基于短期记忆与遗忘系数的用户个性化建模方法研究

陈海燕[1] 徐 峥[2] 张 辉[2]
(华东政法大学计算机科学与技术系 上海 201620)[1] (清华大学公共安全研究院 北京 100084)[2]

摘 要 搜索引擎的一个标准是不同的用户用相同的查询条件检索时,返回的结果相同。为解决准确性问题,个性化搜索引擎被提出,它可以根据用户的不同个性化特征提供不同的搜索结果。然而,现有的方法更注重用户的长时记忆和独立的用户日志文件,从而降低了个性化搜索的有效性。获取用户短时记忆模型来提供准确有效的用户偏好的个性化搜索方法被广泛采用。首先,根据基于查询关键词的相关概念生成短期记忆模型;接着,基于用户的时序有效点击数据生成用户个性化模型;最后,在用户会话中引入了遗忘因子来优化用户个性化模型。实验结果表明,所提出的方法可以较好地表达用户信息需求,较为准确地构建用户的个性化模型。

关键词 个性化搜索,用户模型,语义挖掘,用户语境
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.058

Building User Personalization Model Based on Short Term Memory and Forgetting Factor

CHEN Hai-yan[1] XU Zheng[2] ZHANG Hui[2]
(Department of Computer Science and Technology, East China University of Political Science and Law, Shanghai 201620, China)[1]
(Institute of Public Safety Research, Tsinghua University, Beijing 100084, China)[2]

Abstract One criticism of search engines is that when queries are submitted, the same results are returned to different users. To address the accuratcy problem, personalized search was proposed, since it could provide different search results based upon the preferences of users. However, the existing methods concentrate more on the long-term and independent user profile, thus reducing the effectiveness of personalized search. We introduced an approach that captures the user context to provide accurate preferences of users for effectively personalized search. First, the short-term query context is generated to identify related concepts of the query. Second, the user context is generated based on the click through data of users. Finally, a forgetting factor is introduced to merge the independent user context in a user session, which maintains the evolution of user preferences. The experimental results fully confirm that our approach can successfully represent user context according to individual user information needs.

Keywords Personalized search, User model, Semantic mining, User context

1 引言

随着互联网的发展,人们越来越依赖网络搜索引擎来满足各自的多样化信息需求。统计数据表明,平均每天每个Web用户进行4.28次搜索查询[1]。此外,文献[2]的研究也表明,85%用户使用搜索引擎来获取他们需要的信息。

搜索引擎尽管得到了广泛使用,但仍然存在一些挑战。准确性是最大的问题之一。搜索结果往往不能满足用户的要求,主要是因为不能准确地获取用户的信息需求。用户提交给搜索引擎的查询条件具有以下特点:

(1)查询词较短。为了减轻自身的认知负担,用户经常提交简短的术语来表达他/她的需要。文献[3]对一个流行的搜索引擎进行了研究,结果表明用户查询词的平均长度只有2.35个字。较短的查询信息限制了用户清晰地表达自己的信息需求。

(2)查询词有歧义。用户提交的查询条件往往有多重意义,例如"苹果"可能意味着水果或计算机,这使得搜索引擎返回一些与用户查询意图不相关的结果。

(3)查询词不完整的。有时用户对自己需要的信息并没有具体的概念[4],例如他/她对正在寻找的信息没有相关的背景知识,因此用户难以提交适当的查询词到搜索引擎。

即使提出了最完美的提高搜索引擎的准确性的方案,也是远远不够的,因为搜索引擎的一个标准是不同用户只要提交的查询条件相同,返回结果就相同,尽管这些用户可能有不同的信息需求。各种个性化搜索引擎技术被提出,但效果并不理想[5]。个性化搜索引擎的一个关键问题是如何获取和表达用户的偏好,这对搜索精度有很大的影响。因此,如何精确地描述用户的个性化模型是提高搜索引擎质量的一个重要挑战。

本文提出基于用户实时语境的方法来构建用户模型。用

到稿日期:2015-03-02 返修日期:2015-07-13 本文受国家社会科学基金项目(06BFX051),国家自然科学基金(6130202),上海高校选拔培养优秀青年教师科研专项基金(hzf05046)资助。
陈海燕(1978—),男,博士生,讲师,主要研究方向为计算机网络、数据挖掘、人工智能、信息安全,E-mail:tom_chy@163.com;徐 峥(1984—),男,博士,助理研究员,主要研究方向为公共安全、数据挖掘;张 辉(1969—),男,博士,教授,主要研究方向为公共安全。

户语境作为用户搜索的背景，可以完善和表达用户的实时搜索意图，并确保搜索引擎返回结果的准确性。用户上下文建立在查询文本的基础上，查询文本包含可以缩小用户搜索范围的信息。相应地，本文提出的方法可以分为下面两个步骤：

(1)构建查询语境。当用户提交查询条件时，搜索引擎会返回网络片段[1)]的列表[1]。通过从网络片段中挖掘关键字和它们之间的关系生成查询语境。

(2)构建用户语境。当用户点击搜索引擎返回的结果时，生成用户语境。它是通过提取的概念和更新这些概念再查询文本的权重来实现的。

在本文中，以查询语境为基础来构建用户语境。查询语境是通过用户提交的查询词生成的，它可以作为用户搜索行为的语义背景。与以前的工作不同，我们从网络片段中不仅提取概念，而且还提取它们之间的关系，保证了生成的用户语境能够更准确、更有效地表达用户的真实兴趣。

用户的每一次点击行为可以反映其实时语境。用户语境通过用户的点击来反映用户的兴趣。以往的研究往往需要长期的用户日志文件来完善用户的查询。然而，Shen 等[6]研究发现，用户的短期记忆模型更适合个性化的搜索引擎，因为用户经常搜索的短期信息需求与长期的兴趣不一致。在本文中，通过建立短期语境来完善用户语境。

本文根据用户的点击流来建立用户个性化模型，主要贡献是用户语境根据用户的点击行为，可以实时地表达用户的兴趣。遗忘因子被引入，其是根据用户的点击顺序来构建用户的实时语境。

本文第 2 节介绍个性化搜索的相关工作；第 3 节介绍如何构建查询语境；第 4 节介绍建立用户语境的方法；第 5 节是实验评估和分析的结果；最后总结全文。

2 相关工作

关于如何增大和改善用户的查询搜索，已有全面的研究。传统的用户个性化建模是基于用户长时记忆与用户日志文件的，关于这方面的研究也比较多。用户日志文件一般建立在概念层上，表现了用户的长期兴趣[7,8]。通过对用户发布的查询信息和用户浏览的文档构建概念层次，来对用户进行个性化建模。文献[9]将不同的词条作为向量来构建用户兴趣文件，并对过去的用户偏好进行聚类。文献[10]建立了用户兴趣模型，这个模型来自搜索相关的信息和关于用户的其他信息，包括用户已经阅读的文件。不幸的是，上述大部分工作构建的用户个性化模型都是基于用户浏览的网页或文件，而这些文件受到了搜索引擎效率的影响。另外，用户个性化建模的研究大多忽视了用户实时需求是动态变化的。

另一个与个性化搜索相关的领域是使用上下文作为一个查询词，该查询将上下文作为个性化搜索的背景。这个上下文的产生是建立于查询词条、文件矢量等上的。传统上，一个上下文可以被一个词语向量所代表[11,12]，也可以被这个矢量空间模型所代表[13]。基于上下文，用户的查询通过增加适当词条，然后将其反馈给搜索引擎，这可以改善搜索的效率性[14]。文献[15]中，上下文通过词语权重向量表示，上下文中的每一个元素都来自用户点击文本的一个关键字。基于这个向量，通过余弦函数来发现相似的查询。文献[16]通过概念和概念之间的关系来提高搜索准确度，概念和概念之间的关系是从 Web 片段中提取的。

近三年来，随着社交网络、位置服务等的发展，用户个性化模型在这些领域当中也得到了相应的研究。文献[19]提出了利用基于概率模型的用户模型构建方法，该方法适用于社交网络数据而非搜索引擎。文献[20]提出了一种基于用户模型的在线内容优化方法，其利用用户的操作与反馈，优化网站的内容布局。文献[21]加入用户的时间与空间信息，构建用户的行为模型，区别于个性化模型，行为模型还考虑了用户的位置信息，因此大多适用于基于位置的服务。文献[22]利用主题模型的方法挖掘用户在微博当中的个性化模型。文献[23]研究用户在无线网络服务中的个性化模型，利用用户在自身的无线网络行为来描述用户的关注内容。

3 生成查询语境

查询词一般是指与查询相关的信息[14]。本文将查询词定义为由用户撰写的一段文字(例如，几句话、一个句子、一个段落)。本节主要介绍如何建立查询语境。查询语境的构建包含下列两个基本步骤：1)从查询词 q 的返回结果中提取概念；2)挖掘步骤 1)所提取的概念之间的关系。

3.1 提取概念

从查询词 q 中提取概念的方法是挖掘搜索引擎返回的结果网页，如谷歌，它提供每个返回的搜索结果的 URL。然而，上述方法是不切实际的。原因如下：

(1)费时。虽然谷歌提供每个搜索结果的 URL，但是下载这些网页很耗时。

(2)解析不可行。由于网页数量巨大且网页数量保持高速增长，迅速解析搜索引擎返回的搜索结果是不切实际的；同时，在现在互联网规模下，不同站点具有不同的 HTML 格式，解析不同的站点也是不可行的。

因此，本文利用查询词 q 的搜索片段而不是网页来提取概念。搜索片段由搜索引擎所提供，作为网页的简短摘要，是非常有用的信息资源。一般来说，搜索片段包含一个文本窗口。图 1 显示了由谷歌提供的查询“apple”的一个搜索片段。

Apple - iPhone
iPhone 3G combines three products in one - a revolutionary phone, a widescreen iPod, and a breakthrough Internet device. It features 3G, built-in GPS, ...
www.**apple**.com/iphone/ - 52k - Cached - Similar pages

图 1 由谷歌提供的查询“apple”的一个片段

由于片段中有许多停用词，如介词、代词，因此有必要对片段进行一些预处理，从而减少从片段中提取概念的噪声。考虑到构建查询语境的实时性，不使用一些耗时且依赖于语言的预处理方法，例如词性标注，相反，使用标准的 SMART 废词表来删除废词。

对查询词 q 的每一个搜索片段进行预处理后，开始提取查询 q 的概念。当查询词 q 被提交到 Web 搜索引擎时，通过搜索引擎得到了相应的搜索片段。根据认知科学的理论，如果一个概念频繁出现在查询词 q 的片段中，说明这个概念与 q 密切相关。根据文献[17]中定义的 *support*，从查询词 q 返回的片段中提取概念 c_i：

1) 网络片段：包含搜索引擎返回的标题、摘要和 URL。

$$support(c_i)=sf(c_i)/n \tag{1}$$

其中，n 是搜索引擎返回搜索片段的数量，$sf(c_i)$ 是概念 c_i 的在这些搜索片段中出现的篇频。

为了构建查询语境，要从返回的片段中提取所有的概念，并对提取的概念的 *support* 进行计算，如式(2)所示：

$$w(i,qc)=support(c_i),\text{if } support(c_i)>\alpha \tag{2}$$

表 1 列出了查询词“apple”所提取的概念。搜索片段的数目是 100。

表 1　查询词“apple”挖掘的概念

concept (c_i)	support(c_i)
apples	0.146
news	0.087
iPhone	0.087
aapl	0.078
amp	0.078
Mac	0.078
history	0.078
company	0.078
stock	0.068

显然，*support* 阈值 α 的设定显著影响了查询词的概念向量构建的结果。换句话说，*support* 的阈值不同，可能会导致不同的查询词概念向量。这个因素的影响将在实验部分进行详细的介绍。

3.2　概念间关系的提取

除了概念 c_i 的权重，概念之间的关系也可以从搜索片段中提取。利用信息理论中的公式互信息(Pointwise Mutual Information，PMI)，可以计算概念 c_i 和 c_j 之间的关系：

$$sim(c_i,c_j)=\frac{\log(\frac{n*sf(c_i\cap c_j)}{sf(c_i)*sf(c_j)})}{\log n} \tag{3}$$

其中 $sf(c_i\cap c_j)$ 指的是概念 c_i 和 c_j 共同出现在搜索片段中的频率；$\log n$ 是归一化因子，以确保概念 c_i 和 c_j 的权重范围在 $[0,1]$ 之间。

显然，查询词 q 的概念和概念之间的关系可以构成一个。图 2 显示了由谷歌查询“apple”的概念关系。图 2 中的节点表示查询词 q 提取的概念，链接表示概念 c_i 和 c_j 之间的关系，c_i 和 c_j 之间的链接强度由概念关系的权重决定。

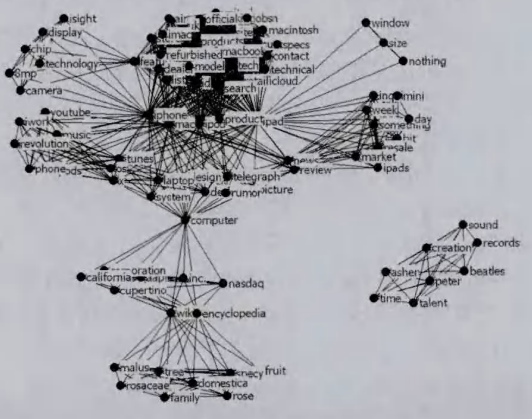

图 2　谷歌查询“apple”的概念关系

与查询词 q 的概念相同，概念之间关系权值的阈值 β 的设定也会影响查询语境构建的结果，将在实验部分对其进行详细的介绍。

4　生成用户语境

第 3 节介绍了如何生成概念语境，概念语境可以在没有任何用户的点击数据中得到。与静态的概念语境不同，用户语境是动态的，是基于用户的点击目标数据获得的。换句话说，用户语境是面向用户的。对于查询词 q，构建用户语境可以分成 3 个阶段。

(1)获取用户语境中的显性概念。显性概念是指出现在用户的点击搜索片段中的概念。例如，当用户搜索查询词“apple”时，如果他/她点击的搜索片段中包含概念“iPhone”，那么“iPhone”是用户的显性概念。

(2)获取用户语境中的隐性概念。隐性概念指的是概念不出现在用户点击的搜索片段中，但用户可能感兴趣。例如，如果用户对概念“iPhone”感兴趣，那么与“iPhone”相关的概念如概念“iPod”可能是用户的隐性概念。

(3)处理用户点击搜索片段的顺序。用户每一次对搜索片段的点击都可以生成一个用户语境。这些语境的合并便是用户在一个搜索会话中的个性化模型。

4.1　获取用户语境中的显性概念

直观地说，在用户点击的搜索片段中出现的概念可以被认为是用户语境中的显性概念。例如，如果一个用户提交查询词“apple”且他/她对概念“iPhone”有兴趣，那么他/她可以点击包含概念“iPhone”的搜索片段。因此，用户语境(user context)可以被表示为一个由显性概念组成的向量：

$$\overrightarrow{uc}=\{w(1,uc),w(2,uc),\cdots,w(i,uc)\},\forall c_i\in\overrightarrow{qc} \tag{4}$$

其中，$w(i,uc)$ 表示在用户语境中的第 i 个概念的权重。

当用户提交查询词 q 到搜索引擎后，一些搜索片段随后被返回给用户。如果用户单击搜索片段 S_j，若概念 C_i 出现在搜索片段 S_j 中，则权重设置为 1，反之为 0，因此，

$$w(i,uc)=\begin{cases}1, & \text{if } c_i\in s_j\\ 0, & \text{others}\end{cases} \tag{5}$$

4.2　生成显性的概念的用户文本

除了出现在用户点击的搜索片段中的概念，其他概念也可能是用户感兴趣的。隐性概念指的是概念没有出现在用户点击的搜索片段中，但用户可能对此概念有兴趣。

从第 3 节介绍的查询语境的概念关系图可以找到用户语境的隐性概念。如果用户对概念 c_i 感兴趣，那么概念关系图中与 c_i 相邻的概念就是隐性概念，用户对这些概念也有可能感兴趣。例如，如果用户提交查询“apple”，他/她感兴趣的是概念“iPhone”，那么他/她会点击包含概念“iPhone”的搜索片段。很显然，在“apple”的查询概念关系图中，用户也有可能对与概念“iPhone”相邻的概念如“stock”和“store”感兴趣。

因此，不仅要计算出现在用户点击的搜索片段中显性概念的权重，还要计算在概念关系图中与显性概念相邻的隐性概念的权重。计算隐性概念 c_i 的权重的直观方法是使用隐性概念 c_i 和显性概念 c_j 之间关系的强度。不幸的是，这种方法是不切实际的，因为一个隐性概念也许关联了许多显性概念。例如，假设“iPhone”和“iPod”是出现在点击的搜索片断中的两个显性概念，但隐性概念“Mac”可以链接到“iPhone”和“iPod”上。在这种情况下，很难选择哪个链接作为隐性概念的权重。为了解决这个问题，提出了 3 种策略来计算用户语境中的隐性概念权重。

(1)策略 1：隐性概念 c_i 的权重由用户点击的搜索片段 s_j

中所有与 c_i 链接的显性概念的权重最大值决定，如式(6)所示：

$$w(i,uc)=\begin{cases}1, & \text{if } c_i\in s_j\\ \max\{w(c_i,c_k)\}, & k\in\{k\mid\forall c_k\in s_j\}\end{cases} \tag{6}$$

(2)策略 2：隐性概念 c_i 的权重由用户点击的搜索片段 s_j 中所有与 c_i 链接的显性概念的权重最小值决定，如式(7)所示：

$$w(i,uc)=\begin{cases}1, & \text{if } c_i\in s_j\\ \min\{w(c_i,c_k)\}, & k\in\{k\mid\forall c_k\in s_j\}\end{cases} \tag{7}$$

(3)策略 3：隐性概念 c_i 的权重由用户点击的搜索片段 s_j 中所有与 c_i 链接的显性概念的权重平均值决定，如式(8)所示：

$$w(i,uc)=\begin{cases}1, & \text{if } c_i\in s_j\\ \sum_k w(c_i,c_k)/n, & k\in\{k\mid\forall c_k\in s_j\}\end{cases} \tag{8}$$

其中，n 是搜索引擎返回搜索片段的数量。

本文将在实验部分讨论哪种策略最适合构建用户语境。

计算完隐性概念的权重，可以在用户语境中增加显性概念和隐性概念。当用户点击搜索片段 s_j 时，出现在搜索片段 s_j 中的概念 c_i 的权重设置为 1，其他的在概念关系图中与概念 c_i 关联的概念权重根据隐性概念的策略来计算。表 2 列出了通过策略 1 计算出的隐性概念的权值。

表 2　根据图 1 的点击片段和图 2 中的概念关系图计算的隐性概念权重

concept c_i	weight
iPhone	1.0
products	1.0
store	0.227
home	0.193
Mac	0.165
stock	0.106
apples	0.091
company	0.077

4.3　基于遗忘因子的用户个性化模型构建

一般而言，对于查询词 q 的搜索引擎返回的结果，用户不止点击一个片段。White 和 Drucker 的一项研究[18]指出，一个查询词的平均点击片段次数为 6.2。因此，如何在一个查询会话中处理点击片段的顺序对构建用户个性化模型是十分重要的。图 3 显示了一个用户用谷歌查询“apple”时点击片段的顺序。

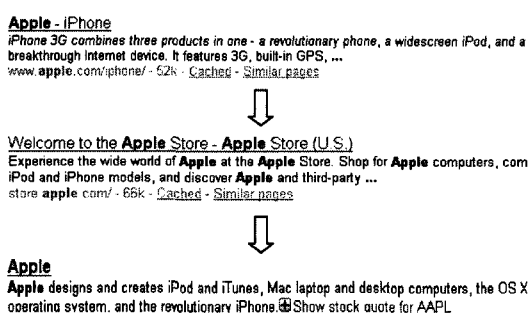

图 3　用户用谷歌查询“apple”时点击片段的顺序

用 4.1 节和 4.2 节的方法可以构建用户每次点击搜索片段后的概念向量，因此在一次查询回话中，只需要合并这些概念向量的顺序就可以构建用户实时的个性化模型。

聚类方法中给出了一组向量的质心向量，质心向量是通过各概念的平均权重获得的向量值。不同于质心向量不考虑向量的顺序，将点击顺序增加到质心向量中来得到一个更合适的用户实时的个性化模型。此外，遗忘因子用于强调最近被点击的文本向量。引入遗忘因子的原因很容易理解，因为用户对最近点击的搜索片段更感兴趣。

用质心向量公式和遗忘因子处理用户点击搜索片段顺序的步骤如下：

(1)在谷歌中查询 q；

(2)$S(q)$作为用户点击的搜索片段的顺序，$s_1,s_2,\cdots,s_n$；

(3)计算用户每一次点击的搜索片段的用户语境 v_i，其中每个片段 $s_i\in S(q)$；

(4)v_q 作为用户实时的个性化模型：

$$v_q=\frac{1}{n}\sum_{i=1}^{n}\frac{i}{n}v_i \tag{9}$$

其中，$\frac{i}{n}$是 i 个点击片段的遗忘因子。例如用户一共点击了 4 个搜索片段，那么第一个点击的搜索片段的遗忘因子为$\frac{1}{4}$，第二个点击的搜索片段的遗忘因子为$\frac{2}{4}$，第三个点击的搜索片段的遗忘因子为$\frac{3}{4}$，第四个点击的搜索片段的遗忘因子为$\frac{4}{4}$。遗忘因子越小，代表由这个搜索片段构成的用户语境对用户的实时个性化模型影响越小。

表 3 列出了按图 3 顺序点击片段的用户语境向量。因为用户对“iPhone”最感兴趣，所以“iPhone”权重最大；而且没有出现在图 3 点击片段中的隐性概念如“industry”也存在于表 3，这是因为这些概念与显性概念相关联(见图 2)。

表 3　按图 3 顺序点击片段的用户语境向量

concept c_i	weight	concept c_i	weight
home	0.241	history	0.131
store	0.882	stock	1.106
news	0.199	industry	0.194
Mac	1.106	company	0.508
aapl	0.479	apples	0.208
iPhone	2.0	products	0.649
amp	0.369		

5　实验和评价

5.1　实验步骤

为了收集点击数据来评估所提出的建立用户个性化模型的方法，邀请我校 50 位计算机科学专业的学生来参与搜索给定测试的查询，测试查询词列于表 4 和表 5 中。当用户提交一个测试查询给编制的网站时，查询内容用于生成用户个性化模型。谷歌提供前 100 名的搜索结果给用户。

在使用表 4 测试的第一个实验(结果将在 5.4 节中介绍)中，50 个用户分为 5 组，每组有 10 个用户。每个组负责搜索不同的查询词，分别为“Apple”、“BMW”、“Java”、“Obama”和“KFC”。之所以采用 3 个不同的查询集合，主要是为了测试在不同语义条件下方法的精度。查询集合 1 中的词没有语义关系，查询集合 2 都是关于汽车的，其中有很强的语义关系。值得注意的是，查询集合 3 都是关于“apple”的，彼此之间具有最强的语义关系。因为“apple”有不明确的含义(例如，“apple computer” 和 “apple pie”)，设置每个用户组进行搜索查询“apple”的不同语义。之所以使用查询集合 3，是因为很多用户可能提交相同的查询，但他们有不同的信息需求(例如用户关注“Mac”或“iPhone”时都可能提交相同的查询词“apple”)。

表 4　用于调整和评估所提方法的查询词

Query set	Set 1	Set 2	Set 3
Query 1	Apple	BMW	Apple (Mac)
Query 2	BMW	Audi	Apple (iPhone)
Query 3	Java	Honda	Apple (AAPL)
Query 4	Obama	Porsche	Apple (stock)
Query 5	KFC	Ferrari	Apple (apples)

在使用表 5 的第二个实验(结果将在第 5.5 节中描述)中,50 个用户也分为 5 组,每组有 10 个用户。与第一个实验中不同的是,每个用户组需要搜索一组查询词而不只是一个。此外,每个用户组被责令特定信息需求(例如"Chanel"、"Gucci"、"Prada"和 "Louis Vuitton")。进行第二个实验的原因是很多用户可能虽然拥有相同的信息需求,但会提出不同的查询,希望验证所提方法在这种情况下的准确性。在这两个实验中,用户组被要求点击与查询相关的搜索片段。点击访问数据被收集来构建每个用户的个性化模型。

表 5　用于评估用户组采样方法的查询词

User Group1(Mac)	macBook	Mac Mini	iMac	Macintosh
User Group2 (Perfume)	Chanel no. 5	Lancôme miracle	Dior me	Burberry touch
User Group3 (Notebook)	Asus	Dell	Lenovo	Acer
User Group4 (Computers)	HP	Dell	Lenovo	Acer
User Group5 (Handbags)	Chanel	Gucci	Prada	Louis Vuitton

5.2　评价方法

在本节中,使用文档聚类方法进行评估。在使用文档聚类方法进行评估时,所有生成的用户个性化模型向量按照初始用户组被分为 5 簇,然后利用 K-means 算法对这些用户进行聚类。评价方法的指标如下:

指标 1　F-measure 值,它是结合了查全率和查准率的指标,计算公式如下:

$$F(i,j)=\frac{2*Recall(i,j)*Precision(i,j)}{Precision(i,j)+Recall(i,j)} \tag{10}$$

指标 2　熵(Entropy),即一个簇有较高的均匀性,计算公式如下:

$$E_j=-\sum_i p_{ij}\log(p_{ij}) \tag{11}$$

聚类数据的总熵的计算公式为每个簇的大小加权每个簇的熵的总和:

$$Entropy=\sum_{j=1}^{m}\frac{n_j*E_j}{n} \tag{12}$$

指标 3　Purity,表示主要成员在集群中的百分比,计算公式如下:

$$Purity=\sum_{j=1}^{m}\frac{n_j}{n}*\max\{Precision(i,j)\} \tag{13}$$

总体来说,F-measure 和 Purity 越大,Entropy 越小,那么聚类的准确度就越高。

5.3　参数 α 与参数 β 对方法的影响

评价参数 α 与参数 β 对方法的影响的步骤如下:

(1)选择第 4.2 节中最佳的策略来计算用户语境的隐含概念权重;

(2)选择最佳的参数 α 来构建用户语境;

(3)选择最佳的参数 β 来构建用户语境。

50 个用户分为 5 个用户组,每组包含 10 个用户。为简单起见,α 的范围设置为{0.05,0.06,0.07,0.08,0.09},β 的范围设置为{0.1,0.2,0.3,0.4}。

表 6 给出了使用 3 种不同策略构建用户语境中的隐性概念所得到的不同的聚类结果。从表中可以看出,策略 1 的结果最好,原因是策略 1 使用概念 c_i 和点击片段 s_j 这两个参数的权重最大。策略 1 类似于信息检索分配方法,这使得它的性能比其它两种策略更好。表 7 给出了使用或不使用隐式概念的聚类结果。从表中可以看出,使用隐性概念比不使用隐性概念所得到的聚类结果要好,其原因可能是隐含的概念可揭示潜在的用户兴趣,这使其成为构建用户文本更精确的方法。因此,本文使用策略 1 进行后续的实验评价。

表 6　不同策略的文档聚类在计算用户文本隐性概念权重时的结果(α=0.05 且 β=0)

	F-measure	Purity	Entropy
Strategy 1	0.859478	0.8422	0.066144
Strategy 2	0.851658	0.8342	0.070009
Strategy 3	0.853471	0.8342	0.069961

表 7　在前述 5.1 节中使用或不使用隐性概念的两个实验中的结果(α=0.05 且 β= 0)

β	0	0.1	0.2	0.3	0.4
F-measure	0.847	0.853	0.829	0.851	0.830
Purity	0.824	0.828	0.804	0.830	0.805
Entropy	0.0723	0.0713	0.0843	0.0737	0.0813

为了尽量减少 α 和 β 之间的干扰,当研究参数 α 对构建用户个性化模型的影响时,β 统一设置为 0。

表 8 列出了不同的评估指标的结果。从表 8 可以看出,当 α 增加时,评价指标单调递减,这意味着 0.05 是建立用户个性化模型最好的参数值。由于构建用户个性化模型的实时性要求,其他比 0.05 小的数值(例如,α=0.04 或 0.03)会增加算法的复杂度,降低算法的实时性,因此,在后续实验评价部分,α 设置为 0.05。

表 8　不同的评价指标结果(β=0)

α	0.05	0.06	0.07	0.08	0.09
F-measure	0.858	0.857	0.855	0.843	0.761
Purity	0.838	0.836	0.832	0.826	0.727
Entropy	0.0678	0.0679	0.0708	0.0756	0.107

表 9 显示了不同 β 下的评价指标结果,在不同的 β 下,评价指标不具有单调趋势。因此,在我们的评估条件下,β 设定为 0.1 是最好的评价指标。

表 9　在不同评价指标下的结果(α=0.05)

β	0	0.1	0.2	0.3	0.4
F-measure	0.847	0.853	0.829	0.851	0.830
Purity	0.824	0.828	0.804	0.830	0.805
Entropy	0.0723	0.0713	0.0843	0.0737	0.0813

5.4　精确度与查询词之间的语义关系

本节评价在不同的语义条件下所提出的方法的性能。实验结果列于表 10,从中可以看出查询词之间的语义关系对所提出的方法有较大的影响。查询集 1 中的查询涉及不同的方面,由于查询之间的弱语义关系,因此查询集 1 的指标较高。与此相反,在查询集 2 中的查询都是关于"car"的,查询词的强语义关系降低了指标的值。不同于查询集 1 和查询集 2,查询集 3 表示同一查询的不同信息需求。由于用户的个性化模型主要用在一些搜索需求比较相近的任务中,较好的指标使我们的方法适合于同一查询的不同信息需求。

表 10　在不同语义条件下的评价指标结果

	Query set 1	Query set 2	Query set 3
F-measure	0.853	0.633	0.433
Purity	0.828	0.641	0.76
Entropy	0.0713	0.131	0.146

5.5　精确度与用户的信息需求

很多用户虽然拥有相同的信息需求，但可能提交不同的查询词，本节测试所提出的方法在这种情况下的准确度。实验结果列于表 11 中。虽然用户提交不同的查询词（例如 iMac、MacBook），但是所提方法可以得到用户的实际信息需求。其原因是相同的信息需求可能会导致用户点击同一个概念。因此，我们的方法可以准确地表达用户的真实信息需求。

表 11　在信息相同查询词却不同的查询情况下的评价指标结果

F-measure	Purity	Entropy
0.45	0.933	0.0516

5.6　对比实验

为了验证本文所提出的用户模型的准确性，除了前面几个实验之外，本节将本文所提出的方法与文献[16]中的方法进行对比。对比实验的基准数据采用传统的 BB 基准数据[24]。对比的实验方法依然采用用户模型聚类，而对比的参数依然是 Precision、Recall、F-measure 。实验对比结果如表 12 所列。从表 12 可以看出本文的方法优于文献[16]的方法，主要原因是在于本文引入了遗忘系数的概念。

表 12　对比实验结果

	Precision	Recall	F-measure
文献[16]的方法	0.521	0.685	0.592
BB 基准值	0.469	0.694	0.559
本文的方法	0.783	0.657	0.715

结束语　搜索引擎的一个标准是不同的用户用相同的查询条件检索时，返回的结果相同。为了解决准确性问题，个性化搜索引擎被提出，它可以根据用户的偏好提供不同的搜索结果。本文研究构建用户语境的有效方法来获取和表达用户的偏好。当用户向搜索引擎提交查询词时，通过从搜索片段中提取概念和概念之间的关系生成查询语境。因此，查询语境可以作为用户搜索行为的背景知识。此外，在谷歌实现用户的有效点击顺序的捕获来构建用户文本。实验结果表明，用户语境可以成功地表达个人用户信息需求。

将来可以从下面几个方向进行扩展：首先，用户语境可以用于协同过滤用户组；其次，用户语境可以用于推荐系统，如推荐相关产品给用户。希望未来能够对这两种场景进行研究。

参 考 文 献

[1] Lau T, Horvitz E. Patterns of search: Analyzing and modeling Web query refinement[C]// Proceedings of the UM. 1999:119-128

[2] Lawrence S. Context in Web Search[J]. IEEE Data Engineering Bulletin, 2000, 23(3):25-32

[3] Jansen M, Spink A, Bateman J, et al. Real Life Information Retrieval: A Study of User Queries on the Web[C]// Proceedings of ACM SIGIR Forum, vol. 32, 1998:5-17

[4] Pandit S, Olston C. Navigation-aided retrieval[C]// Proceedings of the 16th International World Wide Web Conference. 2007: 477-486

[5] Dou Z, Hua R, et al. Evaluating the effectiveness of personalized Web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(8):1178-1190

[6] Shen X, Tan B, Zhai C. Context-sensitive information retrieval using implicit feedback[C]// Proceedings of SIGIR'05. New York, NY, USA: ACM Press, 2005:43-50

[7] Chirita P A, Nejdl W, Paiu R, et al. Using ODP metadata to personalize search[C]// Proceedings of SIGIR '05. New York, NY, USA: ACM Press, 2005:178-185

[8] Trajkova J, Gauch S. Improving ontology-based user profiles[C]// Proceedings of RIAO. 2004:380-389

[9] Sugiyama K, Hatano K, Yoshikawa M. Adaptive Web search based on user profile constructed without any effort from users [C]// Proceedings of WWW'04. New York, NY, USA: ACM Press, 2004:675-684

[10] Teevan J, Dumais S T, Horvitz E. Personalizing search via automated analysis of interests and activities[C]// Proceedings of SIGIR'05. New York, NY, USA: ACM Press, 2005:449-456

[11] Billsus D, Hilbert D, Maynes-Aminzade D. Improving proactive information systems[C]// Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI). 2005:159-166

[12] Henzinger M, Chang B-W, Milch B, et al. Query-free news search[C]// Proceedings of the 12th International World Wide Web Conference. 2003:1-10

[13] Yu C T, Lam K, Salton G. Term weighting in information retrieval using the term precision model[J]. Journal of the ACM, 1982, 29(1):152-170

[14] Kraft R, Chang C C, et al. Searching with contexts[C]// Proceedings of the 15th International World Wide Web Conference. 2006:477-486

[15] Baeza-Yates R A, Hurtado C A, Mendoza M. Query Recommendation Using Query Logs in Search Engines[C]// Proceedings of EDBT Workshop, vol. 3268, 2004:588-596

[16] Leung K W T, Ng W, et al. Personalized Concept-Based Clustering of Search Engine Queries [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(11):1505-1518

[17] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// Proceedings of ACM SIGMOD. 1993

[18] Bilenko M, White W R. Mining the search trails of surfing crowds: identifying relevant websites from user activity[C]// Proceedings of the 17th International Conference on World Wide Web. 2008:51-60

[19] Raghavan V, et al. Modeling Temporal Activity Patterns in Dynamic Social Networks[J]. IEEE Transactions on Computational Social Systems, 2014, 1(1):89-107

[20] Bian J, et al. User Action Interpretation for Online Content Optimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(9):2161-2174

[21] Yang D, et al. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, 45(1):129-142

[22] He Li, Jia Yan, Han Wei-hong, et al. Mining user interest in microblogs with a user-topic model[J]. Communications, 2015, 11(8):131-144

[23] Liu Yao, Wang Shao-xuan, Dey S. Content-Aware Modeling and Enhancing User Experience in Cloud Mobile Rendering and Streaming[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2014, 4(1):43-56

[24] Beeferman D, Berger A. Agglomerative Clustering of a Search Engine Query Log[C]// Proc. of ACM SIGKDD Conference. 2000:407-416