Computer Science and Application 计算机科学与应用, 2011, 1, 128-133
http://dx.doi.org/10.12677/csa.2011.13026 Published Online December 2011 (http://www.hanspub.org/journal/csa)

A Weighted Set-Based Implicit Feedback Algorithm for Search Engine

Hui Zhang*, Yan Chen
State Key Lab. of Software Development Environment, Beihang University, Beijing
Email: {hzhang, chy}@nlsde.buaa.edu.cn
Received: Oct. 7th, 2011; revised: Oct. 29th, 2011; accepted: Nov. 19th, 2011.

Abstract: With the quick development of the Internet, online information recourses are becoming richer and richer. However, traditional search engines are becoming hard to acquire the users' retrieval intention because of the short keywords user inputted, so it is difficult to satisfy the needs of the user. In this paper, we proposed a Weighted Set-based implicit feedback algorithm to improve accuracy of the search engines effectively. Through analyzing the characteristics of the snippets of Web pages search engine returned, we present a kind of Weighted-Set for representing Web page snippets and a method for calculating the weight of elements in this Set. In addition, we design an approach to calculate the intersection between any two Weighted-Sets, which can obtain users' implicit search intention automatically, so that we can improve search accuracy in the way of query expansion. Query expansion experiments on the popular Google search engine show that our algorithm can improve search accuracy effectively.

Keywords: Weighted-Set; Search Engine; Implicit Feedback; Intersection Operation

一种基于带权集合的搜索引擎隐式反馈算法

张　辉*，陈　岩
北京航空航天大学软件开发环境国家重点实验室，北京
Email: {hzhang, chy}@nlsde.buaa.edu.cn
收稿日期：2011 年 10 月 7 日；修回日期：2011 年 10 月 29 日；录用日期：2011 年 11 月 19 日

摘　要：随着 Internet 的迅速发展，网络信息资源开始爆炸式增长。传统的搜索引擎很难从用户输入的检索词中获知其检索意图，只能返回大量匹配结果供用户选择。为了有效的提高搜索引擎的查准率，本文提出了一种基于带权集合的隐式反馈算法。本文通过分析搜索引擎返回结果页面的特点，提出了一种描述网页摘要的带权集合以及相应元素的权重计算方法，并设计了一种带权集合的交集运算方法，通过该运算可以获取用户隐含的检索意图，最后以查询扩展的方式提高搜索引擎的准确性。本文在 Google 搜索引擎上做了本算法的若干实验，验证了本算法的有效性。

关键词：带权集合；搜索引擎；隐式反馈；交集运算

1. 引言

随着互联网上 Web 信息的激增，搜索引擎已成为人们在信息海洋中获取有效信息不可或缺的工具。然而，搜索引擎并未尽如人意，比如当用户输入“PC 价格”时，搜索引擎 Google 会返回“PC 塑料价格”、“PC 电脑价格”等语义截然不同的结果，而且结果总数为 105,000,000 之多[1]，使得用户还需花费许多时间筛选他真正需要的东西，有时也会因找不到理想的结果而感到失望。有研究表明大约有 50%的检索无法得到理想的结果[2]。造成这个问题的原因有两个方面：一方面是人们输入的检索词一般很短，如 Web 检索词的平均个数为 2.4，近一半的用户在检索时只输入 1 个检索词[3]，寥寥几个检索词往往词不达意，不能准确表达用户意图；另一方面，搜索引擎无法从仅有的几个检索词中获知用户真实需求，只能采用关键词匹

配的方式返回大量结果供用户选择[4,5]。

如果搜索引擎具有一定智能并能自动识别用户的检索意图，那么只需把最符合用户意图的检索结果返回即可，从而大大提高检索效率。如何使得搜索引擎能够自动感知用户的搜索意图呢？有关研究人员提出了一种“相关反馈”(Relevance Feedback)方法，并证明了其在信息检索中的有效性与准确性[6-9]。其设计思想是，在信息检索过程中通过用户交互来提高最终的检索效果[9]。相关反馈需要用户明确指出哪些文档是符合本人要求的相关文档，但实际上很少有用户愿意做这种额外的操作[1]。“隐式反馈”已成为当前相关研究的热点[1,9,10]，其主要思想是记录用户在检索信息时的交互行为(如输入的关键词、点击的网页、停留的时间等)，通过分析从中自动获取隐含的用户检索意图。

本文分析了搜索引擎返回结果页面的特点，提出了一种描述网页摘要的带权集合以及相应元素的权重计算方法，并设计了一种带权集合的交集运算方法，通过该运算可以获取用户隐含的检索意图，最后以查询扩展的方式提高搜索引擎的准确性。

本文其余章节安排如下：第 2 章讨论该领域的相关研究；第 3 章详细介绍该算法模型的建立和工作流程；第 4 章对算法的实验结果进行分析；第 5 章算法总结。

2. 相关研究

相关反馈的一个经典算法为 Rocchio 算法[8]，其基本思想可以用下面公式表示

$$\boldsymbol{q}_{opt}=\arg\max_{\boldsymbol{q}}\left[sim\left(\boldsymbol{q},\boldsymbol{C}_r\right)-sim\left(\boldsymbol{q},C_{nr}\right)\right] \quad (1)$$

即最优反馈向量 $\boldsymbol{q}_{opt}$ 与相关结果向量 $\boldsymbol{C}_r$ 相似度最大，与不相关结果向量相似度 C_{nr} 最小。文献[11]是 Rocchio 算法的一个实例，主要基于反馈信息中的文档与初次检索结果中文档之间的距离，利用相关文档与不相关文档作为反馈信息来提高搜索引擎查准率。文献[12]给出了一种利用海量用户点击行为的记录进行网页内容相关性挖掘的方法，在此基础上给出了一种反馈式搜索引擎框架及相关算法；文献[13]以 Web2.0 中用户行为作为研究对象，通过发掘用户反馈方式，提出了用户反馈分值的概念，并通过用户反馈影响搜索结果排名的具体方式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。

以上文章提出的算法均是在具有大量用户访问行为基础上的学习算法，通过提取反馈信息实现对搜索结果排序的优化。本文算法以单个用户在浏览检索结果时的点击行为作为反馈信息，在对反馈信息的描述上，并没有选取当前流行的向量空间模型 VSM[14]，而是采用集合论的集合概念；在对反馈信息的处理上，采用集合交集运算提取不同网页摘要的共性信息，把共性信息作为用户的反馈，这样对提高反馈算法的时效性及准确性起到了重要作用。

3. 算法模型的分析与建立

3.1. 设计思想

3.1.1. 设计分析

人的认知模式表明：用户判断一条信息的相关性比清晰地表达其需求更容易一些[9]。即用户不一定能清楚地表达出所他需要的信息是什么，但是能够容易地判别一条信息是否满足其要求。下面看一个场景：

用户 A 想要了解电脑价格信息，于是他在 Google 搜索引擎上输入关键词“电脑价格”进行查询，搜索引擎返回大量结果项，如图 1(摘取其中两条)所示：

如图 1 所示，搜索引擎返回的结果项一般包含两部分：网页标题与网页摘要。用户在浏览搜索结果项时，往往倾向于查看排名靠前的网页[15]，也就是说如果用户 A 浏览了排名靠前的网页标题和摘要，但最终点击查看了排名靠后的网页，这正好说明排名靠前的网页不是他的最佳选择，而被点击查看的结果项反而能表达 A 的真实意图。

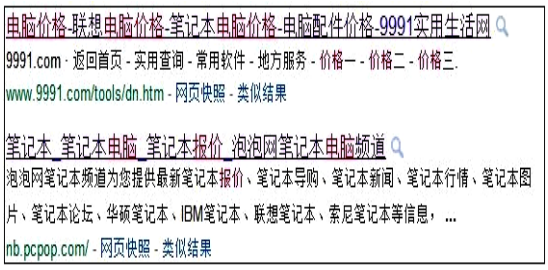

Figure 1. Search engine result items
图 1. 搜索引擎结果项(网页标题与网页摘要)

假定在 Google 返回的前十条结果项中，有 5 条被用户 A 点击浏览，那么这 5 条结果项的共性信息(共同出现的词语)恰恰体现了用户 A 的检索意图。一般情况下，为了得到更多相关信息，用户 A 根据自己的需求不断修正输入的关键词信息，如：“电脑报价”、“电脑厂商报价”等，经过多次的修正查询，终于得到了自己期望的信息。

通过对该场景的分析及对人们浏览行为习惯的总结，可以得到如下结论：

1) 用户浏览检索结果项时对标题的关注度要高于对摘要的关注度，所以查询结果项的标题信息更能体现用户的关注点；

2) 用户点击顺序与查询结果项中所包含的信息重要性成反比[16]，即：被用户查看的查询结果项越靠后，其中包含的信息越可能是用户所需要的东西；

3) 通过对每次被点击的检索结果项做集合交集运算，可以快速提取交集中的共性信息，这些共性信息即为用户的意图，也是对搜索引擎的反馈信息。

3.1.2. 基本假设

为了保证算法的可行性，根据以上得出的结论，在建立本算法模型前，做出如下四个假设：

1) 搜索引擎返回的搜索结果项和检索词具有一定的相关性，并且一次检索能返回多个搜索结果项；

2) 搜索结果项与检索词的相关度大小已进行排序，即越相关的结果项排序越靠前；

3) 搜索结果项包括网页标题和网页摘要两部分信息(如图 1 所示)，这两部分信息能反映该网页的关键内容，并且标题信息的权重要大于摘要信息的权重；

4) 用户点击浏览搜索结果项的行为是用户的有意选择，被选中浏览的结果项内容与用户的检索意图是相符合的，能够体现用户的需求。

在以上四个假设中，实际上搜索引擎都能够满足前三个假设，第四个假设也是符合人们在信息检索时的行为习惯的。

3.1.3. 设计思想

根据 3.1.1.中的分析与 3.1.2.中的假设，针对 Web 搜索引擎，本文设计实现了一种基于带权集合运算的隐式反馈算法，其基本思想为：

用户在浏览搜索引擎返回的结果项时，对自己感兴趣的结果项往往会点击查看具体的网页。因此，该算法可以记录用户的点击行为，并为每个结果项提取关键词并生成对应的集合，考虑被用户点击的结果项排名、关键词词频、关键词所处位置(标题还是摘要)等因素，为每个关键词生成相应的权重，从而为每个搜索结果项生成对应的带权集合。同时，使用带权集合的相交运算从多次点击的结果项中抽取共性信息，然后选择权重较大的共性信息作为新的关键词进行扩展查询，从而缩小搜索结果的范围。如此往复，直到满足用户的检索需求为止。

3.2. 模型符号系统的定义和说明

这里首先定义算法中用到的两个概念，并定义其运算规则。

定义 1 关键词三元组 $K\_G=(k,w,c)$，表示带有权值信息的关键词元组，其中 k 表示关键词，w 表示关键词的权重，c 表示关键词被用户浏览的次数。

设 A, B 分别为任意关键词三元组，a, b 为任意常数，用 $A\to k$ 表示 A 的关键词；$A\to w$ 表示 A 的关键词权重；$A\to c$ 表示 A 的关键词被用户浏览的次数，则有：

$$aA=\left(A\to k, a\left(A\to w\right), a\left(A\to c\right)\right) \tag{2}$$

即：对关键词三元组 A 乘系数 a，三元组中“关键词”项不变，“权重”与“浏览次数”乘相应系数 a。

$$aA+bB=\begin{cases}\left(A\to k, a\left(A\to w\right)+b\left(B\to w\right), a\left(A\to c\right)+b\left(B\to c\right)\right), & \text{若 } A\to k=B\to k\\ \left(\varnothing,0,0\right), & \text{若 } A\to k\neq B\to k\end{cases} \tag{3}$$

即：关键词三元组 A 与 B 做“+”运算，若 A 与 B“关键词”项相同，分别对其“权重”与“浏览次数”项做代数相加元素；否则，结果为空。

推论 1 相等关系：$A=B$当且仅当 $A\to k=B\to k$，$A\to w=B\to w$ 且 $A\to c=B\to c$

推论 2 等价关系：$A\cong B$ 当且仅当 $A\to k=B\to k$

例 1 设关键词三元组：A = (电脑, 0.1, 1), B = (电脑, 0.2, 2), C = (价格, 0.3, 3）。

不妨假设常数: $a=2, b=3$;

则：$aA=$(电脑, 0.2, 2);

$aA+bB=$(电脑, 0.8, 8);

$aA+bC=(\varnothing, 0, 0)$;

定义 2 带权集合：表示带有权重信息的关键词集合，本文指集合元素为关键词三元组的集合。

带权集合元素要满足互异性，互异性指集合中任意元素两两之间均不满足关键词三元组等价关系。

如果 M, N 分别为任意带权集合，带权集合 M, N 的相交运算可表示为 $M+N$，对 $\forall m \in M$， $\forall n \in N$，若 $m \cong n$，则 $m \notin M+N$，$n \notin M+N$，$m+n \in M+N$；否则 $m \in M+N$， $n \in M+N$。

例 2 设关键词三元组：$A=$(电脑, 0.1, 1), $B=$(电脑, 0.2, 2), $C=$(价格, 0.3, 3)。

则：带权集合 $R1=\{A, C\}, R2=\{B, C\}$

$R1$ 与 $R2$ 交集(带权集合的加运算)可表示为：

$R1+R2=\{A+B, C\}=$ {(电脑, 0.3, 3), (价格, 0.3, 3)}。

3.3. 模型的建立

假设用户经过了 m 次的关键词搜索得到了满意的结果，针对第 i 次$(1 \le i \le m)$查询操作，有：

1) 向搜索引擎发送的关键词集合为：

$$\text{Keyword}\_i=\{k1\_i, k2\_i, k3\_i \cdots kn1\_i\} \quad (4)$$

其中 $n1\_i$ 表示第 i 次查询的关键词数；

2) 用户点击的结果项集合为：

$$\text{selectedResult}\_i=\{r1\_i, r2\_i, r3\_i \cdots rn2\_i\} \quad (5)$$

其中 $n2\_i$ 表示第 i 次查询用户点击浏览的结果数；

若 $\forall rj\_i \in \text{selectedResult}\_i \ \ (1 \le j\_i \le n2\_i)$，取对应结果项的带权集合 WeightedSet_i_j，集合中的每个关键词三元组对应着结果项中的一个关键词，对 $\forall A \in \text{WeightedSet}\_i\_j$，则：

$A \to k=$ 关键词名称；

$A \to c=1$；

$$A \to w=\left(\alpha \cdot \log_2 \text{Rank}\_i\_j+\beta\right) \cdot\left(W_{\text{sub}} \cdot F_{\text{sub}}\left(A \to k\right)+W_{\text{sum}} \cdot F_{\text{sum}}\left(A \to k\right)\right) \quad (6)$$

其中：Rank_i_j 为当前结果项的搜索结果排序，α、β 为调节因子，保证权重与排序成反比，并使排序对最终权重的影响控制在一定的范围内。一般 $1 \le \text{Rank}\_i\_j \le 100$，故可取 $\alpha \in(-0.1,0)$，$\beta \in[0,1]$；

W_{sub} 为结果项中网页标题部分所占权重，W_{sum} 为结果项中网页摘要部分所占权重，且有 $W_{\text{sub}}+W_{\text{sum}}=1$，根据经验值一般 $W_{\text{sub}} \in[0.6,0.8]$；$F_{\text{sub}}\left(A \to k\right)$表示 A 中关键词在标题中出现的频数，$F_{\text{sum}}\left(A \to k\right)$表示 A 中关键词在摘要中出现的频数；这样，用户第 i 次查询的关键词带权集合计算公式为：

$$\text{WeightedSet}\_i=\sum_{j=1}^{n2\_i} \text{WeightedSet}\_i\_j \quad (7)$$

3) 选择 WeightedSet_i 中用户点击次数较多，权重较高的关键词作为新的关键词集合 Keyword_i+1。

3.4. 算法流程

该算法具体的算法流程如下：

步骤一 用户通过关键词集合 $K=\{k_1, k_2, k_3 \cdots k_j \cdots\}$ 在搜索引擎中检索信息，其中 k_j 表示用户输入的第 j 个关键词；搜索引擎返回结果项集合 $R=\{r_1, r_2, r_3 \cdots r_i \cdots\}$，其中 $r_i\left(k_{i1}, k_{i2} \cdots k_{ij} \cdots\right)$ 表示搜索引擎返回的第 I 条结果项的特征词(图 2)，K_{ij} 为 r_i 的第 j 个特征词。

步骤二 假如用户查看了结果项集 R 中的 m 条结果项，对每条结果项进行关键词提取，获取 m 个关键词集合：$K_1, K_2 \cdots K_m$；

步骤三 通过对 $K_1, K_2 \cdots K_m$ 集合的带权集合交集运算，获取新的关键词集合 K'(图 3)，即用户选择的 m 条结果项的共性信息，同时计算出集合中关键词的权重。

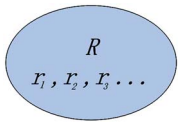

Figure 2. The set of search engine result items
图 2. 检索结果项集合

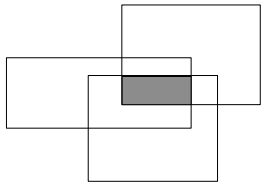

Figure 3. Intersection operation model figure of keywords
图 3. 关键词交集运算模型图

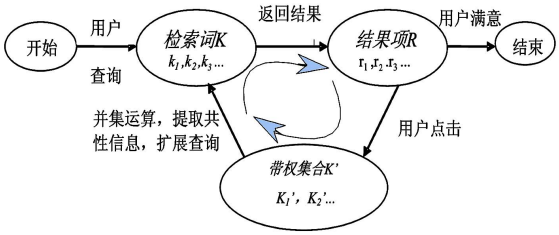

Figure 4. Algorithm process figure
图 4. 算法流程示意图

从步骤三中得到的关键词集合 K' 中选择权重较大的关键词组成新的关键词集合 K，跳转到步骤一重复执行，从而达到快速缩小检索范围的目的，直到满足用户的检索需求为止(图 4)。

4. 算法的检验与参数优化

4.1. 实验一

为了验证基于带权集合的隐式反馈算法的有效性，进行了如下实验：

实验一 选择 50 名不同用户进行测试，每位用户执行 20 次查询任务，查询内容按照他们各自的兴趣自由选择。这 50 名用户平均分成两组，A 组使用 Google 搜索引擎进行查询，B 组使用建立在 Google 搜索引擎基础上的本算法进行查询。

在本次实验中，统计了每位用户在每次查询操作中需要浏览的结果项数、平均每页(每页 10 条)查询结果项的满意数(表 1)以及前 5 页每页结果项的平均满意数(图 5)：

如表 1、图 5 所示，可以看到，基于带权集合的隐式反馈算法显著的提高了查准率，并且减少了用户的浏览次数。

4.2. 实验二

在模型的分析与建立过程中，影响关键词权重的因子主要有三个，即：公式(3)中的 α，β以及 W_{sub}，为了选择合适的影响因子，最大限度的提高搜索引擎的查准率，进行了如下实验：

实验二 选择 50 名用户进行跟踪测试，每位用户执行 20 次查询任务，查询内容按照他们各自的兴趣自由选择。这 50 名平均分成五组(A、B、C、D、E)，每组反馈模型中 α，β，W_{sub} 取值如表 2 所示：

通过本次实验，统计了每位用户每页查询结果的满意数，如表 3、图 6 所示。

Table 1. Experiment 1 results
表 1. 实验一结果

分组	用户达到满意的点击浏览总计次数	每页结果项的满意数
A 组	14	5.8
B 组	11	7.0

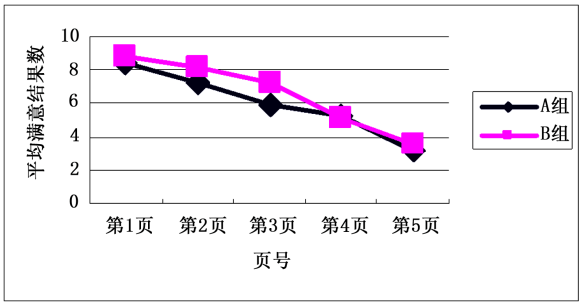

Figure 5. The average number of satisfaction per page of top 5 pages
图 5. 前 5 页每页结果项的平均满意数

Table 2. The value of α, β, W_{sub} of each group in experiment 2
表 2. 实验二每组 α, β, W_{sub} 取值情况

分组	α	β	W_{sub}
A 组	–0.05	1	0.6
B 组	–0.05	1	0.65
C 组	–0.05	1	0.7
D 组	–0.05	1	0.75
E 组	–0.05	1	0.8

Table 3. Experiment 2 results
表 3. 实验二的结果

分组	每页结果的平均满意数
A 组	7.3
B 组	7.6
C 组	7.1
D 组	6.8
E 组	6.8

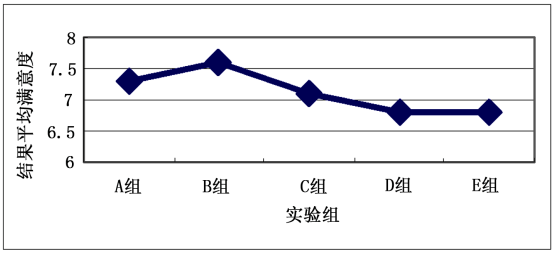

Figure 6. The average number of satisfaction of each group in experiment 2
图 6. 实验二各组每页结果的平均满意数示意图

当W_{sub}不变，选取不同的α，β取值进行比较实验时，结果变化并不明显。

根据实验二，发现选取一定的W_{sub}的取值可以提高算法的查准率；当$W_{sub}=0.65$时，查准率最高；α，β在一定范围内的取值对算法的查准率影响不大，故实际操作中可以取$\alpha=-0.05$，$\beta=1$。

本算法的时间花费主要在共性信息的提取上，即结果项集合的交集运算，假设用户在一次查询中点击了 m 条结果项信息，每条信息的平均特征词为 n 个，对其做交集运算，时间复杂度为 $O(mn)$，由于用户一次查询中点击浏览的结果项 m 条与其特征词数 n 均较小，所以本算法在一次查询中消耗的时间是很小。

5. 总结及未来工作

本文提出了一种基于带权集合的隐式反馈算法，该算法通过实时记录用户的点击行为，通过带权集合交集运算获取用户的查询意图，并自动进行扩展查询，直到用户满意为止，达到了快速提高搜索引擎查准率的目的，给用户带来了便利。

5.1. 算法的创新

本文在处理从搜索引擎结果项中获取共性信息时，提出了一种带权特征词集合的生成算法以及带权集合的相交算法，利用带权集合的相交运算进行快速提取用户隐含的检索意图。此外，本文在估计特征词权重时，将贝叶斯条件概率和最大似然估计有机结合起来，并综合考虑了特征词位置、关键词频率、结果项排名的影响因素，从而提高了搜索引擎的查准率。

5.2. 下一步工作

1) 在对W_{sub}参数优化中，通过参与实验的几组数据进行离散型设计，尚有优化的空间，包括将离散数据连续化，取其极值，得到更优的参数；

2) 如何在进行搜索时有效结合近义词、同义词进行扩展查询，如何更有效地去除一些无意义的虚词等都需要进一步优化；

3) 用户浏览结果项时的滞留时间也是影响权重的重要因素，如果能充分考虑该因素，也能够提高反馈模型的精确度。

参考文献 (References)

[1] D. Kelly, J. Teevan. Implicit feedback for inferring user preference: A bibliography. SIGIR Forum, 2003, 37(2): 18-28.

[2] T. Joachims. Optimizing search engines using clickthrough data. In Proceedings of SIGKDD 2002, 2002: 133-142.

[3] X. H. Shen, B. Tan and C. X. Zhai. Implicit user modeling for personalized search. Bremen: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005.

[4] 周博, 岑荣伟等. 一种基于文档相似度的检索结果重排序方法[J]. 中文信息报, 2010, 25(3): 19-23.

[5] 侯越先, 张鹏, 于瑞国等. 基于内容相关性挖掘的反馈式搜索框架[J]. 天津大学学报, 2008, 41(8): 941-945.

[6] Z.-L. Wu, C.-H. Li. Topic detection in on-line discussion using non-negative matrix factorization. Silicon Valley: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2007: 272-275.

[7] A. Spink, D. Wolfram, M. B. J. Jansen, T. Saracevic. Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology, 2001, 52 (3): 226-234.

[8] G. Salton, C. Buckley. Improving retrieval performance by retrieval feedback. Journal of the American Society for Information Science, 1990, 41(4): 288-297.

[9] C. D. Manning, P. Raghavan and H. Schutze. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[10] J. J. Rocchio. Relevance feedback in information retrieval. The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall Inc., 1971: 313-323.

[11] J. Allan, C. Wade and A. Bolivar. Retrieval and novelty detection at the sentence level. Toronto: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003: 314-321.

[12] 金祖旭, 李敏波. 基于用户反馈的搜索引擎排名算法[J]. 计算机系统应用, 2010, 19(11): 60-65.

[13] T. Moon, L. H. Li, W. Chu, C. Y. Liao, Z. H. Zheng and Y. Chang. Online learning for recency search ranking using real-time user feedback. ACM, New York, 2010.

[14] G. Salton, A. Wong and C. S. Yang. A vector space model for information retrieval. Communications of the ACM, 1975, 18(11): 613-620.

[15] E. Agichtein, E. Brill and S. Dumais. Improving web search ranking by incorporating user behavior information. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006.

[16] T. Joachims, L. Granka, B. Pan, H. Hembrooke and G. Gay. Accurately interpreting clickthrough data as implicit feedback. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005: 154.

