

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2019 年 10 月 31 日 (31.10.2019)



(10) 国际公布号
WO 2019/205378 A1

(51) 国际专利分类号:
G06Q 10/06 (2012.01)

(21) 国际申请号: PCT/CN2018/102127

(22) 国际申请日: 2018 年 8 月 24 日 (24.08.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
2018103823117 2018年4月26日 (26.04.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 李正洋(LI, Zhengyang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 毛小豪(MAO, Xiaohao); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所(普通合伙)(SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

(54) Title: METHOD AND APPARATUS FOR SELECTING INVESTMENT STOCKS BASED ON PUBLIC SENTIMENT FACTOR, AND STORAGE MEDIUM

(54) 发明名称: 基于舆情因子的投资选股方法、装置及存储介质

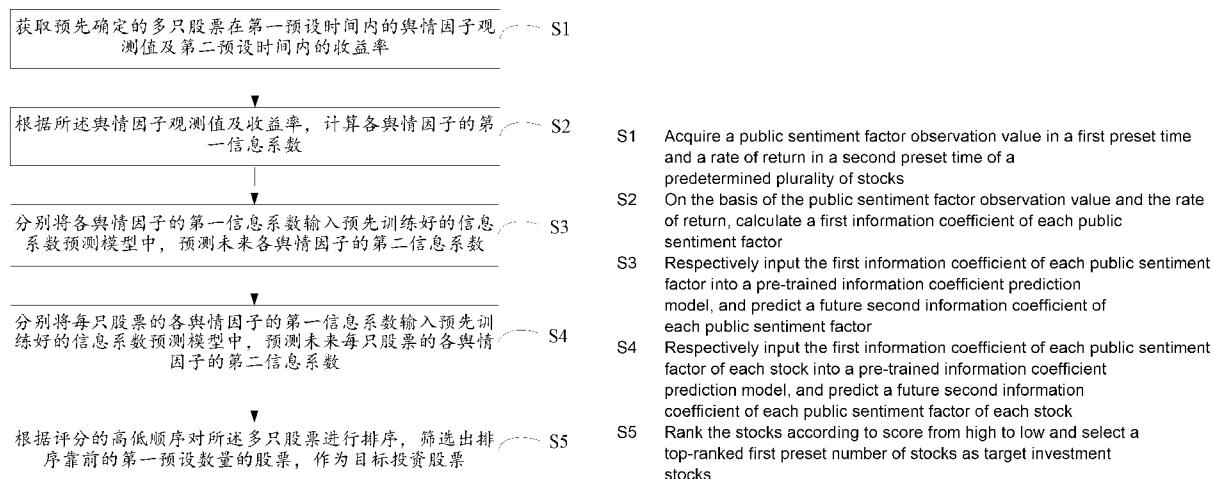


图 3

(57) Abstract: Provided in the present application are a method and an apparatus for selecting investment stocks based on public sentiment factors, the method comprising: acquiring a public sentiment factor observation value in a first preset time and a rate of return in a second preset time of a predetermined plurality of stocks; using the public sentiment factor observation value and the rate of return, calculating a first information coefficient of each public sentiment factor; respectively inputting the first information coefficient of each public sentiment factor into a pre-trained information coefficient prediction model, and predicting a future second information coefficient of each public sentiment factor; respectively calculating a score of the plurality of stocks; and ranking the stocks according to score from high to low and selecting a top-ranked first preset number of stocks as target investment stocks. Also provided in the present application are an electronic apparatus and a storage medium. By means of calculating the score of the stocks, the present application visually displays the impact of the public sentiment factors on the future yield of the stocks.



GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本申请提出一种基于舆情因子的投资选股方法, 该方法包括: 获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率; 利用所述舆情因子观测值及收益率, 计算各舆情因子的第一信息系数; 分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中, 预测未来各舆情因子的第二信息系数; 分别计算所述多只股票的评分; 及, 根据评分的高低顺序对所述多只股票进行排序, 筛选出排序靠前的第一预设数量的股票, 作为目标投资股票。本申请还提出一种电子装置及存储介质。本申请通过计算股票评分, 直观地展示舆情因子对股票未来收益的影响。

基于舆情因子的投资选股方法、装置及存储介质

本申请基于巴黎公约申明享有2018年4月26日递交的申请号为CN 2018103823117、名称为“基于舆情因子的投资选股方法、装置及存储介质”的中国专利申请的优先权，该中国专利申请的整体内容以参考的方式结合在本申请中。

技术领域

本申请涉及金融大数据挖掘领域，尤其涉及一种基于舆情因子的投资选股方法、电子装置及计算机可读存储介质。

背景技术

股票的舆情因子（例如，新闻热度及大众对相应新闻的情绪态度）在一定程度上可以表达为该股票的市场潜在能量，当一只股票的舆情因子异常（例如，当一只股票有重大负面新闻出现或者获得新一轮投资）时，需要即时更改该股票的投资策略。一般都是预先为股票的某个指标设置阈值，然后等权重投资指标满足阈值条件的股票，或者根据专业投资人员市场经验对相应股票设定投资策略。这种方式对做决策投资人员的专业技能要求较高，很容易出错，可能会增加误操作的风险。

发明内容

本申请提供一种基于舆情因子的投资选股方法、电子装置及计算机可读存储介质，其主要目的在于通过计算股票评分，直观地展示舆情因子对股票未来收益的影响，选择评分高的股票作为目标投资股票，减小投资风险、提高投资收益。

为实现上述目的，本申请提供一种基于舆情因子的投资选股方法，该方法包括：

获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；及

- 5 根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

此外，为实现上述目的，本申请还提供一种电子装置，该电子装置包括：存储器、处理器，所述存储器上存储有基于舆情因子的投资选股程序，所述
10 基于舆情因子的投资选股程序被所述处理器执行时实现如下步骤：

获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

15 分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；及

根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

20 此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有基于舆情因子的投资选股程序，所述基于舆情因子的投资选股程序被处理器执行时实现如上所述的基于舆情因子的投资选股方法的步骤。

25 相较于现有技术，本申请提出的基于舆情因子的投资选股方法、电子装置及计算机可读存储介质，根据股票舆情因子观测值及收益率数据，计算股票各舆情因子的第一信息系数，利用第一信息系数及信息系数预测模型预测第二信息系数，反映各舆情因子对股票未来收益的预测能力；根据第二信息
30 系数对各舆情因子进行赋权，实现了对各舆情因子的动态赋权，提高了各舆

情因子的可靠性；根据各舆情因子观测值及相应的权重计算股票评分，选择评分较高的股票作为目标投资股票，提高了股票评分的真实性，直观地展示了舆情因子观测值对股票未来收益的影响，有助于投资者控制风险、提高投资收益。

5

附图说明

图1为本申请电子装置较佳实施例的示意图；

图2为图1中基于舆情因子的投资选股程序的程序模块图；

图3为本申请基于舆情因子的投资选股方法较佳实施例的流程图。

10 本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

15 应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供一种电子装置1。参照图1所示，为本申请电子装置1较佳实施例的示意图。

在本实施例中，该电子装置1包括存储器11、处理器12，网络接口13及通信总线14。其中，通信总线14用于实现这些组件之间的连接通信。

20 网络接口13可以包括标准的有线接口、无线接口（如WI-FI接口）。

存储器11包括至少一种类型的可读存储介质。所述至少一种类型的可读存储介质可为如闪存、硬盘、多媒体卡、卡型存储器等的非易失性存储介质。在一些实施例中，所述可读存储介质可以是所述电子装置1的内部存储单元，例如该电子装置1的硬盘。在另一些实施例中，所述可读存储介质也可以是
25 所述电子装置1的外部存储设备，例如所述电子装置1上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。

在本实施例中，所述存储器11的可读存储介质通常用于存储安装于所述电子装置1的基于舆情因子的投资选股程序10等。所述存储器11还可以用于
30 于暂时地存储已经输出或者将要输出的数据。

处理器 12 在一些实施例中可以是一中央处理器 (Central Processing Unit, CPU), 微处理器或其他数据处理芯片, 用于运行存储器 11 中存储的程序代码或处理数据, 例如执行基于舆情因子的投资选股程序 10 等。

图 1 仅示出了具有组件 11-14 以及基于舆情因子的投资选股程序 10 的电子装置 1, 但是应理解的是, 并不要求实施所有示出的组件, 可以替代的实施更多或者更少的组件。

可选的, 该电子装置 1 还可以包括用户接口, 用户接口可以包括显示器 (Display)、输入单元比如键盘 (Keyboard), 可选的用户接口还可以包括标准的有线接口、无线接口。

10 可选地, 该电子装置 1 还可以包括显示器, 在一些实施例中可以是 LED 显示器、液晶显示器、触控式液晶显示器以及有机发光二极管 (Organic Light-Emitting Diode, OLED) 触摸器等。显示器用于显示在电子装置中处理的信息以及用于显示可视化的用户界面。

15 在图 1 所示的装置实施例中, 作为一种计算机存储介质的存储器 11 中包括基于舆情因子的投资选股程序 10, 处理器 12 执行存储器 11 中存储的基于舆情因子的投资选股程序 10 时实现以下步骤:

获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率;

20 在本实施例中, 股票的舆情因子包括舆情热度因子及舆情情感因子, 故各只股票的舆情因子观测值即为第一预设时间内的舆情热度观测值及舆情情感观测值。

25 假设有预先确定的 M 只股票, 用户可根据需要设置第一预设时间及第二预设时间。可以理解的是, 从预先约定的 M 只股票中选择 N 只目标股票, 作为下一期投资的对象, 选股频率可以是每隔一周进行一次, 也可以是每隔一个月进行一次, 在本实施例中为每隔一周进行一次。假设选股时间为 t 时刻, 相应地, 例如, 第一预设时间可以设置为 t-1 时刻 (即, 选股当期的上一期), 第二预设时间可以设置为 t 时刻 (即, 选股当期), 每两期之间的时间间隔为一周。

30 需要说明的是, 在获取第一预设时间内的各舆情因子观测值及第二预设

时间内的收益率时，对于舆情因子观测值而言，不需要通过收集、处理各只股票相关的舆情信息，而是直接从一些网站上手动提取，或者通过程序对外接口直接提取，例如，从通联数据或新浪股吧提取预先确定的 M 只股票在第一预设时间内的舆情热度观测值及舆情情感观测值。

- 5 对于收益率而言，则需要通过计算得到。在本实施例中，M 只股票在第二预设时间内对应的收益率为周收益率，在计算周收益率时，采用的是 M 只股票的以一周为时间间隔的收盘价，具体地，计算公式为：

$$R = (P_t - P_{t-1}) / P_{t-1}$$

其中，R 为股票在 t 时刻（当期）的周收益率， P_t 为股票在 t 时刻（当期）的收盘价， P_{t-1} 为股票在 t-1 时刻（上一期）的收盘价。

- 10 根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

为了直观地展示出各舆情因子对股票收益的预测能力，需根据 M 只股票在第一预设时间内的各舆情因子观测值及第二预设时间内的收益率，计算 M 只股票在各舆情因子在第二预设时间对应的信息系数（Information Coefficient, IC）。具体地，上述第一信息系数指在 t 时刻（当期）股票各舆情因子的信息

15 系数。

在本实施例中，将斯皮尔曼相关系数（Rank IC）作为股票在各舆情因子的信息系数。Rank IC，即在某时刻，全部股票在某因子上的排序与下一期全部股票收益的排序之间的截面相关系数，计算公式为：

$$\text{Rank IC} = \text{corr}(\text{order}_{t-1}^f - \text{order}_t^r)$$

- 20 其中， $\text{order}_{t-1}^f = (X_1, X_2, \dots, X_M)^T$ ，为 t-1 时刻（上一期）各股票的舆情因子 f 的观测值排名，X 为 t-1 时刻（上一期）舆情因子 f 的观测值； $\text{order}_t^r = (Y_1, Y_2, \dots, Y_M)^T$ ，为 t 时刻（当期）各股票的收益率排名，Y 为 t 时刻（当期）股票的收益率。

利用上述步骤，分别计算 M 只股的舆情热度因子及舆情情感因子在 t 时刻（当期）的 IC 值，即第一 IC 值。

- 25 分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

可以理解的是，各舆情因子的 IC 值能够很好地反映各舆情因子的预测能力，IC 值越大，就说明该因子其在该期对股票收益的预测能力越强。鉴于

本申请的目的在于选择下一期的目标股票，则需要了解各舆情因子在下一期对股票收益的预测能力，也就是各舆情因子在下一期的 IC 值。具体地，上述第二信息系数指股票的各舆情因子在 $t+1$ 时刻（下一期）的 IC 值。

在确定 M 只股票的舆情热度因子及舆情情感因子的第一 IC 值后，将 M 只股票的舆情热度因子及舆情情感因子的第一 IC 值输入预先确定的信息系数预测模型中，预测 M 只股票的舆情热度因子及舆情情感因子在 $t+1$ 时刻（下一期）的 IC 值，即第二 IC 值。

具体地，假设 M 只股票的舆情热度因子及舆情情感因子的第一 IC 值分别为 IC_{At} 、 IC_{Bt} ，将其以 $[IC_{At}, IC_{Bt}]$ 的形式输入信息系数预测模型中，模型输出的结果为 $[IC_{A(t+1)}, IC_{B(t+1)}]$ ，其中， $IC_{A(t+1)}$ 、 $IC_{B(t+1)}$ 分别为该 M 只股票的舆情热度因子及舆情情感因子的第二 IC 值。

在本实施例中，所述预先确定的信息系数预测模型通过训练神经网络得到。鉴于模型的输入数据只是各舆情因子在某个时刻的 IC 值，数据形式比较简单，故选用含有一层隐藏层的三层前馈神经网络进行训练。具体地，该信息系数预测模型的训练步骤包括：

在计算得到预先确定的第二预设数量（例如，3000 只）的股票各舆情因子在第三预设时间（两年）内的每个时刻对应的信息系数后，确定样本数据 $[I_K, I_{K+1}]$ ，其中， I_K 表示各舆情因子在 K 时刻对应的 IC 值， I_{K+1} 表示各舆情因子在 $K+1$ 时刻对应的 IC 值（即，后面提到的实际值）。

具体地，每两个时刻之间的时间间隔可以为一天，也可以为一周。例如，当两个时刻之间的时间间隔为一天时，则样本数据为各舆情因子在每一天的 IC 值；当两个时刻之间的时间间隔为一周时，则样本数据需要采集的是时间间隔为一周的数据，该数据可以是各舆情因子在连续一周的五个交易日的 IC 值的平均值，也可以是取连续一周的五个交易日中最后一个交易日的 IC 值。

基于交叉验证法（cross-validation）将样本数据划分为训练集、评估集和测试集（例如，70% 的样本数据作为训练集，10% 的样本数据作为评估集，20% 的样本数据作为测试集）。

将训练集的样本数据输入至三层前馈神经网络模型，对模型进行训练，初步确定模型参数；所述测试集用于在训练过程中对三层前馈神经网络模型的准确率进行测试，将所述测试集的样本数据输入训练得到的所述三层前馈

神经网络模型中，以对训练得到的所述三层前馈神经网络模型进行测试，当训练得到的所述三层前馈神经网络模型满足预设验证条件（例如，模型预测准确率大于或等于小于预设阈值），则训练完成，将训练完成的三层前馈神经网络模型设置为信息系数预测模型。

5 根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；

具体地，该步骤可以细化为以下步骤：

根据各舆情因子对应的第二信息系数，分别确定所述 M 只股票的各舆情因子对应的权重；根据所述舆情因子观测值及各舆情因子所占的权重，分别
10 计算每只股票的评分。

在 t 时刻，利用信息系数预测模型预测得到 M 只股票的舆情热度因子、舆情情感因子未来一期的信息系数分别为 $IC_{A(t+1)}$ 、 $IC_{B(t+1)}$ 后，需要确定舆情热度因子、舆情情感因子在计算股票评分时分别对应的权重。

在确定各舆情因子对应的权重之前，需判断各舆情因子的因子种类，其中，因子种类包括：正向因子和负向因子。具体地，舆情热度因子为正向因子，表示股票的舆情热度观测值越高，该股票的未来收益会比较高；舆情情感因子既可能是正向因子，又可能是负向因子，需要说明的是，舆情情感观测值的取值范围为 $[-1,1]$ ，舆情情感观测值的绝对值越高，正面(负面)的情感越强烈，当某只股票的舆情情感观测值为负时，对该股票的未来收益存在负
15 面影响，此时，舆情情感因子为负向因子；相反，当某只股票的舆情情感观测值为正时，对该股票的未来收益存在正面影响，此时，舆情情感因子为正向因子。

具体地，对于正向因子，当 $IC_{i(t+1)} > 0$ 时，该因子 i 对应的第一权重 $\omega_{i1} = IC_{i(t+1)}$ ，否则， $\omega_{i1} = 0$ ，表示该因子在未来一期失效；对于负向因子，
25 当 $IC_{i(t+1)} < 0$ 时，该因子 i 对应的第一权重 $\omega_{i1} = -IC_{i(t+1)}$ ，否则， $\omega_{i1} = 0$ ，表示该因子在未来一期失效。

利用上述步骤确定舆情热度因子、舆情情感因子对应的第一权重分别为 ω_{A1} 、 ω_{B1} ，为了便于后续计算股票对应的评分，对股票的舆情热度因子及舆情情感因子对应的第一权重进行归一化，根据归一化结果确定股票的两个因
30 子对应的第二权重 ω_{A2} 、 ω_{B2} ，作为计算股票评分时两个因子对应的权重 ω_A 、

ω_B 。在本实施例中， ω_A 、 ω_B 的计算公式为：

$$\omega_A = \omega_{A2} = \omega_{A1} / (\omega_{A1} + \omega_{B1})$$

$$\omega_B = \omega_{B2} = \omega_{B1} / (\omega_{A1} + \omega_{B1})$$

分别获取 M 只股票的舆情热度观测值 X_A 、舆情情感观测值 X_B 、舆情热度因子 A 对应的权重 ω_A 及舆情情感因子 B 对应的权重 ω_B ，根据预设的计算公式，计算 M 只股票的评分。具体地，股票的评分的计算公式为：

$$S = \sum \omega_i * X_i$$

5 其中，i 为股票的舆情因子， $i = \{A, B\}$ ， ω_i 为股票的舆情因子 i 对应的权重， $\omega_i = \{\omega_A, \omega_B\}$ ， X_i 为 t 时刻股票的各舆情因子的观测值， $X_i = \{X_A, X_B\}$ 。

根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

可以理解的是，利用上述步骤计算得到的股票的评分，在一定程度上可以反映出股票在未来收益的走势，股票的评分越高，其未来收益可能越高。因此，在得到 M 只股票中的每只股票评分后，按照评分高低顺序对 M 只股票进行排序，从 M 只股票中筛选出评分最高的第一预设数量（例如，N 只）的股票，作为待投资的目标股票。

在其他实施例中，确定包含待投资的目标股票的投资组合后，还需要分别确定各只目标股票占投资总额的比重，例如，对目标股票进行等权投资，或者，根据目标股票的评分确定相应的比重，评分高的股票对应的比重较大，评分低的股票对应的比重较小），具体操作可根据投资者的实际需求进行设置。最后，根据目标股票及目标股票对应的比重确定最终投资组合。

上述实施例提出的电子装置，根据股票舆情因子观测值及收益率数据，计算股票各舆情因子的第一信息系数，利用第一信息系数及信息系数预测模型预测第二信息系数，反映各舆情因子对股票未来收益的预测能力；根据第二信息系数对各舆情因子进行赋权，实现了对各舆情因子的动态赋权，提高了各舆情因子的可靠性；根据各舆情因子观测值及相应的权重计算股票评分，选择评分较高的股票作为目标投资股票，提高了股票评分的真实性，直观地展示了舆情因子观测值对股票未来收益的影响，有助于投资者控制风险、提高投资收益。

可选地，在其他的实施例中，基于舆情因子的投资选股程序 10 还可以被分割为一个或者多个模块，一个或者多个模块被存储于存储器 11 中，并由一个或多个处理器 12 所执行，以完成本申请。本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段。参照图 2 所示，为图 1 中基于舆情因子的投资选股程序的程序模块图。在本实施例中，基于舆情因子的投资选股程序 10 可以被分割为：获取模块 110、信息系数计算模块 120、预测模块 130、评分计算模块 140 及选股模块 150。所述模块 110-150 所实现的功能或操作步骤均与上文类似，此处不再详述，示例性地，例如其中：

获取模块 110，用于获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

信息系数计算模块 120，用于根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

预测模块 130，用于分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

评分计算模块 140，用于根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；及

选股模块 150，用于根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

此外，本申请还提供一种基于舆情因子的投资选股方法。参照图 3 所示，为本申请基于舆情因子的投资选股方法较佳实施例的流程图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，基于舆情因子的投资选股方法包括：步骤 S1-S5。

S1、获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

在本实施例中，股票的舆情因子包括舆情热度因子及舆情情感因子，故各只股票的舆情因子观测值即为第一预设时间内的舆情热度观测值及舆情情感观测值。

假设有预先确定的 M 只股票，用户可根据需要设置第一预设时间及第二预设时间。可以理解的是，从预先约定的 M 只股票中选择 N 只目标股票，作

为下一期投资的对象，选股频率可以是每隔一周进行一次，也可以是每隔一个月进行一次，在本实施例中为每隔一周进行一次。假设选股时间为 t 时刻，相应地，例如，第一预设时间可以设置为 $t-1$ 时刻（即，选股当期的上一期），第二预设时间可以设置为 t 时刻（即，选股当期），每两期之间的时间间隔为一周。

需要说明的是，在获取第一预设时间内的各舆情因子观测值及第二预设时间内的收益率时，对于舆情因子观测值而言，不需要通过收集、处理各只股票相关的舆情信息，而是直接从一些网站上手动提取，或者通过程序对外接口直接提取，例如，从通联数据或新浪股吧提取预先确定的 M 只股票在第一预设时间内的舆情热度观测值及舆情情感观测值。

对于收益率而言，则需要通过计算得到。在本实施例中， M 只股票在第二预设时间内对应的收益率为周收益率，在计算周收益率时，采用的是 M 只股票的以一周为时间间隔的收盘价，具体地，计算公式为：

$$R = (P_t - P_{t-1}) / P_{t-1}$$

其中， R 为股票在 t 时刻（当期）的周收益率， P_t 为股票在 t 时刻（当期）的收盘价， P_{t-1} 为股票在 $t-1$ 时刻（上一期）的收盘价。

S2、根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

为了直观地展示出各舆情因子对股票收益的预测能力，需根据 M 只股票在第一预设时间内的各舆情因子观测值及第二预设时间内的收益率，计算 M 只股票在各舆情因子在第二预设时间对应的信息系数（Information Coefficient, IC）。具体地，上述第一信息系数指在 t 时刻（当期）股票各舆情因子的信息系数。

在本实施例中，将斯皮尔曼相关系数（Rank IC）作为股票的各舆情因子的信息系数。Rank IC，即在某时刻，全部股票在某因子上的排序与下一期全部股票收益的排序之间的截面相关系数，计算公式为：

$$\text{Rank IC} = \text{corr}(\text{order}_{t-1}^f - \text{order}_t^f)$$

其中， $\text{order}_{t-1}^f = (X_1, X_2, \dots, X_M)^T$ ，为 $t-1$ 时刻（上一期）各股票的舆情因子 f 的观测值排名， X 为 $t-1$ 时刻（上一期）舆情因子 f 的观测值； $\text{order}_t^f = (Y_1, Y_2, \dots, Y_M)^T$ ，为 t 时刻（当期）各股票的收益率排名， Y 为 t 时

刻（当期）股票的收益率。

利用上述步骤，分别计算 M 只股的舆情热度因子及舆情情感因子在 t 时刻（当期）的 IC 值，即第一 IC 值。

5 S3、分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

可以理解的是，各舆情因子的 IC 值能够很好地反映各舆情因子的预测能力，IC 值越大，就说明该因子其在当期对股票收益的预测能力越强。鉴于本申请的目的在于选择下一期的目标股票，则需要了解各舆情因子在下一期对股票收益的预测能力，也就是各舆情因子在下一期的 IC 值。具体地，上述
10 第二信息系数指股票的各舆情因子在 t+1 时刻（下一期）的 IC 值。

在确定 M 只股票的舆情热度因子及舆情情感因子的第一 IC 值后，将 M 只股票的舆情热度因子及舆情情感因子的第一 IC 值输入预先确定的信息系数预测模型中，预测 M 只股票的舆情热度因子及舆情情感因子在 t+1 时刻（下一期）的 IC 值，即第二 IC 值。

15 具体地，假设 M 只股票的舆情热度因子及舆情情感因子的第一 IC 值分别为 IC_{At} 、 IC_{Bt} ，将其以 $[IC_{At}, IC_{Bt}]$ 的形式输入信息系数预测模型中，模型输出的结果为 $[IC_{A(t+1)}, IC_{B(t+1)}]$ ，其中， $IC_{A(t+1)}$ 、 $IC_{B(t+1)}$ 分别为该 M 只股票的舆情热度因子及舆情情感因子的第二 IC 值。

在本实施例中，所述预先确定的信息系数预测模型通过训练神经网络得到。
20 鉴于模型的输入数据只是各舆情因子在某个时刻的 IC 值，数据形式比较简单，故选用含有一层隐藏层的三层前馈神经网络进行训练。具体地，该信息系数预测模型的训练步骤包括：

分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率，计算各舆情因子在每个时刻对应的信息系数，以获取样本数据；及
25

将样本数据划分为训练集、评估集和测试集，并利用训练集的样本数据对神经网络进行训练，得到信息系数预测模型，利用测试集的样本数据对信息系数预测模型的准确率进行测试，得到最终的信息系数预测模型。

同理，根据 Rank IC 的计算公式计算得到预先确定的第二预设数量（例如，
30 3000 只）的股票在各舆情因子在第三预设时间（两年）内的每个时刻对应的

信息系数后，确定样本数据 $[I_K, I_{K+1}]$ ，其中， I_K 表示各舆情因子在K时刻对应的IC值， I_{K+1} 表示各舆情因子在K+1时刻对应的IC值（即，后面提到的实际值）。

具体地，每两个时刻之间的时间间隔可以为一天，也可以为一周。例如，
5 当两个时刻之间的时间间隔为一天时，则样本数据为各舆情因子在每一天的IC值；当两个时刻之间的时间间隔为一周时，则样本数据需要采集的是时间间隔为一周的数据，该数据可以是各舆情因子在连续一周的五个交易日的IC值的平均值，也可以是取连续一周的五个交易日中最后一个交易日的IC值。

基于交叉验证法（cross-validation）将样本数据划分为训练集、评估集和
10 测试集（例如，70%的样本数据作为训练集，10%的样本数据作为评估集，20%的样本数据作为测试集）。

将训练集的样本数据输入至三层前馈神经网络模型，对模型进行训练，初步确定模型参数；所述测试集用于在训练过程中对三层前馈神经网络模型的准确率进行测试，将所述测试集的样本数据输入训练得到的所述三层前馈
15 神经网络模型中，以对训练得到的所述三层前馈神经网络模型进行测试，当训练得到的所述三层前馈神经网络模型满足预设验证条件（例如，模型预测准确率大于或等于小于预设阈值），则训练完成，将训练完成的三层前馈神经网络模型设置为信息系数预测模型。

鉴于反向传播算法（Back propagation）是目前用来训练人工神经网络
20 （Artificial Neural Network, ANN）的最常用且最有效的算法，因此，本实施例中，基于反向传播法对三层前馈神经网络模型进行训练，具体包括：

将训练集数据输入到模型的输入层，经过隐藏层，最后达到输出层并输出结果；计算模型输出的预测值与样本数据中的实际值之间的误差，并将该误差从输出层向隐藏层反向传播，直至传播到输入层；根据误差调整模型参
25 数的值；不断迭代上述过程，直至收敛。

需要说明的是，为了提高模型训练速度，在训练过程中输入数据时，不会每次只输入一个时刻的数据进行训练，而是每次固定输入batch size=k的样本数据进行训练，其中，k的大小可以根据需要设置，在本实施例中可设置为1024。

30 优选地，在模型训练过程中，为了增加模型的泛化能力和防止过拟合，

例如，将 K 时刻的 IC 值的实际值输入三层前馈网络模型中，输出 K+1 时刻的 IC 值的预测值时，在 K+1 时刻的 IC 值的预测值的基础上增加一个随机噪声值（例如，用从标准正态分布的采样值乘以 0.01 来当作噪声值）。进一步地，在计算 K+1 时刻的 IC 值的预测值（含噪声值）与实际值之间的误差时，
5 利用最小二乘法来最小化预测值（含噪声值）与实际值之间的误差。

需要注意的是，上述基于交叉验证法将样本数据划分为训练集、评估集和测试集的步骤可替换为：基于交叉验证法将样本数据划分为训练集和测试集。且训练集、评估集和测试集中样本数据的数量可根据需要设置，并不限于上述例举的方案。

10 S4、根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；

具体地，该步骤可以细化为以下步骤：

根据各舆情因子对应的第二信息系数，分别确定所述 M 只股票的各舆情因子对应的权重；及

15 根据所述舆情因子观测值及各舆情因子所占的权重，分别计算每只股票的评分。

在 t 时刻，利用信息系数预测模型预测得到 M 只股票的舆情热度因子、舆情情感因子未来一期的信息系数分别为 $IC_{A(t+1)}$ 、 $IC_{B(t+1)}$ 后，需要确定舆情热度因子、舆情情感因子在计算股票评分时分别对应的权重。

20 在确定各舆情因子对应的权重之前，需判断各舆情因子的因子种类，其中，因子种类包括：正向因子和负向因子。具体地，舆情热度因子为正向因子，表示股票的舆情热度观测值越高，该股票的未来收益会比较高；舆情情感因子既可能是正向因子，又可能是负向因子，需要说明的是，舆情情感观测值的取值范围为 $[-1,1]$ ，舆情情感观测值的绝对值越高，正面(负面)的情感越强烈，当某只股票的舆情情感观测值为负时，对该股票的未来收益存在负面影响，此时，舆情情感因子为负向因子；相反，当某只股票的舆情情感观测值为正时，对该股票的未来收益存在正面影响，此时，舆情情感因子为正向因子。

具体地，对于正向因子，当 $IC_{i(t+1)} > 0$ 时，该因子 i 对应的第一权重
30 $\omega_{i1} = IC_{i(t+1)}$ ，否则， $\omega_{i1} = 0$ ，表示该因子在未来一期失效；对于负向因子，

当 $IC_{i(t+1)} < 0$ 时, 该因子 i 对应的第一权重 $\omega_{i1} = -IC_{i(t+1)}$, 否则, $\omega_{i1} = 0$, 表示该因子在未来一期失效。

利用上述步骤确定舆情热度因子、舆情情感因子对应的第一权重分别为 ω_{A1} 、 ω_{B1} , 为了便于后续计算股票对应的评分, 对股票的舆情热度因子及舆情情感因子对应的第一权重进行归一化, 根据归一化结果确定股票的两个因子对应的第二权重 ω_{A2} 、 ω_{B2} , 作为计算股票评分时两个因子对应的权重 ω_A 、 ω_B 。在本实施例中, ω_A 、 ω_B 的计算公式为:

$$\omega_A = \omega_{A2} = \omega_{A1} / (\omega_{A1} + \omega_{B1})$$

$$\omega_B = \omega_{B2} = \omega_{B1} / (\omega_{A1} + \omega_{B1})$$

在其他实施例中, 为了防止所有的因子都失效, 即所有因子对应的权重均为0的情况, 在对各因子的第一权重进行归一化之前, 采用smoothing的方法, 在每个因子的第一权重的基础上加上 ω_0 , 然后再进行归一化处理。此时, ω_A 、 ω_B 的计算公式为:

$$\omega_A = \omega_{A2} = (\omega_{A1} + \omega_0) / (\omega_{A1} + \omega_{B1} + 1)$$

$$\omega_B = \omega_{B2} = (\omega_{B1} + \omega_0) / (\omega_{A1} + \omega_{B1} + 1)$$

$$\omega_0 = 1/c$$

其中, c 为因子总数。在本实施例中, 考虑的舆情因子只有两个: 舆情热度因子及舆情情感因子, 故 $\omega_0 = 1/2$ 。

也就是说, 当舆情热度因子、舆情情感因子都失效时, 采用的就是因子等权重策略, 即每个因子对应的权重都相同。

分别获取 M 只股票的舆情热度观测值 X_A 、舆情情感观测值 X_B 、舆情热度因子 A 对应的权重 ω_A 及舆情情感因子 B 对应的权重 ω_B , 根据预设的计算公式, 计算 M 只股票的评分。具体地, 股票的评分的计算公式为:

$$S = \sum \omega_i * X_i$$

其中, i 为股票的舆情因子, $i = \{A, B\}$, ω_i 为股票的舆情因子 i 对应的权重, $\omega_i = \{\omega_A, \omega_B\}$, X_i 为 t 时刻股票的各舆情因子的观测值, $X_i = \{X_A, X_B\}$ 。

S5、根据评分的高低顺序对所述多只股票进行排序, 筛选出排序靠前的第一预设数量的股票, 作为目标投资股票。

可以理解的是, 利用上述步骤计算得到的股票的评分, 在一定程度上可以反映出股票在未来收益的走势, 股票的评分越高, 其未来收益可能越高。

因此，在得到 M 只股票中的每只股票评分后，按照评分高低顺序对 M 只股票进行排序，从 M 只股票中筛选出评分最高的第一预设数量（例如，N 只）的股票，作为待投资的目标股票。

在其他实施例中，确定包含待投资的目标股票的投资组合后，还需要分别确定各只目标股票占投资总额的比重，例如，对目标股票进行等权投资，或者，根据目标股票的评分确定相应的比重，评分高的股票对应的比重较大，评分低的股票对应的比重较小），具体操作可根据投资者的实际需求进行设置。最后，根据目标股票及目标股票对应的比重确定最终投资组合。

需要说明的是，本申请的方案还适用于对已有投资组合进行优化，分别计算已有投资组合中各股票的评分，选择评分较高的股票作为目标股票，并对目标股票所占的比重进行调整，得到优化后的投资组合。

上述实施例提出的基于舆情因子的投资选股方法，根据股票舆情因子观测值及收益率数据，计算股票各舆情因子的第一信息系数，利用第一信息系数及信息系数预测模型预测第二信息系数，反映各舆情因子对股票未来收益的预测能力；根据第二信息系数对各舆情因子进行赋权，实现了对各舆情因子的动态赋权，提高了各舆情因子的可靠性；根据各舆情因子观测值及相应的权重计算股票评分，选择评分较高的股票作为目标投资股票，提高了股票评分的真实性，直观地展示了舆情因子观测值对股票未来收益的影响，有助于投资者控制风险、提高投资收益。

此外，本申请实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有基于舆情因子的投资选股程序，所述基于舆情因子的投资选股程序被处理器执行时实现如下操作：

获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；及

根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

本申请之计算机可读存储介质的具体实施方式与上述基于舆情因子的投资选股方法的具体实施方式大致相同，在此不再赘述。

5

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，
10 由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很
15 多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本申请各个实施例所述的方法。

20

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权利要求书

1、一种基于舆情因子的投资选股方法，应用于电子装置，其特征在于，该方法包括：

5 获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

10 根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；及

根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

2、根据权利要求 1 所述的基于舆情因子的投资选股方法，其特征在于，所述第一信息系数的计算公式为：

$$\text{Rank IC} = \text{corr}(\text{order}_{t-1}^f - \text{order}_t^r)$$

15 其中， order_{t-1}^f ，为 t-1 时刻各股票的舆情因子 f 的观测值排名， order_t^r ，为 t 时刻各股票的收益率排名。

3、根据权利要求 2 所述的基于舆情因子的投资选股方法，其特征在于，所述预先确定的信息系数预测模型的训练步骤包括：

20 分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率，计算各舆情因子在每个时刻对应的信息系数，以获取样本数据；及

将样本数据划分为训练集、评估集和测试集，并利用训练集的样本数据对神经网络进行训练，得到信息系数预测模型，利用测试集的样本数据对信息系数预测模型的准确率进行测试，得到最终的信息系数预测模型。

25 4、根据权利要求 2 所述的基于舆情因子的投资选股方法，其特征在于，所述“根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分”的步骤包括：

根据各舆情因子对应的第二信息系数，分别确定所述 M 只股票的各舆情因子对应的权重；及

30 根据所述舆情因子观测值及各舆情因子所占的权重，分别计算每只股票

的评分。

5、根据权利要求 4 所述的基于舆情因子的投资选股方法，其特征在于，所述预先确定的信息系数预测模型的训练步骤包括：

5 分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率，计算各舆情因子在每个时刻对应的信息系数，以获取样本数据；及

将样本数据划分为训练集、评估集和测试集，并利用训练集的样本数据对神经网络进行训练，得到信息系数预测模型，利用测试集的样本数据对信息系数预测模型的准确率进行测试，得到最终的信息系数预测模型。

10 6、如权利要求 4 所述的基于舆情因子的投资选股方法，其特征在于，所述各只股票的评分的计算公式为：

$$S = \sum \omega_i * X_i$$

其中，i 为股票的舆情因子， ω_i 为股票的舆情因子 i 对应的权重， X_i 为 t 时刻股票的各舆情因子的观测值。

15 7、根据权利要求 6 所述的基于舆情因子的投资选股方法，其特征在于，所述预先确定的信息系数预测模型的训练步骤包括：

分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率，计算各舆情因子在每个时刻对应的信息系数，以获取样本数据；及

20 将样本数据划分为训练集、评估集和测试集，并利用训练集的样本数据对神经网络进行训练，得到信息系数预测模型，利用测试集的样本数据对信息系数预测模型的准确率进行测试，得到最终的信息系数预测模型。

8、一种电子装置，其特征在于，该电子装置包括：存储器、处理器，所述存储器上存储有基于舆情因子的投资选股程序，所述基于舆情因子的投资选股程序被所述处理器执行时实现以下步骤：

25 获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；及

根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

- 5 9、根据权利要求 8 所述的电子装置，其特征在于，所述第一信息系数的计算公式为：

$$\text{Rank IC} = \text{corr}(\text{order}_{t-1}^f - \text{order}_t^f)$$

其中， order_{t-1}^f ，为 t-1 时刻各股票的舆情因子 f 的观测值排名， order_t^f ，为 t 时刻各股票的收益率排名。

- 10 10、根据权利要求 9 所述的电子装置，其特征在于，所述预先确定的信息系数预测模型的训练步骤包括：

分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率，计算各舆情因子在每个时刻对应的信息系数，以获取样本数据；及

- 15 将样本数据划分为训练集、评估集和测试集，并利用训练集的样本数据对神经网络进行训练，得到信息系数预测模型，利用测试集的样本数据对信息系数预测模型的准确率进行测试，得到最终的信息系数预测模型。

11、根据权利要求 9 所述的电子装置，其特征在于，所述“根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分”的步骤包括：

- 20 根据各舆情因子对应的第二信息系数，分别确定所述 M 只股票的各舆情因子对应的权重；及

根据所述舆情因子观测值及各舆情因子所占的权重，分别计算每只股票的评分。

- 25 12、根据权利要求 11 所述的电子装置，其特征在于，所述预先确定的信息系数预测模型的训练步骤包括：

分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率，计算各舆情因子在每个时刻对应的信息系数，以获取样本数据；及

将样本数据划分为训练集、评估集和测试集，并利用训练集的样本数据

对神经网络进行训练，得到信息系数预测模型，利用测试集的样本数据对信息系数预测模型的准确率进行测试，得到最终的信息系数预测模型。

13、根据权利要求 11 所述的电子装置，其特征在于，所述各只股票的评分的计算公式为：

$$S = \sum \omega_i * X_i$$

5 其中， i 为股票的舆情因子， ω_i 为股票的舆情因子 i 对应的权重， X_i 为 t 时刻股票的各舆情因子的观测值。

14、根据权利要求 13 所述的电子装置，其特征在于，所述预先确定的信息系数预测模型的训练步骤包括：

10 分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率，计算各舆情因子在每个时刻对应的信息系数，以获取样本数据；及

将样本数据划分为训练集、评估集和测试集，并利用训练集的样本数据对神经网络进行训练，得到信息系数预测模型，利用测试集的样本数据对信息系数预测模型的准确率进行测试，得到最终的信息系数预测模型。

15 15、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有基于舆情因子的投资选股程序，所述基于舆情因子的投资选股程序被处理器执行时实现如下步骤：

获取预先确定的多只股票在第一预设时间内的舆情因子观测值及第二预设时间内的收益率；

20 根据所述舆情因子观测值及收益率，计算各舆情因子的第一信息系数；

分别将各舆情因子的第一信息系数输入预先训练好的信息系数预测模型中，预测未来各舆情因子的第二信息系数；

根据各舆情因子观测值及各舆情因子的第二信息系数，分别计算所述多只股票的评分；及

25 根据评分的高低顺序对所述多只股票进行排序，筛选出排序靠前的第一预设数量的股票，作为目标投资股票。

16、根据权利要求 15 所述的计算机可读存储介质，其特征在于，所述第一信息系数的计算公式为：

$$\text{Rank IC} = \text{corr}(\text{order}_{t-1}^f - \text{order}_t^f)$$

其中, order_{t-1}^f , 为 t-1 时刻各股票的舆情因子 f 的观测值排名, order_t^r , 为 t 时刻各股票的收益率排名。

17、根据权利要求 16 所述的计算机可读存储介质, 其特征在于, 所述预先确定的信息系数预测模型的训练步骤包括:

5 分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率, 计算各舆情因子在每个时刻对应的信息系数, 以获取样本数据; 及

将样本数据划分为训练集、评估集和测试集, 并利用训练集的样本数据对神经网络进行训练, 得到信息系数预测模型, 利用测试集的样本数据对信息系数预测模型的准确率进行测试, 得到最终的信息系数预测模型。

18、根据权利要求 16 所述的计算机可读存储介质, 其特征在于, 所述“根据各舆情因子观测值及各舆情因子的第二信息系数, 分别计算所述多只股票的评分”的步骤包括:

15 根据各舆情因子对应的第二信息系数, 分别确定所述 M 只股票的各舆情因子对应的权重; 及

根据所述舆情因子观测值及各舆情因子所占的权重, 分别计算每只股票的评分。

19、如权利要求 18 所述的计算机可读存储介质, 其特征在于, 所述各只股票的评分的计算公式为:

$$S = \sum \omega_i * X_i$$

20 其中, i 为股票的舆情因子, ω_i 为股票的舆情因子 i 对应的权重, X_i 为 t 时刻股票各舆情因子的观测值。

20、根据权利要求 19 所述的计算机可读存储介质, 其特征在于, 所述预先确定的信息系数预测模型的训练步骤包括:

25 分别采集预先确定的第二预设数量的股票在第三预设时间内各舆情因子在每个时刻的历史观测值、及各只股票在每个时刻的历史收益率, 计算各舆情因子在每个时刻对应的信息系数, 以获取样本数据; 及

将样本数据划分为训练集、评估集和测试集, 并利用训练集的样本数据对神经网络进行训练, 得到信息系数预测模型, 利用测试集的样本数据对信息系数预测模型的准确率进行测试, 得到最终的信息系数预测模型。

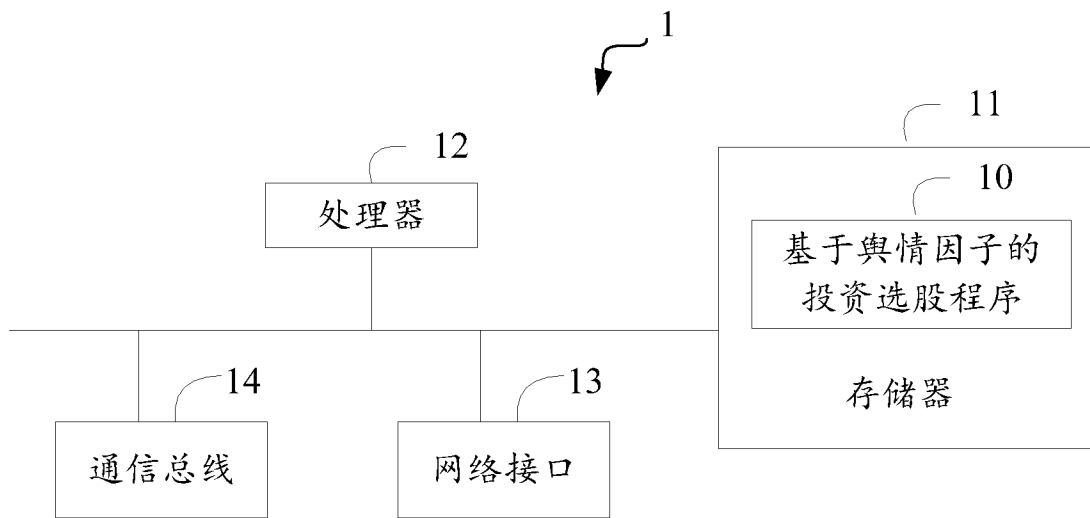


图 1



图 2

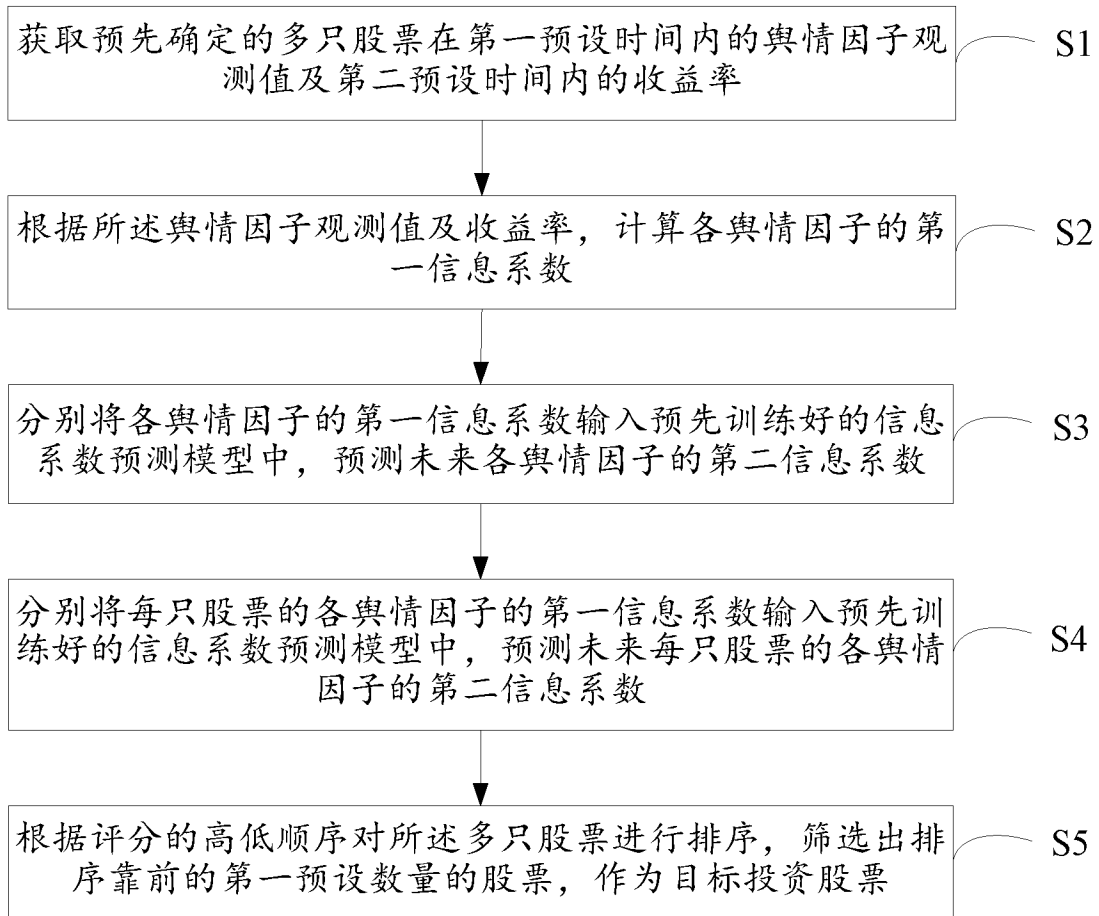


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/102127

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 10/06(2012.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; DWPI; IEEE: 股票, 投资, 舆情, 新闻, 预测, 收益; stock, invest, public sentiment, news, predict, profit

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106022522 A (NANJING UNIVERSITY) 12 October 2016 (2016-10-12) description, paragraphs [0008]-[0028], [0042], and [0049]	1-20
A	CN 105373853 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 02 March 2016 (2016-03-02) entire document	1-20
A	CN 107403017 A (SHANGHAI SHUDAN INFORMATION TECH CO., LTD.) 28 November 2017 (2017-11-28) entire document	1-20

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

19 December 2018

Date of mailing of the international search report

24 January 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/102127

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	106022522	A	12 October 2016	None	
CN	105373853	A	02 March 2016	None	
CN	107403017	A	28 November 2017	None	

国际检索报告

国际申请号

PCT/CN2018/102127

A. 主题的分类

G06Q 10/06 (2012.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

CNABS; CNTXT; CNKI; DWPI; IEEE: 股票, 投资, 舆情, 新闻, 预测, 收益; stock, invest, public sentiment, news, predict, profit

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 106022522 A (南京大学) 2016年 10月 12日 (2016 - 10 - 12) 说明书第[0008]-[0028], [0042], [0049]段	1-20
A	CN 105373853 A (百度在线网络技术北京有限公司) 2016年 3月 2日 (2016 - 03 - 02) 全文	1-20
A	CN 107403017 A (上海数旦信息技术有限公司) 2017年 11月 28日 (2017 - 11 - 28) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2018年 12月 19日

国际检索报告邮寄日期

2019年 1月 24日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10) 62019451

受权官员

李原野

电话号码 62089967

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/102127

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106022522	A	2016年 10月 12日	无	
CN	105373853	A	2016年 3月 2日	无	
CN	107403017	A	2017年 11月 28日	无	