第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)02—0152—03 文献标识码：A 中图分类号：TP312

基于用户评价的查询串与搜索结果特征权重计算

吴春尧，曲文龙，杨炳儒
(北京科技大学信息工程学院，北京 100083)

摘 要：提出了利用大量用户评价结果来进行特征权重的计算方法，用于解决搜索引擎中查询串与搜索结果的相似度分析。该方法完全利用用户对搜索结果的“潜在评价”来进行。用户对输入查询串所做的点击反映了其内部的关联性，该文提出的方法可获取这种关联性，对该问题建立了数学模型，利用 EM 算法解决了特征权重的计算。由于模型的函数比较复杂，难于计算其收敛性，因此，使用了模拟退火算法作为 EM 算法的补充，用于验证算法的收敛性。实验使用百度搜索引擎在竞价广告上进行，提取的测试数据样本为 100 个广告和 144 132 个 query，获得的数据结果显示，所有特征收敛到全局最优解，抽样部分数据获得检索相似准确率为 93.32%，召回率为 87.43%。
关键词：网页排名；特征权重；EM 算法；模拟退火算法

Feature Weight Calculation Between Query and Texts Based on User Evaluation

WU Chunyao, QU Wenlong, YANG Bingru
(Information Engineering College, University of Technology & Science Beijing, Beijing 100083)

【Abstract】This paper proposes a feature evaluation algorithm by using users click in order to analysis similarity between query and texts in search engines. This method gets features from potential evaluation of user search results because user's clicks to search results reflect the inner relation between query and documents in search results. EM algorithm is used to calculate feature weights. It is difficult to know whether the model's function is convergent because of its complexity. So the simulation annealing algorithm validates the model's convergence as the complement of EM algorithm. The experiment is carried out in Baidu's advertisement ranking. The samples have 100 advertisement and 144 132 queries related to these advertisement. The experiment shows its precision is 93.32% and its recall is 87.43%. All features in the experiment are convergent.
【Key words】Page rank; Feature weight; EM algorithm; Simulation annealing algorithm

字符串相关度计算作为模式识别的基本方法在搜索引擎中有很重要的应用。对于中文来说，用户的输入多是一个短串的关键词序列，搜索引擎的网页数据大多是长串的字符串，返回结果与查询串的匹配程度直接关系到用户的体验。字串相关度是通过其特征向量进行计算的，对其特征向量的提取及权重分配对相似度计算效果起着关键的作用，目前常用的是 TF-IDF 方法，这种方法获得的权重评价值是利用现有的语料统计评价出来的。而在搜索引擎等应用中，可以利用用户对搜索结果的点击记录来评价特征权重，用户对查询串返回的结果是否做了点击，说明了搜索引擎返回的结果是否符合用户的查询意图，也一定程度上表明了查询串与查询结果的相关度。

本文利用用户的使用信息来评价查询串和文本串的相关度。使用搜索日志来改善排名已经有很多研究[1~9]，但是，专门作为一种相关度计算方法来处理中文搜索权重的问题研究的并不多，本文对这个问题建立了数据模型，使用EM算法进行参数计算[10~12]，用模拟退火算法实验说明了算法的全局收敛性[13]。

1 数学模型建立

1.1 问题形式化描述

搜索引擎根据用户输入的查询串，在搜索引擎的结果页面内返回搜索结果。假设用户对结果的点击仅仅与查询串有关。建立如下的数学模型：

设引起某个搜索结果A被点击的所有查询串(queries)为$Q=\{q_1 \ldots q_n\}$，W为搜索结果A的所有的词空间，$W=\{w|w$为汉语词$\}$，搜索结果$A=\{w_1,\ldots,w_o|\ w_o \in W\}$，$q_i$是由w组成，$q_i=w_{i1}\ldots w_{ij}\ldots w_{im}$，$w_{ij} \in A$。

用户首先进行了搜索串q_i查询后，使得搜索结果A得到展现B。如果，搜索结果A产生点击C，即A=(B,C)，那么，由q_i引起搜索结果A点击概率$P(A,q_i)=P(B,q_i)*P(C|B,q_i)$，A与查询串$q_i$的相关度为

$$P(C|B,q_i)=\frac{P(A,q_i)}{P(B,q_i)} \tag{1}$$

通过上面的分析可以看出，$P(A,q_i)$为q_i的展现率，极大似然估计为

$$P(A,q_i)=\frac{C(q_i)}{\sum_{k\in Q}C(q_k)} \tag{2}$$

$P(B,q_i)$为q_i搜索后的A的点击率，其极大似然估计为

$$P(B,q_i)=\frac{C(q_i,click)}{C(q_i)} \tag{3}$$

$C(q_i,click)$为q_i搜索后产生点击的次数。

模型的目标，是确定对于搜索结果A的所有包含的词

作者简介：吴春尧(1970－)，男，博士生，主研方向：自然语言处理，搜索引擎数据挖掘；曲文龙，博士生；杨炳儒，博导、教授
收稿日期：2006-01-02 **E-mail：**wuchunyao@baidu.com

$\{w_1,\ldots,w_o \mid w_o \in W\}$与搜索串的相关度特征权重：

$$S=\{s_1,\ldots s_j\ldots,s_o\},\ \sum_{j=1}^{o} s_j{}^2=1 \tag{4}$$

对于所有查询都是单个词的查询情况下，即$q_i=\{q_{i1}\}$，$S=\{s_1,\ldots s_j\ldots,s_o\}$直接用式(2)、式(3)进行估计，然后利用式(1)计算出来。

1.2 多个搜索词模型

在$q_i=w_{i1}\ldots w_{ij}\ldots w_{im}$，$m\geq 1$时情况如下：

(1)m=1 时用户对搜索结果A的点击仅仅是由这个词w_{i1}产生的；

(2)m>1 时，搜索结果 A 的点击是由多个词共同作用的结果，用下面的公式分配权重：

$$F=\{f_{i1},\ldots f_{ij}\ldots,\ f_{im}\},\ \sum_{j=1}^{m} f_{ij}=1 \tag{5}$$

归一化后，得

$$f_{ij}=\frac{f_{ij}}{\sum_{j=1}^{m} weight(f_{ij})} \tag{6}$$

weight 为权重转移函数：

$$weight(w)=\begin{cases} s_j, w_j=w \\ 0, w_j\neq w \end{cases},\ w_j\in A \tag{7}$$

即将搜索结果A内的权重作为f_{ij}查询串的初始权重，再进行归一化处理。

可以看到，模型的目标是要估计搜索结果 A 内词的权重，然而，在进行估计权重时并不知道。EM 算法提供了可以解决这个矛盾的途径。

2 相关参数估计方法

2.1 EM 算法

EM算法是一种统计迭代寻优的方法。假定有两个样本空间A和B，A是完全数据，B是观察数据。h：A→B是从A到B的多对一映射。以$g(\theta \mid b)$表示θ的基于观察数据B的后验分布密度函数，称为观察后验分布。$f(\theta \mid B,C)$表示添加数据C后得到的关于θ的后验分布密度函数，称为添加后验分布。$k(C \mid B,\theta)$表示在给定θ和观察数据B下潜在数据C的条件分布密度函数，目的是计算观察后验分布$g(\theta \mid B)$的总数。记θ^i为第i+1次迭代开始时参数的估计值，则第i+1 次迭代的两步为：

E 步：对 $f(\theta \mid B,C)$或 $\log f(\theta \mid B,C)$求条件期望，从而把 C 积掉。即

$$Q(\theta,\theta^{i})=E_z[\log f((\theta \mid B,C)|B,\theta^{i}]$$

$$=\int \log f(\theta \mid B,\ C)\mid B,\theta^{i})k(C\mid B,\theta^{i})d_z$$

M步：将$Q(\theta,\theta^i)$极大化，找一个点θ^{i+1}，使$Q(\theta^{i+1},\theta^i)=\max Q(\theta,\theta^i)$。即$\theta^{i+1}=\text{argmax}\ \max\ Q(\theta,\theta^i)$。如此迭代$\|\theta^{i+1}-\theta^i\|$。

2.2 全局最优的检验-模拟退化算法

由于问题域的数学函数很复杂，很难进行数学证明计算问题解空间的性质，因此，这里使用实验来验证所得解的全局最优性能。EM 算法本身无法保证所得解为全局最优点，但是，其解一定是个局部最优点。EM 算法的最优值是初值敏感的，这样，就为我们提供了对实验进行验证的可能。

这里借助模拟退火算法的思想对所得的解进行全局最优验证。模拟退火算法属于软件算法的一种，模拟金属退火的过程解决组合最优的问题，其基本思想是利用概率方式随机选取最优初值，进行迭代计算后，获得的最终解与以前的每一步解比较，选取最优值，由此往复，有限步骤内终止。

由于 EM 算法是初值敏感的，因此可以随机地选取不同的值进行迭代，如果在多次迭代后，没有出现不同的特征权重，即可近似认为，所获得的最优解是全局最优的。

3 试验分析

3.1 测试数据

本文使用百度搜索引擎在线使用的竞价广告数据，搜集抽取 100 条竞价广告，提取相关的搜索串记录得到 144 132 条，经过切分处理后，获得的查询串包含关键词的数量。由于特征是通过关键词来表示的，每个查询串包含关键词的数量与算法的收敛和准确性有关，本试验的查询串关键词数量分布见表 1。

表 1 查询串关键词数量分布

%	1	2	3	4	5	6-
比例	28	30.2	21.4	10.4	4.5	6.5

从上表可见，切分成多个串占很大的比例。

3.2 算法设计

对数据分别进行预处理、分词、计算展现率，由此 EM 算法迭代得到最优验证。

(1)预处理

提取某个搜索结果相关的所有信息，包括引起结果展现的查询串和提取搜索结果的词空间。$A=\{w_1,\ldots,w_o|w_o\in W\}$(去掉没有搜索到的词)。遍历所有queries，将其处理成序列，$Q=\{q_1\ldots q_n\}$。

(2)展现率计算

根据式(2)计算所有 query 的展现率：

$$P(A,q_i)=\frac{C(q_i)}{\sum_{k\in Q} C(q_k)}$$

其中，$C(q_i)$为 query 的发生次数，分母为所有次数之和。

(3)EM 算法

E 步(Expectation,计算期望)：

1)第 1 次，随机选择θ^i，即$S=\{s_1,\ldots s_j\ldots,s_o\}$，满足

$$\sum_{j=1}^{o} s_j=1$$

以后，利用M步计算获得θ^{i+1}，作为本次θ^i，这样就获得了模型的期望权重，即θ^i。

2)计算$\|\theta^{i+1}-\theta^i\|$：

$$\|\theta^{i+1}-\theta^i\|=\sqrt{\frac{1}{o}\sum_{j=1}^{o}(s_j^{i+1}-s_j^i)^2} \tag{8}$$

若足够小，则停止，选取结束阈值为 0.000 01；否则，记录本次的θ^i，进入M步。

M 步：

M步(Maximum,最大化目标函数)：获得了θ^i参数，可以计算各个组成词的期望次数，单个词记为 1 次，对于多个词的情况由各个词分担，分担比例按式(4)计算。目标函数的最大化采用式(2)、式(3)计算极大似然估计法。

(4)模拟退火算法

1)随机选取$S=\{s_1,\ldots,s_j,\ldots,s_o\}$的分配比例；

2)EM算法获得θ^i并记录$\|\theta^{i+1}-\theta^i\|$；

3)根据 Metropolis 准则定义转移概率：

$$P_i(i\rightarrow j)=\begin{cases} 1, f(j)\leq f(i) \\ \exp(\frac{f(i)-f(j)}{t}), f(j)\succ f(i) \end{cases}$$

其中 $f(i)=\|\theta^{i+1}-\theta^i\|$。

$P_i(i \rightarrow j)$表示随机选取$S=\{s_1,\ldots s_j\ldots,s_o\}$的可能性，1 表示完全随机，0 表示跟以前一样。t取值为步数的倒数。

3.3 实验数据分析

(1)收敛性分析

实验使用模拟退火算法对 EM 算法的收敛性验证的结果如下：

包含不同广告的查询串关键词个数分布见表 2。

表 2 不同关键词试验的收敛结果

包含关键字	1	2	3	4	5-
广告数量	8	43	37	7	5
最大迭代次数	1	2	3	5	4
平均迭代次数	1	1.1	2.3	2.5	2.2

对 100 个广告测试结果得出，所有的结果都在有限步内收敛到唯一解。这就说明，EM 算法可以在本模型的数据域中收敛到全局最优解。

(2)特征权重评价

对 100 个广告获得的特征权重评价结果进行相似度计算，对 3 432 个不重复关键词随机抽样 200 个关键词进行评定，得出实验结果见表 3。

表 3 特征权重评价方法的试验结果

准确率	召回率	F1
94.45%	88.27%	90.27%

通过分析，位置、相关性的假设、停词和展现问题可能成为结果的影响因素。这些将会在后续研究中得到改进。

4 结论及展望

本文提出了一种经验的方法用于评价字符串相关度计算中的特征向量权重，试验获得了较好的效果，这种方法是利用大量的评价结果获得的，准确性高和时效性强是这个方法的优点，对于有大量评价数据的场合，会有很好的应用。

参考文献

1 Kumar R, Suri P K, Chauhan R K. Search Engines Evaluation[J]. DESIDOC Bulletin of Information Technology, 2005, 25(2): 3-10.

2 Debora D, Stefano L, Panayiotis T. Stability and Similarity of Link Analysis Ranking Algorithms[C]. Proceedings of the 32th International Colloquium on Automata, Languages and Programming, 2005: 717-729

3 Jiang Xuemei, Song Wenguan, Zeng Huajun. Applying Associative Relationship on the Click Through Data to Improve Web Search[C]. Proceedings of the 27th European Conference on IR Research, 2005: 475-486.

4 Ricardo B Y. Applications of Web Query Mining[C]. Proceedings of the 27th European Conference on IR Research, 2005: 7-22.

5 White R W, Jose J M, Ruthven I. Using Top-ranking Sentences to Facilitate Effective Information Access[J]. Journal of the American Society for Information Science and Technology, 2005, 45(10): 1113-1125.

6 Tao Wenxue, Zuo Wanli. Query-sensitive Self-adaptable Web Page Ranking Algorithm[J]. International Conference on Machine Learning and Cybernetics, 2003: 413-418.

7 Diligenti M, Gori M, Maggini M. A Unified Probabilistic Framework for Web Page Scoring Systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(1): 4-16.

8 Xing W, Ghorbani A. Weighted Page Rank Algorithm[C]. Proc. of the 2nd Annual Conference on Communication Networks and Services Research, 2004-05-19: 305-314.

9 Han Jiawei, Chang K C C. Data Mining for Web Intelligence[J]. Computer, 2002, 21(3): 64-70.

10 张士峰. 混合正态分布参数极大似然估计的 EM 算法[J]. 飞行器测控学报, 2004, 4(4): 47-52.

11 李家福, 张亚非. 基于 EM 算法的汉语自动分词方法[J]. 情报学报, 2002, 21(3): 269-272.

12 王 伟, 钟义信, 孙 建, 等.一种基于 EM 非监督训练的自组织分词歧义解决方案[J]. 中文信息学报, 2001, 15(2): 38-44

13 Yang Shuyuan, Wang Min, Jiao Licheng. A Genetic Algorithm Based on Quantum Chromosome[C]. Proceedings of the 7th International Conference on Signal Processing, 2004: 1622-1625.

(上接第 129 页)

比较图 4(a)和图 4(b)可以看出：AEMS 算法的收敛情况比 AMS 稳定得多，受 p 的影响较小。而 AMS 算法极易受 p 的影响。因此，一般在 AMS 算法中 p 不能任意取值，理论上取 p=1/d 时算法效果最佳。而 AEMS 算法由于受 p 的影响很小，可以定期更换路由器的初始标记概率，以避免 AMS 中某些恶意攻击者通过猜测 p 伪造标记包的情况发生。

以上仿真实验的结果较好地验证了第 4 节关于算法稳定性的分析结论。

6 结论

本文提出一种基于反向确认的攻击源追踪模型。该模型继承了 AMS 算法的优点，不需要 AMS 算法所要求的预先具备上游拓扑数据的强假设前提，而是借助上游路由器的力量完成攻击路径的确认。此外，本文提出自适应的边标记算法不仅避免了以前自适应算法的不足，而且通过理论分析和仿真实验证明该算法比 AMS 算法更稳定。

参考文献

1 Stefan S, David W, Anna K, et al. Practical Network Support for IP Traceback[C]. Proc. of ACM SIGCOMM, Sweden, 2000.

2 Song D X, Perrig A. Advanced and Authenticated Marking Schemes for IP Traceback[C]. Proc. of IEEE INFOCOM, USA, 2001.

3 夏春和, 石昀平, 赵沁平. 一种新的攻击源定位算法 NA[J]. 计算机研究与发展, 2004, 41(4): 689-696.

4 Stoica I, Zhang H. Providing Guaranteed Services without Per Flow Management[C]. Proc. of ACM SIGCOMM, USA, 1999.

5 梁 丰, 赵新建, David Y. 通过自适应随机数据包标记实现实时 IP 回溯[J]. 软件学报, 2003, 14(5): 1005-1010.

6 Tao Peng, Leckie C, Kotagiri R. Adjusted Probabilistic Packet Marking for IP Traceback[C]. Proc. of the 2nd IFIP Networking Conference, Italy, 2002: 697-708.

7 李德全, 徐一丁, 苏璞睿, 等. IP 追踪中的自适应包标记[J]. 电子学报, 2004, 32(8): 1334-1337.