

基于概率潜在语义分析的 Web 用户会话识别

高春贞, 吴军华

(南京工业大学 信息科学与工程学院, 江苏 南京 210009)

摘 要: 为提高 Web 用户会话识别的准确性, 给出了一种基于概率潜在语义分析模型(PLSA)和竞争奖罚(CRP)算法的 Web 用户会话识别方法. 其核心内容是利用 PLSA 模型计算出请求页面和每一个活动用户会话的概率, 比较概率判定请求页面应该归属的用户会话, 并采用竞争奖罚算法判别用户会话的结束. 实验结果表明: 基于 PLSA 模型和竞争奖罚算法的用户会话识别方法的识别成功率高于其他常用的会话识别方法.

关键词: 会话识别; 概率潜在语义分析; 竞争奖罚算法; 最大期望值算法; Web 使用挖掘

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)06-0163-04

Web Session Identification Based on Probabilistic Latent Semantic Analysis

GAO Chun-zhen, WU Jun-hua

(College of Information Science and Engineering, Nanjing University of Technology, Nanjing 210009, China)

Abstract: To improve the validity and efficiency of the user session identification, a new approach of user session identification is presented, which is based on the probabilistic latent semantic analysis model and the competitive reward and punishment algorithm. The probabilistic latent semantic analysis model is used to compute the probabilities of a request and each active session. The model identifies which active session a request should belong to in terms of the probabilities. A competitive reward and punishment method is applied to determine the end of sessions. The experimental results show that the average percentage of successfully reconstructed sessions based on PLSA and CRP is higher than other session identification algorithms.

Key words: session identification; probabilistic latent semantic analysis; competitive method; EM; Web usage mining

1 引言

Web 用户会话识别是 Web 使用挖掘中的重要处理步骤之一^[1]. 常用的 Web 用户会话识别方法^[2]: 时间阈值(Timeout)、统计语言模型—— n 元语言文法模型(n -gram)等. 但是这些方法没有根据用户的显式行为挖掘用户潜在的搜索意图.

概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)^[1]方法拥有坚实的数学基础及易于利用的数据生成模型. PLSA 是将隐式变量 $z \in Z = \{z_1, z_2, \dots, z_k\}$ 与可视共现的数据对(比如词汇 $w \in W = \{w_1, w_2, \dots, w_m\}$ 在文档 $d \in D = \{d_1, d_2, \dots, d_n\}$ 中的出现)联系起来的统计模型.

文中提出一种基于概率潜在语义分析模型的 Web 用户会话识别方法. 实验结果表明, 利用该模型得到的 Web 用户会话识别成功率高于其他常用的会话识别方法.

2 Web 用户会话识别的概率潜在语义分析模型

2.1 模型描述

Web 日志数据预处理^[2]后, 得到一个大小为 n 的浏览页面集合 $P = (p_1, p_2, \dots, p_n)$ 和一个大小为 m 的用户会话集合 $S = (s_1, s_2, \dots, s_m)$. 用户会话由一系列 Web 页面组成, 它反映了单个用户的浏览行为. Web 用户会话和页面的共现矩阵 $SP = [W(s_i,$

$p_j) \} m \times n$, 其中 $W(s_i, p_j)$ 表示页面 p_j 在用户会话 s_i 中的权重. 该权重可以是一个二进制位, 表示页面在会话中存在与否, 它也可以是一个反映页面在会话中的存在情况或持续时间的函数. 使用 $Z = (z_1, z_2, \dots, z_k)$ 表示潜在语义(隐式变量)的集合, k 为指定的一个常数. 通过概率潜在语义模型将会话和页面两个同步发生的数据集关联在一起. 用户会话 - 潜在语义 - 页面三者之间的关系如图 1 所示.

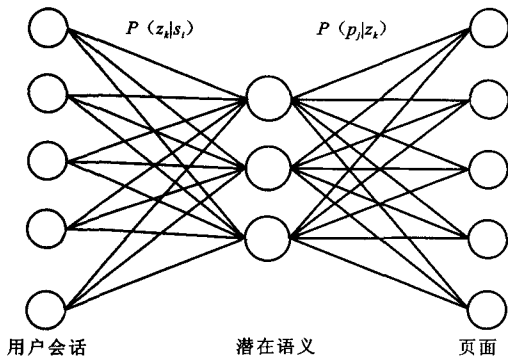


图1 “用户会话 - 潜在语义 - 页面”的关系

概率潜在语义分析假设页面 - 会话之间是条件独立的, 并且潜在语义在会话或页面上分布也是条件独立的. 概率潜在语义分析模型如下:

- $P(s_i)$ 表示从 S 集合中选择一个用户会话的概率;

- $P(z_k)$ 表示潜在语义具有一定的先验知识;

- $P(z_k | s_i)$ 表示会话中潜在语义的概率;

- $P(p_j | z_k)$ 表示页面对潜在语义的贡献度.

经过上述过程获得“会话 - 页面”观测点对 (s_i, p_j) , 潜在语义 z_k 被忽略, 产生“会话 - 页面”的联合概率模型:

$$P(s_i, p_j) = P(s_i) * P(p_j | s_i) \quad (1)$$

其中, $P(p_j | s_i) = \sum_{k=1}^k P(p_j | z_k) * P(z_k | s_i)$ 是所有观测中可能产生的 z_k 值的总和. 由于潜在变量的基数通常比集合中用户会话或页面的数要小很多, 潜在变量 z_k 将成为瓶颈变量, 所以根据贝叶斯公式, 将其箭头方向逆转, 如图 2, 从(a)转换到(b).

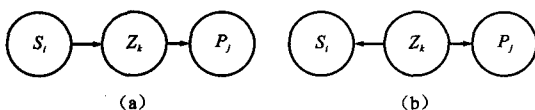


图2 PLSA 算法图示

利用贝叶斯公式可以直接将联合概率转换成

$$P(s_i, p_j) = \sum_{k=1}^k P(z_k) * P(s_i | z_k) * P(p_j | z_k) \quad (2)$$

概率潜在语义分析(PLSA)的核心思想是该模型使得一个不可见的潜在变量集与每一次观测值相关. 这里的观测值为请求页面属于一个活动用户会话的概率. 使用最大期望值 EM(Expectation Maximization)^[3] 算法对潜在语义模型进行拟合. 最大期望值算法(EM)是在一个潜在因素模型中通过最大化似然函数的方法来进行参数估计的著名算法. 在使用随机数初始化潜在变量集之后, EM 算法交替于两个步骤: ①E - 步, 利用当前估计的潜在变量的参数值计算潜在变量的后验概率; ②M - 步, 基于所给后验概率, 更新参数值. 定义似然函数如下:

$$L = (S, P) = \sum_{i=1}^m \sum_{j=1}^n w(s_i, p_j) \log P(s_i, p_j) \quad (3)$$

2.2 模型拟合

(1) E - 步. 根据 $P(z_k)$, $P(s_i | z_k)$ 和 $P(p_j | z_k)$ 的初始值估计期望值:

$$P(z_k | s_i, p_j) = \frac{P(z_k) P(s_i | z_k) P(p_j | z_k)}{\sum_{k'=1}^k P(z_{k'}) P(s_i | z_{k'}) P(p_j | z_{k'})} \quad (4)$$

(2) M - 步. 最大化期望值的似然函数:

$$E(L^c) = \sum_{i=1}^m \sum_{j=1}^n w(s_i, p_j) \sum_{k=1}^k P(z_k | s_i, p_j) \cdot \log P(s_i, p_j) \quad (5)$$

在此似然函数应该满足约束条件:

$$\sum_{k=1}^k \left(\sum_{i=1}^m P(s_i | z_k) - 1 \right) = 0 \quad (6)$$

$$\sum_{k=1}^k \left(\sum_{j=1}^n P(p_j | z_k) - 1 \right) = 0 \quad (7)$$

通过引入拉格朗日因子, 可以从该约束的最大化问题中得到如下等式来进行参数再估计:

$$P(z_k) = \frac{\sum_{i=1}^m \sum_{j=1}^n w(s_i, p_j) P(z_k | s_i, p_j)}{\sum_{i=1}^m \sum_{j=1}^n w(s_i, p_j)} \quad (8)$$

$$P(s_i | z_k) = \frac{\sum_{j=1}^n w(s_i, p_j) P(z_k | s_i, p_j)}{\sum_{i=1}^m \sum_{j=1}^n w(s'_i, p_j) P(z_k | s'_i, p_j)} \quad (9)$$

$$P(p_j | z_k) = \frac{\sum_{i=1}^m w(s_i, p_j) P(z_k | s_i, p_j)}{\sum_{i=1}^m \sum_{j=1}^n w(s_i, p'_j) P(z_k | s_i, p'_j)} \quad (10)$$

重复以上两步的计算,直到似然函数 $L(S, P)$ 期望值的增加量小于一个阈值时停止迭代. 通过 PLSA 计算,能够得到潜在语义变量 z_k 和“会话-页面”的联合概率 $P(s_i, p_j)$ 的局部最优解. EM 算法具有线性的收敛速度,并且简单,容易实现,可以使似然函数达到局部最优^[4].

3 基于 PLSA 的用户会话识别

在用户会话识别中使用 PLSA 模型的一个主要优点是它能生成用来量化 Web 用户会话 S 和潜在语义变量 Z 及 Web 页面 P 和潜在语义变量 Z 之间关系的概率. 使用概率推论法,可以从这些基本的概率值中得出 Web 用户会话 S 和 Web 页面 P 之间的关系概率. 此框架提供了一种灵活的用户识别方法.

PLSA 模型能生成概率 $P(z_k)$, 它是某一个所选择的潜在语义变量的概率值; $P(s_i | z_k)$ 是给定潜在语义变量下的观测一个用户会话的概率; $P(p_j | z_k)$ 是给定潜在语义变量下的访问一个页面的概率. 通过应用贝叶斯法则得到“会话-页面”的联合概率 $P(s_i, p_j)$.

文中利用 PLSA 模型实现交叉用户会话的分离,采用竞争奖罚算法判别用户会话的结束.

3.1 交叉用户会话的析取

基于 PLSA 的用户会话识别框架如图 3.

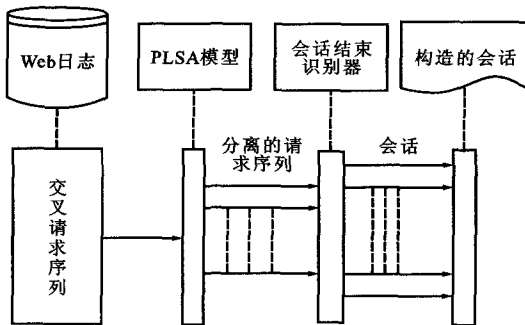


图3 用户会话识别框架模型

使用链表保存分离出来的活动用户会话. 比较基于 PLSA 模型生成的请求页面和每一个活动用户会话的概率 $P(s_i, p_j)$, 具有最大概率的活动用户会话获得请求页面. 如果最大概率小于设定的概率阈

值,重新构造一个新的活动用户会话,当前请求页面是这个新的活动用户会话的第一项.

3.2 用户会话结束的判断

采用竞争奖罚算法判断用户会话何时结束. 竞争奖罚算法为每个新的活动用户会话设定一个能够被更新的能量值. 能量值越小则表明用户会话越接近于结束. 该能量值根据用户请求页面加入到会话的可能性进行更新,如果一个用户会话赢得竞争,它的能量值将会增加 $e^{-a(1-P)}$, P 是用户请求页面加入到这个会话的概率, a 为预设影响因子;反之,如果一个会话竞争失败,它的能量值会减少 e^{-ap} . 当一个会话的能量降低到一定程度时,标志着会话结束.

3.3 算法

基于潜在语义分析模型和竞争奖罚 (PLSA + CRP) 的用户会话识别算法如下:

Input:

Web 日志文件 $L = \{(r_i, t_i), i = 1 \cdots k\}$, r_i 是第 i 项页面请求, t_i 是这个页面请求的时间.

Output:

重构的会话文件

Initialization:

FR(): 奖励函数

FP(): 惩罚函数

energy_{ini}: 初始能量

B: 概率阈值

$S = (s_1, s_2, \dots, s_m)$: 会话链表, 保存活动用户会话, S 集合中的会话 $s_j = \{(r_{j1}, r_{j2}, \dots, r_{jn}), \text{energy}_j\}$, $(r_{j1}, r_{j2}, \dots, r_{jn})$ 是会话 s_j 的请求序列, energy_j 是会话 s_j 的能量.

Begin

For $(r_i) \in L$ do

For $s_j \in S$ do

If $\text{energy}_j \leq 0$ then

Output $\leftarrow s_j$

Remove s_j from S

Else

Compute $P(s_j, r_i)$

// 计算请求 r_i 属于会话 s_j 的概率

End if

End for

$P_{\max} = \max P(s_j, r_i)$

If $P_{\max} > B$ then

Index = the index of the session with $P_{\max}$

$S_{\text{index}} \leftarrow r_i$

$\text{energy}_{\text{index}} = \text{energy}_{\text{index}} + \text{FR}()$

For (s_j) in S , $j \neq \text{index}$ do

```

energyj = energyj - FP()
End for
Else
Snew = {(ri), energyini}
S ← Snew
End if
End for
End

```

4 实验和结果分析

实验所用数据来自于淘宝网的 Web 服务日志数据. 通过对日志数据进行预处理后得到 27459 个用户会话和 594 个浏览页面. 把处理后的数据分为两个数据集 A 和 B. 为潜在语义集合 Z 选择 20 个因素.

利用第一个数据集 A, 通过上面采用的 PLSA 模型得出一组潜在语义变量的值 $Z = (z_1, z_2, \dots, z_{20})$. 然后在这组潜在语义变量值的基础上, 采用基于潜在语义分析模型和竞争奖罚 (PLSA + CRP) 的用户会话识别算法对第二个数据集 B 的请求页面序列进行用户会话识别.

如果一个构造的用户会话和测试会话数据中一个用户会话完全匹配, 或者只存在一个不相符的网页请求, 则该构造的用户会话称为成功的用户会话.

基于用户会话识别的成功率对基于 PLSA + CRP 的用户会话识别方法和其他 3 种常用的用户会话识别方法进行比较. 这三种常用的会话识别方法是: Timeout、Href、n-gram 模型^[5]. 比较结果如图 4 所示. 实验结果表明基于 PLSA + CRP 的用户会话方法的优于其他 3 种方法.

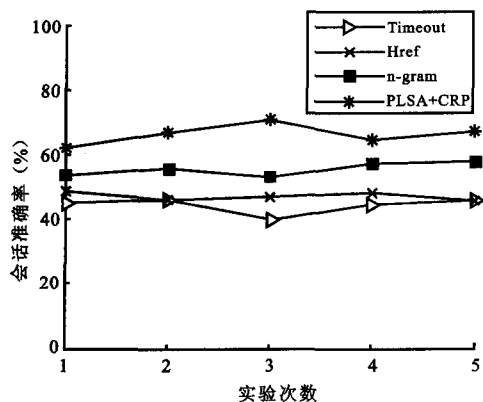


图 4 会话成功识别率比较

5 结束语

文中提出了基于 PLSA 模型和竞争奖罚算法的 Web 用户会话识别方法, 旨在改进传统会话识别技术无法处理用户的潜在搜索意图. 首先利用概率潜在语义模型概率量化 Web 用户会话和潜在语义变量及潜在语义变量和 Web 页面的关系, 得出 Web 用户会话和 Web 页面的概率, 以此作为 Web 用户会话识别的依据; 然后通过比较“会话-页面”的概率, 判定页面应该归属的用户会话; 最后应用竞争奖罚算法判别用户会话的结束. 实验结果表明, 该模型识别用户会话的成功率高于其他常用的会话识别方法.

参考文献:

- [1] Hofmann T. Latent semantic models for collaborative filtering[J]. Transactions on Information Systems, 2004, 22(1): 89-115.
- [2] Spiliopoulou M, Mobasher B, Berendt B, et al. A framework for the evaluation of session reconstruction heuristics in web usage analysis[J]. INFORMS Journal of Computing Special Issue on Mining Web-Based Data for E-Business Applications, 2003, 15(2): 171-190.
- [3] Jin X, Zhou Y, Mobasher B. A unified approach to personalization based on probabilistic latent semantic models of web usage and content[C]//Proceedings of the AAAI 2004 Workshop on Semantic web Personalization. San Jose: AAAI, 2004: 654-658.
- [4] Xu G, Zhang Y, Zhou X. A web recommendation technique based on probabilistic latent semantic analysis[C]//The 6th International Conference on Web Information Systems Engineering. New York: Springer-Verlag, 2005: 15-28.
- [5] Huang X, Peng F, An A, et al. Session boundary detection for association rule learning using n-gram language models[C]//Proceedings of sixteenth conference of the Canadian society for computational Studies of the intelligence. Canada: Halifax, 2003: 237-251.

作者简介:

高春贞 女, (1979-), 硕士研究生. 研究方向为 Web 挖掘、Web 逆向工程.

吴军华 女, (1965-), 副教授. 研究方向为 Web 工程、软件工程.