一种中医药行业搜索引擎的推荐词产生方式①

蔡 勇[1,2], 刘美玲[3], 李 玫[4], 胡 豪[2]

[1](北京师范大学珠海分校 软件研究所, 珠海 519087)
[2](澳门大学 中华医药研究院, 澳门 999078)
[3](北京师范大学珠海分校 图书馆, 珠海 519087)
[4](北京师范大学珠海分校 信息技术学院, 珠海 519087)

摘 要: 随着计算机信息技术的发展, 中医药行业大量的文献资料和数据库资源需要共享到 Internet 网上, 以方便专业人士进行查询搜索, 独特的中医药行业搜索引擎就是顺应这个需求而开发的. 文章中笔者结合自己的经验和体会, 提出了一种根据用户输入的查询词产生相关推荐词的方法. 该推荐词产生方法与其它的搜索引擎如谷歌、雅虎、百度不同, 结合了中医药行业搜素引擎与中医药行业中文分词的特点, 应用一种算法来统计推荐词之间的相关性, 用关系数据表方式对推荐词进行专门存储管理. 实践证明此方法能够及时、准确的生成推荐词集, 行业特征明显, 具有一定的创新性和推广价值.

关键词: 中医药行业; 垂直搜索引擎; 推荐词; 中文分词; 网页权重

A Method to Generate Query Recommendations for Search Engine of the Traditional Chinese Medicine Industry

CAI Yong[1,2], LIU Mei-Ling[3], LI Mei[4], HU Hao[2]

[1](Software Research Institute, Beijing Normal University Zhuhai Campus, Zhuhai 519087, China)
[2](Institute of Chinese Medical Sciences, University of Macau, Macao SAR 999078, China)
[3](Library, Beijing Normal University Zhuhai Campus, Zhuhai 519087, China)
[4](Information Technology College, Beijing Normal University Zhuhai Campus, Zhuhai 519087, China)

Abstract: Along with the development of computer information technology, a large number of Traditional Chinese Medicine(TCM) literatures and database resources need to be shared to the Internet for search by professionals and the unique TCM industry search engine is developed to comply with such requirements. A new method for generating related query recommendations according to the user enquiry is proposed according to the features of the TCM industry, based on the author's experiences and understandings. Different from methods used by other search engines, the new method combines the features of the TCM industry search engine and the Chinese word segmentation of the TCM industry and uses algorithms to calculate the correlations between query recommendations and stores and manages such query recommendations in relation data tables. Practices indicate that such method generates accurate and industry-specific query recommendations promptly, and therefore has certain innovation and promotional value.

Key words: Traditional Chinese Medicine industry; vertical search engine; query recommendation; Chinese word segmentation; Web page weight

中医药行业是中华名族的传统行业, 在几千年的发展过程中, 逐渐形成了完整、独特、系统的理论和诊疗方法. 然而, 当前中医药信息资源庞杂而分散, 有限的信息资源不能够实现共享, 大量有价值的临床、科研信息不能科学地集成, 严重影响到中医药的全面、协调、可持续发展. 中医药行业搜索引擎的开发作为一个新课题, 正是顺应中医药信息化事业的潮流, 为中医药行业人士快速、准确搜索到行业信息提

① 基金项目:国家“十一五”重大科技支撑计划(2008BAI64B02)
收稿时间:2012-09-28;收到修改稿时间:2012-11-04

供了方便, 中医药行业搜索引擎是一个行业垂直搜索引擎, 以“专”、“精”、“深”、“准”为主要特征[1].

伴随着互联网的发展, 互联网上面的中医药行业信息数据量剧增, 中医药行业搜索引擎为了更好的服务于搜索用户, 有必要采取辅助手段来提高搜索效率, 目前已经采用的手段包括支持个性化搜索、分类搜索、查询词智能提示、推荐词自动生成等. 其中推荐词自动生成是其中一种非常重要的辅助手段, 即当用户输入某个关键字时, 搜索结果页面能够出现一系列可能与用户输入的关键字相关的关键字(又称“推荐词”), 点击这些相关关键字, 用户可以进行进一步的检索. 如下图 1、图 2 所示:

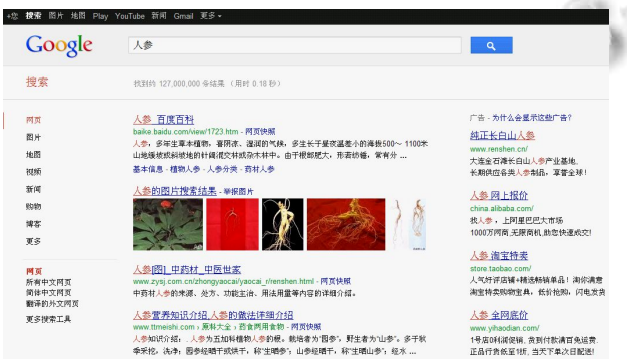

图 1　关键词搜索

图 2　推荐关键词

1　相关研究背景

目前支持这种推荐关键字的搜索引擎系统很多, 如 google.com, baidu.com, sogou.com, bing.com 均提供了智能推荐, 但不同的搜索引擎完成这个功能的方式不同, 通用的搜索引擎由于搜索人群非常广泛, 涉猎的领域非常繁多, 因此在推荐方面大都采用简单高效的推荐计算方式, 常见的方式有:

1.1　关键字语义相似性匹配

根据用户输入的检索关键字, 从系统已经保存的检索关键字库中, 抽取与用户检索关键字语义相似的关键字集合推荐给搜索用户, 如果用户的输入是一个句子, 则需要通过分词分解为多个关键字后再进行相识度匹配[2,3].

此类推荐方式产生的相关关键词基本上包括用户输入的检索关键字内容, 如输入“人参”, 出现“长白山人参”, “人参的功效”等.

1.2　以用户历史查询会话日志为基础形成推荐词

Ricardo Baeza-Yates、Carlos Hurtado 等在 2004 年“Query Recommendation Using Query Logs in Search Engines”一文中提出了一种通过分析用户日志文件来完成推荐词产生的方法[4]. 该方法通过系统日志文件记录任何一个用户输入检索词序列的会话全过程, 通过对会话日志数据采用一些挖掘算法进行分析处理, 当另外一个用户输入检索词后, 后台根据用户检索词, 将系统日志文件分析处理后的结果形成推荐关键词集合呈现出来. 另外 Zhiyong Zhang 和 Olfa Nasraoui 一起在 2006 年 WWW 国际大会上提出了另外一种通过挖掘用户查询日志而产生推荐词的方法[5], 该种方法不仅仅采用了分析日志产生推荐词的方法, 而且还在此基础上如果还没有合适的推荐词集合, 则结合传统的查询词内容相似度方法, 形成推荐词集合呈现出来. 2008 年他们再次在 Applied Soft Computing 杂志中提出了基于用户日志数据挖掘基础上的一种能够动态适应变化趋势的推荐词产生方式[6], 这是一种社会过滤(social filtering)方法, 该种方法能够通过用户自动淘汰过时的推荐词, 产生最新的推荐词, 适用于大型工业级的搜索引擎的推荐系统.

以上通过分析查询会话日志为基础产生的推荐词可能不包括用户输入的关键字内容, 不具有语义相似度匹配的特点, 甚至与用户的关键字完全无关. 利用了数据挖掘技术结合自定义算法, 推荐性相对较强.

1.3　其它的推荐词产生方式

另外还有一些其它的推荐词产生方式, 根据行业的不同而不同, 特别在电子商务网站上面的应用非常广泛, 2006 界世界计算机大会上有专家指出推荐技术将成为电子商务网站(如 Amazon.com, alibaba.com 等)的重要推动技术.

以上一些关键字推荐方式大都能够对搜索用户起到一定的推荐作用, 但应用于中医药行业搜索引擎则存在以下一些需要完善的地方:

① 没有行业特征

由于通用的搜索引擎不针对任何行业, 因此在关键字推荐上面不会考虑任何行业的特征, 也不会考虑搜索用户个人的行业背景, 推荐的关键字大多数是通用的关键字序列.

② 关键字大都与推荐词存在语义相似度匹配的关系

主流的搜索引擎的搜索结果大都提供这种语义相似度匹配推荐, 如输入“人参”, 出现“长白山人参”, “人参的功效”等推荐词. 实际上中医药行业由于很多专业知识是古汉语的原因, 很多的相关关键字之间没有任何语义相似度关系.

③ 推荐词库无法主动生成与即时更新

目前多数的推荐词库的更新速度比较缓慢, 更新过程相对比较被动.

2 中医药行业搜索引擎的主要特征

下面先介绍一下中医药行业搜索引擎的一些主要特点:

① 首先搜索引擎的服务对象大多数为中医药方向的专业人士, 非本行业的网站将不被收录;

② 其次网站与网页的搜索排名计算方法需要采用行业特殊的计算方法, 结合行业内的权威数据与经验算法;

③ 另外搜索引擎除开一个普通静态中文分词库外, 还有一个动态的专属中医药行业分词库. 动态分词库的作用不仅仅用于网页搜索和网页分词, 而且还用于推荐词的产生;

④ 最后中医药行业搜索推荐词的产生, 完全来自于推荐词管理系统, 该系统对推荐词进行动态管理和维护, 专业性非常强.

3 中医药行业推荐词的生成过程

为了生成中医药行业特有的推荐词集合, 我们首先在关系数据库中建立了三张表, 第一张表是专属中医药行业分词主表, 另外一张表为该表的从表. 主表保存所有的专属中医药行业分词, 字段包括: 唯一 ID, 分词, 被查询次数, 审核状态, 创建日期等. 从表保存分词之间的推荐关系, 字段包括: 唯一 ID, 主表分词 ID, 分词 ID, 推荐权重, 更新时间等, 第三张表保存前台用户实际输入的查询词日志.

主表的初始化数据来源于行业特征关键词, 如“阴阳”, “寒热”等, 共约 26,000 多个. 初始化完毕后, 数据可以动态添加、修改、导出, 动态维护该表的途径有两个: 一个是手工新词汇的添加; 另外一个来自于第三张表用户查询词日志记录, 当某个查询词查询次数累计超过一定阀值则被自动添加到该表中, 但最后需要人工审核确认是否是行业特征关键词.

从表的记录源于对网页的分析处理, 首先网页爬虫将爬取回来的网页进行中文分词处理, 分词算法采用逆向最大匹配分词法, 分词库来自于前面提到的两个中文分词库. 接下来将网页文档分词的结果进行处理, 将属于中医药行业分词库的分词序列形成数组;对数组进行迭代程序调用, 通过算法计算同一个文档中主分词与从分词的关系, 累计到从表记录中;对已经在从表中存在关系的分词, 累计其推荐权重值, 对还没有存在关系的分词建立关系, 并初始化推荐权重值.

我们定义计算推荐权重值 $\partial(\lambda,t,p_{(TF)},q_{(TF)})$ 的公式如(1)所示.

$$\partial(\lambda,t,p_{(TF)},q_{(TF)})=\sum_{i=1}^{n}(\lambda*t*(\frac{p_{(TF)}*q_{(TF)}}{p_{(TF)}+q_{(TF)}}))$$
$$p_{(TF)}\geq 1,q_{(TF)}\geq 1,1\leq t>0,\lambda>0 \quad (1)$$

其中 $\partial(\lambda,t,p_{(TF)},q_{(TF)})$ 代表推荐权重值计算函数, 其计算结果是从表中某个分词的最后的推荐权重值, 是动态变化的;

λ 代表某个文档的 PageRank 值, 缺省都为 1, 根据行业权威数据及经验可以进行调整;

其中 t 代表该文档的最后更新时间权重值, 时间越新, 权重越大, 根据表 1 取值.

表 1 相隔年份与 t 值关系表

与当前日期相隔年份	t 值
小于 1 年	1
1~2 年间	0.8
2~3 年间	0.6
2~4 年间	0.4
4~5 年间	0.2
大于 5 年	0.1

其中 $p_{(TF)}$ 代表主分词在该文档的 TF, 而 $q_{(TF)}$ 代表从分词在该文档的 TF, 同一个文档中 $p_{(TF)}$ 和 $q_{(TF)}$ 值都比较大的情况下, 则它们的关联程度越密切.

当从表的推荐权重值计算完成后, 针对用户的查询词而生成推荐词集合就非常简单了, 分为三个步骤:

先将用户输入的查询词进行分词处理, 如果能找出中医药行业专业关键词, 则取出从表中与这些专业关键词相关的推荐词集合, 根据推荐权重值从大到小排序显示出来, 最多 20 个推荐词.

如果分词后无法找到匹配的专业关键词, 则采用参考其它 session 用户输入的关键字序列进行推荐的方

法从用户输入日志表中找到能模糊匹配的关键词, 然后根据时间排序找到相应的 session 序列, 读取 session 序列中下一条查询词, 累计不超过 20 条后进行显示.

如果从日志中仍然无法产生推荐词, 则从分词从表中读取出推荐权重值从大到小排列的 20 个推荐词做为缺省推荐词集合.

4 实验结果

本文的实验以在线中医药行业搜索引擎“博睿搜索”(http://so.brainet.cn/)为环境, 相比 2010 年, 行业特征关键词主表新增了 1811 条, 共 28, 105 条. 以 2011 年的查询日志分析表明, 用户共查询 18, 632, 834 次, 4, 415, 701 个 session, 平均每个用户一次会话(session)提交的查询次数为 4.21 次. 虽然用户查询次数很多, 但总共只查询了 2, 304, 601 个查询词, 其中选中推荐词 420, 440 条, 选中率为 18.24%, 其中每个 session 选中次数为 541, 508, 选中率为 12.26%, 基本达到了预期的推荐效果.

实验也表明, 使用该种推荐词产生方法, 行业特征非常明显, 符合行业专业人士的搜索需求. 以“太阳穴”为搜索词, 各搜索引擎的推荐词结果如下表 2 所示.

表 2 本方法产生的推荐词与其它主流搜索引擎产生的推荐词比较

关键词	“博睿搜索” http://so.brainet.cn	百度 http://www.baidu.com	Yahoo http://search.cn.yahoo.com	搜狗 Sogou http://www.sogou.com	Google http://www.google.com.hk	必应 Bing http://cn.bing.com
太阳穴	太阳人 手太阳 小肠经 足太阳 膀胱经 锁阳 少阳人 阴阳学说 输穴学 经穴学 经穴汇解 铜人腧穴针灸图经	太阳穴疼是怎么回事 太阳穴疼 自体脂肪丰太阳穴 丰太阳穴前后对比图 太阳穴痛是怎么回事 太阳穴胀痛 太阳穴长痘的原因 丰太阳穴 太阳穴长痘痘 太阳穴凹陷	无推荐词	太阳穴疼是怎么回事 太阳穴疼 太阳穴长痘 太阳穴头痛的原因 丰太阳穴 太阳穴凹陷 太阳穴胀痛 女人太阳穴长痣好不好 头痛太阳穴位置胀痛	无推荐词	太阳穴疼是怎么回事 太阳穴胀痛 右边太阳穴疼 自体脂肪丰太阳穴 太阳穴长痘的原因 太阳穴凹陷 左太阳穴疼 丰太阳穴多少钱

5 结论

在实际应用过程中, 我们发现这种推荐词生成方式有一定的优势, 但也存在一些需要改善的地方.

其优势如下:

① 行业特征明显. 本方式由于可以限定特定行业的网站进行爬取, 计算推荐词的关系, 因此可以推荐特定行业的关联推荐词.

② 推荐关键词更加准确. 由于推荐词产生的方法与网页中各个关键词密切相关, 关联词完全来自于因特网本身, 不限于语义相似匹配方式, 网页的内容变化也将导致推荐词的更新, 因此该种方式更能够提供准确的关联关键词.

③ 能产生热门关联关键词. 由于生成的关联词库与网页更新时间直接相关, 因此可以得到行业内检索关键词的最热门的推荐词集合.

需要改进的地方如下:

① 更新计算量大, 需要优化系统性能. 实际应用中, 随着各种网页的不断更新, 计算关联度所使用到的目标分词的密度, 以及各网页页面的权重需要定期进行更新, 相应的, 各检索关键字对应的推荐词的关联度也需要进行定期更新, 而更新需要大量的计算.

② 改进网页抓取内容的过滤能力, 避免重复计算. 实际应用中, 多数的网页包含有广告, 部分网页内容被多处转载, 因此过滤掉广告内容和网页内容去重的工作是提高精准度的重要措施.

6 结语

本推荐词产生方式不仅可以应用于中医药行业搜索引擎, 而且可以应用于其它行业特征明显的搜索引擎. 通过建立行业特征分词库作为推荐基础, 然后利用英特网网页计算关键词之间的推荐关系, 因此推荐

(下转第 202 页)

为了与系统更好的结合，在使用时域差错的模板卷积的技术运动矢量的实验中，采取了系统采集到图片进行恢复. 图 7 是对实际摄像机系统中接收端的某一帧的截取. 由图可得，使用模板卷积方法进行时域的差错掩盖，恢复的图像效果良好.

(a)原图 (b)受损图 (c)还原图(PSNR=28.72)

图 7 系统采集到的图像采用模板卷积法进行时域的差错掩盖

4 结语

通过实验表明，文中提出的基于 RTCP 反馈的自适应分包方法和基于模板卷积的时域差错掩盖计算方法对改善高分辨率图像的无线传输效果具有显著的效果. 采用该技术将实时视频数据传输到接收端，测试得知，在空旷的 50 米内的空间范围内，在图像分辨率为 640×480 时，采集速度稳定在每秒 30 帧，视频较为流畅，丢包现象较少，基本达到了传输要求. 在实际应用中具有很大的商业开发价值.

参考文献

1 Ding XW, Yang ZX, Guo YC. Temporal Error Concealment Technique for MPEG-4 Video.Transactions of Tianjin University, 2006,12(4):291−296.

2 董振亚,张拥军,彭宇行.基于 RTP 的 MPEG-4 视频传输.计算机应用研究,2003,20(7):57−59.

3 徐银辉,周伟,董育宁.一种基于数据隐藏的 H.264 空域差错掩盖的改进方法.广东通信技术,2008,28(3):5−7,11.

4 马鑫,杨小康,宋利.自适应时域差错掩盖方法.中国图象图形学报,2007,12(10):1782−1785.

5 贺贵明,吴元保,蔡朝晖,等.基于内容的视频编码与传输控制技术.武汉:武汉大学出版社,2004.

6 Gonzalez RC, Woods RE.数字图像处理.北京:电子工业出版社,2004.

(上接第 154 页)

词的产生具有一定的客观性，当采样的网页量足够时可以发现一些推荐关键词与检索关键词之间前所未有的关联性，而这些关联性是以前推荐词产生方式无法发现的，因此也具有数据挖掘所追求的特征.

参考文献

1 蔡勇,向婷婷.搜索引擎在中医药行业的实现与应用.辽宁中医药大学学报,2011,1:63−66.

2 Boldi P, Bonchi F, Castillo C. Query suggestions using query-flow graphs. WSCD '09 Proc. of the 2009 workshop on Web Search Click Data:56−63.

3 Abhishek V, Hosanagar K. Keyword generation for search engine advertising using semantic similarity between terms.ICEC '07 Proc. of the ninth international conference on Electronic commerce:89−94.

4 Baeza-Yates R,Hurtado C,Mendoza M.Query recommendation using query logs in search engines.Current Trends in Database Technology.Berlin,Germany:Springer-Verlag, 2004:588−596.

5 Zhang ZY, Nasraoui O. Mining search engine query logs for query recommendation.Proc. of the 15th International Conference on World Wide Web(WWW2006).

6 Zhang ZY, Nasraoui O. Mining search engine query logs for social filtering-based query recommendation. Applied Soft Computing, 2008,8(4):1326−1334.

7 葛玲,蒋宗礼.基于共现词查询的主题爬虫研究.计算机工程,2010,8:286−288.

8 曾广朴,范会联.基于遗传算法的聚焦爬虫搜索策略.计算机工程,2010,11:167−169.

9 金明珠,丁岳伟.基于动态主题库的主题爬虫.计算机应用,2009,12:44−46.

10 张晗,任志国,等.基于主题词关联规则的医学文本数据库数据挖掘的尝试.医学信息学杂志,2008,1:32−35.