第34卷 第10期 计 算 机 工 程 2008年5月
Vol.34 No.10 Computer Engineering May 2008
·软件技术与数据库· 文章编号：1000—3428(2008)10—0043—03 文献标识码：A 中图分类号：TP301.6

基于短信的移动搜索二次排序算法

张 林[1]，郭 兵[1]，张传武[2]，沈 艳[3]
(1. 四川大学计算机学院，成都 610064；2. 西南民族大学电气信息工程学院，成都 610041；
3. 电子科技大学机械电子工程学院，成都 610054)

摘 要：针对基于短信的移动搜索软件平台中的中间软件模块，提出一种能够适应多种搜索引擎的二次排序算法 ISEH。该算法考虑移动终端屏幕小、存储及处理能力弱等特点，对搜索引擎第1次查询返回的结果集从内部相似度和外部热度进行综合评估，得出最终的排序结果。基于Linux平台的算法仿真实验表明，该算法能克服传统搜索引擎海量信息返回、准确度低等缺陷，将查询满意度因子提升到63.57%，并提高了移动搜索效率。

关键词：移动搜索平台；短信；搜索算法；相似度；Linux操作系统

Mobile Search Secondary Ordering Algorithm Based on Short Message

ZHANG Lin[1], GUO Bing[1], ZHANG Chuan-wu[2], SHEN Yan[3]
(1. School of Computer Science & Engineering, Sichuan University, Chengdu 610064;
2. Colleage of Electric Information, Southwest University for Nationalities, Chengdu 610041;
3. School of Mechatronics Engineering, University of Electronic Science & Technology of China, Chengdu 610054)

【Abstract】 A new secondary ordering algorithm is proposed in the middle software module of mobile search platform based on short message, which can cater for various kinds of search engines. According to the characteristics of mobile terminal such as smaller screen, weaker storing and processing ability, based on the internal similarity and external linkage degree, this algorithm totally evaluates the inquired results from search engine, and acquires the final ordering results. Experimental results based on Linux system show that this algorithm overcomes the defects of vast number of returned information and low accuracy of traditional searching engines, increases the search satisfactory factor to 63.57%, and improves the efficiency of mobile search.

【Key words】 mobile search platform; short message; search algorithm; similarity; Linux

1 概述

移动搜索是指用户通过移动终端(如手机、PDA)进行海量信息发布与获取的技术，通过发送关键字到指定的服务器，服务器自动进行相关存储信息的查询、匹配和排序，及时将查询结果返回移动终端。与基于PC的桌面搜索相比，移动搜索的优点包括：

(1)灵活性。桌面搜索终端设备位置固定，移动不灵活，而移动搜索终端可以随时随地进行搜索查询。

(2)信息传播广。中国的手机用户早已突破3亿，远远多于中国的电脑用户，而且手机用户每年还将大幅增加。

因此，移动搜索业务有着巨大的商业价值，也越来越受到人们的关注。

基于短信的移动搜索作为一种新兴的移动搜索手段，与基于WAP的移动搜索相比，更具竞争优势，主要体现在：带宽占用低，费用低廉，方便快捷，能随时随地为移动终端用户提供查询服务。因此，对基于短信的移动搜索软件平台开展研究更具现实意义。

由于移动终端的特殊性，如无线连接带宽有限、显示屏幕小、处理和存储能力弱，如何快速、准确地从海量信息中获取查询信息以满足移动用户的需要成为移动搜索面临的一个主要难题，因此，本文设计了一种新的基于短信的移动搜索软件平台模型，重点解决移动搜索结果的二次排序问题。

2 软件平台模型

基于短信的移动搜索软件平台结构模型如图1所示。

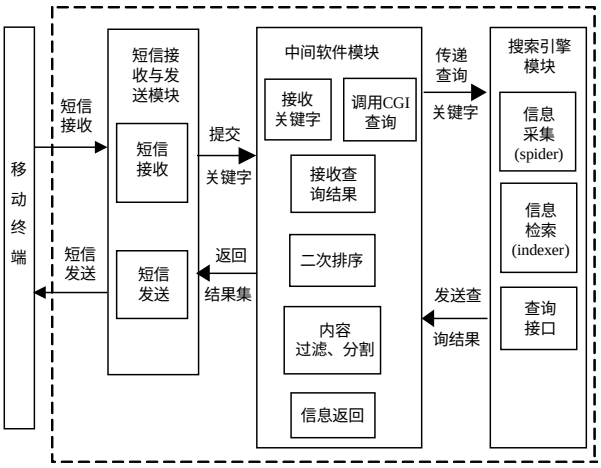

图1 基于短信的移动搜索软件平台

系统采用灵活的3层结构，能适应多种搜索引擎的接口，主要包括以下3个部分：

作者简介：张 林(1981－)，男，硕士研究生，主研方向：嵌入式实时系统；郭 兵，副教授；张传武，教授；沈 艳，副教授
收稿日期：2007-07-08 **E-mail：**christain567@163.com

(1)短信接收与发送模块

利用 GSM/CDMA Modem 设备进行短信的接收与发送。该模块从移动终端发送的短信内容中提取欲查询的关键字，然后提交给中间软件模块。同时接收中间软件模块返回的查询结果集，发送给移动终端。

(2)中间软件模块

该模块对搜索引擎返回的海量信息进行二次排序、分割，使移动终端用户通过编辑短信随时随地、迅速准确地查询信息。通过调用 CGI 接口，该模块将用户提交的查询关键字传递给后台搜索引擎。接收到搜索引擎返回的结果集后，采用一种新的基于内部相似度和外部热度综合的排序算法对结果集进行二次排序，然后将排序值较大的 N 条信息(数量可由用户配置)优先返回给移动终端，避免了查询时海量的信息返回。同时采用短信分页技术，对网页信息进行分页处理。最后将处理完的信息返回给短信接收与发送模块。

(3)搜索引擎模块

该模块主要包括：“ 网络蜘蛛(spider) ”，用于互联网的遍历和网页信息的下载；文档的索引和组织器(indexer)，对网络蜘蛛检索到的网页和相关的描述信息经索引组织后存储在索引库中；查询接口，接收中间软件模块提交的查询关键字，并将结果集按相关度返回给中间软件模块。

基于短信的移动搜索软件平台建立在个性化搜索服务基础之上，采用元搜索引擎[1]进行信息检索，在服务器端组成了一个庞大的数据库服务群提供信息查询服务。但是，由于传统搜索引擎自身存在搜索结果无序性、海量信息内容返回等缺点，同时元搜索引擎的各底层搜索引擎相似度算法各不相同，没有统一的相似度评价标准，因此必须对返回的结果集进行综合相关度排序以统一最终的查询结构。针对上述问题，本文提出了一种基于内部相似度和外部热度分析的ISEH算法，对元搜索引擎返回的结果集进行二次排序，着重提高移动搜索的查准率。

3 面向移动终端的二次排序 ISEH 算法

3.1 基于 Web 文档半结构化特征属性的内部相似度计算

传统向量空间模型将文档和查询式表示成向量形式，从而将信息检索转化为向量空间的向量匹配问题。为便于描述该问题，现给出以下定义，并得出基于 Web 文档半结构化特征属性的内部相似度计算公式。

定义 1 查询向量 指用户提交的查询关键字/词，记为 $\boldsymbol{V}_f$。其中，每一个单字/词代表查询向量的一个分量，记为 V_{fk}，$k \geq 1$。

例如：查询“ 数据挖掘　相似度算法 ”，其中，“ 数据挖掘 ”和“ 相似度 ”为查询向量的 2 个分量。

定义 2 文档 D 一般指文献或文献中的片断，通常指一篇网页中的相关内容集合。

定义 3 索引项 T 指文档中含有能够代表该文档性质的基本语言单位。

定义 4 索引项权重 w_{ik} 表示索引项 T_k 对文档 D_i 的重要程度：

$$w_{ik} = L(i,k) \times G(i)$$

其中，$L(i,k)$ 代表索引项 T_k 在文档 D_i 中的局部权重；$G(i)$ 为索引项 T_k 的全局权重。其计算主要运用 $tf-idf$ 公式。目前有多种 $tf-idf$ 公式，其中一个常用的归一化公式[2]为

$$W_{ik} = tf_{ik} \times \log_n \left(\frac{N}{df_k} + 0.5 \right) \tag{1}$$

其中，tf_{ik} 表示索引项 T_k 在文档 D_i 中出现的次数(即索引项频率)，tf_{ik} 越高，意味着索引项 T_k 对于文档 D_i 越重要；df_k 表示含有索引项 T_k 的文档数量(即索引项的文档频率)，df_k 越高，意味着索引项 T_k 在衡量文档之间相似性方面的作用越低；$N = |D|$，即全部文档的数量，分母为归一化因子。

传统的 $tf-idf$ 公式侧重于从词频上对文档加以区分，并未考虑文档自身的一些内在属性。由于 Web 文档的半结构化特征，文档中包含了丰富的结构体信息，如<title>, <head>, <meta>，而位于这些结构体中的索引项能够在一定程度上反映该文档的主题，因此借助 Web 文档的这种特征，加大出现在这些结构体中的索引项的权值，突出表达该文档的主题。有关定义如表 1 所示。

表 1　半结构化特征属性的索引项权值分配

ID	半结构化特征属性	权值
1	<title>…</title> <head>….</head>	4.0
2	<a herf>…</herf>	3.0
3	<meta>	2.0
4	其他	1.0

定义 5 R_{ik} 表示 Web 文档中索引项的半结构化特征属性权值。其中，$1 \leq i \leq 4$，当索引项处于第 i 个位置时，取其对应的权值。

结合式(1)与定义 5，得出基于 Web 文档的半结构化特征属性的索引项权值计算公式：

$$W_{ik} = tf_{ik} \times \log_n \left(\frac{N}{df_k} + 0.5 \right) \times R_{ik} \tag{2}$$

式(2)优先选取出文档中索引项频率较高、且在关键位置出现的索引项，代表该文档的主题，这有利于查询时匹配相似度，提高查询的准确性。

定义 6 相似度 衡量一篇文档向量与用户查询向量的相近程度，即判断某篇文档是否是用户所需要的，通常用 2 个向量的夹角余弦或 Jaccard 相似度函数计算。本文采用传统空间向量模型(VSM)中的余弦距离公式计算查询向量与文档索引项的相似度，记为 R：

$$R = d(\boldsymbol{V}_f, V_i) = \cos\theta = \frac{\sum_{k=1}^{n} W_{fk} \times W_{ik}}{\sqrt{\left(\sum_{k=1}^{n} W_{fk}^2 \right) \left(\sum_{k=1}^{n} W_{ik}^2 \right)}} \tag{3}$$

其中，$\boldsymbol{V}_f$ 表示查询向量；V_i 表示文档 i 的索引项；W_{ik} 代表 V_i 中索引项的权值；W_{fk} 表示查询向量各个分量的权值。由布尔模型可知，当 $\boldsymbol{V}_f$ 的分量出现在该文档时，W_{fk} 为 1，否则为 0。

3.2 基于网页外部热度计算

定义 7 热度 衡量网页在其主题领域中的重要程度，记为 H。

通常采用链接分析法评估网页的热度。文献[3-4]指出，如果网页 B 存在一条指向网页 A 的超链，则认为 A 得到了 B 的认可；如果有许多网页指向网页 A，则说明 A 在其领域内比较重要，由此得出计算热度的公式：

$$H = PR(A) = (1-d) + d \times \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)} \right) \quad (4)$$

其中，$PR(A)$表示网页 A 的网页级别；$PR(T_i)$表示页面 T_i 指向网页 A 的网页级别；$C(T_i)$表示页面 A 链出的链接数量；d 表示阻尼系数，其取值区间是(0, 1)。通过简单的迭代可以计算出 $PR(A)$，从而得出网页的热度。

3.3 ISEH 算法实现

目前有关搜索引擎的文献认为，搜索引擎结果(排序值)是“ Page Similarity ”与“ PageRank ”因素综合继承的结果。本文基于内部相似度和外部热度分析提出的 ISEH 算法继承了上述算法思想，不仅基于查询向量从内容上对 Web 文档进行了相似度计算，还考虑了网页自身的结构体信息，对于在文档中不同位置出现的索引项进行了加权强调，突出表达该文档的主题。同时，从外部因素考虑网页的热度，提高网页在该领域的权威度，克服了传统搜索引擎只重视网页的 PageRank 值，而忽略了从内容上结合网页结构体信息进行相似度排序的弊端。

计算出相似度与热度后，确定了针对移动搜索的二次排序值：

$$f(R,H) = K_1 \times R + K_2 \times H \quad (5)$$

其中，R 表示基于查询向量的内部相似度权值；H 表示热度值；K_1, K_2 为 R 和 H 的相关系数，$K_1 + K_2 = 1$，取实验获得的经验值。

4 实验与性能分析

4.1 实验平台的搭建

基于短信的移动搜索软件平台搭建在 Linux 操作系统上，所需硬件包括 1 部移动终端、1 个 Modem 和 1 台联网的 PC。实验平台搭建过程如下：

(1)安装搜索引擎

基于 Linux 操作系统搭建搜索引擎 Dpsearch，运行该搜索引擎，将搜索结果保存到 MYSQL 的 SEARCH 数据库中。

(2)配置 GSM Modem

打开服务通信端口，同时初始化该端口，准备接收/发送数据。

(3)安装中间软件模块

实现短信接收/发送模块与搜索引擎的互连与信息的查询。

4.2 搜索性能分析

评价信息搜索性能的主要指标为查全率和查准率。综合考虑查全率和查准率，可以得到新的评估指标——综合评估率 F(满意度因子)，其计算公式如下：

$$F = \frac{precision \times recall \times 2}{precision + recall} \quad (6)$$

在实验中，将 www.google.cn 和 www.baidu.com 设置成元搜索的底层搜索引擎，从 Internet 上获取有关成都城市公共设施建设和发展的3 541篇文档(其中，2 560篇来自google，1 800 篇来自 baidu，并去掉了相同的文档)，将它们人工分为公交信息、铁路信息、航班信息、旅游景点、名小吃共 5 个类别，经特征词约简后分别得到各类别的文档向量空间维数。实验采用的数据集如下：公交信息，535；铁路信息，652；航班信息，409；旅游景点，912；名小吃，1 033。

对 3 541 篇文档做了 2 次对比实验：

(1)利用传统搜索引擎对数据集进行随机查询，统计查全率与查准率，然后计算出满意度因子 F；

(2)仍然采用第 1 次实验的查询向量，借助元搜索引擎对数据集完成第 1 次查询，然后在返回的结果集上使用 ISEH 算法进行二次排序，统计查全率与查准率，计算出返回结果的满意度因子 F。

20 次随机实验的结果如图 2 所示。

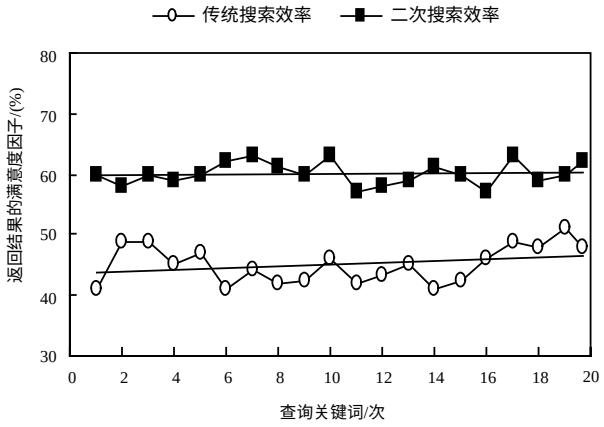

图 2 搜索系统性能对比分析

由图2可见，传统搜索引擎满意度因子为46%，采用ISEH算法进行二次排序后，返回结果的满意度因子提高到63.57%，重点提高了查准率。此外，元搜索引擎 20 次随机查询的平均时间为 0.180 s，在此基础上，平均 0.079 s 就能完成二次排序。因此，ISEH 算法能够适应移动搜索快捷、查准的特点。

5 结束语

随着移动搜索业务的逐步展开，如何利用互联网丰富、全面的信息准确地为移动终端用户服务是现阶段移动搜索有待解决的难题。本文针对基于短信的移动搜索软件平台，提出一种基于内部相似度与外部热度综合排序的 ISEH 算法。用户通过移动终端能迅速、准确地借助互联网进行信息查询，移动搜索效率得到了提高。

参考文献

[1] Howe A E, Dreilinger D. SavvySearch: A MetaSearch Engine That Leams Which Search Engines to Query[J]. AI Magazine, 1997, 18(2): 19-25.

[2] 雷景生, 林东雪, 符浅浅. 基于改进向量空间模型的 Web 信息检索技术研究[J]. 计算机工程, 2005, 31(1): 14-16.

[3] Brin S, Page L. The Anatomy of a Large-scale Hyper-textual Web-search Engine[C]//Proc. of the 7th International World Wide Web Conference. Brisbane, Australia: [s. n.], 1998: 146-164.

[4] Cho Jughoo, Hector G M, Lawrence P. Efficient Crawling Through URL Ordering[C]//Proc. of the 7th International World Wide Web Conference. Brisbane, Australia: [s. n.], 1998: 220-235.