基于用户反馈的搜索引擎排名算法①

金祖旭　李敏波　(复旦大学 软件学院 上海 201203)

摘　要: 以 Web 2.0 中用户行为作为研究对象，通过发掘用户反馈方式，提出用户反馈分值的概念，对用户反馈影响搜索结果排名的具体方法以及相应实现进行研究，提出了一种基于神经网络的网页排序算法。该算法引入 BP 神经网络模型，根据用户反馈分值选择样本训练神经网络。将传统搜索结果输入到经过训练的神经网络进行计算，根据计算出的结果所表示的网页相关性强弱判断后进行二次排序。该算法利用了神经网络具有的模式识别能力，有效地将用户反馈和搜索引擎结合起来，使得搜索结果更加符合用户的搜索要求。

关键词: 搜索引擎; 用户反馈; 神经网络; 排序算法

Ranking Algorithm of Search Engine Based on Users Feedback

JIN Zu-Xu, LI Min-Bo

(Software School, Fudan University, Shanghai 201203, China)

Abstract: This paper used user behavior in Web 2.0 as a research object, explored ways of user feedback, and proposed the concept of user feedback score. It studied the specific methods and corresponding realization for user feedback impacting the final ranking of search results, and presented a sorting algorithm for search results based on neural network. The algorithm used the BP neural network model, select samples to train the neural network based on user feedback score. Traditional search results will be put into the trained neural network to compute, and a new ranking will be made according to relevance of the web page which indicated by the calculated results. This algorithm used the neural network's pattern recognition capabilities, combined user feedback and search engine effectively, making search results more in line with the user' s search request.

Keywords: search engine; user feedback; neural network; ranking algorithm

搜索引擎已经成为大家在工作、学习、娱乐中不可或缺的工具。通过使用搜索引擎，使得我们检索信息的能力获得了极大的提高，成本有效地降低。最早搜索引擎应用于门户网站，获得了极大的成功，如今，它已广泛地应用于各行各业，以它为核心引发了所谓的搜索经济，成为大家关注的焦点。

统计结果表明: 65%-70%的网民仅仅点击搜索结果的第一页，即前 10 条，20%-25%的网民点击搜索结果的第二页，即第 11-20 条，而只有 3%-4%的网民点击第二页以后的搜索结果。因此，一个网站如果要被用户点击到，它在搜索结果中的排序位置是至关重要的。所以，一个好的搜索引擎应该尽可能地将相关性最好的页面排在前面，而这与一个好的排序算法是不可分割的。对于任何一个搜索引擎来说，搜索排名算法无疑是它最为核心的东西。目前搜索排名算法大致可分为两类: 以网页为中心(Web-Centric)和查询为中心(Query-Centric)。

以网页为中心的排序算法主要使用预定的规则通过机器识别来排序。Sergey Brin 等人[1]提出 PageRank 算法开启了链接分析研究的热潮。PageRank 的基本思想是: 如果网页 T 存在一个指向网页 A 的链接，则表明 T 的所有者认为 A 比较重要，

① 收稿时间:2010-03-16;收到修改稿时间:2010-04-19

从而把 T 的一部分重要性得分赋予 A。这个重要性得分的值则由 T 的 PageRank 值 PR(T)和 T 的出链(从 T 链出的链接)数 C(T)决定。具体公式为：PR(T)/C(T)。而对于页面 A，其 PageRank 值 PR (A)的计算如下：$PR(A) = PR(T_1) /C(T_1) +\cdots+ PR(T_n) /C(T_n)$其中，$T_1$，$T_2$，…，$T_n$为含有指向 A 链接的页面。PageRank 在 Google 中的应用获得了巨大的商业成功。随后，出现了一系列 PageRank 算法的改进算法。主题敏感 PageRank 算法(Topic-Sensitive PageRank，TSPR)[2]的核心思想就是通过离线计算，计算出一个 PageRank 向量集合(在 PageRank 算法中，仅计算一个 PageRank 向量)，该集合中的每一个向量与某一主题相关，即计算某个页面关于不同主题的得分。Kleinberg[3] 提出了 HITS(Hypertext-Induced Topic Search)算法，其利用页面之间的引用关系来挖掘隐含在其中的有用信息(如权威性)，具有计算简单且效率高的特点。HITS 算法通过两个评价权值—内容权威度(Authority)和链接权威度(Hub)来对网页质量进行评估。HITS 算法需要大量在线计算，实际使用困难，更大程度上是一种实验性质的尝试。

鉴于目前大多数页面排序算法只分析包含在 Web 页面中的链接，文献[4]提出了 Link Fusion 页面排序算法。在该算法中，将链接分为两类：①Intra-type Links。用于表示同一数据空间中的数据对象关系，多指包含在 Web 页面中的链接。②Inter-type Links。用于表示不同数据空间中数据对象之间的关系，多指用户、查询条件与 Web 页面之间的关系。在链接分析中，同时考虑了 Intra-type Link 和 Inter-type Link 的影响。具体来说，用户和他们提交的查询条件以及用户浏览的 Web 页面分别代表三个数据空间。当用户提交查询请求时、用户浏览 Web 页面时、一个查询参考其他 Web 页面时，这三个不同的数据空间便被联系起来。三种操作(Submit、Browse、Reference)包含了这三个不同数据空间之间的 Inter-type L ink。为了提高用户查询结果的准确性，一些算法通过用户的查询日志(Query Log)来确定用户的偏好，进而找出用户的目的。文献[5]提出了一种新的用于确定用户目标的方法—User-click Behavior，它是利用用户点击数据来提高搜索引擎的效果。将用户提交的查询请求以及用户所点击的查询结果页面记录下来，然后通过对这些数据的分析，获得用户的兴趣以及用户定位信息资源的模式，从而更准确地执行用户的查询请求。文献[6]分析传统 PageRank 算法存在的问题和用户网上浏览页面的习惯，将 Web 内容挖掘的页面相似度引入到 PageRank 算法中，对算法进行改进。改进后的算法可以使页面 PageRank 值因相似度的引入而发生变化，更符合用户的期望。文献[7]通过使用神经网络算法方法，对元搜索引擎的结果进行优化。文献[8]在用户查询方式，查询表达以及查询词三个层次上对用户查询行为进行了分析得到了搜索引擎用户查询的一般。文献[9]通过对用户行为分析，提出了一种自动进行搜索引擎性能评价的方法。此方法能够基于对用户的查询和点击行为的分析自动生成导航类查询测试集合，并对查询对应的标准答案实现自动标注。文献[10]着眼于搜索日志的分析和应用，主要对用户搜索行为模型、搜索行为分类、网页排序算法的优化、异常搜索行为的检测等问题进行研究。

通过查阅国内外研究情况发现，目前搜索引擎的排序算法基本不考虑用户的搜索意图和行为规律以及反馈，而是完全从网页本身出发，基于现有的网页结构给出排序结果。因此，在分析 PageRank 和 Hits 算法的时候不难发现两者共同面临的问题，如主题漂移、无法检测链接的有效性等。由于以网页为中心(Web-Centric)算法内在的缺陷，导致目前大多数商业化搜索的搜索结果都不能使用户满意。

近年来，以查询为中心(Query-Centric)算法逐渐成为学者研究的重点。大部分算法分析搜索日志希望从用户的搜索行为中分析用户搜索的意图、搜索习惯，用以改善搜索结果的排序。其实，仅仅分析搜索日志具有一定的局限性，并不能非常有效的提高搜索结果的准确率。随着 Web2.0 思想的广泛普及，如何让用户在搜索活动中创造内容和分享知识，如何在搜索引擎中体现广大用户的选择和认同，是一个值得研究的课题。本文从用户参与搜索的角度出发，对用户向搜索引擎贡献内容，用户评价影响搜索结果排名的具体方法以及相应实现进行研究，并提出了一种基于神经网络的网页排序算法。该算法利用神经网络具有的模式识别能力，使用神经网络对网页的相关性进行判断，有效的将用户反馈和搜索引擎结合起来，丰富了搜索内容，使得搜索结果更加符合用户的搜索要求，更加人性化。本文探索及实现了此算法，并对实现过程中遇到的问题提出一些改进，希望对将来此类算法的设

计提供一些借鉴和参考的价值。

1　相关知识

1.1 人工神经网络

人工神经网络是一个能够通过已知数据的实验运用来学习和归纳总结的系统。通常来说，一个人工神经元网络是由一个多层神经元结构组成。神经元是一个多输入单输出的信息处理单元。工程上用的神经元模型如图 1 所示。一个处理单元的输入等于其他与之相连接的处理单元施加于它的输入乘以相应的权重之和，神经元的激励函数 F 处理单元当前的活跃状态和输入值求得该处理单元一个新的输出值。

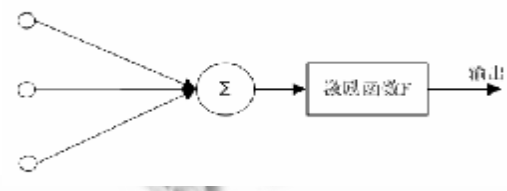

图 1　神经元模型

大量的神经元通过一定的拓扑结构连接起来，就形成了人工神经网络。人工神经网络的信息处理功能是由网络的单元(神经元)的输入输出特性(激励函数特性)，网络的拓扑结构(神经元的连接方式)所决定的。

1.2 用户反馈分析

用户反馈包括显式反馈与隐式反馈两种。隐式反馈是指在用户没有明确参与的情况下，系统通过分析用户行为来得到用户对搜索结果的认同度。显式反馈是指通过用户主动对网页的结果评价来量化用户的认同度，以及通过用户上传自定义的关键词与链接对应关系来增强用户认同度。用户对搜索结果评价可以有两种投票方式，一种是赞同，一种是反对。赞同最多的是用户想要搜索到的结果，反对最多的就是用户最不需要的结果。如果一个链接被用户点击的次数越多，则该链接的质量分值也越高。

本文提出需要通过量化的方法来反应用户对搜索结果的认同度。用户反馈分值是通过对用户反馈的量化提出的一个评价用户对搜索结果认同度的指标。定义搜索结果链接(针对用户关键词 K)的用户反馈分值。

$$Value(K_i,Url_j)=C_1*Explicit(K_i,Url_j)+C_2*Implicit(K_i,Url_j) \quad (1)$$

$$Explicit(K_i,Url_j)=d_1*Aggre(K_i,Url_j)+d_2*Oppse(K_i,Url_j) \quad (2)$$

$$Implicit(K_i,Url_j)=\frac{ClickCount(K_i,Url_j)}{\sum ClickCount(K_i,RUrl)} \quad (3)$$

$Value(K_i,Url_j)$代表用户反馈分值，它包括两部分：显式反馈分值和隐式反馈分值。显式反馈分值 $Explicit(K_i,Url_j)$是用户对搜索结果链接赞同与反对分值的线性组合。隐式分值是指用户点击搜索关键字 Ki 的结果集中链接 Urlj 的次数与用户点击搜索关键字 Ki 的结果集中所有链接 RUrl 总数的比值。

2　基于用户反馈的网页二次排序算法

由于传统的搜索引擎都是通过位置信息返回排名后的搜索结果的，为了提高搜索引擎的查准率，有必要进一步挖掘返回结果除了位置信息以外的其他信息。通过对用户反馈的分析，本章提出了一种基于神经网络的网页排序算法。通过对搜索结果中每条搜索链接引入用户反馈分值，然后根据分值确定神经网络的训练样本。把搜索得到的一次排序结果输入到训练完成的神经网络，利用神经网络的模式识别能力判断网页的相关性，根据相关性排名做出最后的排序。

2.1 神经网络模型

人工神经网络具有自学习、自组织、较好的容错性和优良的非线性逼近能力。在实际应用中，80%～90%的人工神经网络模型是采用误差反传算法或其变化形式的网络模型(简称 BP 网络)。本文的搜索排名算法采用三层 BP 神经网络。三层 BP 神经网络模型如图 2，模型包括输入层，隐藏层，输出层。

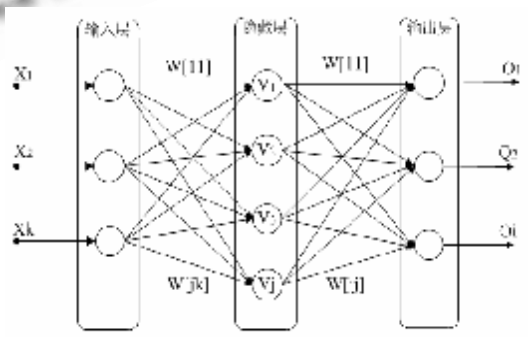

图 2　三层 BP 神经网络模型

2.2 训练算法描述

神经网络通过训练，从样本中学习知识，并且将知识以数值的形式存储于连接权中。BP 算法被称作反

向传播算法，主要思想是从前向后(正向)逐层传播信息；从后向前(反向)逐层传播输出层的误差，间接算出隐含层误差。整个流程如图 3 所示。第一阶段(正向过程)输入信息从输入层向后逐层计算各单元的输出值。第二阶段(反向传播过程)输出误差逐层向前算出隐含层各个单元的误差，并用此误差修正前层的值。在 BP 算法中常采用梯度法修正权值。

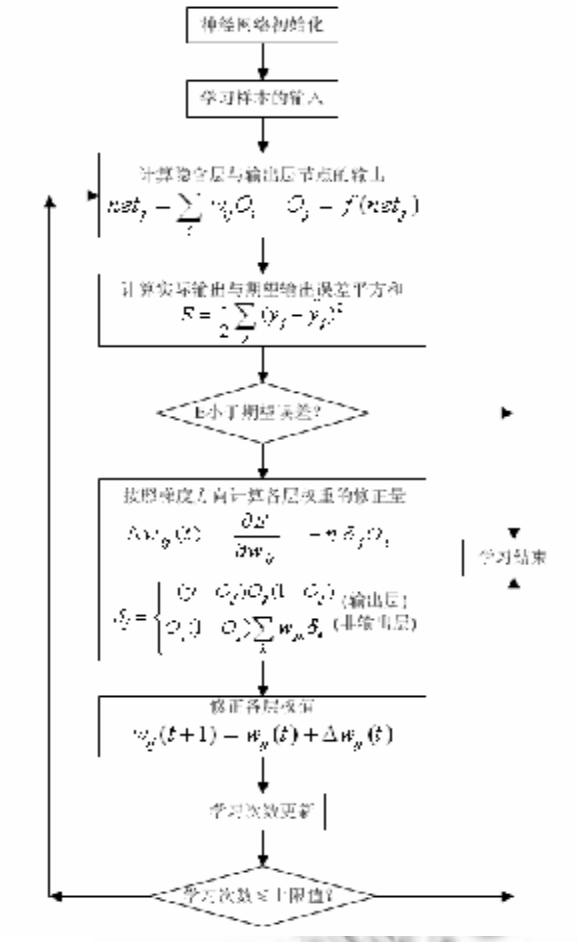

图 3　BP 训练算法流程图

2.3 规范输入输出

神经网络的神经元只能接受二值量化的输入，而网页的相关性信息则是以字符串的形式出现的，因此网页的相关性信息是不能直接输入神经网络进行判断的。本文采取的办法是将页面的相关性信息(如标题、摘要等)进行中文分词、取出关键字等操作，先将字符串转化成一个关键字数组，再由关键字数组抽取建立关键词组模版(图 4)。针对每一条链接对应的关键词，将它与关键词组模版进行对照生成二元向量(图 5)。

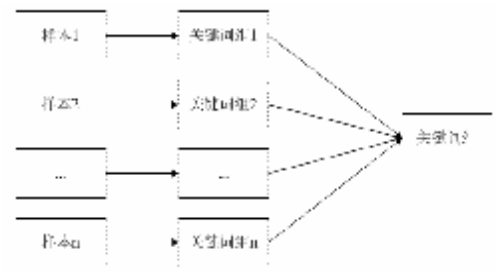

图 4　抽取信息建立关键词组模版

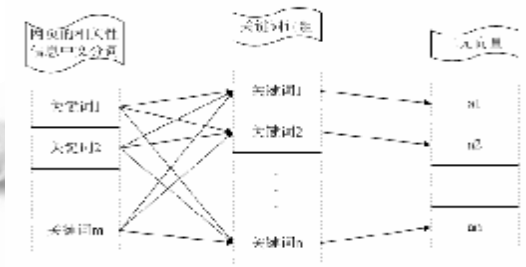

图 5　对照生成二元向量

同样，在神经网络的训练算法中，对于神经网络的输出数据也需要二值化。对于相关性训练样本的输出，定义 y[k]=1；对于非相关性训练样本的输出，定义 y[k]=0，其中 k 代表输出二元向量的长度，即相关性分辨率，k 越大，最后结果排序的区分度越大。

3　算法应用分析

3.1 数据结构设计

通过对数据结构设计加入用户认同度的属性，对每一条链接的结构加入用户赞成数、用户反对数、以及重复度，结构数据表如图 6。

属性名	属性描述	属性类型
linkid	关联 ID	bigint
keyword	关键词	nvarchar200]
linkaddress	链接地址	nvarchar500]
content	链接内容简介	nvarchar500]
agree	用户赞成数	int
oppose	用户反对数	int
overlap	重复度	int

图 6　关键词—链接表结构

通过这个快速算出用户反馈分值。然后根据用户反馈分值选出相关性与非相关性训练样本。本文中选出反馈分值最高的 10 条作为相关性训练样本，反馈分值最低的 5 条作为非相关性训练样本。

3.2 算法应用设计

首先，提取网页一次搜索的结果。对结果信息中文分词确定关键词模版。然后，根据模版与样本的对照，确定训练向量。然后对神经网络进行训练。最后，将所有一次搜索结果都输入到神经网络得到输出，根据输出的相关性强弱进行二次排序。以下展示一个简单的实例应用。

(1) 搜索“三国”得到的结果列表：

网页游戏 热血三国 网游

大型史诗电视连续剧---三国

三国历史

三国演义

三国杀如何玩

(2) 抽取关键词定义关键词组模版(可以有重复)

[三国，历史，游戏，连续剧，三国，演义，电视，网游]

(3) 假设有一个相关性训练样本，通过分词得到以下信息。

[三国，历史，游戏，网游]

对比以上模版，得对应的二值向量。

[1, 1, 1, 0, 1, 0, 0, 1]

同时，假设算法输出中相关性分辨率 k=6，对于相关性训练样本，y[k]=1，即如下向量。

[1, 1, 1, 1, 1, 1]

(4) 将训练样本输入到三层 BP 神经网络进行训练。输入层(Input)元素个数代表数组模版内的关键词个数，隐含层(Hidden)元素的个数代表训练样本的数目，输出层(Output)元素的个数是输出二元向量的长度，即相关性分辨率 k。

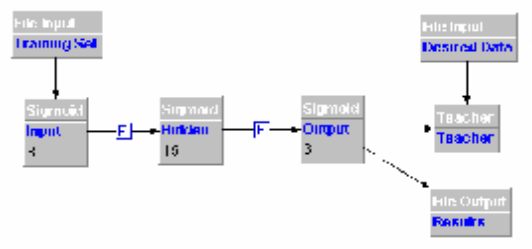

图 7　神经网络的结构图

(5) 训练结束以后，一个具备识别能力的神经网络就构建完成。然后，将一次排序结果对应的二元向量逐个输入到神经网络得出输出向量。输出向量中，其中 1 的元素越多，代表相关性越强，0 的元素越多，代表相关性越弱。根据相关性强弱，对结果重新做出排名。

3.3 应用中涉及到的其它问题

对于一个算法来说，算法运行的性能是至关重要，本文中的算法性能除了算法本身逻辑以外，还牵涉到以下问题：

(1) 中文分词问题。分词速度太慢，即使准确性再高，对于搜索引擎来说也是不可以用的，因为搜索引擎需要处理数以亿计的网页，如果分词耗用的时间过长，会严重影响搜索引擎内容更新的速度。因此对于搜索引擎来说，分词的准确性和速度，二者都需要达到很高的要求。

(2) 算法并行运算问题。算法独立的步骤可以设计成并行运算。神经网络训练过程与各链接的分词可以设计成并列进行。

在实际运用中，本算法的性能还有待进一步的提高。

4　结论

根据以上模型，我们构建了一个新型的搜索引擎，其优点在于能够利用用户反馈对一次搜索结果进行二次排序。本实验邀请 50 名同学参加测试使用。首先，要求 50 名同学对一次排序的结果(即基于网页内容排序)进行满意度投票，每个同学都可以对自己满意的链接结果投票，数目不限。然后，要求其中的 20 名同学对搜索结果做出赞成或反对的评价，系统收集评价信息构建并训练人工神经网络，将一次排序的结果输入到训练完成的神经网络进行二次排序。最后，要求 50 名同学对两次排序结果进行满意度投票。50 名同学的投票情况统计如图 8。在使用了本文算法所得到

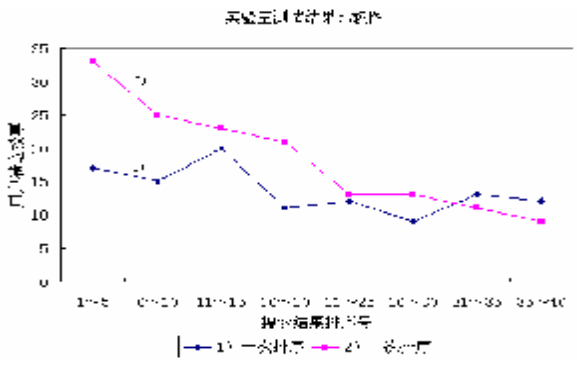

图 8　测试情况统计结果

二次排序的搜索结果中，用户对搜索结果的满意投票大部分集中在前 20 条记录，而说明本算法将用户认同的搜索结果排在了前面，有效的利用用户反馈信息改进搜索结果排名，提高了查准率。

参考文献

1 Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine http://www-db.stanford.edu/~backrub/google 2003-6.

2 Haveliwala, Taher H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Transactions on Knowledge and Data Engineering, 2003,15(4):784－796.

3 Kleinberg Jon M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,46(5):604－632.

4 Xi WS , Zhang BY, Chen Z, Lu YZ, Yan SC, Ma WY, Fox EA. Link fusion: A unified link analysis framework for multi-type interrelated data objects. Thirteenth International World Wide Web Conference Proceedings, 2004. WWW2004, 2004. 319－327.

5 Sun JT, Zeng HJ, Liu H, Lu Y, Chen Z. CubeSVD: a novel approach to personalized Web search. Proceedings of the 14th International Conference on WWW, 2005.

6 李村合,杨春伟.基于 Web 内容挖掘的搜索引擎页面等级改进算法.微计算机应用, 2007,6:571－574.

7 陈默,陈纯,卜佳俊;基于神经网络的元搜索引擎[硕士学位论文].浙江大学, 2006.

8 刘承启,邓庚盛,江婕,徐健锋.基于用户行为分析的搜索引擎研究.计算机与现代化, 2008,9:75－77.

9 刘奕群,岑荣伟,张敏,茹立云,马少平.基于用户行为分析的搜索引擎自动性能评价.软件学报, 2008,19(11):3023－3032.

10 陈红涛,杨放春.基于搜索日志的用户行为研究及应用[博士学位论文].北京邮电大学, 2007.