



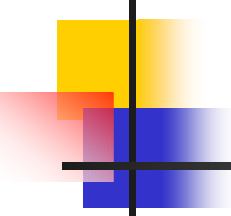
TianWang in Web Track of SEWM2004

何靖 彭波

{hj,pb}@net.pku.edu.cn

北京大学计算机系网络与分布式实验室

2004-11-15



内容提纲

- 问题与分析
- 设计与实现
- 未来的工作

问题与分析

■ TD

- 查找主题相关的关键词资源

名字?

title or anchor text or caption?

■ HP

- 查找指定名字网站的首

■ NP

- 查找指定名

主题相关?

1. 查询词boolean上的ranking结果
2. Rank的因素: similarity, Pr
Pagerank

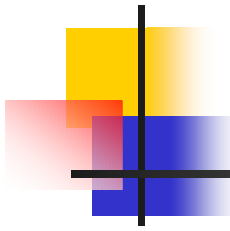
关键词

主页?

什么因素说明一个页面是 home page?

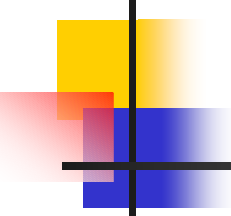
网站结构相关。

什么因素说明一个页面是 entry page?



设计思路

- 主题相关
 - 基本的相关排序方法→baseline
 - Vsm similarity, pagerank, query word proximity,
- 站点入口
 - homepage , entry page
 - url的特征(深度)作为排序一个因素
 - 同站点内相关网页数量作为排序一个因素
 - 综合外部知识(网站目录)
- 名字
 - Phrase/proximity



内容提纲

- 问题与分析
- 设计与实现
- 未来的工作



步骤

- 网站目录数据收集
 - 选取sohu网站目录做为外部知识库,建立网站目录数据库。
- CWT100g数据预处理
 - 链接分析, 提取出anchor text, 建立anchortext数据库
 - 链接分析, 计算pagerank
 - 页面分析和中文分词→包括web page的正文, 对应的anchortxt和网站目录数据。
- 建立倒排索引
 - 带位置数据的全文索引
- 调整排序算法参数, 提交结果
 - 调整排序算法中各个因素的比例参数



TD-RUN1

- 基准的检索系统作为baseline

排序算法使用八级策略

phrase	proxi	title/anchor-text	IR weight/pagerank
--------	-------	-------------------	--------------------

- IR weight

- Tf重要度衰减，按tag加权

- Phrase/proximity

- 简化距离计算，窗口内出现算作临近

- Title/anchor

- Anti-spam (多个title，title过长，anchor错误)

- Pagerank

- 当在title/anchor中出现时，与ir weight线性关系组合(0.6)



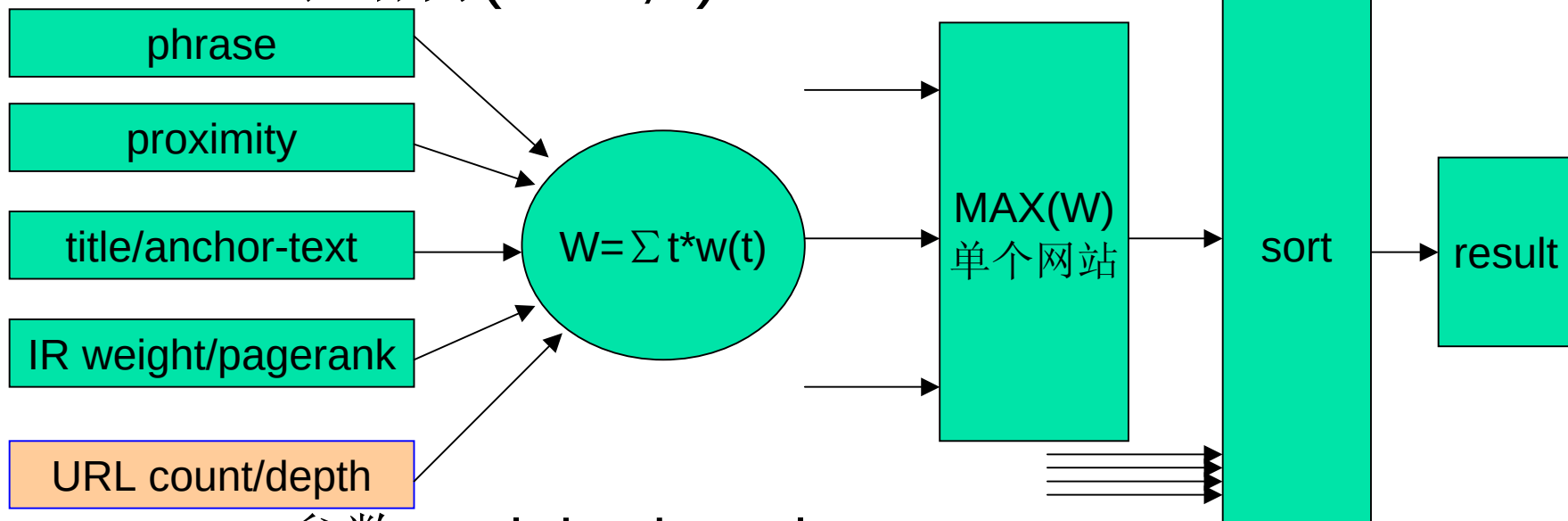
TD-RUN2,3

■ 方法:

- 在**baseline**基础上引入新的排序因素
- 一个站点选择一个关键资源
- 站点相关的网页越多，关键资源的权重越大：
区别不同网站间页面
- 关键资源层深会比相关页面平均层深浅：区
别网站内页面

参数选择

■ 混杂结构(run2,3)

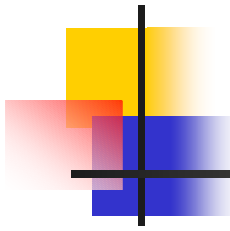


- 参数: url depth, url count
- Run2 (比例小) .vs. run3 (比例大)



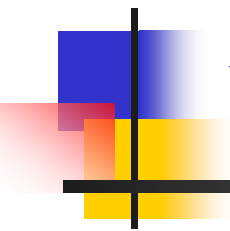
NP-RUN1

- 目标：找出唯一相关页面
 - 步骤：
 - 查找网站目录：HP可能在目录中出现
 - 整个查询串查找：尽量整个串作为短语匹配
 - 查询词分解查找：查询短语过长无法找到
- 以上步骤如果某一步有结果出现，则不进行下一步



未来工作

- 网站内部的局部链接分析（block pagerank/local HITS）
- Homepage/Entrypage从分类角度来做识别



谢谢大家！
