Lucene检索算法的改进

吴云鹏 刘鹏飞 朱旭圻

华南理工大学信息网络工程研究中心

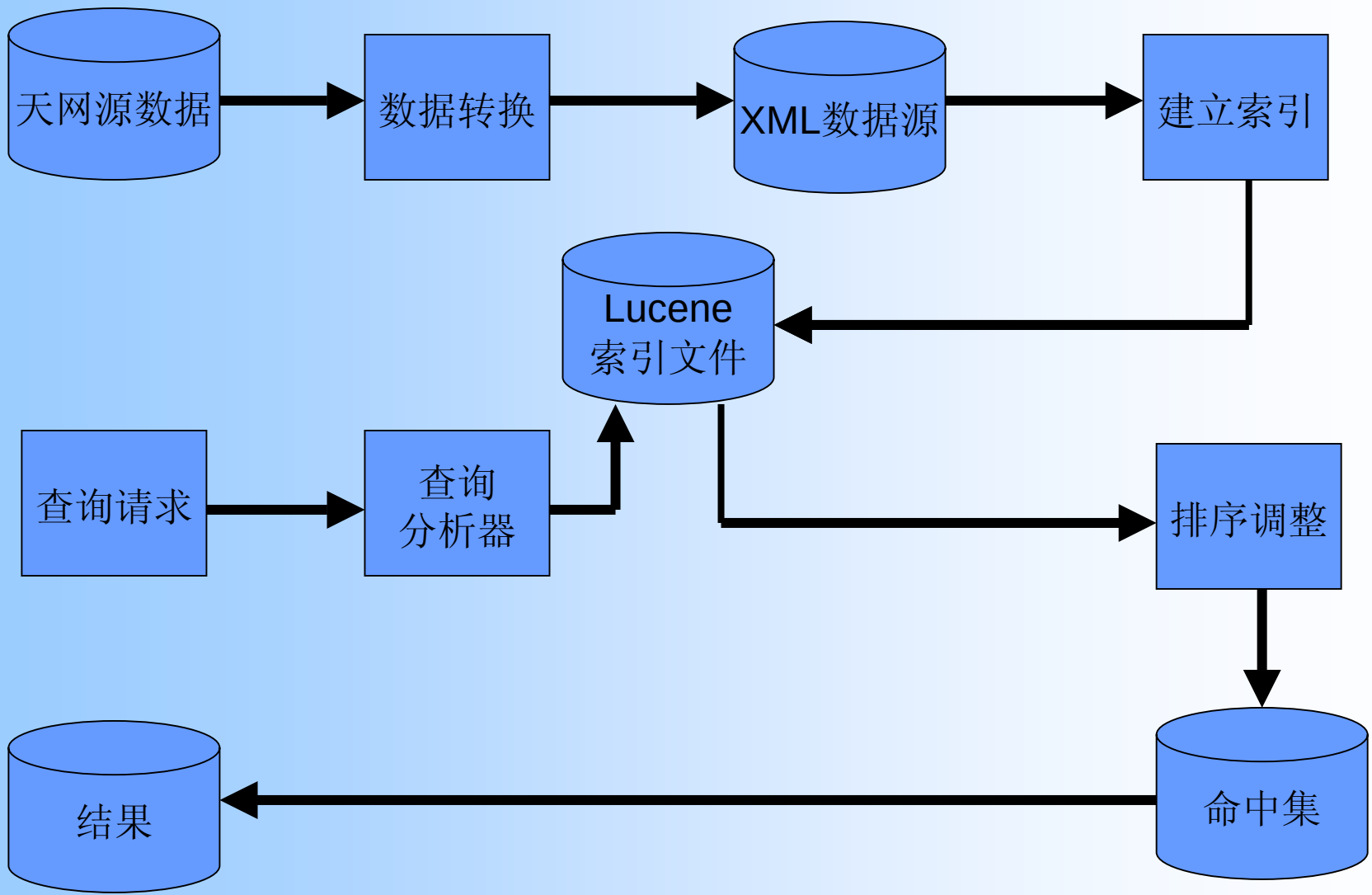
提　　纲

- 检索系统采用的技术
- 系统模型
- 基础排序算法
- 改进的算法
- 存在的问题

检索系统采用的技术

1. 开放源代码的文本检索工具LUCENE
2. 网页分析器，XML转换器
3. 网页权重计算器
4. 二次检索分析器
5. 查询结果缓冲器

系统模型



系统模型的说明

- 数据转换
- XML数据源
- 查询请求

基础排序算法

score_d = sum_t(tf_q * idf_t / norm_q * tf_d * idf_t / norm_d_t)

这里：

- score_d : score for document d
- sum_t : sum for all terms t
- tf_q : the square root of the frequency of t in the query
- tf_d : the square root of the frequency of t in d
- idf_t : log(numDocs/docFreq_t+1) + 1.0
- numDocs : number of documents in index
- docFreq_t : number of documents containing t
- norm_q : sqrt(sum_t((tf_q*idf_t)^2))
- norm_d_t : square root of number of tokens in d in the same field as t

基础排序算法的不足

- 基础排序算法的要点：
  1) 查询词在一个document中位置并不重要
  2) 如果一个document中含有该查询词的次数越多，该得分越高
  3) 一个命中document中，如果除了该查询词之外，其他的词越多，该得分越少

- 不足：
  1) 查询精确度不好
  2) 没有体现网页的重要性

改进的算法

Score_d = k1*OldScore + k2* PrScore + k3* ReScore
+ k4* homePageScore

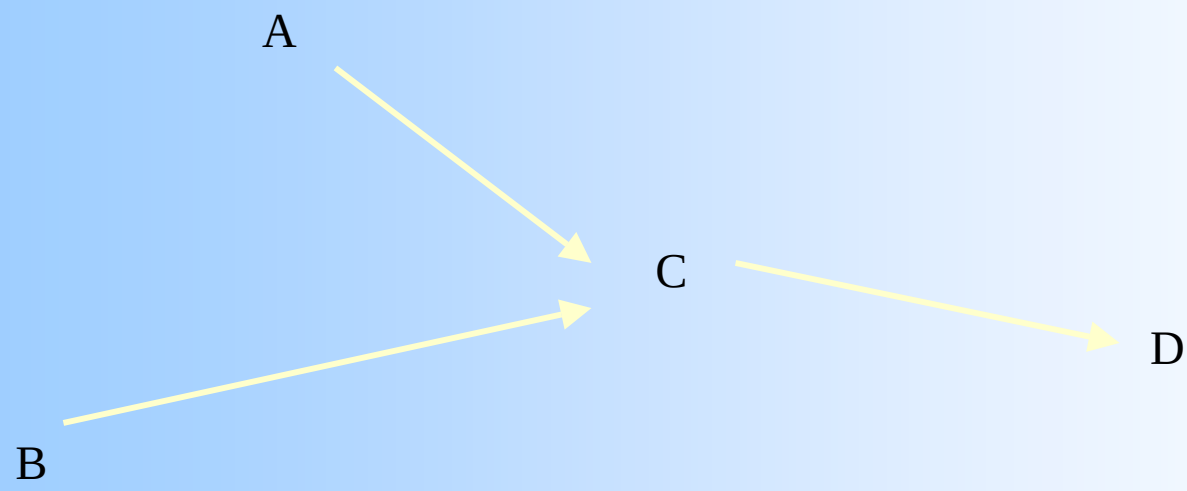
说明：

- Score_d：记录d的得分
- OldScore：由基础排序算法计算出的记录d的得分
- PrScore：记录d的pagerank的得分。
- ReScore：记录d的二次检索的加分。
  ReScore = rescore+(hitNum-1)*increment
- homePageScore：主页的加分
- K1，k2，k3，k4为权重系数

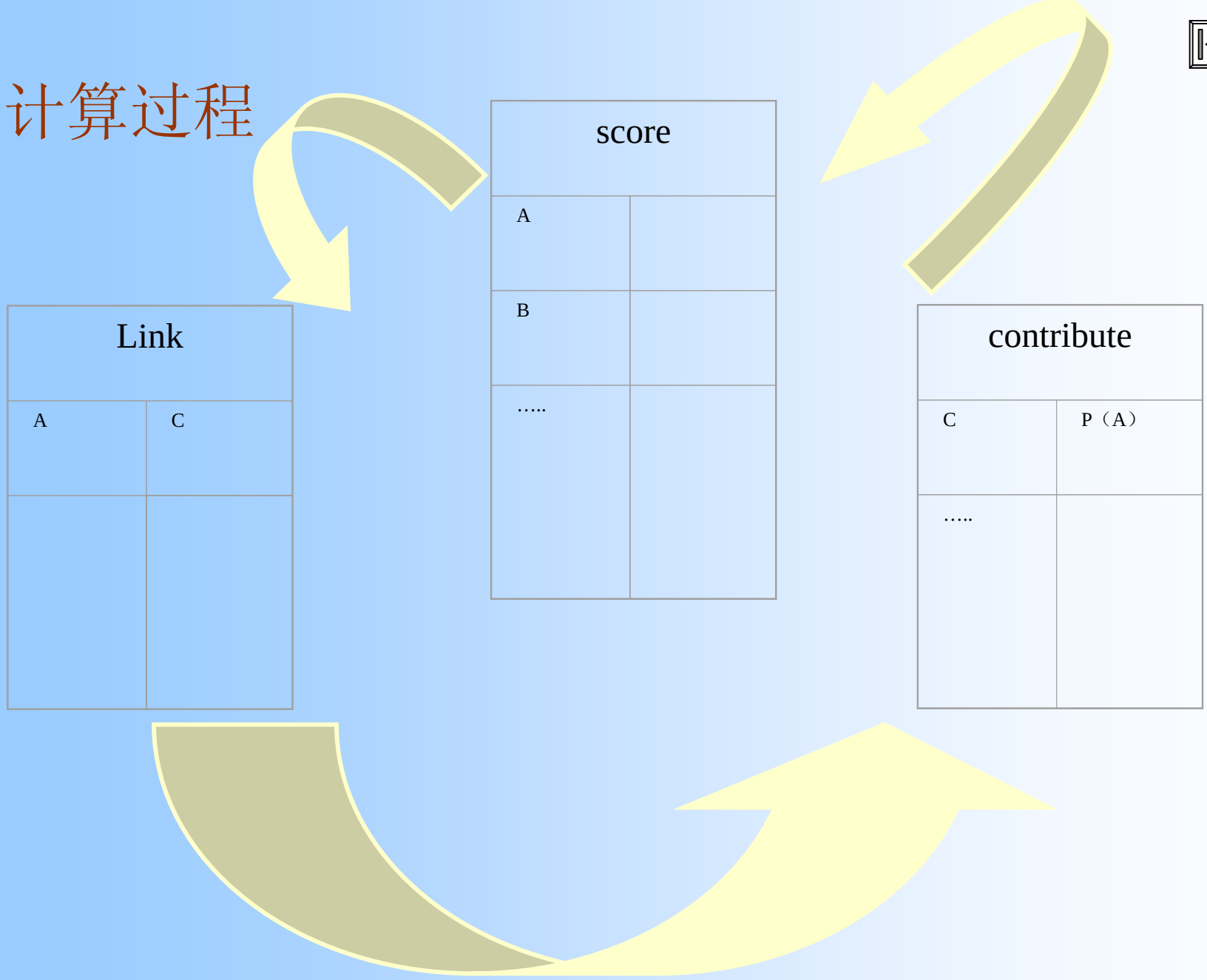
Pagerank的计算

- PR(A) = (1-d) + d(PR(1)/C(1) + ... + PR(n)/C(n))

Pagerank的计算实现



计算过程



实验系统的检索界面


华南木棉教育网全文检索 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
地址(D) http://scutgrid12.scut.edu.cn:8088/search/index.html
转到 链接
天 网 数 据 检 索
每页显示 30 条记录
主题提取查询 主页/指定页面查询（导航查询）
木棉检索 全部 全部
关于主题提取查询的选择：
是否二次检索 ，二次检索命中的基本分为 1.0f 二次检索命中的微调分为 0.1f
（上述选择，第1个选中后，第2,3个才有效）
关于主页/指定页面查询的选择：
是否加入完全匹配，部分匹配 ，完全匹配的加分为 0.5f ，部分匹配的加分为 0.1f
（上述选择，第1个选中后，第2,3个才有效）
两种查询的选择：
是否加入pagerank
主页是否加分 ，主页的加分为 2f 主页的“/”扣分为 1f
title匹配加分 ， title匹配加分为 2f
（上述选择，第2个选中后，第3个才有效）
注意：所有的加分输入框请以f结束输入，否则，可能影响系统稳定。^_^
Copyright (c) 2004 SCUT Network Center All Rights Reserved
华南理工大学 版权所有
Internet
开始
sewm2004
华南木...
Foxmail
15:07

实验结果—导航查询

导航查询：查询词“北京证券 ”

得分调整前

教育网 校内 图片 MP3

搜索：“北京证券” 共有 1093 条结果， 这是第 1 - 30 条。 [搜索用时：2067.0 毫秒]

XML

中安网－北京证券：桦林轮胎
中安网一北京证券：桦林轮胎中安网财经证券北京证券：桦林轮胎 和讯网 1桦林轮胎 桦林轮胎（...发表评论 关闭窗口 打印文章::【相关报道】::北京证券：天山纺织，上海汽车 北京证券：迪马股份，南海发展，一汽四环 北京证券：迪马股份，南京熊猫，南海发展 北京证券...
http://drams.anhuinews.com/ahnews/article/20030317/20030300274976_1.html - 2004-7-17 23:05:45
- 大小：5 k - score:0.99999994 - 网页快照：

北京证券点击网络网址站
北京证券点击网络_网址站_正在进入北京证券.网址：如果浏览器长时间没有响应，请点击这里直接进入中国最好最方便的网站，找网站就到点击网络上网导航...
http://cso.3366.net/www3366net064/www3366net064241/ocp9305.html - 2004-7-17 15:25:15
- 大小：950 byte - score:0.84067273 - 网页快照：

中安网－北京证券：上海石化
中安网一北京证券：上海石化中安网财经证券北京证券：上海石化 和讯网 1上海石化 上海石化（...发表评论 关闭窗口 打印文章::【相关报道】::北京证券：申华控股，蓝星清洗 北京证券：桦林轮胎 北京证券：天山纺织，上海汽车 北京证券...
http://drams.anhuinews.com/ahnews/article/20030318/20030300276710_1.html - 2004-7-17 22:09:14
- 大小：6 k - score:0.8056292 - 网页快照：

中安网北京证券：仕奇实业
中安网北京证券：仕奇实业首页财经新闻证券交易经济信息保险理财彩票园地 所在位置：中安网财经证券北京证券：仕奇实业 和讯网 1仕奇实业 仕奇实业：...发表评论 关闭窗口 打印文章 相关报道：北京证券：长春经开，公用科技 北京证券：长春经开，罗牛山，鑫新股份 北京证券...
http://drams.anhuinews.com/ahnews/article/20021231/20021200207426_1.html - 2004-7-18 4:09:58
- 大小：9 k - score:0.8056292 - 网页快照：

中安网－北京证券：天山纺织，上海汽车
中安网一北京证券：天山纺织，上海汽车中安网财经证券北京证券：天山纺织，上海汽车 和讯网 1天山纺织 ... 发表评论 关闭窗口 打印文章::【相关报道】::北京证券：迪马股份，南海发展，一汽四环 北京证券：迪马股份，南京熊猫，南海发展 北京证券：上海能源，长安汽车，东方热电，光明乳业 北京证券...
http://drams.anhuinews.com/ahnews/article/20030317/20030300274964_1.html - 2004-7-17 21:04:10

得分调整后

教育网 校内 图片 MP3

搜索：“北京证券” 共有 1093 条结果， 这是第 1 - 30 条。 [搜索用时：2314.0 毫秒]

XML

北京证券北京证券北京证券

北京证券function openit another ...js _callpage .htm 北京证券北证首页证券交易理财服务研究成果新闻中心股票资讯分析工具关于北证1北证首页1资讯中心1研发中心1证券交易1服务中心1关于北证document ....我公司各营业网点正在热销泰信先行策略证券...

http://www.bjzq.com.cn/index.asp - 2004-7-18 2:35:03

- 大小：75 k - score:1.0 - 网页快照:

北京证券北京证券北京证券

北京证券function openit another ...js _callpage .htm 北京证券北证首页证券交易理财服务研究成果新闻中心股票资讯分析工具关于北证1北证首页1资讯中心1研发中心1证券交易1服务中心1关于北证document ....我公司各营业网点正在热销泰信先行策略证券...

http://www.bjzq.com.cn/ - 2004-7-17 18:19:15

- 大小：75 k - score:1.0 - 网页快照:

中安网－北京证券：桦林轮胎

中安网—北京证券：桦林轮胎中安网财经证券北京证券：桦林轮胎 和讯网 1桦林轮胎 桦林轮胎（...发表评论 关闭窗口 打印文章::【相关报道】::北京证券：天山纺织，上海汽车 北京证券：迪马股份，南海发展，一汽四环 北京证券：迪马股份，南京熊猫，南海发展 北京证券...

http://drams.anhuinews.com/ahnews/article/20030317/20030300274976_1.html - 2004-7-17 23:05:45

- 大小：5 k - score:0.9922205 - 网页快照:

北京证券北京证券北京证券

北京证券北京证券北证首页证券交易理财服务研究成果新闻中心股票资讯分析工具关于北证1北证首页1资讯中心1研发中心1证券交易1服务中心1关于北证及时报道投资分析行情快讯专题研究行业分析公司研究宏观观点热点透视北证期刊 北京证券a behavior .htc 证券...

http://www.bjzq.com.cn/bstudy/e-zine/index.asp - 2004-7-17 19:40:32

- 大小：25 k - score:0.9876241 - 网页快照:

北京证券北京证券北京证券北京证券

北京证券北京证券北证首页证券交易理财服务研究成果新闻中心股票资讯分析工具关于北证1北证首页1资讯中心1研发中心1证券交易1服务中心1关于北证北京证券个股行情软件下载使用说明风险揭示为方便投资者网上浏...进入Activex

实验结果—主题查询

导航查询：查询词“建筑艺术”

提示：输入更多关键词可以获得更精确的结果 返回首页

新搜索

在结果中搜索 在结果中去除

教育网 校内 图片 MP3

搜索：“建筑艺术” 共有 2758 条结果， 这是第 1 - 30 条。 [搜索用时：27350.0 毫秒]

XML

北京启明星辰信息技术有限公司

北京启明星辰信息技术有限公司function ...plus .gif 中文版ENGLISH 中国建筑艺术概述长城中华民族的纪念碑中国建筑的文化精神及在世界上的地位都城与宫殿 ...充实与总结（明清）坛庙与寺观 儒家思想在宗都性建筑中的映射自然神坛庙与祖先祠祀 佛道寺观陵墓 陵墓艺术的思想内涵 各时代陵墓作品景观楼阁与塔 ...石拱桥 梁桥 木悬臂梁桥 木叠梁拱桥少数民族建筑艺术...

http://home.seechina.com.cn/html/arch2leftdown.html - 2004-7-18 1:50:25

- 大小：16 k - score:1.0 - 网页快照:

[home.seechina.com.cn站内的其它相关信息]

中国传统建筑艺术

中国传统建筑艺术首页文化与艺术：建筑：建筑物中国古塔中国传统建筑艺术介绍中国传统建筑艺术，内容有宫殿、祭祀建筑、佛塔、石窟、园林、王府、民居、桥梁。济南....北京胡同济南.柳埠古迹曲阜.孔庙建筑...

http://www.chinasite.net/%ce%c4%bb%af%d3%eb%d2%d5%ca%f5/%bd%a8%d6%fe/jian-jzhw.htm - 2004-7-17 23:20:11

- 大小：3 k - score:0.9035079 - 网页快照:

建筑
architecture
build
construction
in the construction of

笑话lplen

笑话lplen 情感小筑开心一笑笑话lplen 江苏某日上美术课，内容是建筑艺术。老师提问学生小P 假思索地回答：“胸围，腰围，臀围！”台下一片哗然。 function shutwin window .close return 关闭本窗口...

http://www.reyu.net/myweb2/xh.htm - 2004-7-17 17:13:20

- 大小：1 k - score:0.80812204 - 网页快照:

建筑艺术长廊开平碉楼

建筑艺术长廊开平碉楼 建筑艺术长廊开平碉楼地处珠江三角洲的开平市，素有华侨之乡、建筑之乡和艺术之乡的美誉。20世纪初，为防范劫匪盗贼打家劫舍，...至今保存有1800多座。这些碉楼融入了欧美的建筑艺术风格，形态各异，在建筑...

http://new.cphoto.net/chinese/hij/huzx/02.htm - 2004-7-18 1:05:21

提示：输入更多关键词可以获得更精确的结果 返回首页

新搜索

在结果中搜索 在结果中去除

教育网 校内 图片 MP3

搜索：“建筑艺术” 共有 2773 条结果， 这是第 1 - 30 条。 [搜索用时：384181.0 毫秒]

XML

供求信息招商供求信息斯卡漫淋浴房诚征代理商供求信息...大理石背粘网更多行业展会2004年首届中国国际建筑艺术双年展地点：北京20行业展会第九届中国国际厨房、...上海21行业展会第九届中国国际建筑贸易博览会地点：上海21行业展会第九届中国国际建筑陶瓷、大理石及地点：...10政策法规《建筑材料科技奖管...

http://www.buildnet.cn - 2004-7-17 22:22:35

- 大小：19 k - score:1.0 - 网页快照:

中华读书网读书网讯

中华读书网读书网讯你的位置中华读书网《文化纪念碑的风采??建筑艺术的历史与审美》中国人民大学出版社1999年8月第1版定价：22.00元这是一本专门为非建筑专业的读者鉴赏建筑艺术作品而写的书。作者从东西方建筑艺术产生的独特文化背景出发，帮助读者理解在建筑...

的 of target

http://www.booktide.com/news/20000712/200007120172.html - 2004-7-17 23:30:

- 大小：6 k - score:0.9538894 - 网页快照:

chip .vx vr Math .... 《城镇建设文化参考》是建设部主管、中国建筑文化中心主办的一本传递政策咨询、交流建设经验、... 中国建筑文化中心——系建设部直接管理的全民所有制科研事业单位。其主要职能是：中国建筑文化的研究和传播；陈列国内外建筑发展史，展示古今中外各种风格的建筑...

http://www.chinacon.com.cn/index.htm - 2004-7-17 22:53:37

- 大小：18 k - score:0.87031734 - 网页快照:

[www.chinacon.com.cn站内的其它相关信息]

Untitled Document

Untitled Document 当前位置北京旅游游郊县漫游密云司马台首页漫道雄关人间天险建筑艺术惊险雄秀铁壁铜墙春天之歌服务指南返回简体繁体ENGLISH ...

http://www.beijingwindow.com/n_lvyou/you/jxmy/miyun/smt/index.htm - 2004-7-18 0:07:09

- 大小：4 k - score:0.86599284 - 网页快照:

结论

- Lucene的得分算法，不适合网页搜索
- Pagerank，二次检索，以及主页加分的调整确实优化了查询精确度

存在的问题

- “得分调整算法”需要智能化，自动学习
- 二分词的查询精确度比较低
- IO瓶颈没有解决

谢 谢 大 家!