[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 17/00
G06F 17/30 G06F 9/44

[12] 发明专利申请公开说明书

[21] 申请号 03106807.3

[43] 公开日 2003 年 9 月 17 日

[11] 公开号 CN 1442801A

[22] 申请日 2003.3.3 [21] 申请号 03106807.3

[30] 优先权

[32] 2002. 3. 4 [33] JP [31] 58065/2002

[71] 申请人 精工爱普生株式会社

地址 日本东京都

[72] 发明人 萱原直树

[74] 专利代理机构 中国专利代理(香港)有限公司

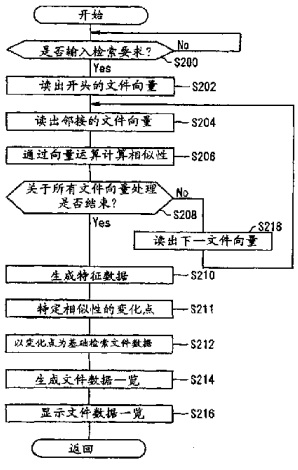
代理人 马铁良 王忠忠

权利要求书 3 页 说明书 19 页 附图 8 页

[54] 发明名称 数据管理以及文件数据检索的装置、方法和程序

[57] 摘要

从文件数据登录 DB44 的文件数据中抽出关于文件数据内容表示相似性时间推移的特征数据，基于所抽出的特征数据特定相似性的变化点，以所特定的变化点为基础从文件数据登录 DB44 中检索文件数据。 在检索中，检索所特定的变化点或属于其附近的文件数据。 由此可提供一种适于从庞大数据中掌握有特征的部分、容易提高抽出可靠性且可即时对应用户要求的数据管理装置。



ISSN1008-4274

1．一种数据管理装置，其用于管理多个数据，其特征在于：包括

特征数据抽出单元，其从上述多个数据抽出关于上述数据内容表示相似性的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点。

2．权利要求1记载的数据管理装置，其特征在于：

上述数据为文件数据。

3．一种文件数据检索装置，其用于从作成时间或更新时间不同的多个文件数据中进行检索，其特征在于：包括

文件数据存储单元，其用于存储上述多个文件数据；特征数据抽出单元，其从上述文件数据存储单元的文件数据抽出关于上述文件数据内容表示相似性时间推移的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点；文件数据检索单元，其以由上述变化点特定单元所特定的变化点为基础，从上述文件数据存储单元中检索上述文件数据。

4．权利要求3记载的文件数据检索装置，其特征在于：

上述文件数据检索单元从上述文件数据存储单元中检索由上述变化点特定单元所特定的变化点或属于其附近的文件数据。

5．权利要求3及4之一记载的文件数据检索装置，其特征在于：

上述变化点特定单元基于由上述特征数据抽出单元所抽出的特征数据，设定允许范围，在上述相似性的时间推移中，特定超过上述允许范围的点作为上述变化点。

6．权利要求3至5之一记载的文件数据检索装置，其特征在于：

上述特征数据抽出单元将上述文件数据存储单元的文件数据按每一规定期间进行区分，按各区分生成合并了属于其区分的文件数据内容的期间文件数据，关于在时间序列上邻接的期间文件数据算出上述相似性，基于算出的相似性生成上述特征数据。

7．权利要求3至5之一记载的文件数据检索装置，其特征在于：

上述特征数据抽出单元将上述文件数据存储单元的文件数据按每一规定期间进行区分，按各区分生成合并了属于其区分的文件数据内容的期间文件数据，关于所生成的期间文件数据相互之间算出上述相似性，基于算出的相似性生成上述特征数据。

8. 权利要求 6 及 7 之一记载的文件数据检索装置，其特征在于：

上述特征数据抽出单元算出表示上述期间文件数据内容特征的文件向量，通过比较算出的文件向量算出上述相似性。

9. 权利要求 8 记载的文件数据检索装置，其特征在于：

上述特征数据抽出单元对上述期间文件数据进行词素解析，按各词素作为上述文件向量生成具有作为向量的量的与上述期间文件数据中其词素出现频度对应的元素的向量。

10. 权利要求 6 至 9 之一记载的文件数据检索装置，其特征在于：

上述特征数据抽出单元从上述各期间文件数据除去在上述各期间文件数据中共同的内容，基于实施了除去的期间文件数据，算出上述相似性。

11. 一种数据管理程序，其用于管理多个数据，其特征在于：使计算机执行作为以下各单元所实现的处理，

特征数据抽出单元，其从上述多个数据抽出关于上述数据内容表示相似性的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点。

12. 权利要求 11 记载的数据管理程序，其特征在于：

上述数据为文件数据。

13. 一种文件数据检索程序，其用于从作成时间或更新时间不同的多个文件数据中进行检索，其特征在于：

针对可利用存储上述多个文件数据用的文件数据存储单元的计算机，使其执行作为以下各单元所实现的处理，

特征数据抽出单元，其从上述文件数据存储单元的文件数据抽出关于上述文件数据内容表示相似性时间推移的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点；文件数据检索单元，其以由上述变化点特定单元所特定的变化点为基础，从上述文件数据存储单元中检索上述文件数据。

14. 一种数据管理方法，其用于管理多个数据，其特征在于：包含

特征数据抽出步骤，其从上述多个数据抽出关于上述数据内容表示相似性的特征数据；变化点特定步骤，其基于由上述特征数据抽出

步骤所抽出的特征数据，特定上述相似性的变化点。

15．权利要求14记载的数据管理方法，其特征在于：

上述数据为文件数据。

16．一种文件数据检索方法，其用于从作成时间或更新时间不同的多个文件数据中进行检索，其特征在于：包含

文件数据存储步骤，其将上述多个文件数据存储到文件数据存储单元；特征数据抽出步骤，其从上述文件数据存储单元的文件数据抽出关于上述文件数据内容表示相似性时间推移的特征数据；变化点特定步骤，其基于由上述特征数据抽出步骤所抽出的特征数据，特定上述相似性的变化点；文件数据检索步骤，其以由上述变化点特定步骤所特定的变化点为基础，从上述文件数据存储单元中检索上述文件数据。

数据管理以及文件数据检索的装置、方法和程序

技术领域

本发明涉及从作成时间或更新时间不同的多个文件数据中进行检索的装置、程序及方法，特别是涉及适于从庞大数据中掌握有特征的部分、并且容易提高抽出可靠性且能即时对应用户要求的数据管理装置、文件数据检索装置、数据管理程序和文件数据检索程序以及数据管理方法和文件数据检索方法。

现有技术

在企业等，有时通过让员工提交业务日志来管理业务的进展情况。多数场合，业务日志形成的报告由一个上司对多个部下所提交的业务日志一个一个过目审阅。

但是，由于职务上的关系等，上司对所提交的所有业务日志不一定每天都能过目。另外，假设即使对所有的业务日志过目，在有限的时间内，能掌握的信息量总是有限的。从而，在审阅的业务日志量庞大时，很难有效管理业务的进展情况。

这种场合，为了有效管理业务进展情况，上司需要从庞大的业务日志有效得到信息。因此，首先研究一下业务日志的性质。业务日志以各员工每天的业务报告为主要内容，因此，关于同一员工提交的业务日志，对作成时间接近的业务日志互相比较时，内容重复的部分多。对内容重复的部分每天进行过目是无效的。因而，如果上司对内容重复的部分只掌握一次，对其后的业务日志只掌握有特征的部分(即有变化的部分)，可以比较有效地得到信息。

作为这一问题的解决方法可以提出如下构成议案，例如，将业务日志作为文件数据累积于文件数据库(以下，数据库只简记为 DB)，可以从文件 DB 中只检索有特征的部分。

迄今为止，作为从多个文件数据中进行检索的技术，例如有特开平 7-325832 号公报中公布的利用单词规格图形时间变化的检索方法(以下，称第 1 现有例)。另外，作为其相关技术，例如有特开平 6-324871 号公报中公布的推理装置(以下，称第 2 现有例)及特开平 5-53814 号公报中公布的事例库检索系统作成支援装置(以下，称第 3 现有例)。

在第 1 现有例中，特征数据抽出部预先从文本信息抽出表示单词用图形时间变化的特征数据。当用户进行检索输入时，输入处理部将用户的检索输入转换为在检索处理部可解释的表现形式、送往检索处理部。检索处理部利用文本信息及特征数据进行检索，检索结果被送往输出处理部，向用户显示。作为特征数据，例如可以使用文本信息中单词出现概率等各种统计量。

由此，可以利用从时间序列文本信息所抽出的特征数据，在特定的领域、期间检索成为话题的单词及信息等，容易进行高质量的趋势、动向分析。

在第 2 现有例中，逻辑向量转换部分别把存入到规则存入部的规则，存入到事例存入部的事例及由推理条件输入部所输入的推理条件转换为逻辑向量的规则向量、事例向量及条件向量。不确定元素附加部向规则向量及事例向量附加不确定元素，分别作为不确定规则向量及不确定事例向量。另外，结果向量运算部把不确定规则向量、不确定事例向量及条件向量的逻辑积作为结果向量。逻辑命题转换部将结果向量转换为不确定逻辑命题。不确定元素除去部从不确定逻辑命题除去不确定元素，作为确定逻辑命题。逻辑命题输出部输出确定逻辑命题。

由此，可以进行推理效率好的知识获得负担少的推理。

第 3 现有例可将事例划分为多部分来检索相似性。向量划分部与子向量相似性计算部相关联。能进行伴随子向量化表现的附加操作。另外，实现了逐渐进行系统性能提高时所使用的变更监控器功能及变更比较功能。

由此，可以提供事例库推理系统建立所需要的作成环境的必须功能。

这样，在第 1 现有例中，是基于表示单词用图形时间变化的特征数据进行检索的，因此，例如可以抽出用户所输入的检索单词使用频度高的文件数据。但是，要在重复内容较多的文件数据群中抽出有特征的部分时，如果在有特征的部分多数使用了特定单词，也能抽出，但是未必多数使用了特定单词。因此，不适于抽出有特征的部分，如在上述业务日志例中所见，从庞大的信息中难以有效得到信息。

另外，将第 2 现有例应用于文件数据检索时，基于专家所建立的

规则进行检索。但是，为了提高抽出的可靠性，需要多累积专家建立的规则，但一般，知识 DB 的规则累积不容易。还有，规则累积需要时间，因此，难以适应用户的要求。

另外，在第 3 现有例中，只在事例属性向量，事例的特征被平均化，漏掉潜在的适合事例时，通过利用子向量、比较部分特征，能够发现潜在的适合事例。但是，这到底是追求高精度进行事例检索的技术，不适于在重复内容较多的文件数据群中抽出有特征的部分，同样，如在上述业务日志例中所见，从庞大的信息中，难以有效得到信息。

这些问题不限于上述业务日志例中所看到的那种文件数据的检索，是在要从庞大信息中有效得到信息的所有场合设想的问题。例如在管理图像数据、音乐数据其它数据时也能发生。

发明内容

于是，本发明是着眼于这种现有技术中存在的未解决的课题而展开的，其目的是提供适于从庞大数据中掌握有特征的部分，容易提高抽出可靠性且能即时对应用户要求的数据管理装置、文件数据检索装置、数据管理程序和文件数据检索程序以及数据管理方法和文件数据检索方法。

发明 1

为达到上述目的，发明 1 的数据管理装置是管理多个数据的装置，其特征在于：包括

特征数据抽出单元，其从上述多个数据抽出关于上述数据内容表示相似性的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点。

如果是这种构成，由特征数据抽出单元从多个数据抽出关于数据内容表示相似性的特征数据，由变化点特定单元基于所抽出的特征数据特定相似性的变化点。从而，用户通过参照所特定的变化点，可以从庞大的数据中比较容易地掌握有特征的部分。

发明 2

进一步，发明 2 的数据管理装置，其特征在于：在发明 1 的数据管理装置中，

上述数据为文件数据。

如果是这种构成，由特征数据抽出单元从多个文件数据抽出关于

文件数据的内容表示相似性的特征数据，由变化点特定单元基于所抽出的特征数据特定相似性的变化点。从而，用户通过参照所特定的变化点，可以从庞大的文件数据中比较容易地掌握有特征的部分。

发明 3

另一方面，为了达到上述目的，发明 3 的文件数据检索装置是从作成时间或更新时间不同的多个文件数据中进行检索的装置，其特征在于：包括

文件数据存储单元，其用于存储上述多个文件数据；特征数据抽出单元，其从上述文件数据存储单元的文件数据抽出关于上述文件数据内容表示相似性时间推移的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点；文件数据检索单元，其以由上述变化点特定单元所特定的变化点为基础，从上述文件数据存储单元中检索上述文件数据。

如果是这种构成，由特征数据抽出单元从文件数据存储单元的文件数据抽出关于文件数据内容表示相似性时间推移的特征数据，由变化点特定单元基于所抽出的特征数据特定相似性的变化点。然后，由文件数据检索单元以所特定的变化点为基础，从文件数据存储单元中检索文件数据。

在这里，文件数据存储单元用所有可能的单元、在所有可能的时期存储文件数据，既可以预先存储文件数据，又可以不预先存储文件数据，而在本装置动作时，通过来自外部的输入等来存储文件数据。以下，在发明 13 的文件数据检索程序中相同。

发明 4

进一步，发明 4 的文件数据检索装置，其特征在于：在发明 3 的文件数据检索装置中，

上述文件数据检索单元从上述文件数据存储单元中检索由上述变化点特定单元所特定的变化点或属于其附近的文件数据。

如果是这种构成，由文件数据检索单元从文件数据存储单元中检索所特定的变化点或属于其附近的文件数据。

发明 5

进一步，发明 5 的文件数据检索装置，其特征在于：在发明 3 及 4 之一的文件数据检索装置中，

上述变化点特定单元基于由上述特征数据抽出单元所抽出的特征数据，设定允许范围，在上述相似性的时间推移中，特定超过上述允许范围的点作为上述变化点。

如果是这种构成，由变化点特定单元基于所抽出的特征数据，设定允许范围，在相似性时间推移中特定超过允许范围的点作为变化点。

发明 6

进一步，发明 6 的文件数据检索装置，其特征在于：在发明 3 至 5 之一的文件数据检索装置中，

上述特征数据抽出单元将上述文件数据存储单元的文件数据按每一规定期间进行区分，按各区分生成合并了属于其区分的文件数据内容的期间文件数据，关于在时间序列上邻接的期间文件数据算出上述相似性，基于算出的相似性生成上述特征数据。

如果是这种构成，由特征数据抽出单元按每一规定期间区分文件数据存储单元的文件数据，按各区分生成期间文件数据。期间文件数据作为合并了属于一个区分的文件数据内容被生成。然后，关于在时间序列上邻接的期间文件数据算出相似性，基于算出的相似性生成特征数据。

发明 7

进一步，发明 7 的文件数据检索装置，其特征在于：在发明 3 至 5 之一的文件数据检索装置中，

上述特征数据抽出单元将上述文件数据存储单元的文件数据按每一规定期间进行区分，按各区分生成合并了属于其区分的文件数据内容的期间文件数据，关于所生成的期间文件数据相互之间算出上述相似性，基于算出的相似性生成上述特征数据。

如果是这种构成，由特征数据抽出单元按每一规定期间区分文件数据存储单元的文件数据，按各区分生成期间文件数据。期间文件数据作为合并了属于一个区分的文件数据内容被生成。然后，关于所生成的期间文件数据相互之间算出相似性，基于算出的相似性，生成特征数据。

发明 8

进一步，发明 8 的文件数据检索装置，其特征在于：在发明 6 及 7

之一的文件数据检索装置中，

上述特征数据抽出单元算出表示上述期间文件数据内容特征的文件向量，通过比较算出的文件向量算出上述相似性。

如果是这种构成，由特征数据抽出单元算出表示期间文件数据内容特征的文件向量，通过比较所算出的文件向量，算出相似性。

发明 9

进一步，发明 9 的文件数据检索装置，其特征在于：在发明 8 的文件数据检索装置中，

上述特征数据抽出单元对上述期间文件数据进行词素解析，按各词素作为上述文件向量生成具有作为向量的量的与上述期间文件数据中其词素出现频度对应的元素的向量。

如果是这种构成，由特征数据抽出单元对期间文件数据进行词素解析，按各词素作为文件向量生成具有作为向量的量的与期间文件数据中其词素出现频度对应的元素的向量。在成为比较对象的期间文件数据间不共同的词素出现在任一文件数据的场合，任一文件数据中包含有特征的部分的可能性大。从而，这样，对应期间文件数据中词素出现频度算出相似性，对从庞大的文件数据中检索有特征的部分是有效的。

发明 10

进一步，发明 10 的文件数据检索装置，其特征在于：在发明 6 至 9 之一的文件数据检索装置中，

上述特征数据抽出单元从上述各期间文件数据除去在上述各期间文件数据中共同的内容，基于实施了除去的期间文件数据，算出上述相似性。

如果是这种构成，由特征数据抽出单元从各期间文件数据除去在各期间文件数据中共同的内容，基于实施了除去的期间文件数据，算出相似性。

发明 11

另一方面，为了达到上述目的，发明 11 的数据管理程序是管理多个数据的程序，其特征在于：使计算机执行作为以下各单元所实现的处理，

特征数据抽出单元，其从上述多个数据抽出关于上述数据内容表

示相似性的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点。

如果是这种构成，由计算机读取程序，当计算机按照所读取的程序执行处理时，则得到与发明 1 的数据管理装置同等的作用。

发明 12

进一步，发明 12 的数据管理程序，其特征在于：在发明 11 的数据管理程序中，

上述数据为文件数据。

如果是这种构成，由计算机读取程序，当计算机按照所读取的程序执行处理时，则得到与发明 2 的数据管理装置同等的作用。

发明 13

另一方面，为了达到上述目的，发明 13 的文件数据检索程序，是用于从作成时间或更新时间不同的多个文件数据中进行检索的程序，其特征在于：

针对可利用存储上述多个文件数据用的文件数据存储单元的计算机，使其执行作为以下各单元所实现的处理，

特征数据抽出单元，其从上述文件数据存储单元的文件数据抽出关于上述文件数据内容表示相似性时间推移的特征数据；变化点特定单元，其基于由上述特征数据抽出单元所抽出的特征数据，特定上述相似性的变化点；文件数据检索单元，其以由上述变化点特定单元所特定的变化点为基础，从上述文件数据存储单元中检索上述文件数据。

如果是这种构成，由计算机读取程序，当计算机按照所读取的程序执行处理时，则可得到与发明 3 的数据管理装置同等的作用。

发明 14

另一方面，为了达到上述目的，发明 14 的数据管理方法是管理多个数据的方法，其特征在于：包含

特征数据抽出步骤，其从上述多个数据抽出关于上述数据内容表示相似性的特征数据；变化点特定步骤，其基于由上述特征数据抽出步骤所抽出的特征数据，特定上述相似性的变化点。

发明 15

进一步，发明 15 的数据管理方法，其特征在于：在发明 14 的数

据管理方法中，

上述数据为文件数据。

发明 16

另一方面，为了达到上述目的，发明 16 的文件数据检索方法是用于从作成时间或更新时间不同的多个文件数据中进行检索的方法，其特征在于：包含

文件数据存储步骤，其将上述多个文件数据存储到文件数据存储单元；特征数据抽出步骤，其从上述文件数据存储单元的文件数据抽出关于上述文件数据内容表示相似性时间推移的特征数据；变化点特定步骤，其基于由上述特征数据抽出步骤所抽出的特征数据，特定上述相似性的变化点；文件数据检索步骤，其以由上述变化点特定步骤所特定的变化点为基础，从上述文件数据存储单元中检索上述文件数据。

附图说明

图 1 是表示应用本发明的计算机 100 构成的框图。

图 2 是表示文件向量计算处理的流程图。

图 3 是表示生成期间文件数据的场合的图。

图 4 是表示生成期间文件数据的场合的图。

图 5 是表示文件向量构成的图。

图 6 是表示文件数据检索处理的流程图。

图 7 是表示特征数据的图。

图 8 是表示相似性时间推移的曲线图。

图 9 是表示相似性时间推移的曲线图。

图 10 是用于说明根据 2 元分析检索文件数据的场合的图。

图 11 是用于说明根据文件向量的轨迹预测特定变化点的场合的图。

符号说明

100	计算机
30	CPU
32	ROM
34	RAM
38	I/F

40 输入装置

42 显示装置

44 文件数据登录 DB

实施方式

以下，参照附图说明本发明的实施方式。图 1 到图 9 是表示本发明相关的数据管理装置、文件数据检索装置、数据管理程序和文件数据检索程序以及数据管理方法和文件数据检索方法实施方式的图。

如图 1 所示，本实施方式是将本发明相关的数据管理装置，文件数据检索装置、数据管理程序和文件数据检索程序以及数据管理方法和文件数据检索方法应用于通过计算机 100 从多个文件数据中检索有特征的文件数据的场合。

首先参照图 1 说明应用本发明的计算机 100 的构成。图 1 是表示应用本发明的计算机 100 构成的框图。

如图 1 所示，计算机 100 由以下部分构成，CPU30，其基于控制程序控制运算和系统整体；ROM32，其将 CPU30 的控制程序等预先存入规定区域；RAM34，其用于存入从 ROM32 等所读出的数据或在 CPU30 运算过程所需要的运算结果；I/F38，其针对部装置通过数据的输入输出，这些的连接通过转送数据用的信号线即总线 39，能相互进行数据交流。

在 I/F38 作为外部装置连接有以下装置，输入装置 40，作为人机接口，由可输入数据的键盘和鼠标等构成；显示装置 42，基于图像信号显示画面；文件数据登录 DB44，存入文件数据。

文件数据登录 DB44 例如按各员工存入与业务日志有关的文件数据。从而，在文件数据登录 DB44 存入作成时间或更新时间不同的多个文件数据。

CPU30 由微处理单元 MPU 等构成，启动存入到 ROM32 规定区域的规定程序，按照其程序分别分时执行图 2 及图 6 的流程图中所示的文件向量计算处理及文件数据检索处理。

首先，参照图 2 详细说明文件向量计算处理。图 2 是表示文件向量计算处理的流程图。

文件向量计算处理是算出文件数据检索所需要的文件向量的处理，当在 CPU30 中执行时，如图 2 所示，首先转移到步骤 S100。

在步骤 S100，判定在文件数据登录 DB44 是否作成了新的文件数据，判定作成了新的文件数据时(Yes：是)，转移到步骤 S102。

在步骤 S102，从文件数据编目 DB44 读出属于自基准时间规定期间(例如 1 个月)的文件数据，转移到步骤 S104，生成合并了所读出的文件数据内容的期间文件数据。在步骤 S104，例如，员工作成文件数据的间隔为 1 日单位、上司审阅文件数据的间隔为 1 个月单位的场合，如图 3(a)、(b)所示，如果是 1 月作成的文件数据，将这些改排为作成时间顺序，通过单纯结合，生成 1 月份的期间文件数据。另外，例如，员工作成文件数据的间隔为 1 个月单位、上司审阅文件数据的间隔同样为 1 个月单位的场合，如图 4 所示，1 月只作成了 1 个文件数据时，将其直接作为 1 月份的期间文件数据；1 月作成了多个文件数据时，通过将这些结合，生成 1 月份的期间文件数据。图 3 及图 4 是表示生成期间文件数据的场合的图。

接着，转移到步骤 S106，将所生成的期间文件数据存入文件数据登录 DB44，转移到步骤 S108，判定关于文件数据登录 DB44 的所有文件数据期间文件数据生成结束了否，判定期间文件数据生成结束时(Yes：是)，转移到步骤 S110。

在步骤 S110，对所有期间文件数据进行词素解析，取得任一期间文件数据中出现的所有种类的词素，转移到步骤 S112，将开头的期间文件数据从文件数据登录 DB44 读出，转移到步骤 S114，按在步骤 S110 所取得的各词素，算出所读出的期间文件数据中其词素的出现频度，转移到步骤 S116，作为文件向量算出具有作为向量的量的与算出的出现频度对应的元素的向量。在这里，参照图 5 说明文件向量的算出方法。图 5 是表示文件向量的构成的图。

首先，如图 5 所示，文件向量可以由下式(1)表现、作为 n 元向量。一般，n 为在对所有期间文件数据进行词素解析时所得到的不重复单词数。由 TFIDF(Term Frequency＆Inverse Document Frequency)求各单词的加权 W。

数式 1

$$\overrightarrow{D}=(W_1,W_2,\cdots,W_n) \qquad \cdots(1)$$

TFIDF 根据下式(2)由在期间文件数据内的单词出现频度(TF:

Term Frequency)与使用了在期间文件数据整体的该单词的期间文件数据数频度的倒数(IDF: Inverse Doxument Frequency)之积来求，数值越大，表示该单词越重要。TF 是常出现的单词重要这一指标，如下式(3)所示，具有当某期间文件数据中单词出现的频度增加则大的性质。IDF 是多个期间文件数据中出现的单词不重要、即特定期间文件数据中出现的单词重要这一指标，如下式(4)～(6)所示，具有当使用某单词的期间文件数据数减少则变大的性质。从而，TFIDF 的值具有针对常出现而在多个期间文件数据中出现的单词(接续词、助词等)或针对只在特定期间文件数据中出现而在其期间文件数据中频度也小的单词则变小、反之，针对特定期间文件数据中以高频度出现的单词则变大的性质。由于 TFIDF，期间文件数据内的单词被数值化，能够以该数值为元素，期间文件数据向量化。

数式 2

$$W(t, d)=TF(t, d)\times IDF(t) \quad \cdots\cdots(2)$$

数式 3

TF(t, d)=在期间文件数据 d 中单词 t 出现的频度 ……(3)

数式 4

$$IDF(t)=\log\left(\frac{D}{DF(t)}\right) \quad \cdots(4)$$

数式 5

DF(t)=在期间文件数据整体中单词 t 出现的期间文件数据数的频度…(5)

数式 6

D=全期间文件数据数 ……(6)

接着，转移到步骤 S118，将算出的文件向量存入到文件数据登录 DB44，转移到步骤 S120，判定关于所有期间文件数据步骤 S112～S118 的处理是否结束，当判定关于所有期间文件数据处理结束时(Yes: 是)，结束一系列的处理，返回到原处理。

另一方面，在步骤 S120，判定关于所有期间文件数据步骤 S112～S118 的处理没结束时(No: 否)，转移到步骤 S122，将下一期间文件数据从文件数据登录 DB44 读出，转移到步骤 S114。

另一方面，在步骤 S108，判定关于文件数据登录 DB44 的所有文件数据期间文件数据的生成没结束时(No：否)，转移到步骤 S124，将属于下一规定期间的文件数据从文件数据登录 DB44 读出，转移到步骤 S104。

再一方面，在步骤 S100，判定在文件数据登录 DB44 没有作成新的文件数据时(No：否)，转移到步骤 S126，判定文件数据登录 DB44 的文件数据是否被更新，判定文件数据被更新时(Yes：是)，转移到步骤 S102，判定未被更新时(No：否)，转移到步骤 S100。

接着，参照图 6 详细说明文件数据检索处理。图 6 是表示文件数据检索处理的流程图。

文件数据检索处理是特定关于在时间序列上邻接的期间文件数据相似性变化点、从文件数据登录 DB44 中检索属于所特定的变化点的文件数据的处理，当在 CPU30 中执行时，如图 6 所示，首先转移到步骤 S200。

在步骤 S200，判定输入了来自用户的检索要求否，判定输入了检索要求时(Yes：是)，转移到步骤 S202，判定没有输入时(No：否)，在步骤 S100 待机，直到输入检索要求为止。另外，这里所说的检索要求并非检索关键字或文章，而是对计算机 100 提出应检索的要求。

在步骤 S202，将开头的期间文件数据的文件向量从文件数据登录 DB44 读出，转移到步骤 S204，将与所读出的文件向量相关的期间文件数据中在时间序列上邻接的期间文件数据(在时间上与新的一方邻接的期间文件数据)的文件向量从文件数据登录 DB44 读出，转移到步骤 S206。

在步骤 S206，通过使用所读出的 2 个文件向量进行向量运算，算出与这些相关的期间文件数据的相似性。基于向量运算的相似性的计算被称为向量检索技术，通过反映单词的重要性进行数值化的 TFIDF 和由此计算向量化了的文件相似性的向量空间模型来实现。例如，以所读出的 2 个文件向量为文件向量 D_1、D_2 的场合，相似性可以由下式(7)作为文件向量 D_1、D_2 之间形成的角的余弦值(0～1)算出。

数式 7

$$\text{向量的余弦（角度）} = \frac{\overrightarrow{D_1}\,\overrightarrow{D_2}}{|\overrightarrow{D_1}||\overrightarrow{D_2}|}$$

可以只在双方都不是0的场合进行计算
如果某一方为0则无须计算

$$= \frac{W_{m1}W_{q1}+W_{m2}W_{q2}+\cdots+W_{mn}W_{qn}}{\sqrt{W_{m1}^{\ 2}+W_{m2}^{\ 2}+\cdots+W_{mn}^{\ 2}}\sqrt{W_{q1}^{\ 2}+W_{q2}^{\ 2}+\cdots+W_{qn}^{\ 2}}} \quad \cdots(7)$$

接着转移到步骤 S208，判定关于所有文件向量步骤 S204、S206 的处理结束了否，判定关于所有文件向量处理结束时(Yes：是)，转移到步骤 S210。

在步骤 S210，基于在步骤 S206 所算出的 1 或多个期间文件数据的相似性，生成关于这些期间文件数据内容表示相似性时间推移的特征数据。如图 7 所示，以图 3 的例为对象的场合，特征数据被生成作为文件向量之间形成的角的余弦值(0～1)。图 7 是表示特征数据的图。

接着，转移到步骤 S211，基于所生成的特征数据，特定相似性的变化点。具体说，基于所生成的特征数据，设定允许范围，特定在相似性时间推移中超过允许范围的点作为变化点。例如，如图 8 所示，可以根据相似性的平均值及分散求形成临界线的 2 条水平线、设定由这些临界线所围起的区域作为允许范围。这种场合，因为期间文件数据 P_x 的相似性超过了该允许范围，所以将其特定作为变化点。另外，例如，如图 9 所示，也可以根据相似性的平均值及分散求沿相似性推移曲线的 2 条近似曲线、设定由这些近似曲线所围起的区域作为允许范围。这种场合，同样，因为期间文件数据 P_x 的相似性超过了该允许范围，所以将其特定作为变化点。图 8 及图 9 是表示相似性时间推移的曲线图。

接着，转移到步骤 S212，从文件数据登录 DB44 中检索所特定的变化点或属于其附近的文件数据。在图 3 的例中，例如，在 10 月和 11 月之间存在相似性变化点的场合，可知从 10 月转移到 11 月时，业务内容发生了变化，因此，可以按照小的日期顺序检索 11 月业务日志的文件数据。

接着，转移到步骤 S214，将由检索所抽出的文件数据改排为相似性高的顺序，生成文件数据一览，转移到步骤 S216，将所生成的文件数据一览显示于显示装置 42，结束一系列处理，返回到原处理。

另一方面，在步骤S208，判定关于所有文件向量步骤S204、S206的处理没结束时（No：否），转移到步骤S218，将下一期间文件数据的文件向量从文件数据登录DB44读出，转移到步骤S204。

下面，说明本实施方式的动作。

在某企业等，通过让员工提交业务日志来管理业务的进展情况。由业务日志形成的报告由一个上司对多个部下所提交的业务日志一一过目审阅。各员工作成记载了每天业务情况的业务日报作为文件数据，将所作成的文件数据添附于邮件，寄送给上司同时登录于文件数据登录DB44。

首先说明根据各员工所作成的文件数据作成文件向量的场合。

当作成文件数据，经步骤S100～S106，属于自基准时间规定期间（例如1个月）的文件数据被从文件数据登录DB44读出，生成合并了所读出的文件数据内容的期间文件数据，所生成的期间文件数据被存入文件数据登录DB44。然后，反复经过步骤S102、S104，进行关于文件数据登录DB44的所有文件数据期间文件数据的生成和存入。

当生成关于所有文件数据期间文件数据，经步骤S110，对所有期间文件数据进行词素解析，取得任一期间文件数据中出现的所有种类词素。接着，经步骤S112～S118，开头的期间文件数据被从文件数据登录DB44读出，按所取得的各词素算出所读出的期间文件数据中其词素的出现频度，作为文件向量算出具有作为向量的量的与所算出的出现频度对应的元素的向量。然后，反复经过步骤S114～S118，进行关于所有期间文件数据出现频度的算出及文件向量的算出和存入。

下面，说明上司审阅各员工所寄送来的文件数据的场合。

上司在审阅文件数据前首先输入检索要求。当检索要求被输入，经步骤S200～S206，开关的期间文件数据的文件向量被从文件数据登录DB44读出，与所读出的文件向量相关的期间文件数据中在时间序列上邻接的期间文件数据的文件向量被从文件数据登录DB44读出，通过使用所读出的2个文件向量进行向量运算，算出与这些相关的期间文件数据的相似性。然后，反复经过步骤S204、S206，进行关于所有文件向量邻接文件向量的读出和相似性的算出。

当算出关于所有文件向量相似性，经步骤S210、S211，基于所算出的1或多个期间文件数据的相似性，生成关于这些期间文件数据内

容表示相似性时间推移的特征数据，基于所生成的特征数据，特定相似性的变化点。接着，经步骤212，从文件数据登录DB44中检索所特定的变化点或属于其附近的文件数据。其结果，当该文件数据被抽出，经步骤S214、S216，由检索所抽出的文件数据按相似性高的顺序被改排，生成文件数据一览，所生成的文件数据一览被显示于显示装置42。

作为检索结果显示于显示装置42的文件数据为相似性的变化点或属于其附近的文件数据，即是认为业务内容有变化时的文件数据，因此，上司对所有文件数据过目困难的场合，可以从由检索所抽出的文件数据优先审阅，由此，审阅的业务日志量即使庞大的场合，也可以有效管理业务的进展情况。

这样，在本实施方式中，从文件数据登录DB44的文件数据抽出关于文件数据内容表示相似性时间推移的特征数据，基于所抽出的特征数据特定相似性的变化点，以所特定的变化点为基础，从文件数据登录DB44中检索文件数据。

由此，用户通过参照由检索所抽出的文件数据，可以从庞大的文件数据中比较容易地掌握有特征的部分，另外，因为从多个文件数据抽出特征数据，所以与累积专家所建立的规则的场合相比，容易提高抽出的可靠性，而且可以比较适应用户的要求。

进一步，在本实施方式中，从文件数据登录DB44中检索所特定的变化点或属于其附近的文件数据。

由此，因为变化点或属于其附近的文件数据被检索，所以用户可以从宠大的文件数据中更加容易地掌握有特征的部分。

进一步，在本实施方式中，基于所抽出的特征数据，设定允许范围，特定在相似性的时间推移中超过允许范围的点作为变化点。

由此，可以统一进行变化点的特定，因此，比较容易特定变化点。

进一步，在本实施方式中，按每一规定期间区分文件数据登录DB44的文件数据，按各区分生成合并了属于其区分的文件数据内容的期间文件数据，关于在时间序列上邻接的期间文件数据算出相似性，基于所算出的相似性生成特征数据。

由此，在时间序列上看文件数据的关系时，用户可以比较容易地掌握有特征的部分。

进一步，在本实施方式中，对期间文件数据进行词素解析，按各

词素作为文件向量生成具有作为向量的量的与期间文件数据中其词素出现频度对应的元素的向量。

由此，对应期间文件数据中词素出频度算出相似性，因此，能够以比较结合实际情况的形式算出相似性，用户可以从庞大的文件数据中更加容易地掌握有特征的部分。

进一步，在本实施方式中，在文件向量的角度计算中，如上式（7）所示，只计算相同元数之间的加权 W 不是「0」的部分。

由此，可以使计算省略化。

在上述实施方式中，文件数据登录 DB44 对应发明 3、4、6、13 或 16 的文件数据存储单元，步骤 S210 对应发明 1、3、5、6、8、9、11 或 13 的特征数据抽出单元，或者对应发明 14 或 16 的特征数据抽出步骤。另外，步骤 S211 对应发明 1、3 到 5、11 或 13 的变化点特定单元，或者对应发明 14 或 16 的变化点特定步骤，步骤 S212 对应发明 3、4 或 13 的文件数据检索单元，或者对应发明 16 的文件数据检索步骤。

另外，在上述实施方式中，构成为按每一规定期间区分文件数据登录 DB44 的文件数据，按各区分生成合并了属于其区分的文件数据内容的期间文件数据，关于在时间序列上邻接的期间文件数据算出相似性，基于所算出的相似性生成特征数据。但不限于此，如图 10 所示，也可以构成为按每一规定期间区分文件数据登录 DB44 的文件数据，按各区分生成合并了属于其区分的文件数据内容的期间文件数据，关于所生成的期间文件数据的相互算出相似性，基于所算出的相似性生成特征数据。图 10 是用于说明根据 2 元分析检索文件数据的场合的图。

如上述实施方式，只在邻接期间互相比较，缓慢变化的场合才可能纳入稳定状态。作为分析方法，虽然多少需要些成本，但当进行图 10 所示那样的 2 元分析，也可以检出缓慢的变化。当然，不限于进行 2 元分析，将此想法展开，也可以进行 3 元以上的多元分析。

由此，按每一规定期间互相看文件数据的关系时，用户可以比较容易地掌握有特征的部分。

在这种场合，文件数据登录 DB44 对应发明 7 的文件数据存储单元，步骤 S210 对应发明 7 的特征数据抽出单元。

另外，在上述实施方式中，构成为按每一规定期间区分文件数据登录 DB44 的文件数据，按各区分生成合并了属于其区分的文件数据内

容的期间文件数据，基于所生成的期间文件数据算出相似性，但不限于此，也可以构成为从各期间文件数据除去在各期间文件数据中共同的内容，基于实施了除去的期间文件数据，算出相似性。

由此，因为除去共同的内容后，算出相似性，因此，能够以比较结合实际情况的形式算出相似性，用户可以从庞大的文件数据中更容易地掌握有特征的部分。

在这种场合，步骤 S210 对应发明 10 的特征数据抽出单元。

另外，在上述实施方式中，构成为基于所抽出的特征数据，设定允许范围，特定在相似性时间推移中超过允许范围的点作为变化点，但不限于此，如图 11 所示，也可以构成为关于各期间文件数据的文件向量预测多元向量空间中的轨迹，设定预测范围，特定超过预测范围的文件向量作为变化点。图 11 是用于说明根据文件向量的轨迹预测特定变化点的场合的图。

另外，在上述实施方式中，在执行图 2 及图 6 的流程图中所示的处理时，都是说明了关于执行预先存入到 ROM32 的控制程序的场合，但不限于此，也可以从存储了示有这些顺序的程序的存储媒体，将其程序读入到 RAM34 来进行执行。

在这里，所谓存储媒体是 RAM、ROM 等半导体存储媒体；FD、HD 等磁性存储型存储媒体；CD、CDV、LD、DVD 等光学读取式存储媒体；MO 等磁性存储型/光学读取式存储媒体，不管电子的、磁性的、光学的等读取方法，只要是用计算机可读取的存储媒体都包含在内，包含所有的存储媒体。

另外，如图 1 所示，在上述实施方式中，将本发明相关的数据管理装置、文件数据检索装置、数据管理程序和文件数据检索程序以及数据管理方法和文件数据检索方法通过计算机 100 应用于从多个文件数据中检索有特征的文件数据的场合，但不限于此，在不脱离本发明主旨的范围内，也可以应用于其它场合。例如在因特网等其它网络中，也可以应用作为从多个文件数据中检索有特征的文件数据的检索服务。

发明效果

如以上说明，如果根据本发明相关的权利要求 1 或 2 记载的数据管理装置，用户通过参照所特定的变化点，可以从庞大的数据中比较

容易地掌握有特征的部分。另外，因为从多个数据抽出的特征数据，因此，与累积专家所建立的规则的场合相比，容易提高抽出的可靠性，而且，能够比较适应用户的要求。

进一步，如果根据本发明相关的权利要求2记载的数据管理装置，用户通过参照所特定的变化点，可以从庞大的文件数据中比较容易地掌握有特征的部分。

另一方面，如果根据本发明相关的权利要求3到10记载的文件数据检索装置，用户通过参照由检索所抽出的文件数据，可以从庞大的文件数据中比较容易地掌握有特征的部分。另外，因为从多个文件数据抽出特征数据，因此，与累积专家所建立的规则的场合相比，容易提高抽出的可靠性，而且能够适应用户的要求。

进一步，如果根据本发明相关的权利要求4记载的文件数据检索装置，因为变化点或属于其附近的文件数据被检索，因此，用户可以从庞大的文件数据中更加容易地掌握有特征的部分。

进一步，如果根据本发明相关的权利要求5记载的文件数据检索装置，因为能够统一地进行变化点的特定，因此，比较容易特定变化点。

进一步，如果根据本发明相关的权利要求6记载的文件数据检索装置，在时间序列上看文件数据的关系时，用户可以比较容易地掌握有特征的部分。

进一步，如果根据本发明相关的权利要求7记载的文件数据检索装置，按每一规定期间相互看文件数据的关系时，用户可以比较容易地掌握有特征的部分。

进一步，如果根据本发明相关的权利要求9记载的文件数据检索装置，因为对应期间文件数据中的词素出现频度算出相似性，因此能够以比较结合实际情况的形式算出相似性，用户可以从庞大的文件数据中更容易地掌握有特征的部分。

进一步，如果根据本发明相关的权利要求10记载的文件数据检索装置，因为除去共同的内容后算出相似性，因此能够以比较结合实际情况的形式算出相似性，用户可以从庞大的文件数据中更容易地掌握有特征的部分。

另一方面，如果根据本发明相关的权利要求11或12记载的数据

管理程序，得到的效果与权利要求 1 记载的数据管理装置相同。

进一步，如果根据本发明相关的权利要求 12 记载的数据管理程序，得到的效果与权利要求 2 记载的数据管理装置也相同。

另一方面，如果根据发明相关的权利要求 13 记载的文件数据检索程序，得到的效果与权利要求 3 记载的数据管理装置相同。

另一方面，如果根据本发明相关的权利要求 14 或 15 记载的数据管理方法，得到的效果与权利要求 1 记载的数据管理装置相同。

进一步，如果根据本发明相关的权利要求 15 记载的数据管理方法，得到的效果与权利要求 2 记载的数据管理装置也相同。

另一方面，如果根据本发明相关的权利要求 16 记载的文件数据检索方法，得到的效果与权利要求 3 记载的数据管理装置相同。

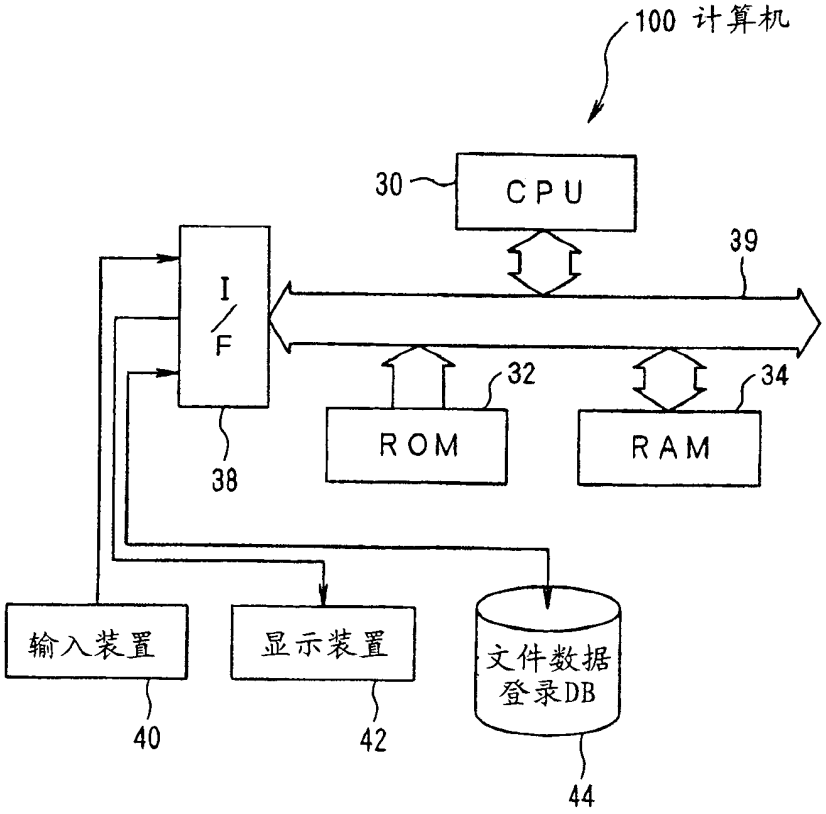


图 1

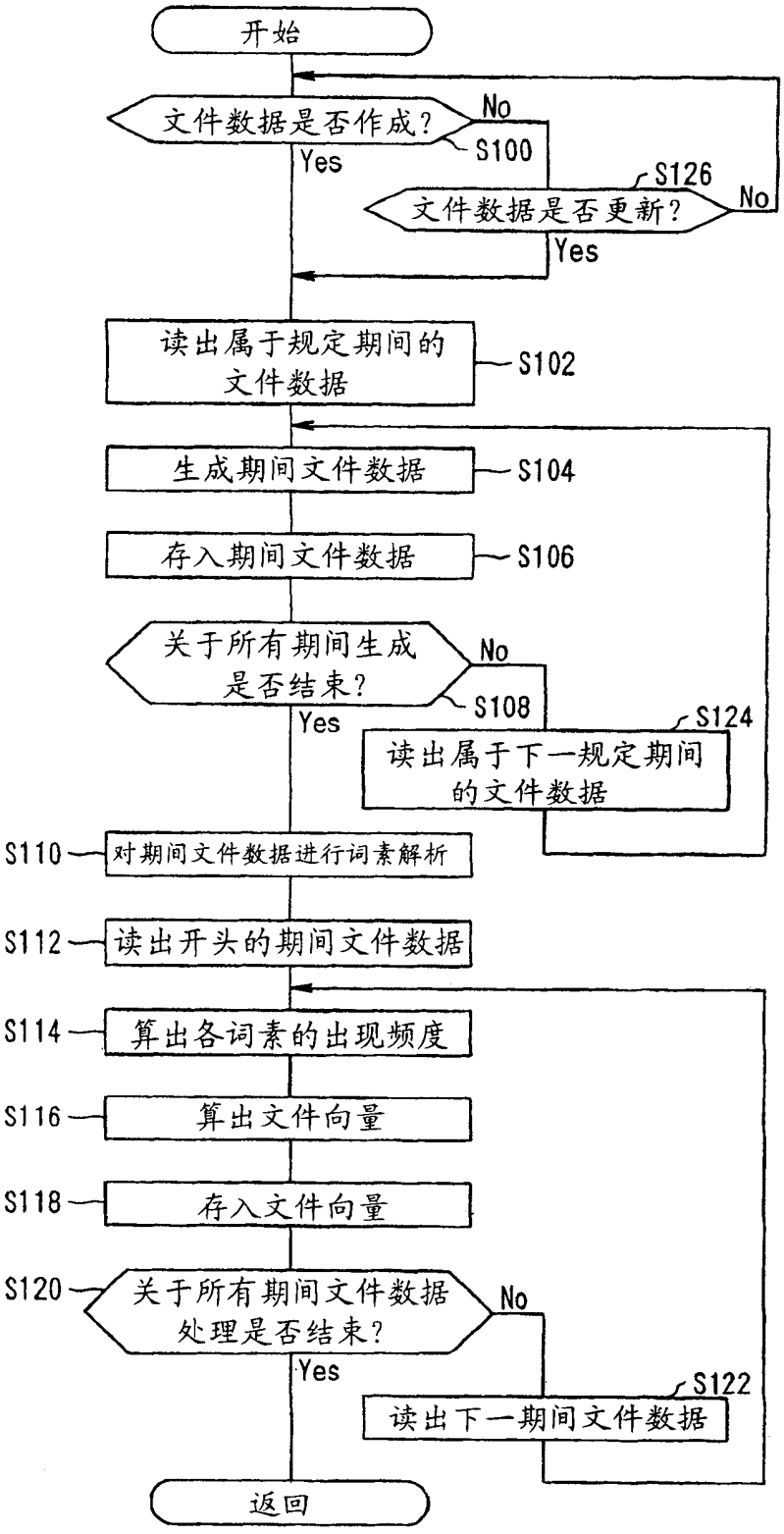

开始
文件数据是否作成？
No
S100
Yes
S126
文件数据是否更新？
No
Yes
读出属于规定期间的文件数据
S102
生成期间文件数据
S104
存入期间文件数据
S106
关于所有期间生成是否结束？
No
S108
Yes
S124
读出属于下一规定期间的文件数据
S110
对期间文件数据进行词素解析
S112
读出开头的期间文件数据
S114
算出各词素的出现频度
S116
算出文件向量
S118
存入文件向量
S120
关于所有期间文件数据处理是否结束？
No
Yes
S122
读出下一期间文件数据
返回

图 2

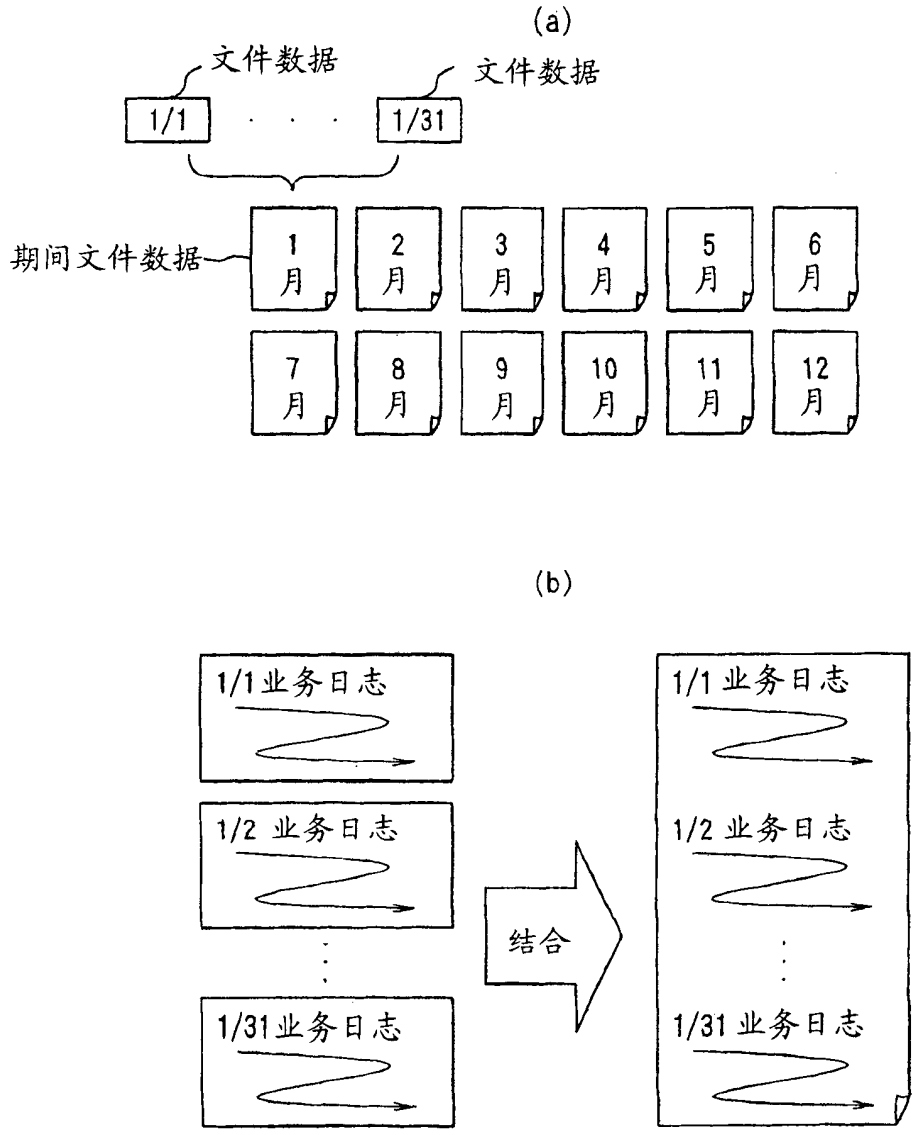

(a)
文件数据
文件数据
1/1
1/31
期间文件数据
1月
2月
3月
4月
5月
6月
7月
8月
9月
10月
11月
12月
(b)
1/1业务日志
1/2 业务日志
1/31业务日志
结合
1/1 业务日志
1/2 业务日志
1/31 业务日志

图 3

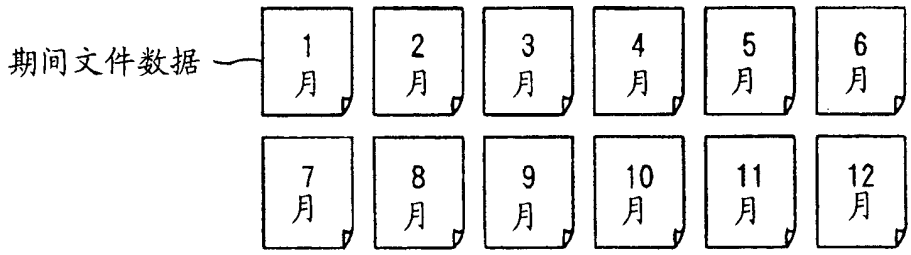

期间文件数据
1月
2月
3月
4月
5月
6月
7月
8月
9月
10月
11月
12月

图 4

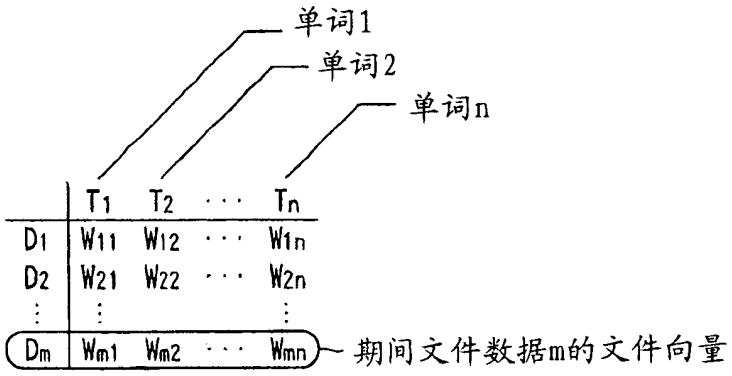

单词1
单词2
单词n
T1 T2 ··· Tn
D1 W11 W12 ··· W1n
D2 W21 W22 ··· W2n
Dm Wm1 Wm2 ··· Wmn
期间文件数据m的文件向量

图 5

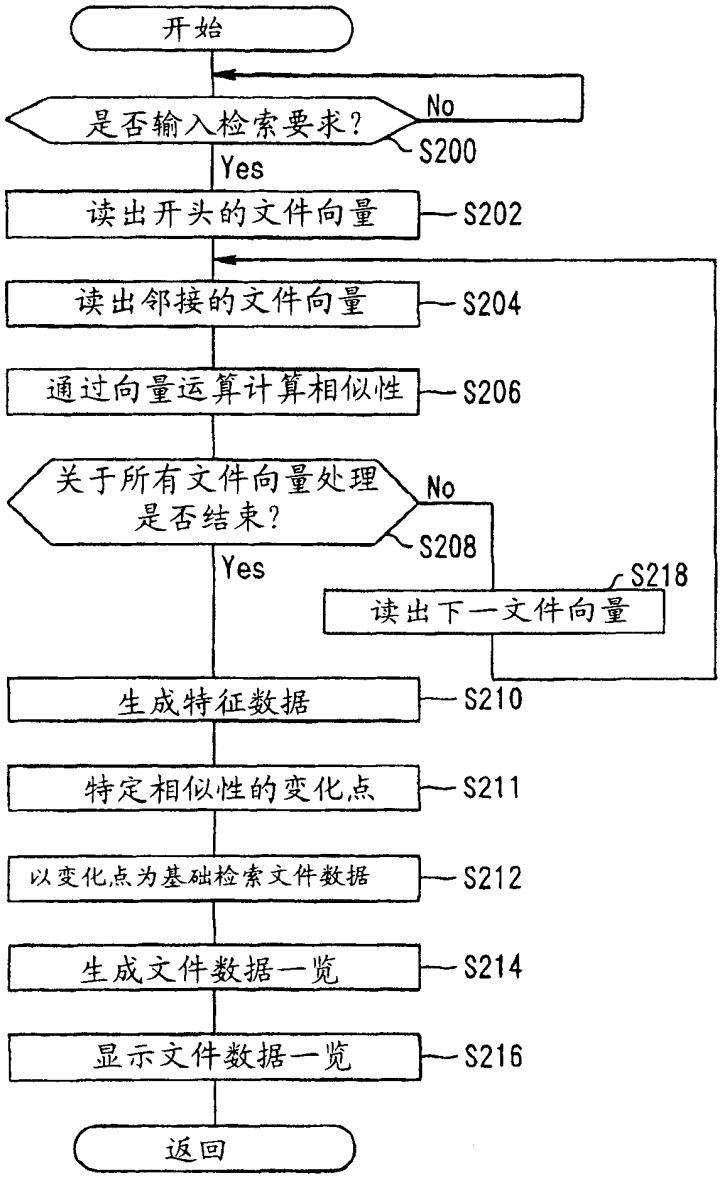

开始
是否输入检索要求?
No
S200
Yes
读出开头的文件向量
S202
读出邻接的文件向量
S204
通过向量运算计算相似性
S206
关于所有文件向量处理是否结束?
No
S208
Yes
S218
读出下一文件向量
生成特征数据
S210
特定相似性的变化点
S211
以变化点为基础检索文件数据
S212
生成文件数据一览
S214
显示文件数据一览
S216
返回

图 6

1月⇔2月	0.124477
2月⇔3月	0.148288
3月⇔4月	0.096936
4月⇔5月	0.176061
5月⇔6月	0.176194
6月⇔7月	0.124569
7月⇔8月	0.088018
8月⇔9月	0.146267
9月⇔10月	0.141868
10月⇔11月	0.164984
11月⇔12月	0.163609

图 7

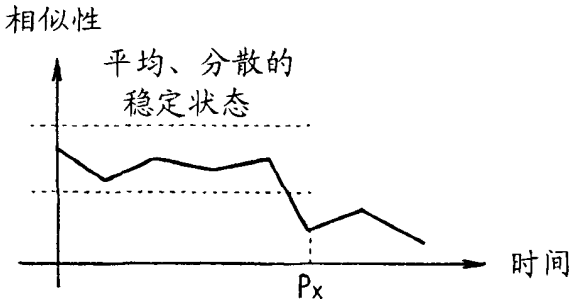


图 8

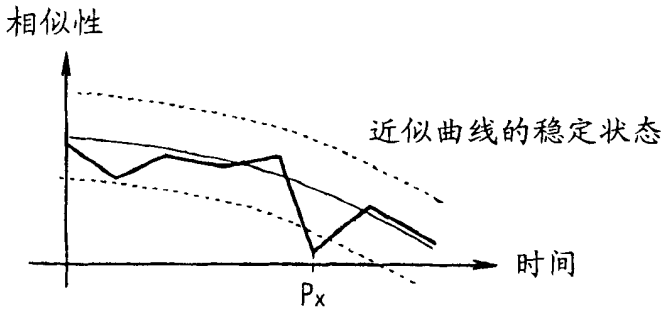


图 9

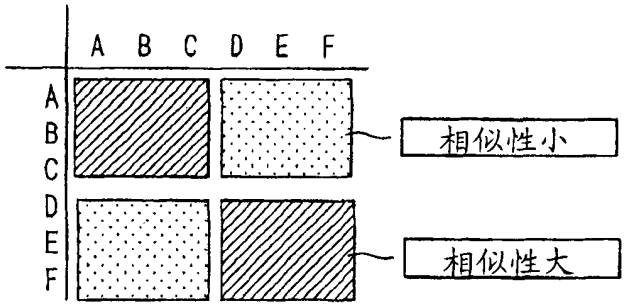


图 10

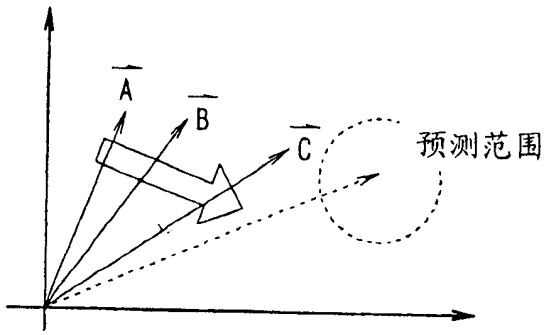


图 11