别对这 100 个查询进行性能预测,实验结果如图 3 所示.

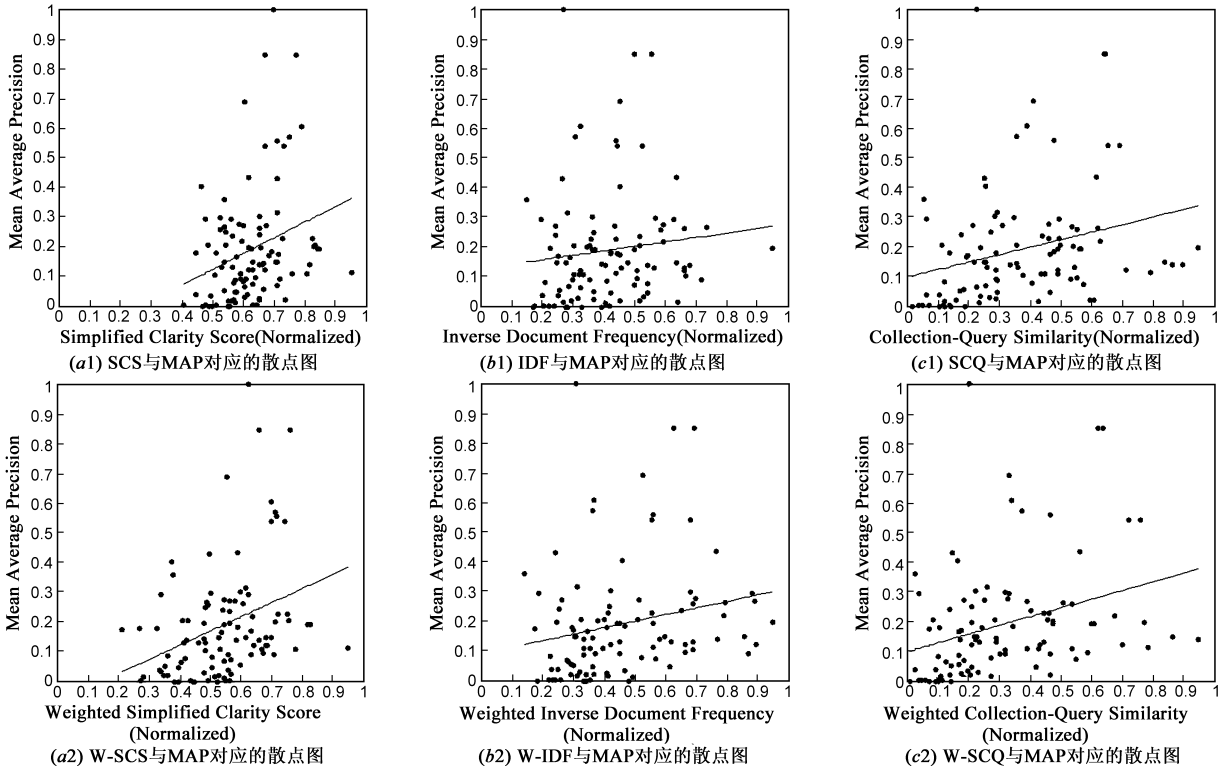

图3　实验结果

图 3 分为 6 个子图,依次为 SCS,W-SCS,IDF,W-IDF,SCQ,W-SCQ 与 MAP 进行比较的散点图,每个性能预测值都经过了归一化(Normalization),以使性能预测值位于 0 与 1 之间.图中的直线为根据 100 个点(对应 100 个查询)经过 1 次曲线拟合得出的直线,该直线和 X 轴的交角越接近 45 度说明性能预测的效果越好.可以看出经过语义图加权的性能预测模型的性能预测效果都优于未经语义图加权的模型.

查询性能预测技术不关心性能预测值本身,只关心性能预测值在不同查询间的排序,因此为了更准确地衡量性能预测模型的效果,还使用了 Spearman's rank test 对实验结果进行检验.表 1 列出了每个模型的性能预测值和实际检索结果的相关度 R 和 P-Value.从表 1 可以看出,每个经过加权的性能预测模型和实际检索结果的相关度都高于未经加权的模型.相关度最高提升了 46.32%,最低也提升了 9.27%.根据 P-Value,所有 6 个模型的性能预测值和实际检索结果的相关性都是统计显著的($P < 0.05$),但经过加权的性能预测模型的 P-Value 明显低于未经加权的模型.

表 1　性能预测值和实际检索性能的 Spearman's rank test 结果

模型	相关度(R)	P-Value
SCS	0.3194	0.0012
W-SCS	0.3925 (+22.89%)	5.3747×10^{-5}
IDF	0.2215	0.0268
W-IDF	0.3241(+46.32%)	0.0010
SCQ	0.3969	4.3350×10^{-5}
W-SCQ	0.4337 (+9.27%)	6.5633×10^{-6}

5　结束语

查询性能预测技术可以根据查询本身的特性对查询的相对性能做出预测,该技术可以有效地规避"差的查询"(bad queries),积极构造"好的查询"(good queries),因此被广泛应用于冗长查询削减和查询扩展等环境中.本文针对传统查询预测技术忽略查询词间语义关系的问题,提出了一种查询语义图辅助的查询性能预

测模型,可以有效地提升传统模型的预测精确度.

本文还存在两个方面的问题.首先是核心词的确定上,本文采用了模型自识别机制,尽管没有额外的计算消耗,核心词的识别精确度也可以接受,但能否在降低些许速度的基础上大幅度提高核心词识别的精确度?其次,本文默认一个查询中只有一个核心词,这个假设是否合理?以上两个问题还需要在下一步工作中进行深入地研究和讨论.

参考文献

[1] 韩立新,陈贵海,谢立.一个面向 Internet 的个性化信息检索系统模型[J].电子学报,2002,30 (2):240 – 244.
Han Li-xin,Chen Gui-hai,Xie Li.A model of personalized information retrieval systems for Internet applications[J]. Acta Electronica Sinica,2002,30(2):240 – 244.(in Chinese)

[2] 韩立新.对搜索引擎中评分方法的研究[J].电子学报,2005,33(11):2094 – 2096.
Han Li-xin.A study on the ranking method of search engines [J]. Acta Electronica Sinica,2005,33(11):2094 – 2096.(in Chinese)

[3] Carmel D,Yom-Tov E,Darlow A,et al.What makes a query difficult? [A]. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C].New York:ACM,2006.390 – 397.

[4] Zhou Y,Croft W B.Ranking robustness:a novel framework to predict query performance[A].Proceedings of the 15th ACM International Conference on Information and Knowledge Management[C].New York:ACM,2006.567 – 574.

[5] Zhou Y,et al.Query performance prediction in web search environments[A].Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C].New York:ACM,2007.543 – 550.

[6] Plachouras V,Cacheda F,et al.University of glasgow at the web Track:dynamic application of hyperlink analysis using the query scope[A].Proceedings of the 12th Text Retrieval Conference (TREC 2003)[C]. Gaithersburg: NIST,2003.248 – 254.

[7] Cronen-Townsend S,Zhou Y,Croft W B.Predicting query performance[A]. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C].New York:ACM,2002.299 – 306.

[8] He B,Ounis I.Inferring query performance using pre-retrieval predictors[A].Proceedings of the 11th Symposium on String Processing and Information Retrieval [C]. Berlin: Springer, 2004.229 – 248.

[9] Salton G,Wong A,Yang CS.A vector space model for information retrieval [J]. Communications of the ACM, 1975, 18 (11):613 – 620.

[10] Zhao Y,Scholer F,Tsegay Y.Effective pre-retrieval query performance prediction using similarity and variability evidence[A].Proceedings of the 30th European Conference on Information Retrieval[C].Berlin:Springer,2008.52 – 64.

[11] Kumaran G,Carvalho V R.Reducing long queries using query quality predictors[A].Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval[C].New York:ACM,2009.564 – 571.

[12] 彭波,李晓明.搜索引擎倒排文件的一种分块组织技术[J].电子学报,2005,33 (2):358 – 362.
Peng Bo,Li Xiao-ming.A blocking inverted file structure for search engine[J].Acta Electronica Sinica,2005,33(2):358 – 362.(in Chinese)

[13] 黄震华,向阳,林琛.有效降低分布式 SKYLINE 查询网络传输代价[J].电子学报,2010,38 (4):848 – 852.
HUANG Zhen-hua,XIANG Yang,LIN Chen.Efficiently reducing transferring cost of skyline queries over distributed networks[J].Acta Electronica Sinica,2010,38 (4):848 – 852.(in Chinese)

[14] WordNet[OL].http://wordnet.princeton.edu/.2010.

[15] Patwardhan S,Pedersen T.Using WordNet-based context vectors to estimate the semantic relatedness of concepts[A].Proceedings of the Workshop on Making Sense of Sense:Bringing Psycholinguistics and Computational Linguistics Together[C]. East Stroudsburg:ACL.2006.1 – 8.

[16] Wikipedia[OL].http://www.wikipedia.org/.2010.

[17] Bendersky M,Croft W B.Discovering key concepts in verbose queries[A]. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C].New York:ACM,2008.491 – 498.

[18] Robertson S E,Walker S,Jones S,et al.Okapi at TREC-3 [A].The Third Text Retrieval Conference (TREC-3)[C]. Gaithersburg:NIST,1995.109 – 126.

作者简介

乔亚男 男,1979 年 2 月出生,陕西西安人.西安交通大学电信学院计算机科学与技术系博士研究生,从事信息检索、机器学习和语义 Web 等方面的研究工作.
E-mail:new_siberia@163.com

齐 勇 男,1957 年 12 月出生,陕西西安人.工学博士,西安交通大学计算机系教授,博士生导师,主要从事操作系统、分布式系统和普适计算等方面的研究.
E-mail:qiy@mail.xjtu.edu.cn