(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108764007 A
(43)申请公布日 2018.11.06

(21)申请号 201810138534.9

(22)申请日 2018.02.10

(71)申请人 集智学园(北京)科技有限公司
地址 100000 北京市海淀区学院南路12号院57号楼118室B-019号

(72)发明人 张江 陈孟园 龚力 张倩

(74)专利代理机构 江苏爱信律师事务所 32241
代理人 唐小红

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/20(2006.01)
G06F 17/27(2006.01)

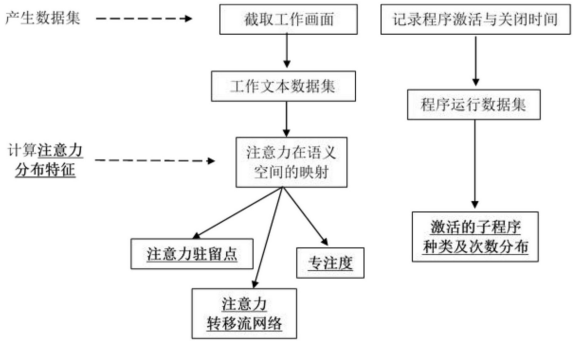
权利要求书1页 说明书6页 附图1页

(54)发明名称

基于OCR与文本分析技术对注意力的测量方法

(57)摘要

本发明公开了基于OCR与文本分析技术对注意力的测量方法,通过OCR等技术获取被观测者的行为数据,用如特征选择、相似性度量等文本分析技术计算出被观测者注意力映射在语义空间中的概念,间接测量注意力的分布与转移方法。传统的心理学与神经脑科学专注于测量注意力这种认知能力的强弱,而本发明所指的注意力属于管理科学范畴,是指人们关注一个主题、事件的持久尺度。本发明方法在个人管理、企业管理、推荐系统等方面有极大的用处。



CN 108764007 A

1.基于OCR与文本分析技术对注意力的测量方法,其特征在于,包括以下步骤:

步骤1.捕获视觉范围:

步骤2.OCR技术处理:

使用OCR技术处理得到的捕获画面,识别画面中的文字,生成文本集;

步骤3.获取数据集:

步骤4.计算注意力分布特征。

2.如权利要求1所述的方法,其特征在于,所述步骤1包括:

1-1)捕获的视觉范围的随截取频率的变化而有显著的变化;以每天处于工作状态的8小时截取,固定每1分钟截取一次工作画面;

1-2)编写截取个人工作画面的代码,每小时累计捕获60帧画面,每天累计捕获480帧工作画面。

3.如权利要求1所述的方法,其特征在于,所述步骤2包括:

3-1)对文本集进行分词、去除噪声等处理生成文本集;

3-2)使用去除噪声的分词集,对每个画面的文本做一次特征提取,提取关键语义信息,记录时间与语义信息,生成语义数据集;

3-3)获取当天电子设备的激活进程以及进程的激活时间和关闭时间,构成程序数据集。

4.如权利要求1所述的方法,其特征在于,所述步骤4包括:

4-1)将激活程序与标签对应,计算在某特定时间内激活的子程序标签分布及次数分布;

4-2)获取注意力转移流网络,记录语义发生转移的时刻以及该时刻的关键词,根据时间的先后顺序,生成注意力转移流网络;

4-3)计算注意力驻留时间,注意力转移流网络中的节点A所在时间与节点B所在时间相减,得节点A的驻留时间;长期观测下,形成长期驻留点在时间上的变化情况;

4-4)计算专注度,记录每天各个时段的专注度,生成当天的专注度曲线,进一步的,在长时间观测下,形成长期的专注度曲线。

基于OCR与文本分析技术对注意力的测量方法

技术领域

[0001] 本发明属于注意力测量领域，主要致力于研究语义空间中注意力分布与转移，具体涉及基于OCR与文本分析技术测量语义空间中的注意力分布与转移。

背景技术

[0002] 什么是注意力，注意的是人脑意识被外在或内在事物占据的过程。而注意力是指人的心理活动指向和集中于某种事物的能力。注意力经济的鼻祖Michael H.Goldhaber提到，当今社会是一个信息极大丰富甚至泛滥的社会，相对于过剩的信息，人们的注意力资源是稀缺的，随着信息时代的发展，有价值的不是信息，而是注意力。人们越来越注重时间的管理，希望将有限的注意力放在更多有意义的事情上，市场上一些关于时间管理的书持续热销，各类时间管理的手机电脑软件层出不穷，对注意力的度量在信息时代有着重大的意义。

[0003] 近年来，科学家们从多个方面出发，开发了注意力的测量技术。一方面是对注意能力，例如，心理学专家将注意力测量技术认为是一种认知能力测试，注意力测试的方法有注意力测试图，测试表，测试习题等等；此外，对注意力的测量可能最好通过眼球本身连同头部的运动来评定，当你眼睛盯着某一处时，你的注意力很大的概率就在该处，由此，一些可以监控头部和眼球运动的可穿戴设备被称为注意力监控器，如眼动仪；注意力神经生理学家通过研究脑电波变化及人体的脑电波反馈，开发出脑电头盔计算注意力。另一方面，扩大注意过程的时间尺度，空间尺度，从由注意过程产生的一系列关注的主题、事件入手测量注意力。例如，网易云音乐软件通过记录人对音乐的注意行为，形成了独特的推荐系统；Rescue TimeAPP会统计用户使用App的时间，根据分类，判断用户分别花了多少时间在开发、设计、聊天、娱乐和其他上面，用户根据统计可以获得注意力在APP上的分配。ihour是一款用户录入任务，记录任务执行时长的时间管理软件，用户主动将自己所专注的事情记录在ihour上，希望获得自己对每件事情的投入程度。这两方面的注意力测量，为人类自身的注意力训练、与自我管理约束提供了很大的帮助。

[0004] 如今，大大小小的电子设备都会安装一个屏幕，是为了获取人类注意力而设计的，而人类现在80％以上的清醒时间，都是在盯着各种各样的屏幕看。准确的说，屏幕中的内容获得了人的注意。基于此种情况，本发明提出了基于OCR与文本分析技术的个人注意力测量方法，从较大的时间和空间尺度出发，结合上述两方面的测量角度，对现有的注意力测量方式提出技术上的完善与改进，获取被观测者关注的主题在语义空间中的表达，最终得到实时的、长期的注意力的分布、流动状态。

[0005] 百度百科对语义空间的定义是，语言意义的世界。一般来说，信息是意义和符号的统一体，内在的意义只有通过一定的外在形式(动作、表情、文字、音声、图画、影像等符号)才能表达出来。因此，每一种符号体系在广义上都是传达意义的语言，它们所表达的意义构成了特定的语义空间。据此，本发明从外在注意的角度出发，将关注主题的内在意义提取出来，通过该意义表现注意力在语义空间上的变化。

[0006] OCR技术(Optical Character Recognition),即光学字符识别,是指电子设备检查纸上打印的字符,通过检测暗、亮的模式确定其形状,然后用字符识别方法将形状翻译成计算机文字的过程。OCR识别系统的输入为不同格式的图像,输出为计算机文字,其目的很简单,是要把影像作转换,使影像内的图形继续保存、有表格则表格内资料及影像内的文字,一律变成计算机文字,使能达到影像资料的储存量减少、识别出的文字可再使用及分析,也可节省因键盘输入的人力与时间。目前的OCR技术比较成熟,对于证件的识别准确率一般可以达到99%。画质清晰的普通图像的文字识别准确度也可达到85%以上。个人视觉所涉及到的浏览的电脑网页、手机屏幕等信息,使用OCR技术,可以轻松的转化为计算机可识别的文字,提取出被观测者关注的主题在语义空间的表达。

[0007] 文本分析技术,采用自然语言处理(NLP:Natural Language Processing)和分析方法将文本内容转换成数据,建立它的数学模型,对文本进行科学的抽象,用以描述和代替文本。使计算机能够通过对这种模型的计算和操作来实现对文本的识别。文本分析一般由三步组成,解析数据,搜索检索,文本挖掘。目前文本分析被广泛用于客户体验、客户洞察、数据分析等方面。从语义空间角度出发,使用文本分析技术,将语义信息作为数据集,结合时间信息,可以获得各个时间段的注意力映射在语义空间中的概念,获得注意力驻留点以及注意力的流动等信息,从而实现对注意力间接的测量。

发明内容

[0008] 本发明从外在注意与语义空间的角度入手,每天固定时间间隔(秒/分钟)对被观测者所操作的电子设备截取一帧个人浏览图像,使用OCR技术,将图像转化为可分析的文本集。使用文本分析技术从杂乱无章的文本集提取出语义信息,与时间结合构成分析注意力的语义数据集。同时,本发明将记录每一个子程序被激活的次数以及时间,构成程序数据集,从子程序的性质中可以获得有关个人注意力在时间和空间分布上宏观的一些性质。两个数据集交相呼应,反映注意力的分布特征。例如,每个人在某一个特定时间内激活的子程序次数分布、学习程序和聊天程序的注意力分配等。从数据集中,本方法可以得到,注意力在语义空间的映射,注意力驻留点、注意力随时间的转移。根据每天生成的个人注意力结构特征,可以积累形成对个人注意力特征的长期观测,从而分析出个人在工作时的注意力结构等更多长期注意力特征。

[0009] 本发明提出基于OCR与文本分析技术对注意力的测量方法,包括以下步骤:

[0010] 步骤1.捕获视觉范围

[0011] 1-1)捕获的视觉范围的随截取频率的变化而有显著的变化,此处,我们以每天处于工作状态的8小时截取,固定每1分钟截取一次工作画面为例;

[0012] 1-2)编写截取个人工作画面的代码,每小时累计捕获60帧画面,每天累计捕获480帧工作画面。

[0013] 步骤2.OCR技术处理

[0014] 使用OCR技术处理得到的捕获画面,识别画面中的文字,生成文本集;

[0015] 步骤3.获取数据集

[0016] 3-1)对文本集进行分词、去除噪声等处理生成文本集;

[0017] 3-2)使用去除噪声的分词集,对每个画面的文本做一次特征提取,提取关键语义

信息,记录时间与语义信息,生成语义数据集;

[0018] 3-3)获取当天电子设备的激活进程以及进程的激活时间和关闭时间,构成程序数据集。

[0019] 步骤4.计算注意力分布特征

[0020] 4-1)将激活程序与标签对应,计算在某特定时间内激活的子程序标签分布及次数分布;

[0021] 4-2)获取注意力转移流网络,记录语义发生转移的时刻以及该时刻的关键词,根据时间的先后顺序,生成注意力转移流网络;

[0022] 4-3)计算注意力驻留时间,注意力转移流网络中的节点A所在时间与节点B所在时间相减,可得节点A的驻留时间;长期观测下,可形成长期驻留点在时间上的变化情况;

[0023] 4-4)计算专注度,记录每天各个时段的专注度,生成当天的专注度曲线,进一步的,在长时间观测下,形成长期的专注度曲线。

[0024] 有益效果

[0025] 1、相较于使用眼动仪、心理学等方法测试注意认知能力。本方法主要测量更为宽泛的注意力,测量的是由注意产生的关注的主题、事物在时间上的分布与转移情况。本测量方法无需穿戴大型机械设备,不对身体造成辐射、不对个人的工作造成影响,不限制个人的活动范围,观测方便;

[0026] 2、相较于Rescue Time,Ihour等时间管理软件,本方法测量的是由注意产生的关注的主题、事物在语义空间上的映射,而不是关注的事物本身,涵盖的能反映注意力特征的信息更加全面,无需手动录入,能够实现自动记录。

[0027] 3、方便对被观测者的注意力行为形成长期的监测,较之于传统的注意力测量技术,能够观测到一些短期注意力测量观测不到的注意力变化。

附图说明

[0028] 图1为本发明注意力方法流程示意图;

[0029] 图2为DBOW模型示意图。

具体实施方式

[0030] 下面结合附图对本发明的技术方案进行详细说明:

[0031] 本发明的思路是收集被观测对象的电子设备使用信息,具体包括程序与使用界面的信息,使用统计方法与OCR技术处理使用信息,这些信息间接的表示了被观测对象注意力信息在语义空间上的映射,根据这些信息,最终得到注意力的分布特征。

[0032] 本发明方法的基本流程如图1所示,具体包括以下步骤:

[0033] 步骤1.捕获视觉范围

[0034] 使用Python处理图像的模块Pillow、Time、selenium定时(例如,一分钟)截取被观测者当前使用的电子屏幕;每天的观测时间段为工作时间,上午9:00—12:00,下午13:00-17:00,共计8个小时,每小时累计捕获60帧画面,每天累计捕获480帧工作画面,将所有画面存储至同一个文件夹下。图2是一帧截取的工作画面示例。

[0035] 步骤2.OCR技术处理

[0036] 存储一天的工作屏幕,ApiOCR是百度AI文字识别提供的API服务,API

[0037] (Application Programming Interface,应用程序编程接口)是一些预先定义的函数,目的是提供应用程序与开发人员基于某软件或硬件得以访问一组例程的能力。我们使用ApiOCR逐个将截取的屏幕识别为文本。识别的图中的文字与时间对应生成表2,将表2的信息在Python中存储成字典式文本集。

[0038] 表1生成文本集

[0039]

时间	文本
2018-1-12 16:35	{'Jupyter try1 Last Checkpoint3 小时前(autosaved)\nFile Edit View\nCell Kemel\nv M Run I C Code\nln [2]: from aip import Aipocr\n 你的 APPID AK SK\nPTD=\'18665425\nl KEY =6GTG1U2TAEXG30LMQAB48MPT\nSECRET KEY =Mvap9mvjoznf0gze9pichragytyqiirb\n"client Aipocr(APP ID, API KEY, SECRET KEY)"\nln [25]: from PIL import

[0040]

	Imagegrab\nln [13]: import image\nln [14]: #Pillow bacakage\nfrom PIL import Image\nln\n"ge. open(""D: /pythonword/tt.jpg"\n"format, im size, im mode"\nln [22]: import pyscreenshot\nln [26]: import pyscreenshot\nH fullscreen\nscreenshot-pyscreenshot. grab()\nscreenshot show(\npart of the screen\n"screenshot-pyscreenshot. grab(bbox=(10, 10, 500, 500))"\nscreenshot show)\npyscreenshot. grab_ to file(screenshot. png)\nln [24]: im Imagegrab grab()\nOserror\nTraceback(most recent call last)\n<ipython-input-24-3f241b5e0575> in <module>O)\n'}
.....	.....

[0041] 步骤3.获得数据集

[0042] 3-1)使用python的NTLK模块对文本集首先进行分句处理,再进行分词处理,同时

过滤掉无用的信息。如,统计一天之内的分词词频,去掉一些只在少数工作画面中出现的低频词;去掉不携带任何信息的助词、连接词等停用词,除噪后为文本集;停用词举例:{about、above、according、accordingly、across、actually、after、afterwards、again、against、ain't、all};

[0043] 3-2)文本集中的信息反应了人脑的注意对象,也即占意物,提取占意物的语义信息生成语义数据集,使用去噪后的分词集,对每个画面的文本做一次特征选择(Feature Selection)。文本特征选择,指的是从原有的特征(所有文本)中提取出少量的,具有代表性的特征,但特征的类型没有变化,原来是文本词汇集合,特征提取后仍是词汇,不过数量大大减少。通过文本特征选择提取出分词集中的关键词,分析工作者的注意力。使用TF-IDF特征选择算法,具体步骤如下:

[0044] $$词频(TF)=\frac{某个词在文章中出现的次数}{本文章总次数} \quad (1)$$

[0045] $$逆文档频率(IDF)=log_{10}\left(\frac{语料库的文档总数}{包含该词的文档数+1}\right) \quad (2)$$

[0046] TF-IDF=词频(TF)×逆文档频率(IDF) (3)

[0047] 计算出文档中每个词的TF-IDF的值,然后按照降序排列,取前面的几个词作为特征属性。这里由于只取前K大的,作为注意力映射在语义空间的概念。

[0048] 3-3)特征提取出的关键词被认为是每分钟作者的注意力所在,我们称之为注意点,类似地,统计每天480张画面的分词集,使用特征提取关键词,这些关键词为注意力在语义空间上的表达;

[0049] 3-4)每隔固定的时间(例如,1分钟)使用windows PowerShell调用process on API获取现有的程序进程、进程的电脑内存占用情况。按照时间,统计电脑在观测时间的程序运行的时间长度、不同程序运行时的电脑内存占用情况,不同时间运行的程序个数。

[0050] 步骤4.计算注意力分布特征

[0051] 4-1)从程序数据集中,将激活程序与程序标签数据库对应,如工作类、娱乐类、学习类、聊天类等对应,计算在一天的时间轴上,某特定时间段激活程序的标签分布。计算在某特定时间内激活的子程序种类及次数分布;

[0052] 表2激活程序标签数据库

[0053]

程序	标签
Python	编程、工作
微信(wechat)	聊天
R	编程、工作
绝地求生	游戏
腾讯视频	娱乐
Word	工作
……	……

[0054] 4-2)每天共产生480个分词集,使用Doc2vec做文本相似度分析,若两个文本相似度低,认为在这个时间内注意力发生了转移生成注意力转移流网络。记录注意点发生转移的时刻以及该时刻的关键词,根据时间的先后顺序,生成注意力转移流网络。网络的节点为

注意点,节点A与节点B之间的有向连边,意思为上一时刻的注意力点A,下一时刻转移到了注意力点B;Doc2vec算法原理如下,得到

[0055] Sentence/Document的向量表示,Doc2Vec也有两种模型,分别为:Distributed

[0056] Memory(DM)和Distributed Bag ofWords(DBOW),DM模型在给定上下文和文档向量的情况下预测单词的概率,DBOW模型在给定文档向量的情况下预测文档中一组随机单词的概率。这里本发明使用DBOW模型,该模型的输入是文档的向量,预测的是该文档中随机抽样的词.

[0057] 4-3)计算注意力驻留时间,注意力转移流网络中的节点i所在时间与节点i -1所在时间相减,可得节点i -1的驻留时间;通常情况下,我们计算驻留时间的公式如下,

[0058] $Focus=top10\{max(time(node_i)-time(node_{i-1}))\}i\in(1,2,3,.....n)$ (4)

[0059] 长期观测下,可形成长期驻留点在时间上的变化情况;

[0060] 4-4)计算专注度,专注度的计算公式为:

[0061] $(1-n/N)\times100\%$ (5)

[0062] 其中n为每个时间段包含的网络节点数,N为每个时间段产生的分词集个数。例如,计算每个小时的专注度,此时N为60,若早上9:00-10:00产生了10个节点,则专注度为90%。记录每天各个时段的专注度,生成当天的专注度曲线,进一步的,在长时间观测下,形成长期的专注度曲线。

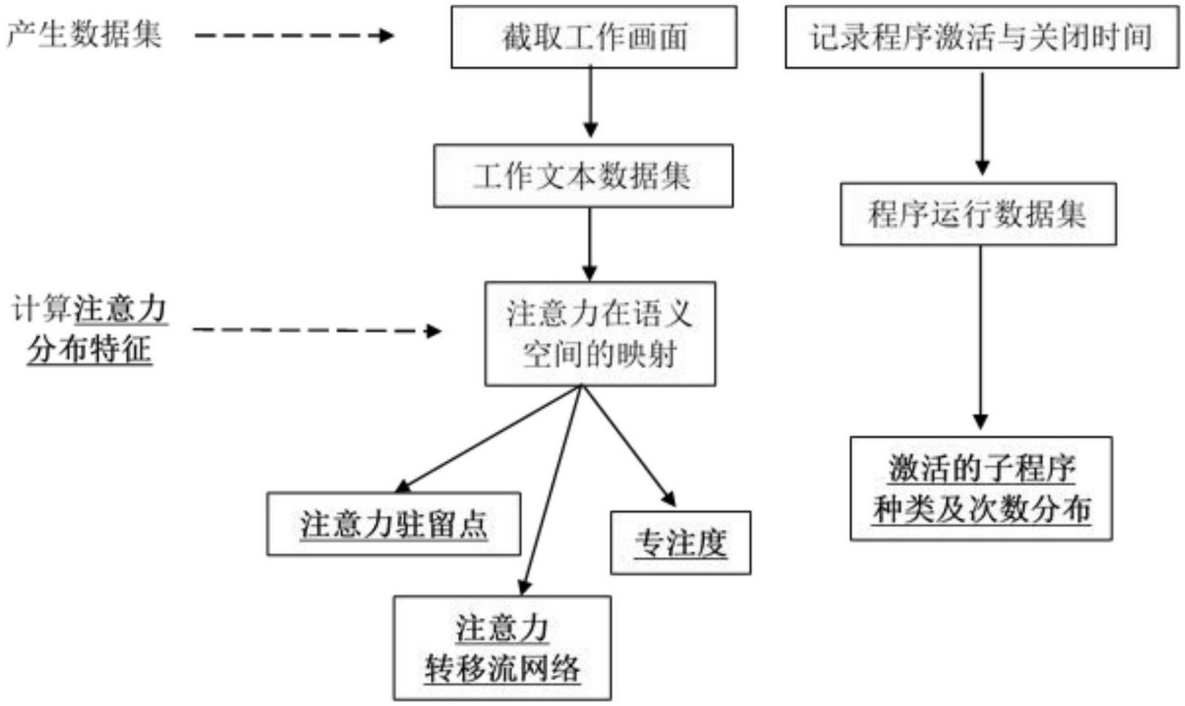

产生数据集
截取工作画面
记录程序激活与关闭时间
工作文本数据集
程序运行数据集
计算注意力分布特征
注意力在语义空间的映射
注意力驻留点
专注度
注意力转移流网络
激活的子程序种类及次数分布

图1

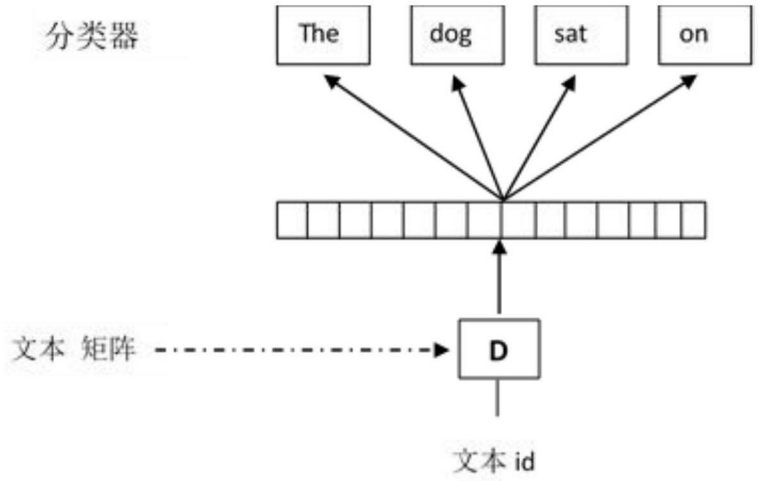

分类器
The
dog
sat
on
文本 矩阵
D
文本 id

图2