服务器日志挖掘在电力业务系统功能推荐中的应用①

胡扬波, 陈咏秋, 周红林

(江苏电力信息技术有限公司, 南京 210024)

摘 要: 提出了一种基于服务器日志挖掘的电力业务系统功能推荐服务, 首先从电力业务系统服务器日志中获取用户日志数据, 然后对含有“脏”数据的用户日志数据进行预处理, 以适应数据挖掘与处理; 接着由待处理的数据计算用户访问兴趣度, 并基于改进的 K 均值聚类算法将用户访问兴趣度数据集划分为多个具有相近兴趣度的用户集合, 最终为用户提供功能个性化推荐服务. 实验结果证明该方法在实现电力业务系统信息推荐方面具有较好的效果.

关键词: 用户日志挖掘; 电力业务系统; 业务功能推荐

Application of Server Logs Mining to Functional Recommender Service of Electric Power Business Systems

HU Yang-Bo, CHEN Yong-Qiu, ZHOU Hong-Lin

(Jiangsu Electric Power Information Technology Co. Ltd, Nanjing 210024, China)

Abstract: This paper proposes a functional recommendation service of the electric power system based on server logs mining. First of all, we get the user log data from server los, and then preprocess those log data with "dirty" data. Secondly, we calculate interest-measure of each user pairs by the processed data sets, and we divide data set of interest-measure of each user pairs into multiple classes with similar interest-measure based on improved K-means clustering algorithm. Finally, personalized functional recommendation method is provided to each user. The experimental results prove the effectiveness of our method in electric power business system.

Key words: server logs mining; electric power business system; functional recommendation

在电力行业, 坚强智能电网的迅速发展使信息通信技术正以前所未有的广度、深度与电网生产、企业管理快速融合, 信息通信系统已经成为智能电网的“中枢神经”, 支撑新一代电网生产和管理发展. 江苏电力公司在 2008 年建成了统一权限管理平台, 目前该平台已经集成了江苏电力各业务系统的功能, 全面管理江苏电力所有信息化用户. 截止到 2014 年 1 月, 电力信息公司一体化平台已接入系统 209 个, 总功能为 19708 个, 授权信息数已达到 1777405 条, 已产生日志 3 亿多条. 数据来源主要是统一权限中的功能使用日志, 数据量每年增长量为 1 亿条左右. 面对如此庞大的日志数据和纷繁复杂的数据关系, 利用数据挖掘和分析技术, 综合大量日志数据分析用户的偏好和需求, 为用户提供个性化推荐服务[1], 从而节省用户搜索业务功能的时间, 提高用户对电力业务系统的满意度.

本文致力于通过日志挖掘技术实现电力业务系统功能个性化推荐服务的研究. 日志挖掘就是从大量的日志数据、文档和活动中发现用户感兴趣的、潜在信息的过程[2], 其将传统的数据挖掘技术与日志挖掘结合起来, 以数据挖掘、文本挖掘、多媒体挖掘为基础, 并覆盖了多个研究领域, 包括计算机网络、数据库与数据仓库、人工智能、统计学、概率理论、可视化等[3]. 一般而言, 日志挖掘可以分为三类: 日志结构挖掘、日志内容挖掘以及日志使用挖掘[4], 决定了日志挖掘研究

① 收稿时间:2014-06-30;收到修改稿时间:2014-08-14

的多样性. 聚类是目前日志挖掘常用的方法[5-7], 本文在现有 K 均值算法的基础上, 提出一种改进的 K 均值聚类算法, 设计了电力业务系统功能推荐的实现方案. 该方案的功能是针对电力信息用户提供个性化的功能能模块信息主动推荐, 帮助用户更快获取有用信息, 提高工作效率.

1 系统整体流程

电力公司服务器日志是用来记录用户访问活动信息, 是获取用户访问活动情况的首要数据来源. 服务器日志的基本信息包括用户访问请求时间, 用户访问某项功能菜单的次数、持续时间及该菜单长度等. 图 1 为本系统的整体流程图, 具体过程如下: 首先从电力业务系统服务器日志中获取用户日志数据, 然后对用户日志数据进行清理、识别等, 为数据挖掘与分析奠定基础; 接着由待处理的用户日志数据集计算用户访问兴趣度, 并基于改进的 K 均值聚类, 将用户访问兴趣度数据集划分为多个具有相近兴趣度的类, 分析用户的偏好和需求, 为用户提供个性化业务功能推荐服务, 最终使系统更加安全、可靠、稳定和友好.

在上述系统实现过程中, 采用基于改进的 K 均值聚类算法, 考虑如下: 由于现有的 K 均值算法初始点是建立在随机选取的基础之上的, 如果初始化点选择不好, 一般很难跳出局部最优, 而且产生的最终聚类结果也会很差. 因此, 本文通过选取周围密度最大 k 个点作为初始化点, 该方式不仅能够有效地解决 K 均值初始化点选择问题, 而且也能有效地降低孤立点对 K 均值算法的影响, 具体算法如 3.2 节所示.

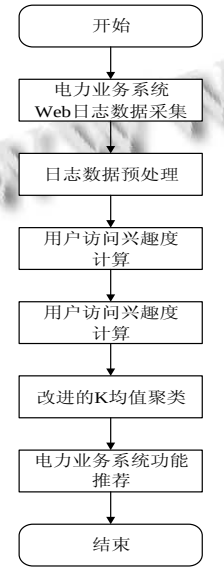

图 1 系统整体流程图

2 日志数据预处理

数据预处理是系统实施有效挖掘算法的前提, 在日志挖掘中具有非常重要的作用. 主要表现在: 数据来源于多个数据源的未被加工、高维、冗余、含有噪音且非均匀分布的复杂数据, 在数据模型、含义、模式、结构和语义上存在不一致性和冲突, 因此数据预处理是日志挖掘质量保障的关键.

数据预处理一般包括数据清理、用户识别、会话识别、路径补全等[7], 具体过程如图 2 所示.

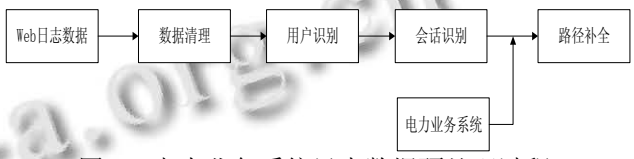

图 2 电力业务系统日志数据预处理过程

1) 数据清理: 主要是清除 Web 服务器日志文件中无关数据项的过程. 一般包括数据合并、删除无关数据、处理代理访问、规范 URL 等.

2) 用户识别: 主要是识别访问电力业务系统的独立用户. 当用户转换使用浏览器或直接输入 URL 时, 一般视为多个用户; 而同一操作系统、同一 IP、同一浏览器访问网站, 且浏览页面集合相同, 一般视为同一个用户.

3) 会话识别: 把用户访问页面的行为划分成多个互相独立的浏览序列, 通过研究这些序列, 获得用户在站点中的浏览爱好或是访问模式.

4) 路径补全: 主要是对访问路径边行修补, 使不完整的访路径变得完整. 在日志引用不完整情况下, 可利用电力业务系统站点的拓扑结构信息填充路径.

3 电力业务系统功能推荐

经过数据预处理环节后, 现在需要对这些日志数据进行挖掘, 本文首先建立用户访问兴趣度来衡量用户对某项业务功能的兴趣度, 然后通过聚类将按照用户共同的访问习惯进行聚类, 最后基于聚类结果完成针对电力信息用户的业务功能个性化推荐.

3.1 用户访问兴趣度

为了能恰当地反映用户对电力业务系统中某项功能关注的程度, 即获得用户访问兴趣度, 需要适当地设定用户访问业务功能的权重. 一般电力公司服务器日志中显示用户访问了某个业务功能, 就认为用户对此业务功能有一定的兴趣. 本文采用有两种度量方法

衡量用户对某项电力业务功能的兴趣度，即“频率系数”和“持续时间系数”[8,9].

1) 访问频率系数 $F(i)$

若某个用户频繁地访问某项电力业务功能，就说明该用户对该功能比较感兴趣. 因此，本文用 $F(i)$表示“频率系数”，表示业务功能 i 在一次会话中被访问次数占用户本次会话的业务功能访问总次数的比重，用式(1)表示:

$$F(i) = \frac{V(i)}{\sum_{i_m \in s_j} V(i_m)} \tag{1}$$

上式中，$V(i)$表示业务功能 i 在一次会话中被访问次数.

2) 访问持续时间系数 $C(i)$

用户访问某项电力业务功能的持续时间的是用户在浏览该业务功能时所花费的时间，在这里计算访问持续时间用同一会话中两次相邻访问请求的间隔时间，但是这种方法不能得到在一次会话中的最后一个业务功能的访问持续时间，因此，对于最后一个业务功能的访问持续时间，本文采用此次会话的平均持续时间来估算. 用户在一个业务功能上消耗的时间越长，则说明用户对该业务功能越感兴趣. 用户浏览业务功能的时间长短也与该业务功能长度有关，因此需要将业务功能的长度与“持续时间”联系起来. C(i)表示“持续时间系数”，是浏览业务功能 i 单位长度上的所耗时间与本次会话所有的浏览网页单位长度上的所耗的最长时间的比值，如式(2)所示:

$$C(i) = \frac{T(i) / L(i)}{\max_{i_m \in s_j}(T(i_m) / L(i_m))} \tag{2}$$

上式中，$T(i)$表示用户浏览业务功能 i 的时间，$L(i)$表示用户所浏览业务功能的长度.

综上，可以看出：用户访问兴趣度与用户访问频率和用户访问持续时间有关，本文假定这两种方法具有相同的重要性，且 $F(i)$和 $C(i)$都是经过标准化的值，因此，为了更好更恰当地反映用户的兴趣度，必须考虑在频率和持续时间同时增大的时候用户的兴趣度随之增大. 本文综合 $F(i)$和 $C(i)$，表示在某一会话中用户对某个业务功能 i 的兴趣度 $I(i)$，$I(i)$的定义表达式如(3)所示:

$$I(i) = \frac{2 \cdot F(i) \cdot C(i)}{F(i) + C(i)} \tag{3}$$

上式表明，兴趣度 $I(i)$只有在用户访问频率系数 $F(i)$和持续时间系数 $C(i)$同时较高的情况下才会具有较高的值，换言之，若用户频繁访问某一电力业务功能，并且浏览这个业务功能的时间也较较长，则表明用户对该业务功能比较感兴趣.

有些业务功能被用户访问的次数比较少，不能反映用户的兴趣度，所以在进行电力信息用户业务功能兴趣度计算时需要设定一个最小的阈值，这样可以减少访问总次数比较少的业务功能，提高系统的处理速度，同时也提高业务功能推荐的可用性.

3.2 改进的 K 均值的用户聚类算法

对 3.1 节中得到的用户兴趣度数据集进行聚类，聚类结果则为多个用户类(簇)，每个类(簇)中的用户访问习惯相近，本文用“用户访问模式”来表示. 用户访问模式是用来描述具有相同浏览访问特征的用户组. 由于多个不同用户在其访问期间可能有相同的兴趣，用户访问模式能有效获得这些用户共同的兴趣或共同的业务需求. 此外，用户访问模式也能将不同兴趣的用户区分开来. 本文采用 K 均值聚类算法获得用户访问模式，该算法具体的实现过程如算法 1 所示.

算法 1 基于改进的 K 均值的用户聚类算法

输入: 用户兴趣度数据集

输出: 用户访问模式

1. 确定要生成的类的数目 k;

2. 按照公式(4)选取 k 个对象作为聚类中心点 $C=\{c_1, c_2, \ldots, c_k\}$，并设置初始迭代次数 r=1，式中 σ 选取 0.5，选取密度最大的前 k 个数据作为初始聚类中心点，公式(4)如下所示:

$$density(x_i) = \sum_{j=1}^{n} e^{\frac{d(x_i - x_j)^2}{2\sigma^2}} \tag{4}$$

3. 对于数据集中其它每个对象 d_i，则根据它们与各个聚类的聚类中心点 c_j 的距离，分别将它们分配给与其最近的、最相似的类，即分至与其具有最小距离的聚类中心点的类中，形成 k 个类;

4. 重新计算每个类的聚类中心点，使用的距离为欧几里德距离，公式如式(5)所示:

$$dis(d_i, c_j) = \sqrt{\sum_{k=1}^{n} (d_{ik} - c_{jk})^2} \tag{5}$$

5. 若所有聚类中心点达到稳定，则结束聚类；否则 r=r+1，跳至步骤(3)，重复执行，直至聚类中心点不再发生变化.

算法 1 中，本文首先针对现有的 K 均值算法初始点选择问题，通过迭代的方式选取密度最大的前 k 个数据点作为初始化聚类中心，该方式能够有效地解决初始点敏感的问题，而且降低了孤立点对 K 均值算法的影响；再依次计算初始类子集中每一个对象到各个种

子点的距离, 并根据计算结果将数据对象逐个分派到其最近均值的类中去, 然后重新计算接受新对象的类和失去对象类的均值, 如此重复, 直到各类再无元素进出. 聚类得到的 k 个用户类具有如下特点: 用户类本身比较紧凑, 而各用户类之间尽可能的分开, 聚类结果得到用户类集合 $C=\{c_1, c_2,..., c_k\}$, 其中每个类 c_i 是具有共同业务系统操作习惯的用户集合.

3.3 用户业务功能推荐

3.2 节基于聚类算法找到的具有共同业务系统操作习惯的用户集合, 本节主要通过这些集合根据业务功能的访问频度进行匹配, 计算每个业务功能的推荐度, 按推荐度从大到小进行排序, 完成用户个性化推荐.

用户业务功能推荐主要以以下形式形成业务功能推荐集: 根据每一个类 c_i 的共同的用户访问兴趣度, 构建面向用户的业务系统推荐集合, 选取 Top-5 个业务系统, 当用户登录到该业务系统时, 将这 5 个业务系统推荐以快捷方式的形式推荐给该类用户.

4 实验结果

为了验证改进的 K 均值用户业务功能聚类算法的有效性, 本文使用了江苏电力信息技术有限公司的统一框架平台中的用户行为数据集, 并选取 2014 年 4 月 1 日-2014 年 4 月 30 日期间的用户行为日志, 具体包括 1200 个用户、4000 个业务系统功能菜单以及 120,522 条用户业务系统点击记录. 运行环境为 Win7 系统, 主频 2.8GHz, 内存 2G, 硬盘 500G, 程序使用 Java 语言实现. 本文根据业务人员所处的部门设置了 9 类用户, 分别为: 财务、规划、建设、运行、检修、营销、信息、办公、审计, 并与 K 均值聚类算法进行比较, 实验中分别设置了 200、400、600、800、1000 以及 1200 个用户作为比较对象, 聚类结果如图 3 所示.

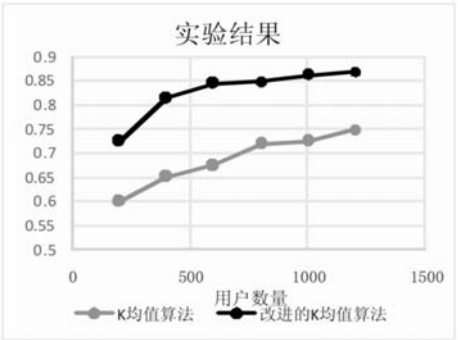

图 3 改进的 K-means 算法和 K-means 算法的聚类准确度比较

在图 3 中, 横轴为选取的用户个数, 纵轴为聚类结果的准确度. 聚类结果准确度在[0, 1]之间, 结果越大则说明聚类效果越好. 从上图中可以看出采用改进的 K-means 算法, 聚类准确度区间为[0.725, 0.87], 而一般的 K-means 算法的准确度区间为[0.6, 0.75], 两个算法的准确度均随着用户个数的增加而增加, 但改进的 K-means 算法的要比一般的 K-means 算法的性能更佳, 因此使用改进的 K-means 算法要比 K-means 算法实现的电力信息用户聚类效率高, 将该算法用在电力业务系统功能推荐上进一步提高了用户工作效率及其满意度.

5 结语

基于服务器日志挖掘的电力业务系统功能推荐服务的研究, 一方面协助用户很快的找到感兴趣的业务功能, 另一方面能够帮助业务系统内容和结构的个性化完善. 论文首先建立用户访问兴趣度计算模型, 接着通过改进的 K-means 算法对用户进行聚类, 然后实现用户业务功能推荐, 最后通过实验验证算法的有效性. 当前, 国内的业务系统推荐的实践仍处在快速发展的时期, 所有还需要继续研究更智能、更优化的业务系统推荐技术.

参考文献

1 易明.基于 Web 挖掘的电子商务个性化推荐机理与方法研究[学位论文].武汉:华中科技大学,2006.

2 韩家炜,孟小峰,王静,李盛恩.Web 挖掘研究.计算机研究与发展,2001,38(4):405–414.

3 丁一.基于 Web 挖掘的个性化推荐服务研究[学位论文].武汉:华中科技大学,2004.

4 Perkowitz M, Etzioni O. Adaptive sites: Automatically learning from user access patterns. 6th Int. World Wide Web Conf. Santa Clara, California. 1997.

5 Zhou B, Hui S, Chang K. An intelligent reconunender System using sequential web access patterns. Cybernetics and Intelligent Systems, 2004, 1: 393–398.

6 Mobasher B, Dai H, Luo T, et al. Effeetive personalization based on association rule discovery from web usage data. Workshop on Web Information and Data Management. Atlanta, Georgia, USA. 2001. 9–15.

7 Phatak D, Mulvaney R. Clustering for personalized mobile Web usage. IEEE FUZZ'02. Honolulu, Hl, USA. 2002. 705–710.

8 Madria SK, Bhowmiek NgWK, et al. Research issues in Web data mining. 1st Int Conf. on Data Warehousing and Knowledge Discovery(Dawak'99). Florence, Italy. 1999. 303–312.

9 Eirinaki M, Vazirgiannis M. Web mining for Web personalization. ACM Trans. on Internet Technology, 2003, 3(1): 1–27.