Web信息检索结果个性化排序模型

毕 鹏

(山东大学齐鲁软件学院 济南250000)

摘 要 本文讨论了如何从网页点击次数的统计数据中获得用户对网页中包含信息的评价。在考虑了网页内容，时间等因素对信息价值的影响后，给出了一种基于用户评价的对信息检索结果个性化排序的模型。模型根据用户浏览网页时的行为和用户的特征信息，预测用户对信息的需求，智能地对信息检索结果进行个性化的排序。模型实现简单，可以应用于多数信息检索系统，为用户提供个性化的信息服务。

关键词 奇异值分解，个性化，点击量，个人特征

A Personalize Sort Model for Search Results of Information Search System

Bi Peng

(Software College, Shandong University, Shandong 250000, China)

Abstract This paper discusses how to obtain customers' evaluation of the information from the Click through. After considering the effect of time and other factors to the value of the information, we provide a personalize sort model, which is based on the customers' evaluation. The model can forecast customers' requirement and sort the search results intelligently. The realization of the model is easy and it can be applied to most of Information Search System to provide personalize information search service.

Keywords Singular value decomposition, Personal, Click through, Personal profile

1 引言

伴随着 Internet 的飞速发展，互联网上每天发布的信息量正在飞速地增长。准确地把用户需要的信息从海量的数据中检索出来有着重大的意义。搜索引擎是从互联网中检索数据的有效工具。但是，由于大型搜索引擎覆盖大量的数据，而且多数用户的需求存在模糊性，使得检索结果仍包含大量无用数据，给用户造成不便。对检索结果进行一定的排序，突出重要数据的地位，可以帮助用户更快捷地得到有用的信息。

传统的检索结果排序方法有，按网页中包含的关键字的数目排序、PageRank[1]等，较为成功的方法为 PageRank。PageRank 通过计算网页中超级链接的数量评价网页的重要性，它通过网络自身的结构体现网页包含信息的重要程度，取得了一定成功。但是信息检索是一种个体行为，检索结果应受用户年龄、经验、文化程度等因素的影响，不同用户对同一主题检索希望得到的结果是不同的。所以，为用户提供高质量的信息检索服务，应考虑到用户的个性，这样才能高效地为每个用户提供准确的检索结果。传统的算法已无法满足人们日益增长的对信息个性化的需求。

本文探讨了一种基于用户信息对检索结果排序的方法。

2 用户需求预测模型

经过关键字简单筛选的信息记录需要进一步的处理才能满足用户的需求。个性化的排序应考虑网页包含的信息与用户检索内容的相关程度和网页包含信息的价值。可以通过对网页内容特征与用户查询请求的差异程度来评估网页包含的信息满足用户需求的程度；并通过其他用户对该网页的评价来模糊的推测该网页的质量。

2.1 网页内容评价

可以用于对网页内容评价的模型有：布尔逻辑模型、向量空间模型、概率模型、神经网络模型等。矩阵的奇异值分解方法(singular value decomposition, SVD)是一种实践证明有效的通过上下文提取信息的方法，已经大量用于网页的采集，所以我们选用了矩阵的奇异值分解方法。在此方法中，首先把网页中的关键字表示为一个 n 维向量，然后把相关网页的集合组织成一个矩阵 X 称为文档矩阵。SVD 分解是一种矩阵的正交分解方法，它把文档矩阵分解为三个矩阵。

$$X=U\Sigma V^T$$

U 和 V 为正交的矩阵，Σ 为一对角矩阵。该方法用 U 表示各个关键字的关系，V 表示网页间的关

毕 鹏 本科生，主要从事神经网络机器学习等方面的研究。

系。为减少计算量，可用近似矩阵 X_k 替代 X 进行分析得到公式：

$X_K=U_K\Sigma KV^TK \quad (U_k^TU_k=I;V_K^TV_k=I)$

U_K 是上式中 U 的前 K 列，V_K 是上式中 V 的前 K 列，Σ_K 包含 X 的前 k 个最大的奇异值。同样，用户的查询表示为一个 K 维向量 Q_p（表示一个虚文档），通过 Q_p 与 V 进行相似匹配。Q_p 的求法：

$Q_p=O^TU_K\Sigma_K^{-1}$

通过计算 Q_p 与 V 向量的内积可得到一个反映用户的查询与网页内容相关程度的向量。

利用以上算法可得到每个网页与用户查询的相似度，从而得到用户对网页内容的评价。

2.2 网页包含信息价值的评价依据

人们在考虑同一问题时往往具有不同的出发点，具有相同背景的人具有相同的出发点更容易达成共识，对同一问题感兴趣。所以，可以按照不同的标注对用户进行分类，通过记录以往同类用户对该信息的评价来预测当前用户对信息的反应。

2.3 获取评价的方法

获取用户评价的方法有两种：1 显式地询问用户。2 通过用户的行为预测。方法1能够得到准确的答复，但是会增加用户的负担，引起用户厌烦。方法2虽然存在噪声，但是对于粗略的估计不会产生太大的影响。所以我们的模型采用了方法2。

浏览次数（以后称点击量）是反映用户对网页所包含信息的评价的重要数据，而且它易于获得，用它作为获取用户评价的标准很合适。为了获取用户的评价，我们把网页分为四类。

1）用户感兴趣的：用户浏览两次以上的网页。

2）用户可能感兴趣的：用户只浏览过一次的网页。

3）用户可能不感兴趣的：

Journal of Computer Science&Technology
Weloome to Journal of Computer Science&Technology 2003-2004 Publicatio
College of Computer Science and Technology
Introduction Introduction of the College Research on computer science…
Welloome to The Department of Computer Science
Back The Department of Computer Science Computer science department…

图1

如图1所示，检索结果中第一条和第三条记录已被点击，我们有理由相信用户已经看到了第二条记录并且决定不去点击它。根据人的视觉原理和记录显示的宽度来估计点击记录上下用户可以看到的记录的条数。如图中的记录二的记录会被定义为用户可能不感兴趣的。

4）用户可能没看到的：剩余的纪录就是用户可能没看到的。

因为 URL 唯一标示网页，所以可以通过记录 URL 来记录用户访问的网页。

2.4 预测算法

用户被按照标准分为 n 类$\{u_1,u_2,u_3,\cdots,u_n\}$，一天之中网页 A 被 u_1类型用户的点击量(θ)的计算方法为：

$\theta=a_1*x_1+a_2*x_2+a_3*x_3+a_4*x_4$

其中 x_1,x_2,x_3,x_4为一天中评价该网页为感兴趣的、可能感兴趣的、可能不感兴趣的以及可能没看到的用户的数量。a_1,a_2,a_3,a_4为常数，可通过手工设置。

时间是对信息价值影响最大的因素，所以在对网页包含信息价值评价时应考虑时间因素的影响。信息的价值会随着时间而发生变化。随着时间的推移，信息可能会失去其过去的价值，例如：当一则新闻的累计访问量达到最大时，该新闻一般已失去了价值，所以在计算点击量时应该考虑时间对信息价值的影响。考虑了时间因素的点击量计算方法如下：

$\varepsilon=\Sigma\Phi(\theta_i,t) \quad (i=1,2,3,\cdots,n)$

ε—为一网页的累计点击量，Φ 是由网页内容特性确定的随时间变化的函数，t＝当前时间-网页创建的时间。

全部用户被按 n 种标注进行 N 次分类，(例如标注为年龄，兴趣等，按年龄分为儿童，少年，青年等)分为 $A_i(i=1,2,3,\cdots,n)$，$B_i(i=1,2,3,\cdots,m)\cdots N_i(i=1,2,3,\cdots,j)$。用户个人特征被记为向量$(n_1,n_2,n_3,\cdots,n_n)$，每个分量属于一种分类方式的一个元素，如果 $n_1\in\{B_X\}$且 B_X 对应的累计点击量为 m，则对应于 n 个属性得到一个用户点击量的向量 $p(m_1,m_2,m_3,\cdots,m_n)$。($m_n$ 通过上文中的公式 $\varepsilon=\Sigma\Phi(\theta_i,t)$计算)通过向量 P 与用户个人特征向量的内积得到一个标量即为用户评价度，记为 c。c 就是通过以往同类用户评价而得到的当前用户对该网页可能感兴趣的一个度量。

c 的计算公式为：$c=f(wp+b)$，其中 f 是一个线性函数，w 是一个权值向量(由用户个人特征决定)，b 是偏置值。权值向量 w 和偏置值 b 可通过对用户行为的学习得到。对搜索结果中的每条记录，计算相应的与用户查询的相似度和用户评价度，通过比较它们和的大小对检索结果进行排序。

3 系统的设计与实现

为了减少信息检索服务器的系统负荷，系统采

用了分布式模型。在服务器数据库的页面记录中设置记录用户评价度的数据结构和对文档矩阵的支持。

3.1 客户端的实现

agent 是一种工作在客户端具有自主性，机动性，智能性，合作性的软件形式。客户端采用这种形式，以增强灵活性和智能性。

客户端以 Plug-in 的形式与浏览器系统结合。用户首次登录系统时，系统向用户询问用户基本信息并记录在本地文件中。用户提出查询时，服务器通过关键字查找数据，并生成相应的文档矩阵。服务器返回数据后，客户端计算查询与文档的相似度，根据用户信息计算用户评价度，对 URL 进行排序。在用户浏览过程中，记录用户点击动作和浏览的 URL，并把数据存入临时文件中。系统退出时把数据发送到服务器。服务器每天根据收集的数据计算网页的点击量。

3.2 客户端用户界面的设计

与传统的搜索引擎界面不同，为了用户浏览方便，网页被分三类显示：用户已经浏览的；用户不感兴趣的；用户未浏览的和新页面。这样用户在第二次查找时可以过滤记录，用户就可以更轻松地找到需要的页面。

4 实验结果及改进

我们通过互联网收集了十个主题的5000余个网页进行实验，并在 Windows Xp 和 Internet Explore6.0平台上以 com 组件的形式实现了一个模型系统。实验性的设定 $a_1=5, a_2=3, a_3=2, a_4=0$，经过对系统一个月的使用，通过对50个人的使用情况的统计得到以下数据，如表1。

表1

用户希望网页在第一条的比率	38.7%
用户希望网页在前十条的比率	83.4
用户希望网页在后十条的比率	0.3%

实验表明，系统能够体现用户的期望。虽然实验规模比较小，但仍表明系统有良好的性能。客户端排序网页的平均延时为0.05秒，所以用户不会有任何的感觉。

实验表明的问题：在计算网页的点击量时需要一定的计算量，串行计算点击量的方法使用时间较长。改进方案，可以并行计算点击量缩短计算点击量的时间。

结论 本文提出的个性化信息检索，可以根据用户的喜好，为用户提供有效的、个性化的信息。这样，不仅可以大大节省用户的时间，提高用户的效率，还可以充分地利用网络资源，提高网络的可用性。

参 考 文 献

1 Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Stanford Digital Libraries Working Paper, 1998

2 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. In: Proc. of the 7th Intl. World Wide Web Conf., Brisbane, Australia, 1998

3 Hinrich S. Automatic word sense discrimination. Computational Linguistics, 1998, 24(1): 97～123

4 Mitchell T M. 机器学习[M]. 北京：机械工业出版社，2003

（上接第26页）

多评价和比较分类器性能的方法，例如交叉验证法、刀切法、自助法、贝叶斯模型选择法等。每种方法都基于一些假定，这些假设与最终评价并无明显的关系。

4 独立于算法的机器学习

除上述的各种具体分类算法外，我们还要特别提到一种广义的算法——独立于算法的机器学习。它包括两层含义：首先，其数学基础不依赖于所采用的特定分类器和特定学习算法；其次，这些技术可与各种不同的学习算法组合使用，或者为它们提供应用指导，以改善分类器的总体性能。以上罗列的各种评价和比较分类器性能的方法（交叉验证法等）、偏差和方差法、重采样方法（Bagging、AdaBoost 法等）、组合分类器法等均属于独立于算法的机器学习[4]。

结束语 本文总结和跟踪了各种常用和最新的文本分类的技术和算法，分析了评价与比较分类器性能的定理和方法，简述了独立于算法的机器学习。今后的研究工作将着重围绕新算法的提出和旧算法的改进展开，并将结合计算语言学和本体论知识，将现有模型中常用的词空间拓展为概念空间来进行文本自动分类。

参 考 文 献

1 史忠植. 知识发现. 清华大学出版社，2002.1

2 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现. 计算机应用研究，2001(18)

3 Duda R O, Hart P E. Stork D G. 模式分类. 机械工业出版社，2003.9

4 黄萱菁，吴立德. 独立于语种的文本分类方法. In: 2000 Intl Conf. on Multilingual Information Processing, 2000

5 王永庆. 人工智能原理与方法. 西安交通大学出版社，1998.5