基于概率潜在语义分析的 Web 用户聚类

俞 辉,景海峰
YU Hui,JING Hai-feng

中国石油大学 计算机与通信工程学院,山东 东营 257061
Institute of Computer & Communication Engineering,China University of Petroleum,Dongying,Shandong 257061,China

YU Hui,JING Hai-feng.Web user clustering based on Probabilistic Latent Semantic Analysis.Computer Engineering and Applications,2008,44(23):160-162.

Abstract: Knowledge of Web user clustering can improve the efficiency of information searching and personalized service.Firstly, session-page matrix can been constructed by analyzing a great deal of log.Then,based on information theory,the local weight and global weight are considered in calculation of weight in session-page matrix.With usage of probabilistic latent semantic analysis, the conditional probability of latent variable Z to page P is transformed the conditional probability of latent variable Z to session S,then the transformed results are used in similarity calculation.The k-medoids algorithm is adopted to further improve clursting result.Experiment results validate validity and limitation of this algorithm.

Key words: Web log;preprocessing;Web user;Probabilistic Latent Semantic Analysis(PLSA);clustering

摘 要:Web 用户聚类知识可以为改进信息搜索效率和提供个性化服务提供帮助。通过对海量日志记录分析，构建会话-页面矩阵;根据信息论理论,在会话-页面矩阵中权值计算中考虑局部和全局权值贡献;利用概率潜在语义分析将隐式变量 Z 对页面 P 的条件概率转换为隐式变量 Z 对会话 S 的条件概率,然后在聚类分析中以此作为相似度计算依据。聚类算法采用了基于距离的 k-medoids 算法,以进一步改善聚类精度。实验结果验证了该算法的有效性和局限性。

关键词:Web 日志;预处理;Web 用户;概率潜在语义分析;聚类
DOI:10.3778/j.issn.1002-8331.2008.23.049 **文章编号**:1002-8331(2008)23-0160-03 **文献标识码**:A **中图分类号**:TP391

随着互联网的发展和网络用户的增加,如何根据用户喜好的不同提供页面过滤、网页推荐等个性化服务,不但有助于提高信息搜索效率缓解网络拥塞，而且也是 Internet 向智能化发展的方向[1-2]。在 Internet 上,用户按照自己的兴趣进行访问，这种兴趣度可以通过用户在 Web 站点上的浏览行为体现出来，通过对具有相似访问兴趣的用户群进行聚类,发现用户的类属模式,利用这些聚类知识可以为站点结构改进,网页推荐服务以及电子商务中发掘潜在客户等应用提供决策依据。

本文提出基于概率潜在语义分析 PLSA(Probabilistic Latent Semantic Analysis)的 Web 用户聚类算法,首先对 Web 日志进行数据预处理,将日志记录处理成用户会话形式;然后根据用户会话构建会话-页面矩阵 ***SP***;利用概率潜在语义分析将独立隐式变量 Z 与共现数据对——如页面 P 在用户会话 S 中的出现——联系成概率统计模型的特点,将隐式变量 Z 对页面 P 的条件概率转换为隐式变量 Z 对会话 S 的条件概率,以此为依据进行聚类分析。在计算过程中,根据信息论的原理设计矩阵权值计算方法,权值由局部权值和全局权值组成。计算方法着重考虑了页面对会话的类别区分能力;聚类方法采用了基于距离的 k-medoids 算法,以进一步改善聚类精度。

1 相关概念

1.1 概率潜在语义分析

概率潜在语义分析起源于自然语言处理研究,用于分析文档的潜在语义[3-4]。它的基本思想是对于给定文档集 $D=\{d_1,d_2,\cdots,d_m\}$和词集 $W=\{w_1,w_2,\cdots,w_n\}$以及文档和词的共现矩阵 $\boldsymbol{A}=|a_{ij}|_{n\times m}$,其中 a_{ij} 代表不同词 w_j 在文档 d_i 中的权值。使用 $Z=\{z_1,z_2,\cdots,z_k\}$ 表示潜在语义的集合,k 为指定的一个常数。概率潜在语义分析假设词-文档对之间是条件独立的，并且潜在语义在文档或词上分布也是条件独立的。在上面假设的前提下,可使用下列公式来表示词-文档的条件概率:$p(w_j|d_i)=\sum_{k=1}^{k}p(w_j|z_k)p(z_k|d_i)$。上式中的 $p(w_j|z_k)$为潜在语义在词上的分布概率,也可以解释为词对潜在语义的贡献度,通过对 $p(w_j|z_k)$排序可以得到潜在语义的一个直观的词的表示。$p(z_k|d_i)$表示文档中的潜在语义分布概率。

概率潜在语义分析使用最大期望 EM(Expectation Maximization)算法对潜在语义模型进行拟合[5-6]。在使用随机数初始化之后,交替实施 E 步骤和 M 步骤进行迭代计算。在 E 步骤中计算何一个(d_i,w_j)对产生潜在语义 z_k 的先验概率:

作者简介:俞辉(1974-),男,讲师,主要从事数据挖掘、网络集成的研究;景海峰(1981-),男,助教,主要从事计算机应用、网络集成的研究。
收稿日期:2007-10-15 **修回日期**:2008-01-28

$$p(z_k|d_i,w_j)=\frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{k}P(w_j|z_l)P(z_l|d_i)}$$

在 M 步骤中，使用下列公式对模型重新估计：

$$P(w_j|z_k)=\frac{\sum_{i=1}^{n}a(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{j=1}^{m}\sum_{i=1}^{n}a(d_i,w_j)P(z_k|d_i,w_j)}$$

$$P(z_k|d_i)=\frac{\sum_{j=1}^{m}a(d_i,w_j)P(z_k|d_i,w_j)}{a(d_i)}$$

当 L 期望值的增加量小于一个阈值时停止迭代，此时得到一个最优解：

$$E(L)=\sum_{i=1}^{n}\sum_{j=1}^{m}a(d_i,w_j)\sum_{l=1}^{k}P(z_l|d_i,w_j)\log[P(w_j|z_k)P(z_k|d_i)]$$

其中 $a(d_i,w_j)$ 代表词-文档矩阵权值 a_{ij}。对于会话-页面矩阵，利用此模型可以得到 $p(z_k|s_i)$、$p(z_k|p_j)$，其中 s_i、p_j 分别代表会话 i 和页面 j。

1.2　k-mediods 算法[7-8]

k-means 算法是常用的基于划分的聚类算法，该算法对异常数据较敏感。而 k-mediods 算法利用中心点来作为聚类中心以代替 k-means 算法中以均值作为聚类中心，作为聚类中心的中心点受到异常数据或极端数据的影响较少，从而可以根据各元素与各中心点之间的距离之和最小化的原则，应用划分方法。同 k-means 聚类算法相比，k-medoids 聚类算法在处理异常数据和噪声数据方面更为鲁棒。k-medoids 算法步骤：

(1)从 n 个数据元素任意选择 k 个元素作为初始聚类(中心)代表；

(2)循环步骤(3)到步骤(4)直到每个聚类不再发生变化为止；

(3)依据每个聚类的中心代表元素，以及各元素与这些中心元素间距离，并根据最小距离重新对相应元素进行划分；

(4)任意选择一个非中心元素 O_i，计算其与中心元素 O_j 交换的成本 S。若 S 为负值则交换 O_i 与 O_j 以构成新聚类的 k 个中心元素。

k-medoids 聚类算法的基本思想就是首先任意为每个聚类找到一个代表元素(medoids)为确定 n 个数据元素的 k 个聚类，每个元素需根据它们与这些聚类中心的距离分别将它们归属到各相应聚类中。如果替换一个聚类中心能够改善所获聚类质量，那么就用元素替换聚类中心。这里将利用一个基于各元素与其聚类中心间距离的成本函数来对聚类质量进行评估。为了确定任一个非聚类中心元素 O_i 是否可以替换当前一个聚类中心 O_j，需要根据情况对非聚类中心元素 P 进行检查。

2　基于概率潜在语义分析的 Web 用户聚类算法描述

基于概率潜在语义分析的 Web 用户聚类算法由两部分组成：数据处理和聚类分析，下面分别描述这两过程。

2.1　数据处理

算法以 Web 日志作为数据处理的基础，Web 日志按时间顺序记录了不同用户对站点的访问。通过数据预处理可以将日志数据整理成用户会话的形式，即在规定的超时时间限制内，用户对站点连续、完整访问的序列形式。预处理过程一般包括 4 个步骤：数据清理、用户识别、会话识别、路径补偿。在本文中还应根据各页面访问时间，计算出页面的停留时间。因此最后用户会话整理成为以访问页面和停留时间组成的二元组为元素的序列。

构造会话-页面矩阵：$\boldsymbol{SP}=|a_{ij}|_{m\times n}$，其中 a_{ij} 为非负值，表示第 j 个页面在第 i 个会话中出现的权重；m、n 分别表示会话数和页面总数。不同的会话对应矩阵 $\boldsymbol{SP}$ 不同的行，每一个页面则对应矩阵 $\boldsymbol{SP}$ 的一列。通常权值要考虑来自两方面的贡献，即局部权值和全局权值。按照信息论的理论可知，如果某一页面在一个会话中出现的频率越高而在其他会话很少出现，则此页面具有较好的类别区分能力，适合用来分类。因为在会话中某一页面可以多次出现并且每次停留的时间不一定相同，直观上页面点击次数的多少和停留时间长短都反映了用户的兴趣度，因此定义局部权值 $lw_{ij}=\frac{n_j}{\sum_k n_k}\times\frac{t_j}{\sum_k t_k}$，$n_j$ 代表页面 j 在会话 i 中的出现次数；t_j 代表页面 j 在会话 i 中的停留时间；$\sum_k n_k$ 代表会话 i 中的所有页面出现的总次数；$\sum_k t_k$ 代表会话 i 中的所有页面总的停留时间。同时；由于页面在不同会话中出现的频率也反映了用户的兴趣度，定义全局权值 $gw_{ij}=\log\frac{\sum_k s_k}{\sum_k s_{kj}}$，其中 $\sum_k s_k$ 代表会话总数 $\sum_k s_{kj}$，代表含有页面 j 的会话总数。总权值是上述两权值之积，为了消除量纲的影响，还要对总权值进行归一化处理，所以权值 a 的计算方法如下式：$a_{ij}=\frac{lw_{ij}\times gw_{ij}}{\sum_j lw_{ij}\times gw_{ij}}$。

2.2　聚类分析

对会话-页面矩阵利用 PLSA 模型分析，不但可以得到隐含变量 z_k 在页面 p_j 已知的条件下的条件概率 $p(z_k|p_j)$，可以得到隐含变量 z_k 在会话 s_i 已知的条件下的条件概率 $p(z_k|s_i)$，这样可以构建会话-隐含变量向量 $\boldsymbol{dl}_i=(s_{i,1},\cdots,s_{i,k-1},s_{i,k})$，其中 $s_{i,k}$ 代表隐含变量 z_k 在会话 s_i 已知的条件下的条件概率 $p(z_k|s_i)$，向量反映了会话和隐含变量的关系。利用此向量可以计算两会话的相似度，设计相似度计算公式如下：

$$sim(d_i,d_j)=\boldsymbol{dl}_i,\boldsymbol{dl}_j/(\|\boldsymbol{dl}_i\|_2\cdot\|\boldsymbol{dl}_j\|_2)$$

其中 $(\boldsymbol{dl}_i,\boldsymbol{dl}_j)=\sum_{m=1}^{k}c_{i,m}c_{j,m}$，$\|\boldsymbol{dl}_i\|_2=\sqrt{\sum_{m=1}^{k}c_{i,m}^2}$，在聚类算法的选择中，本文选用基于距离的 k-medoids 算法，与 k-means 算法比较，该算法选用聚类中位置最靠中心的点做参考点，从而消除了 k-means 算法因采用质心做参考点而导致对孤立点敏感的缺点。同时，该类算法无需事先给出聚类的个数；可以发现非球形的聚类和大小差别很大的聚类；聚类结果与数据输入次序无关并且结果稳定、鲁棒性好。算法在迭代中利用评价函数来选择聚类中心。

2.3　算法步骤

输入值：Web 日志、阈值 μ、隐式因子数 k。

输出值:用户聚类结果 $DC=\{DC_1,DC_2,\cdots,DC_l\}$和相应的聚类中心 $Cid=\{Cid_1,Cid_2,\cdots,Cid_l\}$,其中 l 表示聚类数,DC_i 内包含所属的代表用户的会话。

步骤 1 Web 日志数据预处理。

步骤 2 构造出会话-页面矩阵 ***SP***,计算矩阵各权值。

步骤 3 利用 PLSA 和会话-页面矩阵 ***SP***,求得每个会话的会话-隐含变量向量 ***dl***。

步骤 4 随机选择 1 个向量初始化聚类 DC_1,使 $DC_1=\{dl_1\}$,$Cid_1=\{dl_1\}$。

步骤 5 对每一个 ***dl***,计算其与每个聚类中心点的相似度 $sim(\boldsymbol{dl},Cid_j)$。

步骤 6 如果 $sim(\boldsymbol{dl},Cid_t)=\max_j(sim(\boldsymbol{dl},Cid_j))>\mu$,则将 ***dl*** 插入 DC_t,并检查更新 Cid_t,否则 ***dl*** 将成为新类并成为新类中心。

步骤 7 如果类中心点不再改变或没有未归类的 ***dl*** 则停止,否则重复步骤 5、6。

步骤 8 输出用户聚类结果 $DC=\{DC_1,DC_2,\cdots,DC_l\}$和相应的聚类中心 $Cid=\{Cid_1,Cid_2,\cdots,Cid_l\}$。

3 实验分析

实验选取的 Web 日志文件大小为 14.2 M,包含 103 245 条记录,经预处理共识别出 239 条用户会话,共涉及 373 个不同的页面,分别在不同的控制参数下对算法进行测试得到的结果图如图 1、2 所示。

首先可以看出该算法在一定控制参数下具有较高的聚类准确率,同时在图 1 表明,隐含因子数对聚类准确性有一定影响,当隐含因子数过小时,无法利用有效的隐含语义信息进行聚类;当过大时,隐含语义提供的识别能力降低,同样降低了聚类精度。图 2 表明阈值对聚类准确率的影响,因为阈值大小直接影响着用户的归属,图中表明当阈值取 0.75 时聚类精度较高。

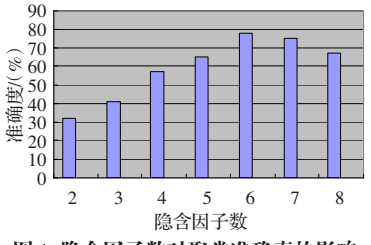

图 1 隐含因子数对聚类准确率的影响

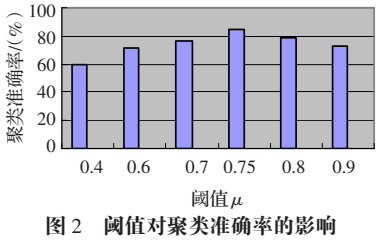

图 2 阈值对聚类准确率的影响

4 结束语

本文提出基于概率潜在语义分析的 Web 用户聚类算法,首先对 Web 日志进行数据预处理,将 Web 日志转换为会话-页面矩阵;根据信息论的原理设计矩阵权值计算方法,权值由局部权值和全局权值组成,着重考虑了会话中各元素对用户分类的能力。然后利用概率潜在语义分析 PLSA 将独立隐式变量 Z 与共现数据对——如页面 P 在用户会话 S 中的出现——联系成概率统计模型的特点,将隐式变量 Z 对页面 P 的条件概率转换为隐式变量 Z 对会话 S 的条件概率,以此作为聚类分析中相似度计算依据,聚类算法采用了基于距离的 k-medoids 算法,以改善聚类精度。实验结果验证此算法的有效性,同时本算法受一定参数影响,考虑到提高用户聚类的准确性,下一步应当优化会话-页面矩阵权值的计算方法同时还应当参考网页内容信息,希望本文工作可以对相关研究提供借鉴。

参考文献:

[1] Hofmann T.Latent semantic models for collaborative filtering[J].ACM Transactions on Information Systems,2004,22(1):89-115.

[2] 王实,高文.路径聚类:在 Web 站点中的知识发现[J].计算机研究与发展,2001,38(4):482-486.

[3] Xiao J.Measuring similarity of interests for clustering Web-users[C]//Proceedings of the 12th Australasian Database Conference(ADC2001).Queensland,Australia:ACS Tnc,2001.

[4] Xu G,Zhang Y,Zhou X.A latent usage approach for clustering Web transaction and building user profile[C]//The First International Conference on Advanced Data Mining and Applications(ADMA 2005).Wuhan,China:Springer,2005.

[5] Zhang Y,Xu G,Zhou X.A latent usage approach for clustering Web transaction and building user profile[C]//ADMA,2005:31-42.

[6] Jin X,Zhon 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(SWP'04),San Jose,2004.

[7] Cohn D,Chang H.Learning to probabilistically identify authoritative[C]//Proc of the 17th International Conference on Machine Learning.San Francisco,CA:Morgan Kaufmann,2000.

[8] Hofmann T.Unsupervised learning by probabilistic latent semantic analysis[J].Machine Learning Journal,2001,42(1):177-196.

(上接 159 页)

法在运行时间、聚类个数、聚类结果等方面作比较,并证明本文提出的方法能有效地保证运行效率及聚类个数的有效性,且能够得到更令人满意的结果记录。

参考文献:

[1] Zhao D J,Lee D J,Luo Qiong.A meta-search method with clustering and term correlation[C]//LNCS 2973:Proc of DASFAA,2004:543-553.

[2] Howe A E,Dreilinger D.Savvy search:a metasearch engine that learns which search engines to query[J].AI Magazine,1997,18:19-25.

[3] Radovanovi M,Ivanovi M.CatS:a classification-powered meta-search engine[J].Advances in Web Intelligence and Data Mining(SCI),2006,23:191-200.

[4] 王勇.基于模糊聚类的 Web 使用模式挖掘研究[D].重庆:重庆大学,2004.

[5] 高新波.模糊聚类分析及其应用[M].西安:西安电子科技出版社,2004.

[6] 钱夕元,邵志清.模糊 ISODATA 聚类分析算法的实现及其应用研究[J].计算机工程与应用,2004,40(15):70-71.

[7] 邵峰晶,于忠清.数据挖掘-原理与算法[M].北京:中国水利水电出版社,2003.