用户行为聚类的搜索引擎算法与实现①

蔡　岳　袁津生 (北京林业大学 信息学院 北京 100083)

摘　要： 提出一种基于用户行为聚类的搜索引擎算法。该算法从用户行为日志中挖掘用户意图，并根据用户的反馈信息定位用户意图信息，提升了查询的准确率，有效地解决了传统的全文检索式搜索引擎查询具有二义性词时的不足，并通过实践验证了算法的可行性。

关键词： 用户行为；用户意图；聚类；搜索引擎；二义性

User Activities Clustering of Search Engine

CAI Yue, YUAN Jin-Sheng

(Department of Information, Beijing Forestry University, Beijing 100083, China)

Abstract: This paper proposes an algorithm and an instance of the search engine based on clustering users' activities. The algorithm mines users' concepts from activities log and uses users' feedback to identify the concepts. Both of them improve the accuracy of the engine. It resolves the ambiguity problem of full-text retrieval search engine. At last, the paper gives a case to implement the algorithm.

Keywords: user activity; clustering; search engine; concept; ambiguity

1　引言

随着网络的不断发展，网络中所承载的信息量以惊人的速度递增，传统的搜索引擎带给用户的体验越来越差，例如百度、谷歌等，主要是通过对已抓取网页的全文进行分词，并以此进行索引。当用户输入关键字来进行搜索的时候，搜索引擎会在索引中查找包含用户输入关键字的网页，并按照一定的排序规则对其进行排序，按照优先级由高到低的顺序返回给用户。这类搜索引擎存在的一个重要问题是缺乏对二义性的词的分辨，例如“苹果”，搜索引擎很难知道用户是想获取关于苹果公司的信息还是一种水果的信息。而基于用户意图的搜索引擎恰恰弥补了这一点。

本文提出了一种基于用户行为聚类的搜索引擎算法，并根据 Sogou 搜索引擎(www.sogou.com)提供的 2006 年 8 月 1 日的用户行为日志来测试算法，验证算法的可行性。

2　相关工作

用户意图需要反映出用户的实际需求，即对某一类信息的兴趣。在本算法中，用户意图是由若干词组成，举例来说，如果用户对苹果公司感兴趣，那么依据本算法，用户意图最终可能表示为“苹果，mac，ipod”。并不是所有的词都会被用来作为组成用户意图的词，要确定哪些词会用来表示用户意图，需要根据用户输入的查询词来确定。

定义 1. 一组可以表示用户对某方面感兴趣的词的集合，叫做用户意图。

尽管基于用户意图的搜素引擎出现较晚，但在国内外已经成为了热点课题。其中一个很重要的问题是如何获得用户意图。

在张玉连等[1]的算法中，利用鼠标钩子来获得用户意图，并在其中考虑到各种用户行为，例如保存网页、浏览时间等等。这种方法的优点是利用各种用户行为来提取用户意图，能更精确地提取意图，并保证查询的准确率；缺点是需要用户使用指定的程序，在实际应用中很难要求用户安装特定的浏览器。

在肖卓程[2]等的算法中，需要用户的注册后才能使用搜索引擎，在注册的时候需要提供一些个人信息。

① 收稿时间:2009-08-11;收到修改稿时间:2009-09-16

这种方法的好处是在多数情况下，用户的个人信息能够反映用户的兴趣点，这有利于提取用户意图；但问题是，这与现在流行的商业搜索引擎相比，使用过程有些繁琐，而且用户个人信息属于隐私，这种收集用户隐私的行为对于用户来说是不安全的。

在 Kenneth Wai-Ting Leung[3]等的算法中提到，根据网页中所包含重要词的位置来计算网页相似度，据此来计算用户的兴趣点，从而提取用户意图。这种算法不要求改变用户使用流行搜索引擎已养成的习惯，查询准确率很高；缺点是算法复杂度过高，距离实践尚有一定的差距。

本文提出了一种基于用户行为聚类的搜索引擎算法，该算法在准确率方面与之前提到的算法的准确率有所下降，但具有更好的实用性，可应用到实践中。

3 算法描述

算法中涉及到几个关键的部分：用户意图计算、相似度计算和聚类算法，第 3.1 节至第 3.3 节分别阐述了这三部分的计算方法，在此之后，第 3.4 节描述了整个算法的流程。

3.1 用户意图计算

在 Kenneth Wai-Ting Leung 等的算法中，用于表示用户意图的词，其支持度要大于某个特定的值。词 t_i 的支持度由以下公式计算：

$$s(t_i)=\frac{f(t_i)}{n}\cdot|t_i| \tag{1}$$

其中，n 为网页的总数；$f(t_i)$ 为词 t_i 的频率，这里的频率为词 t_i 出现的网页的数量；$|t_i|$ 是词 $|t_i|$ 的长度。一般来说，支持度应大于 0.03。Kenneth Wai-Ting Leung 等的算法需要搜索引擎收录大量的网页，再从中提取用户意图词，这种算法的计算量比较大，并且，当用户的意图可以用一些出现频率高的词来表示的时候，这种方法比较准确。但当用户的意图是一些词频较低的词时，这种算法会导致低词频的词不会出现在用户意图词的集合中，也就无法正确的表示用户意图。

一般来说，用户输入的查询词，是用户认为应当大量出现在其关注信息中的词，或者与关注信息极为密切的词，而这些词，其支持度必然会比较高。在本文中提到的用户意图词的提取算法中，将根据用户输入的查询词，在用户点击的网页中查找大于等于查询词词频的词，作为表示用户意图的词。

$$s(t_i)=\frac{f(t_i)}{n}\cdot\log\frac{f(t_i)}{f(t_q)} \tag{2}$$

其中，$f(t_i)$ 为词 t_i 的频率，$f(t_q)$ 为查询词 q 在网页中的频率。当 $s(t_i)\geq 0$ 时，词 t_i 作为用户意图词。

3.2 相似度计算

本文采用的相似度计算公式来自 Chan[4]对 Beeferman 相似度公式的改进公式。Beeferman 提出了一种与内容无关的聚类算法。该算法的计算依赖于图 1 所示的偶图。

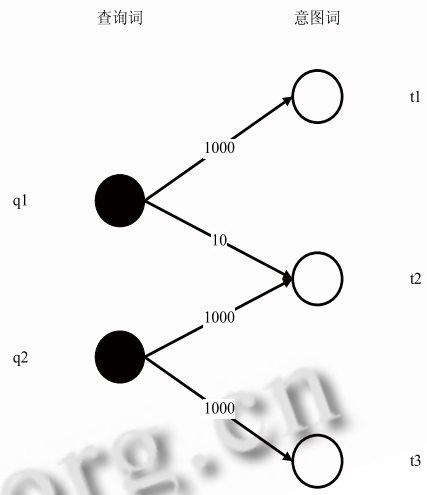

图 1 查询-意图偶图

其中左侧的结点为用户输入的查询词，右侧为用来表示意图的词。如果用户在输入某个查询词 q1，并选取了包含 t1 的网页，则在 q1 和 t1 之间建立一条连接，连接数为 1；若 q1 和 t1 之间已存在连接，则连接数加 1。

为了体现个性化搜索，图中左侧的查询结点需要按用户来进行区分，假设两个用户 A 和 B 输入同一个查询词“苹果”，用户 A 希望了解苹果公司的信息，用户 B 希望了解水果方面的知识。此时，用户 A 和用户 B 虽然输入的查询词相同，但由于意图不同，应考虑为两个不同的查询词。查询-意图偶图则变为图 2 所示的用户-查询-意图偶图。

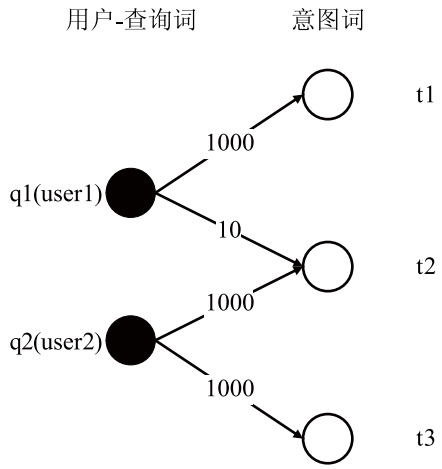

图 2　用户-查询-意图偶图

词 t_i 和词 t_j 的相似度由以下公式计算：

$$sim_L(t_i,t_j)=\begin{cases}\dfrac{\left|L(t_i t_j)\right|}{\left|L(t_i)\cup L(t_j)\right|} & \left|L(t_i)\cup L(t_j)\right|>0 \\ 0 & otherwise\end{cases} \quad (3)$$

其中 $\left|L(t_i t_j)\right|$ 表示词 t_i 和词 t_j 同时连接同一结点的连接数，$L(t_i)$ 表示词 t_i 的所有的连接数，$L(t_j)$ 表示词 t_j 的所有的连接数。$\left|L(t_i)\cup L(t_j)\right|$ 表示词 t_i 和词 t_j 所包含的全部的连接数。

3.3 聚类算法

假设存在偶图 (Q,T)，其中 Q 为用户输入的查询词 q_i 的集合，T 为用户选取的查询结果所包含的意图词 t_i 的集合。Beeferman 聚类过程如下：

① 获得用户-查询-意图偶图 (Q,T)

② 根据公式(3)，合并相似度最大的两个查询结点 q_i 和 q_j

③ 根据公式(3)，合并相似度最大的两个结果结点 t_i 和 t_j

④ 除非达到终止条件，否则重复步骤②和③

其中，终止条件需要根据实际的聚类情况来定。终止条件的设置通常是指定一个最小相似度，即当全部结点间的相似度均小于最小相似度时，停止聚类，以防止过度聚类。本文中设最小相似度为 0.5。

3.4 算法流程实现

本文使用的数据来自 Sogou 搜索引擎提供的 2008 年 6 月 1 日的用户行为日志。日志的格式为：

用户 ID\t[查询词]\t 排名\t 选中的 URL

本算法将以此来构建用户-查询-意图偶图，利用 Beeferman 改进算法进行聚类，最重获得用户意图表。算法流程如下所示：

① 初始化一个空的用户-查询-意图偶图。

② 根据查询词，在选中的 URL 所对应的网页中，利用公式(2)，提取用户意图词。

③ 若存在记录 $user_1-q_1-url$，则在偶图中插入左结点 $user_1-q_1$，在偶图中插入若干右结点 t_i，其中 t_i 为选中的 url 所对应的网页中包含的用户意图词。

④ 重复②和③，直至日志中的全部记录遍历完毕。

⑤ 根据公式(3)进行聚类，获得聚类后的用户-查询-意图表。

⑥ 对聚类后的用户-查询-意图表中的左侧结点重命名，删除原名中的用户名信息，获得用户意图表。

获得用户意图表后，搜索引擎在接收到用户输入的查询词时，将按照以下流程进行查询服务。

① 接收用户提交的查询词 q。

② 在用户意图表中搜索包含 q 的左结点。

③ 若存在多个包含 q 的左结点，跳转至④，否则跳转至⑤。

④ 将各个左结点的名称反馈给用户，并提示信息，供用户选择。

⑤ 优先返回包含用户意图的查询结果。

4　实验结果

本节针对上述提出的搜索引擎算法和实现，利用 Sogou 引擎的数据[5,6]来验证其实际的查询结果的准确性。实验将选取一组特定的查询词，这些查询词将分别使用百度、谷歌和上述搜索引擎来进行查询，查询结果选取前 100 条记录。实验中选取的查询词如下所示。

表 1　测试查询词

查询词	查询目的
微软	微软公司及其相关产品
苹果	苹果公司及其产品
	水果信息
功夫	关于电影功夫的相关信息
搜狐	关于搜狐 CEO 张朝阳的相关信息

下图所示为三个搜索引擎的查询准确率的比较，其中横轴代表查询词，纵轴代表准确率。

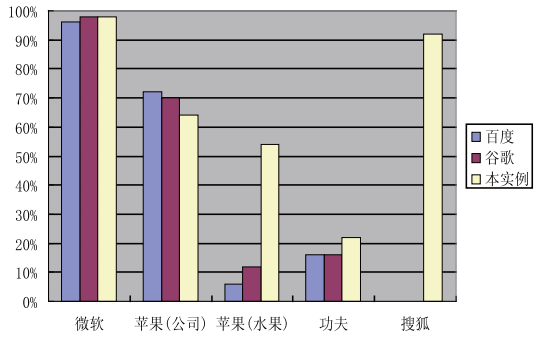

图 3 查询准确率图

实验选取这个五个词是有特定意图的。首先，“ 微软 ”这个词不存在二义性，用来检验搜索引擎对意义唯一的词查询准确度；“ 苹果 ”和“ 功夫 ”都是具有二义性的词，用来检验搜索引擎对二义性词的准确度；“ 搜狐 ”这个词是用来检测一种特殊情况，当用户输入的查询词可能不包含于用户的查询意图的情况。表 2 的实验结果表明，以上这些词中，包含二义性的词的查询准确率有所提高，“ 苹果 ”这个词尤为明显。而“ 功夫 ”之所以没有太多的提高，主要由于电影《功夫》上映时间较早，而数据是来自 2006 年，因此搜索“ 功夫 ”的用户并不多，导致准确率不高。第五个查询词“ 搜狐 ”的查询结果要好于百度和谷歌的查询结果，这主要由于反馈选择步骤，将用户意图定位在不包含“ 搜狐 ”二字的结果中了。

5 结语

本文提出的基于用户意图聚类的搜索引擎算法是可行的，并且与大多数的聚类算法相比，其计算量较小，能符合实际应用的需求，并且与全文检索式的搜索引擎相比，在准确率方面有了一定得提升。

参考文献

1 张玉连,李彦威,王权,原福永.搜索引擎查询日志的聚类.计算机工程, 2009,35(1):43－45.

2 肖卓程,荆金华.基于用户兴趣的搜索引擎.计算机应用与软件, 2007,24(9):134－136.

3 Leung KWT, Ng W, Lee DL. Personalized Concept-Based Clustering of Search Engine Queries. IEEE Transactions on knowledge and data engineering, November 2008,20(11).

4 Church K, Gale W, Hanks P, Hindle D. Using Statistics in Lexical Analysis. Bell Laboratories and Oxford University Press.

5 陈红涛,杨放春,陈磊.基于大规模中文搜索引擎的搜索日志挖掘.计算机应用与研究, 2008,25(6):1663－1665.

6 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.