对搜索引擎中评分方法的研究

韩立新

(河海大学计算机科学与技术系,江苏南京 210024; 南京大学计算机软件新技术国家重点实验室,江苏南京 210093; 南京大学数学系,江苏南京 210093)

摘 要: 针对搜索引擎评分较为困难的问题,文中提出了一种评分方法. 该方法使用协同过滤技术,在同一兴趣组中各用户所提供的搜索结果集的基础上,采用文中提出的并行关联规则算法对各用户的局部有向图进行预处理,找出兴趣组中各成员都感兴趣的页面. 然后对这些页面的内容和超链接附近出现的文本以及链接结构进行分析. 计算权威页面和引导页面,以找到虽不包括在检索结果中,但相关的页面. 此外,在对所获得的页面进行评价时,除考虑 Web 页自身的链接结构和兴趣组中查询用户对页面的评价,还考虑兴趣组中其它成员对页面的评价和所有成员对页面的使用情况等因素,从而使推荐给用户的页面排序更加合理.

关键词: 信息检索; 搜索引擎; 数据挖掘; 协同过滤

中图分类号: TP391; TP393 文献标识码: A 文章编号: 0372-2112 (2005) 11-2094-03

A Study on the Ranking Method of Search Engines

HAN Li-xin

(*Department of Computer Science & Technology, Hohai University, Nanjing, Jiangsu 210024, China;*
State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210093, China;
Department of Mathematics, Nanjing University, Nanjing, Jiangsu 210093, China)

Abstract: Currently it is difficult for search engine to rank effectively. This paper proposes a ranking method of search engines. The method applies collaborative filtering based on the retrieved results from the users in the same community. A parallel algorithm for mining association rules is described to preprocess all users' local directed graphs to find the commonly interesting pages for the users in the same community. Web pages contents, hyperlink structures and the associated texts are then analyzed. Authority pages and hub pages are recognized to discover the related results not found by the search engines. In addition, the evaluation of the web pages is based on not only the hyperlink structures and the query user's evaluation, but also the evaluation of other users in the same community and the usage of the pages by all users. As a result, the ranking method of the search engine is reasonable and effective.

Key words: information retrieval; search engine; data mining; collaborative filtering

1 引言

目前网络搜索引擎普遍存在对用户的查询请求返回过多的查询结果,而如何从这些查询结果中发现有用的信息是迫切需要解决的问题. 此外,由于检索出过多的相关网页,因此多数 Web 上的用户使用搜索引擎时更关注排序靠前的搜索结果. 这些都可能与评分问题有关.

Cornell 大学的 Jon M. Kleinberg 等人研制了 Clever 系统[1],并在该系统中提出了一组算法对超媒体环境的链接结构进行研究,并从中自动抽取出超媒体某些类型的信息. 纽约州立大学 Binghamton 分校的 M. Cutler 等人研制了 Webor 系统. 该系统主要研究如何将 HTML 文件的结构用于改进检索的性能和效果[2]. 文献[3]讨论了自适应 Web 站点的概念,它通过对访问存取模式的学习来半自动地改进 Web 站点的组织和表示. Letizia[4]是通过记录用户的浏览行为来挖掘相关与该用户兴趣的页面.

我们认为上述系统主要存在以下不足: (1)没有很好地利用协同过滤技术; (2)综合考虑页面间的链接结构和页面自身内容的语义信息各自的优点不够.

为解决上述问题,我们提出了 RM(Ranking Method)方法. 该方法主要是利用协同过滤技术对查询预处理所返回的搜索结果集进行进一步的处理,找出更多和更好的页面,并对其进行更好的排序,以方便用户浏览页面.

2 评分方法 RM

RM 方法的具体步骤如下:

收稿日期: 2003-03-11; 修回日期: 2005-07-19
基金项目: 国家自然科学基金(No. 60073029); 中国博士后科学基金(No. 2005037720)

步骤 **1**　用户所提供搜索结果集的表示. 我们用带标号的有向图来表示用户所提供的搜索结果集, 其中搜索结果集中的页面看作图中的结点, 页面间的链接看作边. 标号表示结点相对应页面的权值.

步骤 **2**　生成兴趣组中各用户都感兴趣的页面. 我们首先将各用户所提供的搜索结果集构成局部有向图. 在此我们提出并行关联规则算法 Findinterestedpage 来对各用户的局部有向图进行预处理, 找出兴趣组中各用户浏览较多的页面. Findinterestedpage 算法的时间复杂度为 $O(\max(m_i * \log_2 m_i + \text{count}_i * r_i + t_i))$.

Findinterestedpage 算法的处理步骤如下:

输入: 各用户的局部有向图

输出: 频繁使用的结点集

{从第 i 个用户的局部有向图中找出使用频率较多的结点组成局部频繁项目集 LL_1;

count= 1;

While LL_{count}不是最大局部频繁项目集

{count= count+ 1;

由发现的所有局部频繁项目集 $LL_{\text{count}-1}$经过修剪后生成局部候选项目集的集合 C_{count};

根据第 i 个用户的局部有向图中的搜索结果, 对 C_{count}中的局部候选项目集进行支持度计数;

生成局部频繁项目集 LL_{count};

}

发送消息告知其它用户自己生成的最大局部频繁项目集;

汇总由其它用户发送回来自己所需的这些最大局部频繁项目集的支持度, 获得全局支持度计数;

生成全局频繁项目集 L_{count};

相互交换各自生成的全局频繁项目集 L_{count}, 并去除重复的全局频繁项目集, 获得频繁使用的结点集;

}

各用户可以各自先生成局部频繁项目集, 直至获得最大局部频繁项目集后进行同步, 这样做避免 Count Distribution[5], Data Distribution[5], FDM[6] 等并行关联规则算法必须在每一趟的末尾进行同步的弊端, 从而减少各处理机相互等待的时间. 此外, 由于每一个全局频繁项目集至少对一个用户是局部频繁的, 因此我们只是对各用户的最大局部频繁项目集进行全局支持度计数, 而不是象 Count Distribution 那样必须对各用户所有的局部候选项目集进行全局支持度计数, 从而减少交换的支持度计数数目, 进而减少网络流量. 再者, Findinterestedpage 与这些算法适用在不同的应用领域中. Findinterestedpage 主要是从 Web 上的各用户的局部有向图中找出频繁使用的结点集, 而这些算法主要是从数据库中挖掘出有用的数据.

步骤 **3**　对频繁使用的结点集中的结点相对应页面进行处理. 我们将处理分为两类: 一是根据页面中出现的链接标记的特点, 从返回结果的页面中获得超链接, 以供下面对可用结点进行扩充以及计算结点的入度、出度时使用; 二是依据页面的语法特点, 获得一些启发式规则来对这些页面中的信息进行抽取, 从中获得页面中的关键词.

步骤 **4**　获得可用结点, 并计算可用结点的语义权重. 通过使用余弦相似性公式, 计算兴趣组中该用户的用户概要 (user profile) 和可用结点相对应页面的关键词集间的相似度, 获得它们的语义权重 w_s, 并且找出相似度大于阈值的页面所对应的结点作为可用结点.

步骤 **5**　计算可用结点的综合评价权重. 通过使用余弦相似性公式, 计算公共用户概要和可用结点相对应页面的关键词集间的相似度, 获得兴趣组中各用户对这些页面的综合评价权重 w_a. 在这里的公共用户概要是按照在兴趣组中的重要性对各用户的用户概要中的关键词进行加权来构造兴趣组中的公共用户概要.

步骤 **6**　利用获得的可用结点相对应页面中的超链接, 对可用结点进行扩充. 其主要思想是采用扩充与可用结点距离较近的相关结点, 以此保证扩充的质量. 经过上述扩充生成兴趣组的全局有向图.

步骤 **7**　计算扩充的可用结点的语义权重. 求出扩充的可用结点相对应页面的关键词集和兴趣组中该用户的用户概要间的相似度, 获得它们的语义权重 w_s.

步骤 **8**　计算扩充的可用结点的综合评价权重. 通过使用余弦相似性公式, 计算公共用户概要和扩充的可用结点相对应页面的关键词集间的相似度, 从而获得兴趣组中各用户对这些页面的综合评价权重 w_a.

步骤 **9**　对于兴趣组中的全局有向图进行分析, 找出好的引导页面和权威页面. 对于权威页面, 我们给这些权威页面增加较多的链接权重 w_l, 权威页面的链接权重 w_l 是该页面相对应结点的入度数进行处理后获得的结果. 而对于引导页面, 我们对这些引导页面降低其链接权重 w_l, 引导页面的链接权重 w_l 是该页面相对应结点的出度数进行处理后获得的结果. 从而能找到一些基于关键字检索的搜索引擎所不能找到的好页面.

步骤 **10**　我们对兴趣组中的全局有向图中结点相对应页面中的超链接附近出现的文本进行分析. 如果该用户的用户概要和公共用户概要中的关键字在链接附近出现, 那么就意味着该链接指向相关内容的可能性高于无关键字的页面. 所以需要增加这些链接所指向页面的语义权重 w_s.

步骤 **11**　计算全局有向图中结点相对应页面的使用情况. 页面的使用情况 u 可以表示成如下形式: $u= c_1 * y + c_2 * t + c_3 * n$. 其中: y 表示用户最后一次浏览该页面的日期, t 表示用户阅读该页面的时间之和, n 表示同一兴趣组中使用该页面的用户数. c_1, c_2, c_3 分别表示 y, t, n 重要性的权值.

步骤 **12**　按照每个页面的链接权重 w_l, 综合评价权重 w_a, 语义权重 w_s, 页面的使用情况 u 之和所构成页面的权重 w 对所有页面进行排序, 以便重新构成全局有向图.

3　与国内外同类工作的进一步比较

在文献[1]中, 由于 HITS 算法只考虑页面间的链接而没有利用任何文本信息来确定链接的重要性, 这导致当对某些比较特殊的话题进行查询时, 由于一些更普遍的话题的链接更多, 因此返回的结果往往是这些更普遍的话题. 而本文提出

的评分方法 RM 不仅充分利用链接结构的优点,找出好的引导页面和权威页面,而且考虑了较好页面自身内容的语义信息,同时对可能包含在链接中的解释信息加以利用,从而可以适用于各种情况.

在文献[2]中,Webor 系统使用的算法基本上还是基于文本内容的检索算法.它使用 HTML 文档的结构和超链接的一些解释信息来改进检索性能,然而对链接结构的利用相当有限.虽然它能在一组 HTML 文档中选择较为合适的页面,但是它只是对这个文档集合中的页面进行评价,找到的也只是这个集合中较好的页面.而 RM 方法既充分利用链接结构的优点,又考虑页面自身内容的语义信息.此外,还使用多种措施来完成对页面的评价.因此能推荐给用户一些不是原有集合中的页面,可是又是用户需要的页面.

在文献[3]中,主要是提供给 Web 站点管理员一个易于导航的索引页,以便改进 Web 站点的组织和表示.而 RM 方法却是为用户提供可满足用户需要,易于浏览的页面集合.在文献[3]中,注重考虑页面间的链接结构,而对页面自身内容的语义信息考虑不多.然而 RM 方法既考虑页面间的链接结构,又考虑页面自身内容的语义信息.

在文献[4]中,Letizia 系统仅仅使用基于内容的过滤,而 RM 方法既是基于内容的过滤,又是基于协同的过滤.

4 结束语

本文提出评分方法 RM.该方法通过提出一系列新的步骤,综合运用信息检索、协同过滤和数据挖掘等方面的技术,从而使推荐给用户的页面评分更加合理.

参考文献:

[1] CHAKRABARTI Soumen, et al. Mining the Web's link structure[J]. IEEE Computer, 1999, 32(8): 60-67.

[2] CUTLER Michal, et al. A new study on using HTML structures to improve retrieval[A]. 1999 11th IEEE International Conference on Tools with Artificial Intelligence[C]. Chicago, Illinois, USA: IEEE Computer Society. 1999. 406-409.

[3] PERKOWITZ Mike, ETZIONI Oren. Towards adaptive Web sites: Conceptual framework and case study[J]. Artif Intell, 2000, 118(1-2): 245-275.

[4] LIEBERMAN Henry. Letizia: An agent that assists web browsing[A]. 1995 4th International Joint Conference on Artificial Intelligence[C]. Montréal, Québec, Canada: AAAI Press, 1995. 924-929.

[5] AGRAWAL Rakesh, et al. Parallel mining of association rules[J]. IEEE Trans Knowl Data Eng, 1996, 8(6): 962-969.

[6] CHEUNG David Wai Lok, et al. A fast distributed algorithm for mining association rules[A]. 1996 4th International Conference on Parallel and Distributed Information Systems[C]. Miami Beach, Florida, USA, 1996. 31-43.

作者简介:

韩立新 男,1967 年 5 月生于江苏省南京市,博士,硕士生导师,主要研究方向为信息检索等. E-mail: lixinhan2002@yahoo.com.cn.