(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 106445988 A
(43)申请公布日 2017.02.22

(21)申请号 201610382955.7

(22)申请日 2016.06.01

(71)申请人 上海坤士合生信息科技有限公司
地址 201203 上海市浦东新区中国(上海)自由贸易试验区芳春路400号1幢301-254室

(72)发明人 程明强 蒋朦 曹国梁 耿志贤

(74)专利代理机构 上海大邦律师事务所 31252
代理人 郜少毅

(51)Int.Cl.
G06F 17/30(2006.01)

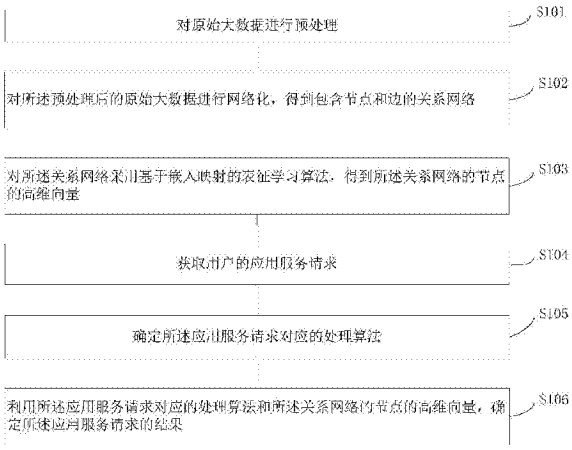
权利要求书3页 说明书19页 附图3页

(54)发明名称

一种大数据的智能处理方法和系统

(57)摘要

本发明实施例提供了一种大数据的智能处理方法和系统，该系统包括数据结构化模块用于对原始大数据进行预处理，以及对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络；表征学习模块用于对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量；应用算法模块用于获取用户的应用服务请求，确定所述应用服务请求对应的处理算法，以及利用所述应用服务请求对应的处理算法和所述表征学习模块得到的所述关系网络的节点的高维向量，确定所述应用服务请求的结果。本发明实施例所述的系统可以有效地提取大数据中的特征信息，并且统一用高维向量的形式进行表示，计算效率高，准确性高，对用户请求响应灵敏，可为多种应用服务提供统一有效地处理方法。



CN 106445988 A

1.一种大数据的智能处理系统，其特征在于，包括：

数据结构化模块，用于对原始大数据进行预处理，以及对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络；

表征学习模块：用于对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量；

应用算法模块：用于获取用户的应用服务请求；确定所述应用服务请求对应的处理算法，以及利用所述应用服务请求对应的处理算法和所述表征学习模块得到的所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

2.根据权利要求1所述的系统，其特征在于，所述关系网络中包含高维关系网络，则所述表征学习模块具体用于对所述高维关系网络进行嵌入映射，得到所述高维关系网络的节点的高维向量。

3.根据权利要求1所述的系统，其特征在于，所述关系网络中包含语义网络，则所述表征学习模块具体用于对所述语义网络进行嵌入映射，得到所述语义网络的节点的高维向量。

4.根据权利要求2或3所述的系统，其特征在于，所述关系网络中包含二维关系网络，则所述表征学习模块具体用于对所述二维关系网络进行嵌入映射，得到所述二维关系网络的节点的高维向量。

5.根据权利要求1所述的系统，其特征在于，所述原始大数据包括行为数据、属性数据和文本数据。

6.根据权利要求1或5所述的系统，其特征在于，所述数据结构化模块具体用于对所述预处理后的原始大数据中的行为数据进行网络化，得到包含节点和边的行为网络；

对所述预处理后的原始大数据中的属性数据进行网络化，得到包含节点和边的属性网络；以及，

对所述预处理后的原始大数据中的文本数据进行网络化，得到包含节点和边的语义网络；

其中，所述行为网络、所述属性网络和所述语义网络共同组成了所述关系网络。

7.根据权利要求1所述的系统，其特征在于，所述数据结构化模块具体用于对所述原始大数据进行数据分析和清理。

8.根据权利要求1所述的系统，其特征在于，所述应用算法模块具体用于利用所述关系网络中的部分节点的高维向量，以及所述应用服务请求对应的处理算法，确定所述应用服务请求的结果。

9.一种大数据的智能处理系统，其特征在于，包括：

获取模块，用于获取用户的应用服务请求以及由原始大数据转化而来的关系网络的节点的高维向量；

确定模块，用于确定所述应用服务请求对应的处理算法，利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

10.根据权利要求9所述的系统，其特征在于，所述由原始大数据转化而来的关系网络为：由所述原始大数据经过预处理之后进行网络化所得到的关系网络。

11.一种大数据的智能处理方法，其特征在于，包括：

对原始大数据进行预处理；

对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络；

对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量；

获取用户的应用服务请求；

确定所述应用服务请求对应的处理算法；

利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

12.根据权利要求11所述的方法，其特征在于，所述关系网络中包含高维关系网络，则所述对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量，包括：

对所述高维关系网络进行嵌入映射，得到所述高维关系网络的节点的高维向量。

13.根据权利要求11所述的方法，其特征在于，所述关系网络中包含语义网络，则所述对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量，包括：

对所述语义网络进行嵌入映射，得到所述语义网络的节点的高维向量。

14.根据权利要求12或13所述的方法，其特征在于，所述关系网络中包含二维关系网络，则所述对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量，包括：

对所述二维关系网络进行嵌入映射，得到所述二维关系网络的节点的高维向量。

15.根据权利要求11所述的方法，其特征在于，所述原始大数据包括行为数据、属性数据和文本数据。

16.根据权利要求11或15所述的方法，其特征在于，所述对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络，包括：

对所述预处理后的原始大数据中的行为数据进行网络化，得到包含节点和边的行为网络；

对所述预处理后的原始大数据中的属性数据进行网络化，得到包含节点和边的属性网络；以及，

对所述预处理后的原始大数据中的文本数据进行网络化，得到包含节点和边的语义网络；

所述行为网络、所述属性网络和所述语义网络共同组成了所述关系网络。

17.根据权利要求11所述的方法，其特征在于，所述对原始大数据进行预处理包括对所述原始大数据进行数据分析和清理。

18.根据权利要求11所述的方法，其特征在于，所述利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果，包括：

利用所述关系网络中的部分节点的高维向量，以及所述应用服务请求对应的处理算法，确定所述应用服务请求的结果。

19.一种大数据的智能处理方法，其特征在于，包括：

获取用户的应用服务请求以及由原始大数据转化而来的关系网络的节点的高维向量；

确定所述应用服务请求对应的处理算法；

利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

20.根据权利要求19所述的方法，其特征在于，所述由原始大数据转化而来的关系网络为：由所述原始大数据经过预处理之后进行网络化所得到的关系网络。

一种大数据的智能处理方法和系统

技术领域

[0001] 本发明实施例涉及计算机技术领域，尤其涉及一种大数据的智能处理方法和系统。

背景技术

[0002] 保险行业正因科技进步而发生巨大的改变，大数据的广泛应用改变了保险公司实现服务的方式。现有的保险业网站和软件通常收集了海量数据，蕴藏着大量有用信息，包括用户的个人信息、消费习惯等。只有充分利用保险业大数据，才能在风险定价、产品设计、营销策略、客户服务、风险管控等诸多方面适应大数据时代的要求。

[0003] 当前在保险业行业中，通常采用数据库系统对保险业数据进行存储和管理。数据库系统中通常采用表格的方式存储数据，表格中会存在大量的关系数据和文本信息，存储的数据的格式也可以是多种多样的。比如，用户的个人简介和产品的描述信息通常在数据库中用文本字符串的形式进行存储，而用户的年龄和产品价格通常采用非负数字的形式进行存储。虽然当前的数据处理技术能够对格式化的数字和类别等数值进行提取和匹配，但是对文本等非结构化数据却无法从中提取出有用的特征信息。

[0004] 常见的保险业业务包括保险业数据的产品精准推荐、购险用户分类和欺诈骗保检测等。在保险业营销服务中，要么是让用户通过搜索获取保险产品进而购买，要么采用流行度推荐、关联规则推荐和协同过滤推荐等方法来给用户主动推荐保险产品。其中，流行度推荐是指给用户推荐当前最流行的保险产品，缺点是缺乏个性化考虑，准确性低。关联规则推荐是通过数据分析，学习用户购买兴趣与自身特征和产品特征之间的规则，例如40岁以上的女性更易购买健康类保险，推荐的准确性也不高。协同过滤推荐是基于一个基本假设，对相似的保险产品有过兴趣的用户此后会购买相似保险产品，被相似用户购买的产品此后还会被相似的用户购买，这种推荐在单一用户的行为很少时，存在数据稀疏度高，无法进行有效计算和推荐。

[0005] 在进行购险用户分类时，由于用户类别可以描述用户的生活习惯、交友习惯、消费习惯等，不同的类别需要提取不同的用户特征。通常采用的方式是从用户的消费记录中提取诸如用户月收入、月花销、年度的收入标准差、年度的花销标准差等特征，通过标记大量的用户类别标签，训练监督学习模型，对测试用户进行分类。这种方法既需要依靠经验提取大量特征，更需要收集大量的标记数据，会造成代价高、准确度差等问题。

[0006] 欺诈骗保检测，即判断某用户的申保行为是否是欺诈行为，最核心任务是收集用户在申报行为中的特征。现有的欺诈骗保检测系统主要是从包括用户个人信息、所申保的保险产品信息、申保流程信息等中提取大量的数值统计结果，同时对其中一部分用户进行标注，利用人力判断是否是欺诈用户，继而训练监督学习模型，对申保行为进行分类。然而，该系统需要依靠经验提取特征并收集标记数据，造成无法有效实施。

[0007] 由此可见，现有保险业大数据的智能处理系统至少具有如下缺点：1)现有保险业数据技术缺乏对非结构化数据的分析，丢失了大量有效信息，影响保险业业务的分析结果；

2)现有的保险业推荐系统、购险用户分类系统和欺诈骗保检测系统等过分依赖于人力的特征提取，准确性低、计算效率差，对用户请求响应缓慢，影响用户体验；3)不同的保险业服务通常采用不同的数据处理和特征提取方法，造成大量的冗余数据处理，并且不同服务的数据单元的特征不相兼容。

发明内容

[0008] 本发明实施例的目的在于提供一种大数据的智能处理方法和系统，能够从多种大数据源中有效地提取特征信息，无需人工参与，并且计算效率高，准确性高，对用户请求响应灵敏，可为多种应用服务提供统一有效地处理方法。

[0009] 本发明实施例采用的技术方案如下：

[0010] 本发明实施例系统了一种大数据的智能处理系统，该系统包括数据结构化模块、表征学习模块和应用算法模块；

[0011] 其中，所述数据结构化模，用于对原始大数据进行预处理，以及对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络；

[0012] 所述表征学习模块用于对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量；

[0013] 所述应用算法模块用于获取用户的应用服务请求；确定所述应用服务请求对应的处理算法，以及利用所述应用服务请求对应的处理算法和所述表征学习模块得到的所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

[0014] 可选地，所述关系网络中包含高维关系网络，则所述表征学习模块具体用于对所述高维关系网络进行嵌入映射，得到所述高维关系网络的节点的高维向量。

[0015] 可选地，所述关系网络中包含语义网络，则所述表征学习模块具体用于对所述语义网络进行嵌入映射，得到所述语义网络的节点的高维向量。

[0016] 可选地，所述数据结构化模块具体用于对所述预处理后的原始大数据中的行为数据进行网络化，得到包含节点和边的行为网络；

[0017] 对所述预处理后的原始大数据中的属性数据进行网络化，得到包含节点和边的属性网络；以及，

[0018] 对所述预处理后的原始大数据中的文本数据进行网络化，得到包含节点和边的语义网络；

[0019] 其中，所述行为网络、所述属性网络和所述语义网络共同组成了所述关系网络。

[0020] 本发明实施例还提供了一种大数据的智能处理方法，该方法包括：

[0021] 对原始大数据进行预处理；

[0022] 对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络；

[0023] 对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量；

[0024] 获取用户的应用服务请求；

[0025] 确定所述应用服务请求对应的处理算法；

[0026] 利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

[0027] 可选地，所述关系网络中包含高维关系网络，则所述对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量，包括：对所述高维关系网络进行嵌入映射，得到所述高维关系网络的节点的高维向量。

[0028] 可选地，所述关系网络中包含语义网络，则所述对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量，包括：对所述语义网络进行嵌入映射，得到所述语义网络的节点的高维向量。

[0029] 可选地，所述对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络，包括：对所述预处理后的原始大数据中的行为数据进行网络化，得到包含节点和边的行为网络；

[0030] 对所述预处理后的原始大数据中的属性数据进行网络化，得到包含节点和边的属性网络；以及，

[0031] 对所述预处理后的原始大数据中的文本数据进行网络化，得到包含节点和边的语义网络；

[0032] 所述行为网络、所述属性网络和所述语义网络共同组成了所述关系网络。

[0033] 本发明实施例还提供了一种大数据的智能处理方法，包括：

[0034] 获取用户的应用服务请求以及由原始大数据转化而来的关系网络的节点的高维向量；

[0035] 确定所述应用服务请求对应的处理算法；

[0036] 利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

[0037] 可选地，所述由原始大数据转化而来的关系网络为：由所述原始大数据经过预处理之后进行网络化所得到的关系网络。

[0038] 本发明实施例的技术方案具有以下优点：所述数据结构化模块能够对原始大数据进行预处理以及网络化，使得所述原始大数据转化为网络数据或者结构数据，从而所述表征学习模块可以利用网络数据的表征学习算法，来实现对数据的快速、统一的特征提取，并以高维向量的形式进行表示；所述应用算法模块可以根据用户的应用服务请求，确定对应的处理算法，并利用所述表征学习模块提取到的以向量形式表示的特征进行计算，确定处理结果。不同于现有技术，本发明实施例中整个特征提取的过程无需人的参与，利用基于嵌入映射的表征学习算法自动完成，计算效率高；特征提取的过程中还极大地保留了原始大数据中的结构信息(即有效信息)，从而提高了进行分类或预测等任务的准确性；不仅如此，由于采用了基于嵌入映射的表征学习算法，使得从原始大数据中挖掘到的数据特征统可以统一由高维向量的形式进行表示，从而本发明实施例中的系统不仅限于为某个特定的应用服务，可以为多种应用服务提供统一有效地处理方法。

附图说明

[0039] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0040] 图1为本发明实施例提供的一种大数据的智能处理方法的流程图；

[0041] 图2为一种行为网络的结构示意图；

[0042] 图3为本发明实施例提供的又一种大数据的智能处理方法的流程图；

[0043] 图4为本发明实施例提供的一种大数据的智能处理系统的结构组成示意图；

[0044] 图5为本发明实施例提供的又一种大数据的智能处理系统的结构组成示意图。

具体实施方式

[0045] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0046] 为了更好的解释本发明实施例，在对本发明实施例进行描述之前，对相关概念进行解释。

[0047] 数据单元是指表示关系数据时不可分的基本单元，比如某个“客户或者用户”，某个“年龄段”，某一个“产品”，某一种“产品分类”等。这些基本单元在生活中是有实体的。与数据单元相对的是非数据单元，是指客户关系、客户对产品的行为、产品同属某一系列等组成这些数据单元的结构。

[0048] 行为数据是指用户对产品发生行为所产生的数据，例如用户购买、退订或评价某保险产品而产生的数据。行为数据描述了两个或多个数据单元之间的关系，通常描述的是“用户”与“产品”之间的关系。

[0049] 属性数据是指用户、产品等数据单元和其属性之间的关系，例如用户的年龄、产品的种类等。属性数据描述了数据单元与其属性的关系，通常描述的是“用户”与其属性，或“产品”与其属性”之间的关系。

[0050] 文本数据是指含有词汇或短语的文本。可以以词汇或者短语作为数据单元。

[0051] 结构化数据是指能够用数据或统一的结构加以表示的数据，如数字或者符号，存储在数据库里可以用二维表结构来逻辑表达实现。

[0052] 非结构化数据，相对于结构化数据而言，是指无法用数字或统一的结构表示的数据，不方便用数据库二维逻辑表来表现，例如文本、图像、声音、网页、各类报表等。

[0053] 高维关系是指该关系牵涉到多个数据单元(或者指网络中的多个节点)，是多个数据单元的交互。二维关系是仅有两个数据单元的交互。购买行为在信息富足的情况下是高维关系的行为，通常可能包括用户、产品、购买地点和购买方式等，但如果信息收集不完全，有可能只是二维关系的行为，比如仅含有用户和产品。传统数据处理系统仅能够考虑到二维关系的行为，但无法处理高维关系的行为。而高维关系的行为产生的高维关系的数据在当前的各个领域中是普遍存在的。

[0054] 此外，随着网络技术的发展，使得非结构化数据的数量日趋增大。这时，仅能够对结构化数据进行管理和分析的数据处理系统的局限性暴露地越来越明显。不仅如此，在很多行业中，不仅限于保险业中，对大数据的特征提取仍需要利用专家，无法仅靠计算机来完成。对大数据进行处理的系统还普遍存在准确性低、计算效率差，对用户请求响应缓慢等一系列问题。

[0055] 为了解决上述问题，本发明实施例提供了一种大数据的智能处理方法，如图1所示，所述方法包括：

[0056] S101：对原始大数据进行预处理。

[0057] 原始大数据可以是通过各个网站或者应用程序(Application，APP)收集而来的，因而可能包括行为数据、属性数据等结构数据，也可能包括文本数据等非结构化数据，并且数据的格式也可能是多种多样的。因此，在对数据提取特征或者利用数据提供服务之前，可以先对原始大数据进行预处理。数据预处理的方法包括数据清理、数据集成、数据变换、数据分析和数据归约等。

[0058] 可选地，在本发明实施例中，对原始大数据进行预处理可以是，对所述原始大数据进行数据分析和清理，即对原始大数据进行统计分析，去除不合规或错误的数据内容，可以是将非法数据格式进行过滤，例如去除理应为浮点数、却填充为字符串型的价格等数值，还可以是将时间或者单位进行统一，也可以是对缺失的指进行填写、光滑噪声数据等，从而可以将大数据的格式标准化，清除异常数据，纠正错误或者清除重复数据等。

[0059] S102：对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络。

[0060] 所述关系网络中的节点，是由所述预处理后的原始大数据中的数据单元转化而来，所述关系网络中的边，用于表示所述网络中节点与节点之间的关系。

[0061] 大数据通常是以表格的形式进行存储，然而这种传统的数据存储方式，无法对数据进行统一地大规模的存储和管理，且会丢失大量的文本数据中所含有的语义信息(该语义信息是有用信息，对于向用户提供准确的应用服务至关重要)，最重要的是，碎片化的表格存储方式，无法方便快捷地被后续应用服务进行访问和利用，无法满足实现频度高、响应速度快的应用服务的需求。

[0062] 本发明实施例中，通过对原始大数据进行网络化，可以将表格中的大数据或海量数据转化为关系网络，有效解决了上述问题。首先，将预处理后的原始大数据网络化之后，可以采用节点和边的方式统一处理这些数据，大大缩减了数据存储和管理的成本。其次，针对预处理后的原始大数据中的词汇和短语等文本数据，将其进行网络化，构建出语义网络，保留了文本中的语义信息，以便后续可以有效利用，提高应用服务的准确性。此外，将预处理后的原始大数据表示为包含节点和边的关系网络之后，就可以利用网络数据的表征学习算法，来实现对数据的快速、统一的特征提取，从而做到快速响应不同应用服务请求。

[0063] 可选地，所述预处理后的原始大数据可以包括行为数据、属性数据和文本数据，则对所述预处理后的原始大数据进行网络化可以包括：对所述预处理后的原始大数据中的行为数据进行网络化，例如将购买、评价等行为数据转化为行为网络；或者，还可以包括对所述预处理后的原始大数据中的属性数据进行网络化，例如将年龄、价格等属性信息转化为属性网络；又或者还可以包括将对所述预处理后的原始大数据中的文本数据进行网络化，例如将产品介绍或者评价内容等文本数据转化为以词和短语为节点的语义网络。则所述行为网络、所述属性网络和所述语义网络共同组成了所述关系网络。

[0064] S103：对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量。

[0065] 表征学习是机器学习和数据挖掘中核心的研究问题之一。在本发明实施例中，通过对所述关系网络采用基于嵌入映射的表征学习算法，将所述关系网络中的节点，例如用

户、产品和短语等，统一用维度较高的向量来进行表示，并保留了原始大数据中的结构信息。其中，每个向量可以表示所述关系网络中的一个节点，该向量中的一个维度表示了该节点的一个特征。所述关系网络中节点与节点之间的关系(或者说边)，转化为节点的高维向量与节点的高维向量之间的相似度，如果节点1与节点2之间存在关系(即在所述关系网络中通过边连接)，则节点1的高维向量与节点2的高维向量之间的相似度高，反之，则相似度低。

[0066] 通过上述表征学习的方式，避免了现有技术中依赖于专家经验的人工特征提取方式，实现了以大数据为驱动而得到的符合数据规律的特征，并特征以向量的形式表示之后，使得后续可以直接应用于多种任务，包括分类、聚类、预测等。

[0067] 进一步地，采用基于嵌入映射的表征学习算法，能够尽可能地保留所述关系网络中的结构信息，并且针对不同的网络可以保留不同的结构信息。例如，对于“用户-产品”的行为网络，可以保留购买行为信息，使得向量中相似特征表示的用户具有相似的购买习惯，相似特征表示的产品具有相似的购买人群，比如可以选择高维向量中的50维向量来保存“购买行为关系”这种结构信息，使得存在“购买行为关系”这种结构的两个节点(用户与产品)对应的高维向量之间的向量相似度高，还可以选择高维向量中的另外的50维向量来保存“相似购买倾向”这种结果信息，使得存在“相似购买倾向”这种结构的两个节点这种结构的两个节点(用户与用户)对应的高维向量之间的向量相似度高。由此可知，这将会大大提升后期应用服务对应的分类和预测等任务的准确性，解决了现有技术中无法有效提取数据中的结构信息，丢失了大量有效信息的问题。

[0068] 此外，常见的学习方法是利用矩阵或张量分解获取节点的高维表示，然而这类方法往往面临复杂度过高(立方级别)的问题，无法广泛应用于海量数据的工业化场景中，并且计算效率也不高。而在本发明实施例中，采用嵌入映射的学校方法，该方法采用了负采样技术(Negative Sampling)，针对大量数据进行合理比例地采样学习，从而保证了学习工程能够用较少的时间达到较好的学习结果。并且通过将所述关系网络用高维向量进行表示之后，不仅可以可以缩短学习的时间，还可以大大的提高计算效率，快速响应用户的请求。

[0069] 表征学习算法的实现，除了利用基于嵌入映射，还有其他方式，比如奇异值分解、非负矩阵分解等，但这这些方法仅限于二维关系网络，并且计算速度也非常的缓慢。本发明实施例中，考虑到目前无论是保险行业、金融行业、购物和电商等等的应用场景中，收集到的大数据越来越趋于多样化，利用本发明实施例的技术处理后得到的关系网络，往往不仅限于二维关系网络，在绝大部分情况下是高维关系网络。数据的规模往往也相当的大，因此选用基于嵌入映射的表征学习算法，不仅可以应用于二维关系网络和多维关系网络，而且可以实现计算速度的加速，大大缩短计算时间，快速响应应用需求。

[0070] 具体地，可以采用“嵌入映射”的表征学习算法，利用范畴论中的“态射”实现保结构映射的降维“嵌入”来实现表征学习。即针对所述关系网络中的数据，通过保留所述关系网络中的结构信息的学习算法，获得节点的高维向量表示。

[0071] S104:获取用户的应用服务请求。

[0072] 用户在浏览网页，使用某个APP，或者点击某个操作界面的某个功能按钮等情况下，都有可能触发应用服务请求，因此可以获取该应用服务请求，以确定后续应该采用的相关算法。

[0073] S105：确定所述应用服务请求对应的处理算法。

[0074] S106：利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

[0075] 可以将应用层的服务定义为排序、分类、聚类、预测、关联分析和异常检测等任务，这些任务可以用特定的处理算法完成，根据表征学习后所得到的高维向量，利用上述任务对应的处理算法(即应用服务请求对应的处理算法)，就可以获得准确、高效的解决方案，并返回给用户。

[0076] 具体地，可以预先指定或者获取所述应用服务请求与处理算法之间的对应关系，例如当应用服务请求是产品推荐时，可以知道推荐产品实际就是进行预测，预测得到的用户最可能购买的一系列产品，处理算法即计算用户节点的高维向量与产品节点的高维向量的相似程度，那么如果预先指定或者获取该应用服务请求与该处理算法的对应关系，那么在收到该应用服务请求后，就可以确定所述应用服务请求对应的处理算法是计算用户节点的高维向量与产品节点的高维向量的相似程度。最后，利用用户节点的高维向量和产品节点的高维向量，进行相似度计算，就可以得到与用户相似度最高的一些列产品，即得到所述应用服务请求的结果。

[0077] 在本发明实施例中，通过预处理后的原始大数据进行网络化，得到包含节点和边的关系网络，并对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量，即实现了对原始大数据的特征提取，且整个过程不需要依靠专家的经验，无需人的参与，利用基于嵌入映射的表征学习算法自动完成，计算效率高。不同于现有技术，在本发明实施例中特征提取的过程中还极大地保留了有效信息，从而提高了后续的分类或预测等任务的准确性。进一步地，在本发明实施例中，由于数据的特征统一由高维向量的形式进行表示，使得可以根据应用服务请求，来确定处理算法，从而利用以高维向量进行表示的特征，来确定所述应用服务请求的结果，本发明实施例所述的大数据的智能处理方法，不仅限于某个特定的应用服务，可以为多种应用服务提供统一有效地处理方法。

[0078] 需要注意的是，本发明实施例所述的大数据的智能处理方法，不仅仅可以应用于保险业领域，还可以应用于其他领域，例如应用于金融领域、购物消费领域等，尤其适用于对包含结构数据和非结构数据进行处理的情况，以及需要处理高维关系的数据的场合，较现有技术将具有明显的优势。

[0079] 需要说明的是，S106中，利用所述关系网络的节点的高维向量，确定所述应用服务请求的结果时，可以利用所述关系网络的所有节点的高维向量去确定所述应用服务请求的结果；还可以只利用所述关系网络的部分的高维向量，去确定所述应用服务请求的结果。具体地，可以只利用与所述应用服务请求相关的节点去确定所述应用服务请求的结果。例如，当应用服务请求是产品推荐时，可以只利用产品节点的高维向量和用户节点的高维向量进行计算。

[0080] 可选地，在步骤S102中，具体该如何对行为数据进行网络化、对文本数据进行网络化或者对属性数据进行网络化，可以参照以下方式。

[0081] 1、对所述预处理后的原始大数据中的行为数据进行网络化

[0082] 具体地，行为数据描述了两个或多个数据单元之间的关系，对行为数据进行网络化是指将该关系表示成网络的边，数据单元表示为网络的节点。该网络可以是二维关系网

络，还可以是高维关系网络，相应地，对行为数据进行网络化时可以将关系表示成二维的边或高维的边。即将购买、退订或评价等行为表示为网络的的边。其中，二维的边是指边上含有两个节点，高维的边是指边上含有多个节点。

[0083] 举例来说：简略的用户的行为数据可以表示为“用户-产品”的二维关系形式。此外，用户行为可能还具有丰富的上下文信息，可以将上下文信息节点化后形成多元关系图，如“用户-产品-评价”的三维关系图。以张先生对保险产品A进行购买的行为为例，张先生购买该给予该保险产品A的评价为：价格虽然昂贵，但还是值得的。对上述数据进行行为数据网络化可以得到如图2所示的行为网络。在图2中，将“张先生”和“保险产品A”表示为该行为网络的节点，购买行为构成了上述两个节点之间的边。此外，评价的短语或词语——“昂贵”以及“值得的”，表示为网络的节点，该部分其实属于对文本数据进行网络化，将在后面的描述中进行详细解释。由此形成了“用户-产品-评价”的行为网络，也即三维关系网络。

[0084] 2、对所述预处理后的原始大数据中的文本数据进行网络化

[0085] 对文本数据进行网络化就是将词汇或短语组成的数据单元表示为网络的节点，从而将文本构建成以词汇或短语为节点的关系网络。网络中以词汇或短语组成的节点之间的边，描述了它们出现在句子或文档中的频度。例如，如果“昂贵”和“值得的”这两个短语共同出在了3个句子之中，则“昂贵”和“值得的”可以作为关系网络的两个节点，它们之间可以存在边进行连接，边的权重可以设置为3；如果网络中“昂贵”和“真便宜”从未在句子中共同出现，则这两个节点之间不存在边进行连接。另外，这些以词汇或短语组成的节点和其他节点(如用户、产品)等形成的边，属于行为数据，描述了两个或多个数据单元之间的关系。

[0086] 以上述张先生对保险产品A进行购买和评价为例，可以将评价内容等文本数据进行结构化，即进行分词、短语抽取、类别标注、情感分析等等，从而将自然语言表述成可以处理的数据结构。具体地，根据“价格虽然昂贵，但还是值得的”，可以获知“昂贵”和“值得的”是核心词汇，并且“昂贵”描述了产品在“价格”层面的特征，“值得的”反映了用户积极的购买心态和情感。因而在对该文本数据进行网络化时，将“昂贵”和“值得的”表示为网络的节点，这两个节点与其它节点，如用户以及产品形成的边，属于行为数据。

[0087] 由此可知，对文本数据进行网络化，不仅实现了对非结构化数据的分析，而且可以将词汇或短语等与行为数据进行关联，保留了一定的有用信息。

[0088] 3、对所述预处理后的原始大数据中的属性数据进行网络化

[0089] 属性数据描述了数据单元与其属性的关系，对属性数据进行网络化是指将该关系表示成网络的边，将数据单元表示为网络的节点。属性数据既可以是类别信息，例如健康险或者旅游险，还可以是年龄或者价格等数值信息。从而对属性数据进行网络化，可以是将类别信息表示为网络的节点，将年龄、价格等属性信息中的数值信息进行分区间后，进行节点化表示。

[0090] 例如，年龄为25岁的张先生，购买了价格为2000的保险产品。在这个例子中，可以将某个包含25岁的年龄区间表示为节点，如可以将年龄在24-30岁之间的青年表示为节点“青壮年”；可以将某个包含价格为2000的数值的价格区间表示为节点，如将价格在1000-5000之间表示为节点“入门级保险产品”。通过上述处理后，最终转化为“用户-年龄层”和“产品-价格区间”的属性网络。

[0091] 可选地，对所述预处理后的原始大数据进行网络化之后，可以对所述关系网络的

节点和边进行格式规整的大规模存储和管理，以方便后续的特征提取和使用。因此，在S102之后，还可以包括：

[0092] S102'：将所述关系网络的节点和边保存在数据库中。

[0093] 例如，所述数据库中可以存储两种表格用以分别保存所述关系网络的节点和边，保存节点信息的表格中每行是节点的ID、名称和查询频次等。保存边信息的表格中的每行是边的ID、相关节点的ID和产生时间等。对所述预处理后的原始大数据进行网络化之后，实际上将所有的网络化处理前的数据都转变为了结构化数据。在实际应用中，对于结构化数据进行管理(Structured Data Management)，存在若干种数据管理技术，比如分布式存储、云数据库、NOSQL数据库(非关系型数据库)和移动数据库等。例如BaseX、MongoDB和No2DB是分别依托Java、C++和C#语言开发成为流行的三种NO-SQL数据库；MySQL和HBase是常用数据库软件；网络关系存储中AllegroGraph、DEX、Neo4j和FlockDB是依托于SPARQL、Java和Scala的图形数据库。

[0094] 可选地，在实现步骤S103时，由于所述关系网络既可能包扩语义网络，也可能包括属性网络和行为网络。它们既可能属于同构关系网络、也可能属于二维关系网络，还可能属于高维关系网络。因此，对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量可以包括：对所述关系网络中的高维关系网络进行嵌入映射，得到所述高维关系网络的节点的高维向量；或者，对所述关系网络中的二维关系网络进行嵌入映射，得到所述二维关系网络的节点的高维向量；或者，对所述关系网络中的语义网络进行嵌入映射，得到所述语义网络的节点的高维向量；或者对所述关系网络中的同构网络进行行嵌入映射，得到所述同构网络的节点的高维向量。

[0095] 一、对所述语义网络进行嵌入映射(Text Embedding)

[0096] 利用嵌入映射的方法，将语义网络中的词和短语形式的节点表示为高维向量，并且通过嵌入映射后，使得节点中表示相近的词或者短语的节点的高维向量相似度很高，即使得相近的词与短语具有相似的语义。

[0097] 具体地，可以基于Skip-gram模型的词嵌入映射方法，通过学习词的向量表示，来达到精准预测临近词语的目的。最有效地学习目标(即最大化的目标函数)为：隐藏在句子中某个词语后，通过给定的句中临近的其他词语，可以得到最适合的被隐藏的词的向量。在自然语态下，能够填进隐藏的词语所在空缺的词语之间是具有相似语义的，则在进行嵌入映射时，使得它们的向量的相似度很高。

[0098] 简而言之，语义网络的嵌入映射最大化条件概率的目标函数是给定临近节点(相连接的节点)的向量，预测目标节点的向量，使得与给定的一些节点相连接的节点之间具有相似的向量。还可以进行进一步拓展，融入词、短语和短语类别等多种元素，实现语义层面的表征学习。

[0099] 选定训练的文本的上下文信息的规模c，也即窗口大小，将当前词w_t作为输入，将临近的单位元作为输出层的训练模型的最大化的目标函数为：

[0100] $$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} log\, p(w_{t+j}/w_t) \qquad (1)$$

[0101] 其中，w_i指代文本中的第i个词语。

[0102] 通过该最大化该目标函数，学习得到每一个词语的向量表示$w_{(i)}$，使得给定向量

$w_{(t)}$和位置t时，学习该目标函数就可以得到位置(t+j)的向量会和实际文档中该位置的词的向量的相似度很高(概率被最大化)，使得相近的词与短语具有相似的语义，让词语的语义能够被保留下来。

[0103] 例如，语义网络中出现“今天”、“中午”、“吃了”这几个临近的词语，可能来自原始大数据中的文本信息“今天中午吃了米饭”和“今天中午吃了白饭”。采用本发明实施例的方法，此时“白饭”、“米饭”的向量就是$w_{(t)}$，“今天”、“中午”、“吃了”的向量就是$w_{(t+j)}$，也就是$w_{(t-3)}$，$w_{(t-2)}$，$w_{(t-1)}$，通过基于嵌入映射的表征学习算法，得到“白饭”和“米饭”对应的向量的相似度很高，即“白饭”和“米饭”这两个语或者短语具有相似语义。而在现有技术中，由于“白饭”和“米饭”是两个不同词语，则认为“白饭”和“米饭”是不同的，无法保留语义信息。

[0104] 二、对所述二维关系网络进行嵌入映射(Bipartite Network Embedding)

[0105] 二维关系网络是指网络中每一条边都对应两个节点，并且网络中的节点只有两类，比如“用户-产品”就是一种二维关系网络。

[0106] 对所述二维关系网络进行嵌入映射是指利用嵌入映射的方法，将具有二维关系(如用户-产品、用户-年龄、产品-价格等)的行为网络和属性网络中的节点(如用户、产品、年龄层、价格层等节点)表示为高维向量。

[0107] 于语义网络的嵌入映射一样，二维关系网络的嵌入映射，最大化条件概率的目标函数是给定临近节点(相连接的节点)的向量，预测目标节点的向量，使得与给定的一些节点v_j相连接的节点v_i之间具有相似的向量。

[0108] 假设二维关系网络中含有A类节点和B类节点。则通过该最大化该目标函数，可以在给定B类节点v_j时，得出的与v_j相连接的节点的向量，会和A类节点v_i的向量相似，即条件概率最大化。

[0109] 可以定义由B类节点中的v_j能够产生A类节点的v_i表示的条件概率为：

[0110] $$p(v_i/v_j)=\frac{exp(\vec{u}_i^T\cdot\vec{u}_j)}{\sum_{i'\in A}exp(\vec{u}_i^T\cdot\vec{u}_j)} \qquad (2)$$

[0111] 其中u_i是v_i的高维向量，u_j是v_j的高维向量。

[0112] 以“用户-产品”组成的二维关系网络为例，假设A类节点表示用户，B类节点表示产品，那么通过上述方式，可以在给定某个产品的情况下，预测出有哪些用户可能会买，或者说可以计算得到用户购买该产品的概率是多少。

[0113] 举例来说，对数据进行网络化之后，存在二维关系网络为：用户A-产品C，用户A-产品D，用户B-产品C。那么目标函数为：给定“产品C”节点时，通过改变(学习)“用户A”节点和“用户B”节点对应的向量，让与“产品C”节点相连接的所有节点的向量既与“用户A”节点的向量相似，又与“用户B”节点的向量相似，于是“用户A”节点的向量和“用户B”节点的向量相似。通过上述方式，成功的保存了网络中的结构信息，大大提升了后续解决相应问题的准确性。

[0114] 三、对所述高维关系网络进行嵌入映射(Tensor Network Embedding)

[0115] 高维关系网络是指网络中有边是对应三个节点的，例如图2所示的“用户-产品-评价”网络属于高维关系网络。高维关系(High-order Relation)也是数据中常见的，比如评价行为同时涉及用户、产品和评价文本，因而需要用张量而非矩阵、三元关系而非简单的二部图来表示这样的行为数据。

[0116] 对所述高维关系网络进行嵌入映射是指利用嵌入映射的方法，将具有高维关系(如用户-产品-评价)的行为网络和属性网络中的节点表示为高维向量。

[0117] 于语义网络的嵌入映射一样，高维关系网络的嵌入映射，最大化条件概率的目标函数是给定临近节点(相连接的节点)的向量，预测目标节点的向量，使得与给定的一些节点相连接的节点之间具有相似的向量。

[0118] 要实现高维关系网络的嵌入映射，需要更新目标函数，可以有两种处理方法。一种是每采样一次多元关系，更新相关联节点的向量表示，那么最大化的目标函数如下：

[0119] $$L_1 = -\sum_{j=1}^{|S|}\sum_{r_{m/j}\in A_j}\lambda_{m,j}\,d\left(\hat{P}_1(\cdot\,|r_{m/j}),P_1(\cdot\,|r_{m/j})\right) \quad (3)$$

[0120] 其中，S是节点的集合，$A_{(j)}$是指与j节点相关联的高维关系集合，$r_{(m/j)}$是指其中的一个高维关系，m是该高维关系的编号，$\lambda_{m,/j}$是该高维关系的权重，P_1是给定该高维关系时所关联节点的概率，L_1是对于每一个节点j，最大化其所关联高维关系中节点两两之间向量的相似度。

[0121] 另一种是采样多元关系时，分裂成若干个二元关系，并更新关联节点的向量表示，最大化目标函数如下：

[0122] $$L_2 = -\sum_{r_m\in\tilde{A}}\lambda'_m\,d\left(\hat{P}_2(\cdot\,|r_m),P_2(\cdot\,|r_m)\right) \quad (4)$$

[0123] 其中，$\tilde{A}$是高维关系拆分成多个二维关系后所有二维关系的集合，r_m是第m个二维关系，λ_m是第m个二维关系的权重，P2是给定该高维关系时所关联节点的概率，L_2是对于每一个拆分后的二维关系，最大化该关系的两个节点之间向量相似度。

[0124] 举例来说，假设对数据进行网络化之后的高维关系网络为：用户A-产品C-购买地点E，用户A-产品C-购买地点F，用户B-产品C-购买地点E。

[0125] 那么目标函数就是给定“产品C”节点和“购买地E”节点后，与它们相关联(即通过边相连)的节点的向量相似，从而让“用户A”节点的向量和“用户B”节点的向量相似。当然，我们可能会遍历每一种给定信息，如给定“用户A”节点和“产品C”节点后，让“购买地点E”节点和“购买地点F”节点对应的向量相似。

[0126] 如果采用最大目标函数L_1，即给定某个关系的其它节点(如产品C和购买地E)，让被隐藏的一个节点被学习(如用户节点)。

[0127] 如果采用最大目标函数L_2，即把高维关系分裂成A-C、C-E、A-E、A-C、A-F、C-F等9个二维关系，然后调用二维关系的嵌入映射实现。

[0128] 由上述对语义网络进行嵌入映射、对二维网络进行嵌入映射以及对高维网络进行嵌入映射可知，通过对所述关系网络采用基于嵌入映射的表征学习算法，可以将关系网络的节点统一用维度较高的向量来进行表示，向量的每一个维度代表了该节点的特征，实现了原始大数据的特征提取。且由于高维向量中并保留了原始大数据中的结构信息，如语义信息、购买行为信息等，大大提升后期应用服务对应的分类和预测等任务的准确性。而且本发明实施例中的基于嵌入映射的表征学习算法，还可以应用于高维关系的数据，适用于各种复杂的应用环境，且计算速度很快，可以快速响应应用需求。

[0129] 可选地，在实现步骤S105-S106时，可以将应用服务请求转化为排序、分类、聚类、预测、关联分析和异常检测等任务，这些任务可以用特定的处理算法完成，可以预先指定或

者获取这些任务与处理算法之间的对应关系(即所述应用服务请求与处理算法之间的对应关系),从而当获取到应用服务请求时,可以知道采用何种处理算法。为了更好的理解本发明实施例,了解到这些任务与何种处理算法对应,如何用处理算法完成的,本发明实施例将对相关内容做详细的介绍。

[0130] 1、排序(Ranking)任务

[0131] 排序任务往往基于某种特定的相似度实现,通常涉及到所述关系网络的节点的相似度计算,包括皮尔森关联度(Pearson Correlation)和余弦相似度(Cosine Similarity)等。

[0132] 举例来说,当应用服务请求需要解决的问题是,给定某个产品,列出与之在被购买方面最相似的产品时,可以将该问题转化为排序任务。

[0133] 处理算法:我们可以通过执行S101-S103得到的高维向量中,找到该产品节点的高维向量u_i,则问题转化为求出与u_i相似度最高的一系列产品节点。由于每一个产品节点都具有一个高维向量表示,通常为K维(K通常为200到500之间的数字),因而可以通过求向量的数量积来得到节点之间的相似度。最终该问题转化为求与向量u_i在数量积上最大的一系列向量。通过上述算法就实现了排序任务或者说得到了应用服务请求的结果。

[0134] 2、分类(Classification)任务

[0135] 分类任务包括二分类和多分类,支撑向量机(Support Vector Machine)和逻辑回归(Logistic Regression)等监督学习算法能够有效解决分类任务;

[0136] 例如,应用服务请求需要解决的问题可能为给定大量用户,根据年龄层、收入区间等信息确定用户类别。然而实际应用中,数据中往往会存在信息缺失,如何将未知年龄、收入等信息的用户分类到正确的年龄层和收入区间,是一个重要的问题。可以将该问题转化为分类任务。

[0137] 处理算法:通过表征学习能够得到用户、年龄层、收入区间等节点的高维向量,那么仅需计算用户节点的高维向量与年龄层节点的高维向量的相似度,以及计算用户节点的高维向量与收入区间节点的高维向量的相似度,选取与用户节点的高维向量相似度最高的年龄层节点和收入区间节点即可。就可以将该用户分类到正确的年龄层和收入区间。

[0138] 3、聚类(Clustering)任务

[0139] 聚类任务往往用最近邻、谱聚类等非监督学习算法完成。

[0140] 例如,应用服务请求需要解决的问题可能为:给定大量用户,在未知类别的情况下,把用户根据购买行为习惯聚成K类,以便可以对同一类用户制定同样的策略。可以将该问题转化为聚类任务。

[0141] 处理算法:可以根据用户的高维特征表示,采用K-means或者KNN的算法快速实现聚类。通常聚类问题的难点在于如何降低结构化信息的维度,该维度高达用户的数量,即节点的数量N,然而嵌入映射已经成功的将维度降低到K。

[0142] 4、预测(Prediction)任务

[0143] 预测任务通常是利用矩阵分解(Matrix Factorization)或张量分解(Tensor Factorization),实现对矩阵和高维张量的填充,从而预测数据中的缺失值(Missing Value)。

[0144] 举例来说,应用服务请求需要解决的问题可能为:预测某用户将来是否会购买某

产品。事实上，推荐问题可以转化为预测问题，即给出预测得到的用户最可能购买的一系列产品。

[0145] 处理算法：我们可以通过本发明实施例所述的方法，得到给定用户节点高维向量和产品节点的高维向量，通过计算给定用户节点高维向量与产品节点的高维向量的相似度，可以将与用户节点相似度最高的产品推荐给该给定用户。

[0146] 5、关联分析(Correlation Analysis)任务

[0147] 应用服务请求需要解决的问题可能为：判断用户的年龄层、收入区间与产品的价格区间是否有关联关系。

[0148] 处理算法：通过本发明实施例所述的方法，可以得到年龄层节点、收入区间节点和价格区间节点的高维向量，因而通过快速的计算它们之间的相似度，就可以了解不同用户属性(用户的年龄层和收入)与产品属性(产品的价格区间)之间的关联关系和关联的强度。

[0149] 6、异常检测(Outlier Detection)任务

[0150] 应用服务请求需要解决的问题可能为：判断某用户是否是其所在的用户群中的异常用户，如欺诈用户等。

[0151] 处理算法：通过本发明实施例所述的方法，可以得到所有用户节点的高维向量，通过计算当前用户节点的高维向量与其他用户节点的高维向量之间的相似度，如果相似度很大，可以认为当前用户是异常用户。

[0152] 可选地，在执行步骤S101-S103之后，即完成了对原始大数据进行数据挖掘，得到统一的高维向量表示的数据特征之后，如果原始大数据有更新，可以只对更新的数据执行步骤S101-S103，不必再对所有数据再执行一次S101-S103。

[0153] 可选地，可以是在数据有更新的情况下，就对新数据执行步骤S101-S103以实现对新数据的数据挖掘，也可以是在新数据积累到一定数量时才执行，或者可以定期对新数据执行步骤S101-S103。

[0154] 本发明实施例还提供了一种大数据的智能处理方法，如图2所示，该方法包括：

[0155] S301：获取用户的应用服务请求以及由原始大数据转化而来的关系网络的节点的高维向量。

[0156] 可选地，所述由原始大数据转化而来的关系网络为：由所述原始大数据经过预处理之后进行网络化所得到的关系网络。

[0157] 在本发明实施例中，可以直接获取用高维向量表示的特征，从而不需要利用原始大数据进行特征挖掘。利用原始大数据进行特征挖掘的过程可以是在其它装置上完成的，本发明实施例在此不作限制。利用原始大数据进行挖掘的特征挖掘的过程可以参考S101-S103，本发明实施例在此不做赘述。

[0158] S302：确定所述应用服务请求对应的处理算法。

[0159] S303：利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

[0160] S302和S303的具体实现方式可以参考S105-S106。

[0161] 在本发明实施例中，直接获取用高维向量表示的特征，利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。本发明实施例所述的大数据的智能处理方法，不仅限于某个特定的应用服务，可以为多种应用服

务提供统一有效地处理方法。

[0162] 对应于图1所述的方法实施例，本发明还提供了一种大数据的智能处理系统，如图4所示，包括数据结构化模块401、表征学习模块402和应用算法模块403。

[0163] 所述数据结构化模块401，用于对原始大数据进行预处理，以及对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络。其中，所述关系网络中的节点，由所述预处理后的原始大数据中的数据单元转化而来，所述关系网络中的边，用于表示所述网络中节点与节点之间的关系。通过对原始大数据进行网络化，可以将表格中的大数据或海量数据转化为关系网络，从而可以采用节点和边的方式统一处理这些数据，大大缩减了数据存储和管理的成本。其次，针对预处理后的原始大数据中的词汇和短语等文本数据，将其进行网络化，构建出语义网络，保留了文本中的语义信息，以便后续可以有效利用，提高应用服务的准确性。此外，将预处理后的原始大数据表示为包含节点和边的关系网络之后，就可以利用网络数据的表征学习算法，来实现对数据的快速、统一的特征提取，从而做到快速响应不同应用服务请求。

[0164] 所述表征学习模块402，用于对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量。所述表征学习模块402通过对所述关系网络采用基于嵌入映射的表征学习算法，将所述关系网络中的节点，例如用户、产品和短语等，统一用维度较高的向量来进行表示，其中，每个向量可以表示所述关系网络中的一个节点，该向量中的一个维度表示了该节点的一个特征。所述关系网络中节点与节点之间的关系(或者说边)，转化为节点的高维向量与节点的高维向量之间的相似度，从而保留了原始大数据中的结构信息，大大提升后期应用服务对应的分类和预测等任务的准确性。

[0165] 应用算法模块403，用于获取用户的应用服务请求；确定所述应用服务请求对应的处理算法，以及利用所述应用服务请求对应的处理算法和所述表征学习模块402得到的所述关系网络的节点的高维向量，确定所述应用服务请求的结果。也即，在所述表征学习模块402将大数据中的特征用高维向量的形式统一表示之后，应用算法模块403可以利用这些统一用高维向量表示的特征，去提供各种应用服务的解决方案或者说返回应用服务需要解决的问题结果。

[0166] 在本发明实施例中，所述数据结构化模块401用于对原始大数据进行预处理以及网络化，从而所述表征学习模块402可以利用网络数据的表征学习算法，来实现对数据的快速、统一的特征提取，所述应用算法模块403可以根据用户的应用服务请求，确定对应的处理算法，并利用所述表征学习模块402提取到的以向量形式表示的特征进行计算，得到处理结果返回给用户。不同于现有技术，本发明实施例中整个特征提取的过程无需人的参与，利用基于嵌入映射的表征学习算法自动完成，计算效率高；特征提取的过程中还极大地保留了原始大数据中的结构信息(即有效信息)，从而提高了进行分类或预测等任务的准确性；不仅如此，由于采用了基于嵌入映射的表征学习算法，使得从原始大数据中挖掘到的数据特征统可以统一由高维向量的形式进行表示，从而本发明实施例中的系统不仅限于为某个特定的应用服务，可以为多种应用服务提供统一有效地处理方法。

[0167] 需要注意的是，本发明实施例所述的大数据的智能处理方法，不仅仅可以应用于保险业领域，还可以应用于其他领域，例如应用于金融领域、购物消费领域等，尤其适用于对包含结构数据和非结构数据进行处理的情况，以及需要处理高维关系的数据的场合，较

现有技术将具有明显的优势。

[0168] 可选地，由于所述关系网络既可能包扩语义网络，也可能包括属性网络和行为网络。它们既可能属于同构关系网络、也可能属于二维关系网络，还可能属于高维关系网络。因此，所述表征学习模块402可以具体用于对所述关系网络中的高维关系网络进行嵌入映射，得到所述高维关系网络的节点的高维向量；或者，具体用于对所述关系网络中的二维关系网络进行嵌入映射，得到所述二维关系网络的节点的高维向量；或者，具体用于对所述关系网络中的语义网络进行嵌入映射，得到所述语义网络的节点的高维向量；或者具体用于对所述关系网络中的同构网络进行行嵌入映射，得到所述同构网络的节点的高维向量。

[0169] 可选地，在本发明实施例中，所述原始大数据可以是通过各个网站或者APP收集而来的，可能包括行为数据、属性数据等结构数据，也可能包括文本数据等非结构化数据，本发明实施例在此不做限定。

[0170] 所述数据结构化模块401对原始大数据进行预处理可以是，对所述原始大数据进行数据分析和清理，即对原始大数据进行统计分析，去除不合规或错误的数据内容，可以是将非法数据格式进行过滤，例如去除理应为浮点数、却填充为字符串型的价格等数值，还可以是将时间或者单位进行统一，也可以是对缺失的指进行填写、光滑噪声数据等，从而可以将大数据的格式标准化，清除异常数据，纠正错误或者清除重复数据等。

[0171] 可选地，所述预处理后的原始大数据可以包括行为数据、属性数据和文本数据，则所述数据处理模块对所述预处理后的原始大数据进行网络化可以包括：对所述预处理后的原始大数据中的行为数据进行网络化，例如将购买、评价等行为数据转化为行为网络；或者，还可以包括对所述预处理后的原始大数据中的属性数据进行网络化，例如将年龄、价格等属性信息转化为属性网络；又或者还可以包括将对所述预处理后的原始大数据中的文本数据进行网络化，例如将产品介绍或者评价内容等文本数据转化为以词和短语为节点的语义网络。则所述行为网络、所述属性网络和所述语义网络共同组成了所述关系网络。

[0172] 可选地，所述应用算法模块403利用所述关系网络的节点的高维向量，确定所述应用服务请求的结果时，可以利用所述关系网络的所有节点的高维向量去确定所述应用服务请求的结果；还可以只利用所述关系网络的部分的高维向量，去确定所述应用服务请求的结果。具体地，可以只利用与所述应用服务请求相关的节点去确定所述应用服务请求的结果。例如，当应用服务请求是产品推荐时，可以只利用产品节点的高维向量和用户节点的高维向量进行计算。

[0173] 需要说明的是，在本发明实施例中，各个模块的具体实现，可以参考方法实施例的描述，例如关于具体如何进行基于嵌入映射的表征学习算法，可以参考方法实施例的描述，本发明实施例在此不做赘述。

[0174] 本发明实施例所述的系统，可以以软件或者程序的形式，在一台或者多台计算机或者服务器实现，本发明实施例在此不做限定。

[0175] 为了更好的理解本发明实施例，以将本发明实施例所述的大数据的智能处理系统应用于保险业为例进行详细说明。

[0176] 用户在个人计算机(personal computer，PC)或者移动端进行完善个人信息、查看保险细则、购险、退险或者建立社交关系等操作时，可以通过服务器收集上述操作信息，形成原始大数据，所述原始大数据可以以表格的形式存储在数据库中。本发明实施例所述的

系统可以获取上述原始大数据。

[0177] 例如，通过收集操作信息，数据库中可能保存了如表1所示的用户个人信息表、如表2所示的产品信息表、如表3所示的购险行为表以及如表4所示的退险行为表。

[0178] 表1用户个人信息表

[0179]

用户 ID	姓名	性别	年龄	职业	自我介绍	……
用户 1	张先生	男	36	律师	平常工作很忙、出差很多	……
用户 2	赵先生	男	40	销售	出差很多、常过度劳累	……
用户 3	王女士	女	45	无	身体不好，离异有一子	……
用户 4	孙女士	女	39	文职	爱好旅游，出差很多	……
用户 5	李女士	女	56	退休	身体不好	……
……	……	……	……	……	……	……

[0180] 表2产品信息表

[0181]

险名	类别	价格	售险公司	产品介绍	……
险A	车险	……	……	保费低、理赔方便	……
险B	寿险	……	……	终身寿险、投保年龄范围广	……
险C	健康险	……	……	重大疾病赔付金额高	……
……	……	……	……	……	……

[0182] 表3购险行为表

[0183]

用户ID	险名	购险地点(GPS)	购买金额	用户评价
用户1	险A	XX公司	……	购买方便:)
用户2	险C	XX企业	……	总在外面还是买个险好
用户3	险B	1.765	……	……
用户4	险A	XX路	……	给爱车加个险！
用户5	险B	XX小区	……	……
……	……	……	……	……

[0184] 表4退险行为表

[0185]

用户ID	险名	退险地点(GPS)	退险金额	退险理由
用户3	险B	XX街(家中)	……	……
……	……	……	……	……

[0186] 首先，所述系统中的数据结构化模块可以对上述数据进行数据分析和清洗。以对表3所示的构险行为表中的数据进行数据进行数据分析和清洗为例。数据分析指通过数据统计和关联获取更多的信息，所述数据结构化模块可以将“工作地点”、“家中”、“营销点附近”等信息补充在地理位置信息上。数据清洗是指将非法的数值去除、乃至将非法的数据记录去除，例如当“购险地点”为实数时，所述数据结构化模块可以隐藏该数值；当表3中记录

的“用户ID”或“险名”的数值非法时，所述数据结构化模块可以去除该购险记录。表5为表3中数据经过所述数据结构化模块进行数据分析和清洗后的结果。

[0187] 表5经过数据分析和清洗后的购险行为表

[0188]

用户ID	险名	购险地点(GPS)	购买金额	用户评价
用户1	险A	XX公司【工作地点】	……	购买方便:)
用户2	险C	XX企业【工作地点】	……	总在外面还是买个险好
用户3	险B	【缺失】	……	……
用户4	险A	XX路【某营销点附近】	……	给爱车加个险！
用户5	险B	XX小区【家中】	……	……
……	……	……	……	……

[0189] 接下来，可以对经过数据分析和清洗后原始大数据进行网络化，得到包含节点和边的关系网络。通过上述表格可知，原始大数据中存在大量的文本信，因此所述数据结构化模块可以对文本数据进行网络化，得到由短语或者词语组成的节点，以及节点之间的边，即得到包含节点和边的语义网络。后续表征学习模块可以对该语义网络利用表征学习方法，学习其中的语义信息。例如，利用分词工具可以将表1至表4中经过数据分析和清洗后的文本数据提取出来，得到“文档-短语”形式的文本数据如表6所示，表6中每个短语可以表示为语义网络中的一个节点。短语组成的节点之间，如果共同出现在句子或文档中，则它们之间可以存在边进行连接，边的权重由它们共同出现在句子或文档中的频度决定。如“旅游”节点与“出差很多”节点之间有边连接，“出差很多”节点与“过度劳累”节点之间有边连接。

[0190] 表6

[0191]

平常 工作很忙 出差很多
出差很多 常 过度劳累
身体不好 离异 有一子
爱好旅游 出差很多
身体不好
保费低 理赔方便
终身寿险 投保 年龄范围广
重大疾病 赔付金额高
购买方便
总在外面 还是 买个险 好
给 爱车 加个险
价格太高 不合适

[0192] 此外，可以将表中的内容进行网络化转化为关系网络：如将表1的内容转化为“用户ID-性别”、“用户ID-年龄段”、“用户ID-职业”和“用户ID-自我介绍短语”等多个二维关系网络；将表2的内容转化为“险名-类别”、“险名-价格区间”、“险名-售险公司”和“险名-产品介绍短语”等多个二维关系网络；将表3的内容转化为“用户ID-险名-购险地点-金额区间-评价短语”的高维关系网络，将表4的内容转化为“用户ID-险名-退险地点-金额区间-退险

理由短语”的高维关系网络。

[0193] 最后形成的关系网络中，既包含了上述的语义网络，也包含了由表1-表4的内容转化而来的多个高维关系网络和二维关系网络，这些高维关系网络和二维关系网络中即有属性网络又有行为网络；关系网络中以用户ID、用户属性、产品属性、地点、短语等为节点，以它们之间的交互/关系作为所述关系网络的边。

[0194] 需要注意的是，关系网络中允许节点重叠，上述二维关系网络和高维关系网络可以用“用户ID”、“险名”、“短语”等融合成含有多种类别节点的关系网络，即多源异构网络。在所述数据结构化模块将原始大数据转化为关系网络之后，表征学习模块可以对所述关系网络中的数据进行表征学习。假定高维向量的维度数量为K(K通常取值在200到500之间)，表征学习的结果是将关系网络中的节点(如短语节点、用户节点、用户属性节点、产品节点等)表示为多个高维向量，高维向量中保留了该关系网络中的关联关系(即边)。

[0195] 通过前面的分析可知，在本发明实施例中，所述关系网络中包括了语义网络、二维关系网络和高维关系网络。则所述表征学习模块可以对所述语义网络采用基于嵌入映射的表征学习算法，具体地可以是：将“旅游”、“出差很多”、”过度劳累”、“身体不好”、“重大疾病”等语义网络中的节点表示为高维向量，如$u=[u_1,u_2,\cdots,u_K]$，并且通过表征学习算法，可以挖掘出“旅游”节点的向量与“出差很多”节点的向量相似，“出差很多”节点的向量与“过度劳累”节点的向量的向量相似，“过度劳累”节点的向量、“身体不好”节点的向量与“重大疾病”节点的向量相似。从而保留了网络中数据的结构信息。

[0196] 所述表征学习模块可以对二维关系网络采用基于嵌入映射的表征学习算法，所述二维关系网络的表征学习结果可以是：将“用户ID”、“用户属性”、“产品ID(险名)”、“产品属性”等节点表示为高维向量，通过使得属性相似的用户节点的向量相似度高，属性相似的产品节点的向量相似度高，保留了所述关系网络中的结构信息，最终使得出行较多的用户节点之间、类别相同的产品节点之间有相似的向量。

[0197] 所述表征学习模块可以对高维关系网络采用基于嵌入映射的表征学习算法，所述高维关系网络的表征学习结果可以是：将“用户”、“险名”、“地点”、“评价短语”等节点表示为高维向量，使得具有相似购买、退险习惯的用户节点的向量相似度高，购买、退订用户相似的产品节点的向量相似度高，保留了关系网络中的结构信息。

[0198] 在本发明实施例中，所述表征学习模块可以基于嵌入映射(Embedding)、结合Skip-gram和Negative Sampling实现，保证了算法的计算复杂度低，算法可扩展性强。

[0199] 在本发明实施例中，将关系网络的每一个节点统一用高维向量进行表示，并保留了关系网络中的结构信息，在后续的应用服务中针对不同的任务需求，应用算法模块可以调用其中部分节点的高维向量进行计算，计算复杂度低。

[0200] 例如，假设用户的应用服务请求需要解决的问题为保险产品推荐，我们可以用应用算法模块实现保险产品推荐。保险产品推荐是给定用户，寻找与该用户在购买行为上最相似、退订行为上最不同的产品。则与该应用服务请求对应的处理算法为：利用余弦相似度(Cosine similarity)等向量相似度计算方法，选择用户节点的向量和产品节点的向量，计算用户节点的向量和产品节点的向量的相似度。比如，在表征学习时，我们可以在用户节点的向量和产品节点的向量中通过第1至100维向量来保存“购险行为”信息，即如果用户A购买了产品A，则用户A节点的向量的第1至100维与产品A节点的向量的第1至100维相似；我们

还可以在用户节点的向量和产品节点的向量中通过第101至200维向量来保存“退险行为”信息，即如果用户A退订了产品B，则用户A节点的向量的第101至200维与产品B节点的向量的第101至200维相似。因此，如果需要给用户A推荐产品，则是寻找与用户A节点的向量的第1至100维向量相似，第101至200维向量不相似的产品节点的向量。

[0201] 同理，所述应用算法模块还可以利用经过表征学习得到的高维向量，实现用户类别的分类和欺诈骗保用户的检测等，本发明实施例在此不再赘述。

[0202] 对应于图3所述的大数据的智能处理方法，本发明实施例提供了一种一种大数据的智能处理系统，如图5所示，该系统可以包括：

[0203] 获取模块501，用于获取用户的应用服务请求以及由原始大数据转化而来的关系网络的节点的高维向量。

[0204] 确定模块502，用于确定所述应用服务请求对应的处理算法，利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。

[0205] 可选地，所述由原始大数据转化而来的关系网络为：由所述原始大数据经过预处理之后进行网络化所得到的关系网络。

[0206] 在本发明实施例中，所述获取模块501可以直接获取用高维向量表示的特征，从而所述确定模块利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果。本发明实施例所述的大数据的智能处理系统，不仅限于某个特定的应用服务，可以为多种应用服务提供统一有效地处理方法。

[0207] 以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0208] 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各实施方式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。基于这样的理解，上述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，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0209] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

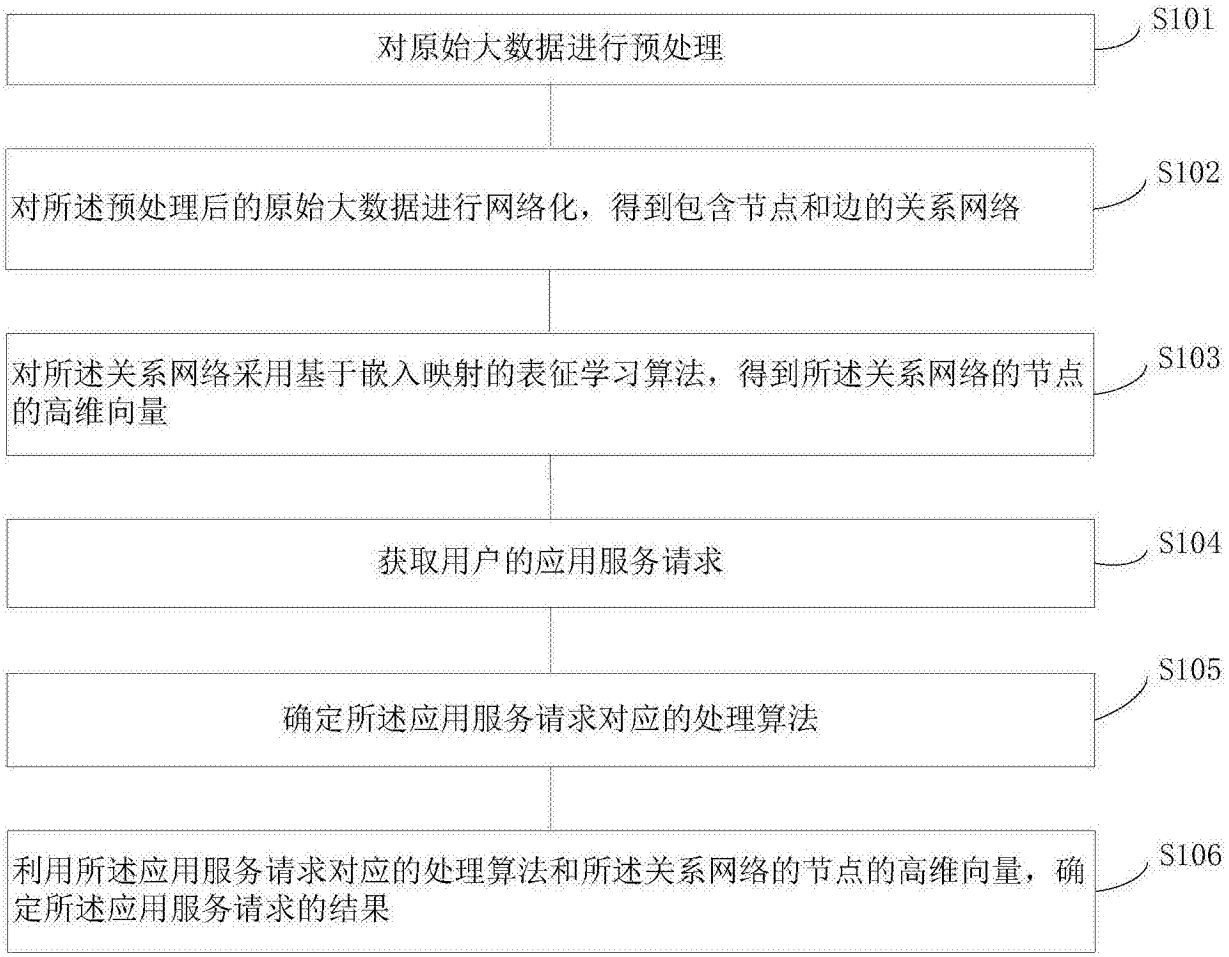

对原始大数据进行预处理
S101
对所述预处理后的原始大数据进行网络化，得到包含节点和边的关系网络
S102
对所述关系网络采用基于嵌入映射的表征学习算法，得到所述关系网络的节点的高维向量
S103
获取用户的应用服务请求
S104
确定所述应用服务请求对应的处理算法
S105
利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果
S106

图1

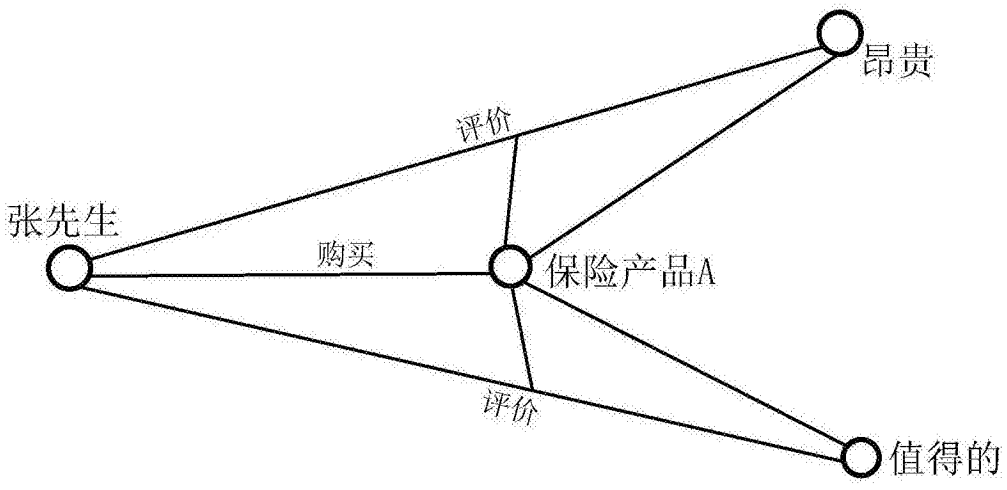

昂贵
评价
张先生
购买
保险产品A
评价
值得的

图2

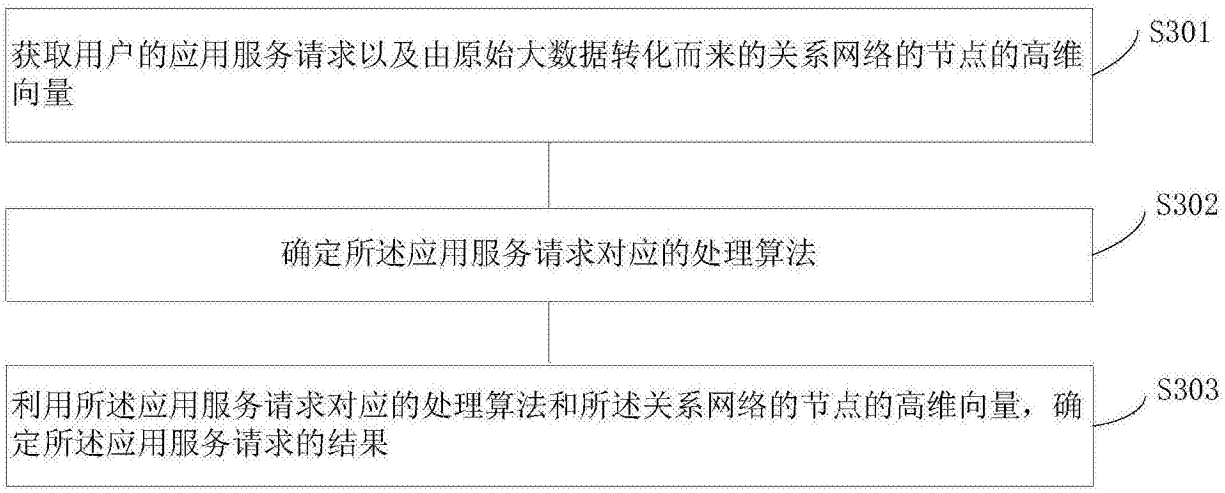

获取用户的应用服务请求以及由原始大数据转化而来的关系网络的节点的高维向量
S301
确定所述应用服务请求对应的处理算法
S302
利用所述应用服务请求对应的处理算法和所述关系网络的节点的高维向量，确定所述应用服务请求的结果
S303

图3

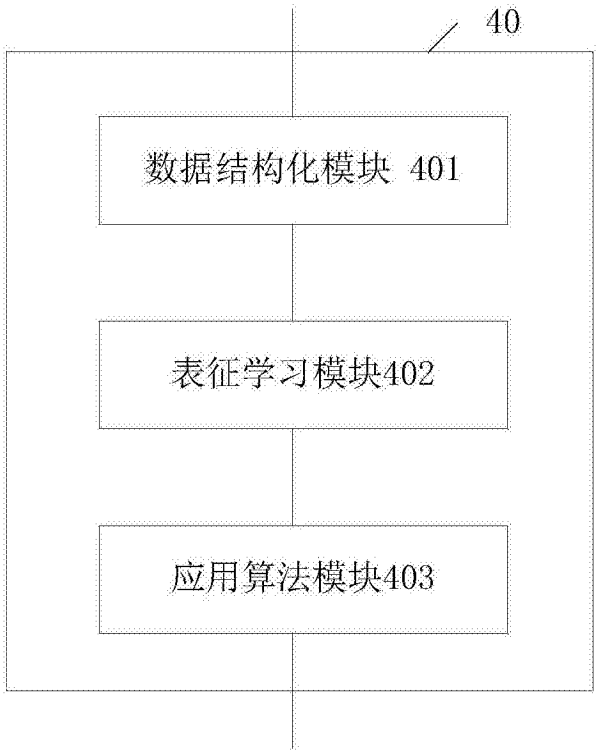

40
数据结构化模块 401
表征学习模块402
应用算法模块403

图4

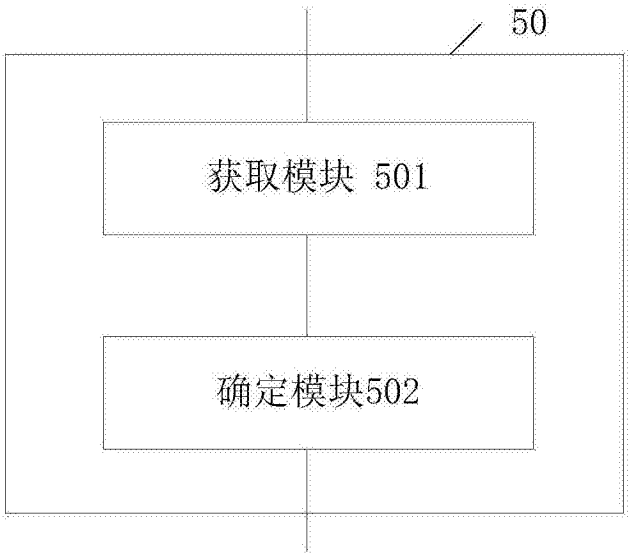


图5