# 第三章

本文研究基于搜索经验的查询推荐方法。在此前对复杂搜索的研究中，作者所在实验室提出了时间树理论来帮助用户记录搜索过程，并对复杂搜索任务进行管理，以及对复杂搜索过程进行可视化。本章首先研究时间树记录的用户的搜索过程中是否蕴含了搜索经验，为研究基于搜索经验的查询推荐方法提供理论基础。

## 3.1 时间树与用户的搜索经验

在时间树理论中，每一个搜索任务都可以被表示成一棵时间树。在时间树中，查询和点击被表示为树的节点，这些节点则被组织成为一种时序溯源的树型结构（Relative Chronological Source-tracking Tree ，RCST）。

在本研究中，作者将时间树的RCST结构定义为一个四元组N的集合{N<C, F, L, R>}，其中C和F也是一个四元组N，C代表某一个查询或者点击节点，F代表该节点的唯一父节点，若F为空，则C节点为时间树的根节点。L代表当前节点C所处的纵向层数，R代表当前节点C处于它所有兄弟节点从上之下的第几个位置。

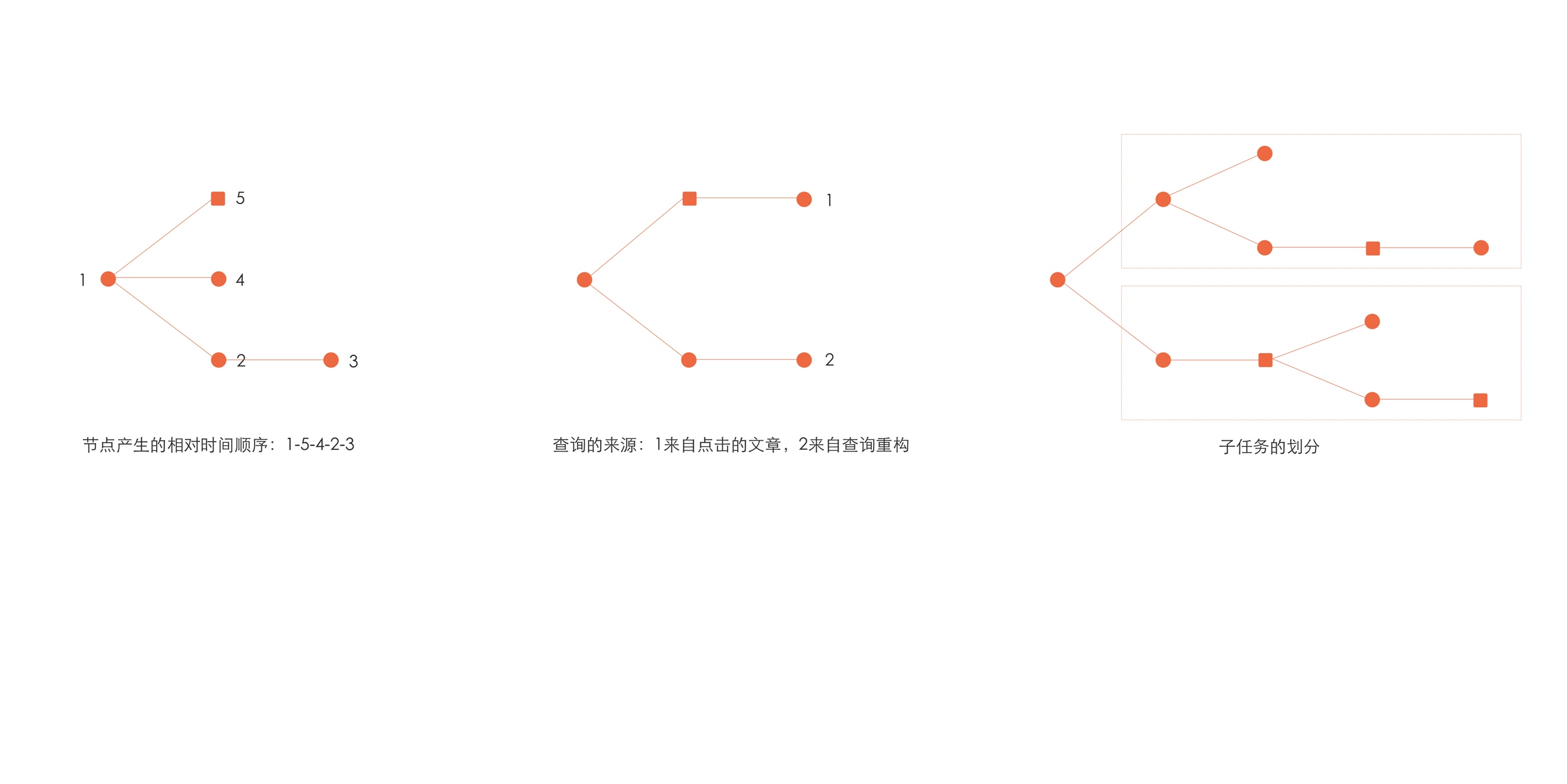
基于RCST结构，时间树能够帮助用户理解查询与点击之间的关系。例如，用户可以根据节点的位置关系来判别节点间的相对时间顺序；也可以得到查询节点的来源；用户还可以根据时间树的RCST结构对复杂搜索任务进行子任务的划分。如图1所示。

图1 时间树RCST结构示意图

时间树能够记录用户的搜索过程，由此引出的一个问题是，时间树中是否同时记录了用户的搜索经验。本章通过建立用户搜索经验模型并设计用户复杂搜索过程验证实验，回答时间树中是否蕴含了用户搜索经验的问题。

为了回答时间树中是否蕴含用户搜索经验的问题，首先需要对用户的搜索经验进行模型化的描述。为了提出用户搜索经验模型，作者首先对用户的复杂搜索过程进行了观察，以分析用户在进行复杂搜索的过程中会产生哪些方面的经验。最基本也最容易观察到的是，用户的查询和点击会按照时间顺序先后进行，而不可能同时进行两个查询或点击行为，因此用户的查询和点击一定具有时间先后顺序。其次，用户进行的某个查询，一定是出于某种原因，例如查询词来自点击结果，或者查询词来自用户当前所进行任务的任务需求，或者查询词来自突然迸发的灵感。最后，用户在进行复杂搜索任务的过程中，由于任务的复杂性，用户常常需要将任务划分为各个小的子任务，并在逐一完成各个子任务的过程中完成复杂搜索任务。子任务的划分对复杂搜索任务的完成有很重要的意义，它反映了了用户对搜索任务理解的完整度。更好地对子任务进行划分，有助于用户更清晰地完成复杂搜索任务。

但用户在进行搜索的过程中，每一个子任务并不是连贯完成的，用户常常会因为各种干扰因素而发生子任务的切换，例如用户在进行子任务A的过程中点击了一篇文章，文章中提到了一个关键词C，而这个关键词是属于子任务B中的内容，用户如果搜索了关键词C，便产生了一次由子任务A到子任务B的切换，而此时，子任务A尚没有被完成，在未来的某个时间户很可能会重新对子任务A进行搜索。因此，子任务往往散落点，用在用户搜索过程中的各个时间点上，因此，将离散杂乱的搜索过程划分为各个子任务的过程，也应该是用户产生的搜索经验中很重要的一点。

综上所述，作者观察到了在用户搜索过程中产生的时间先后顺序，查询原因以及子任务的划分三个方面的搜索经验。基于这一观察，作者提出用户搜索经验模型。

## 3.2 用户搜索经验模型

本研究中，作者将搜索经验模型定义为一个三元组E<TE, CE, THE>，其中TE，CE和THE分别表示用户在进行复杂搜索过程中的时间经验，因果经验以及主题经验。其中时间经验为用户在搜索过程中进行查询和点击的时间先后关系，例如用户首先查询了某个词，进行了某些点击，又查询了某些词。因果经验为用户所进行查询的查询来源，例如用户从某篇文章中看到了某个关键词获得了查询灵感，或者用户是进行了一次查询重构。主题经验为用户对复杂搜索任务所进行的子任务划分。用户在进行一次复杂搜索任务的过程中，经常会将复杂的搜索任务划分为几个子话题、子任务。这些子任务提供了一种复杂搜索任务的清晰结构展现，反映了用户对其所进行的复杂搜索任务的规划。

时间经验，因果经验，主题经验共同组成了用户在复杂搜索过程中的搜索经验模型，但是这三种用户的搜索经验都无法被直观地观测到或是直接进行衡量。为了解决这一问题，本研究使用一种侧面的评价标准，来衡量用户的搜 索经验。作者提出一种合理的推断：如果用户在搜索过程中产生了搜索经验E<TE, CE, THE>，并且时间树能够记录这种搜索经验，那么用户在依据时间树回忆自己的搜索过程时，一定能够分别满足时间一致性，因果一致性，主题一致性。这三种一致性的概念由Habermas等人在2000年提出[1]，他们指出，好的回想应当具有这三种性质。其中时间一致性指用户能够准确回忆起自己在搜索过程中查询与点击的先后顺序。因果一致性指用户能够准确回忆起搜索中的某个查询的查询动机。主题一致性指用户能够合理地对任务进行子任务划分，以及对搜索过程中的某一个阶段，能够判断其中的查询与点击分别属于哪些子任务。这三种一致性与搜索经验模型中的TE, CE, THE相对应。如果时间树能够很好地帮助用户记录搜素经验E<TE, CE, THE>，那么用户在依据时间树进行搜索过程回想时一定能够获得更好的时间一致性，因果一致性以及主题一致性，反之，如果时间树不能很好地记录用户的搜索经验，那么用户在依据时间树进行搜索过程的回想时一定不能获得很好的时间一致性，因果一致性和主题一致性。

因此本章从用户对复杂任务搜索过程的回想的角度设计实验，依据用户在对复杂搜索过程进行回想时在时间一致性，因果一致性以及主题一致性上的表现来回答时间树中时间经验，因果经验以及主题经验的蕴含性。

## 3.3 复杂搜索过程验证实验

本节实验的主要目的是以用户对复杂搜索过程进行回想为载体，通过考察用户回想过程中的时间一致性，因果一致性与主题一致性来回答时间树在记录用户搜索过程的同时是否也记录了用户的搜索经验的问题。

### 3.3.1 复杂搜索环境

为了针对上述目的进行实验的设计，首先需要设计搜索场景以使用户的搜索过程满足复杂搜索的定义。作者采用Singer等提出的复杂搜索任务的定义[2]来定义复杂搜索任务——“一个搜索任务如果需要进行以下几种行为中的一种，就可以被称为是复杂的：汇总，探索和组合。”在有了复杂搜索任务的定义作为理论基础后，作者参考了White等提出的用户可能需要在其中进行复杂搜索的三种典型场景[3]：1）用户对搜索任务领域的知识匮乏；2）哦搜索任务本身需要浏览及探索；3）系统对信息的索引不够充分。以这三种复杂搜索的典型场景为原则，作者设计复杂搜索任务，使实验参与者进行的搜索过程满足复杂搜索进行的条件。

### 3.3.2 复杂搜索任务设计

本实验设计采用学术搜索引擎进行，在学术搜索引擎中，文章常常不会被全文索引，因此自然符合第三种复杂搜索场景。所以在设计复杂搜索任务时只需要考虑使其符合前两种场景即可。

我们共设计了两类共4个任务，前两个搜索任务满足被称为第一类任务，符合第二种复杂搜索场景，是那些“需要浏览和探索”的任务。为了最大化用户浏览和探索的程度，第一类任务的布置以及参与者的完成过程都使用参与者的母语汉语进行。后两个搜索任务被称为第二类任务，符合第一种复杂搜索场景，是那些用户“缺少知识或者上下文而不能很好地组织查询”的任务。为了增加组织查询的困难程度，第二类任务的布置和完成都使用参与者的非母语英语进行。对于第一类任务，参与者被要求根据任务要求完成搜索报告，两个第一类任务的任务描述如下：

任务一：搜索相关论文，了解化疗中的常用药物，药物机理，使用方法，适用情况及副 作用等，尤其针对癌症治疗，了解化疗药物的治疗手段，组合方式等。(不仅仅要了解支持化学疗法的药物，还要求了解对化学疗法进行辅助的药物，如缓解化学疗法副作用的药物等。)

任务二：完成一篇报告，内容包括PM2.5 的概念、来源，PM2.5 的危害以及危害原因。找到中国 PM2.5 污染最为严重的几个城市，并分析其形成原因。研究降低 PM2.5 的方法，并分析其原理以及其实际实施可能性(包括但不局限于利弊)。

对于第二类任务，参与者不被直接告知搜索任务的具体描述，而是首先被要求阅读一些阅读材料，在两天以后，参与者被告知需要围绕之前阅读的材料凭记忆进行搜索，并最终完成一份与阅读材料相关的文章标题清单。第二类任务的阅读材料使用了下面两篇文章的摘要：

任务三：S. Shunmuga Krishnan, Ramesh K. Sitaraman: Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-Experimental Designs. IEEE/ACM Trans. Netw. 21(6): 2001-2014 (2013)

任务四：Hanqiang Cheng, Yu-Li Liang, Xinyu Xing, Xue Liu, Richard Han, Qin Lv, Shivakant Mishra: Efficient misbehaving user detection in online video chat services. WSDM 2012: 23-32

### 3.3.3 实验过程设计

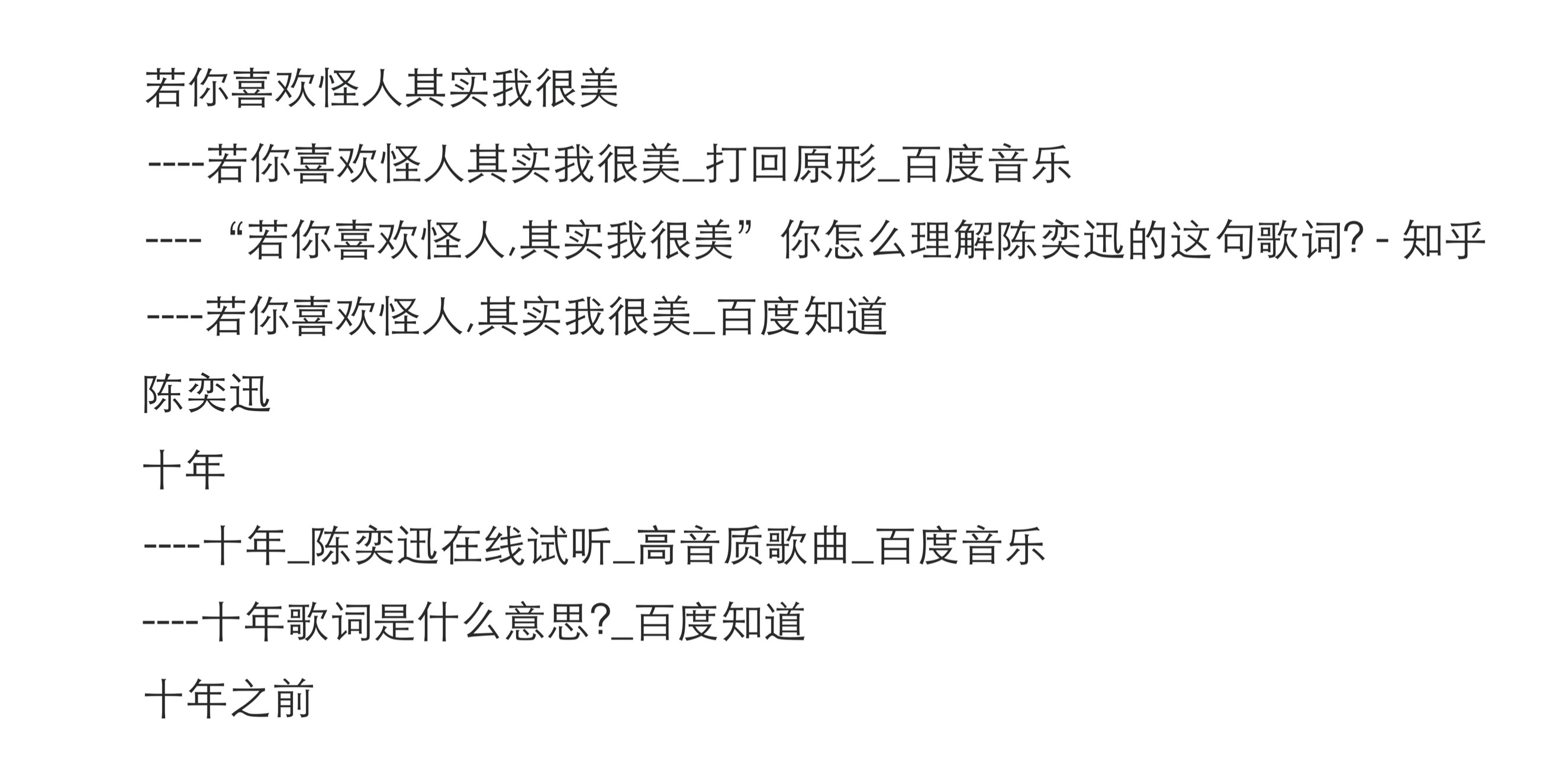
本实验旨在通过考察用户在对复杂搜索过程回想中的时间一致性，因果一致性和主题一致性来回答时间树中是否蕴含用户搜索经验的问题。为了客观衡量用户使用时间树对复杂搜索过程回想过程中的三种一致性，作者采用一种用户搜索过程的时间线记录（Timeline SearchLog，以下简称SearchLog）形式与时间树进行对比。SearchLog形式如图2所示。

图2 SearchLog示意图

参与者被要求完成3.3.2节中所述的4个复杂搜索任务。每一位参与者分别从两类任务中随机分配一个任务使用时间树进行搜索过程的记录与回想，另一个任务则使用SearchLog进行搜索过程的记录与回想。并且最终保证每个任务分别有4个参与者使用时间树和SearchLog进行搜索过程的记录与回想。