[19] 中华人民共和国国家知识产权局

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

[12] 发明专利说明书

专利号 ZL 200610088580.X

[45] 授权公告日 2009 年 7 月 8 日

[11] 授权公告号 CN 100511232C

[22] 申请日 2006.6.6
[21] 申请号 200610088580.X
[30] 优先权
[32] 2005. 6. 7 [33] JP [31] 167347/2005
[73] 专利权人 佳能株式会社
地址 日本东京都
[72] 发明人 户岛英一郎
[56] 参考文献
CN1336604A 2002.2.20
CN1157673C 2004.7.14
CN1135485C 2004.1.21
CN1612154A 2005.5.4
US2004/0243601A1 2004.12.2
CN1503193A 2004.6.9
审查员 董 刚

[74] 专利代理机构 北京怡丰知识产权代理有限公司
代理人 于振强

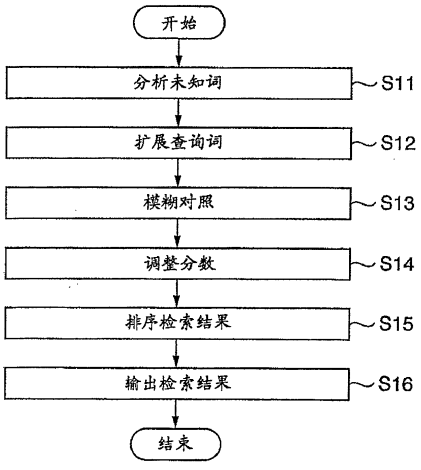
权利要求书 3 页 说明书 14 页 附图 15 页

[54] 发明名称

文档检索装置及方法

[57] 摘要

本发明提供一种文档检索装置及方法，根据检索查询词和该检索查询词的扩展查询词，对检索对象的文档数据进行检索，抽取与所述检索查询词和所述扩展查询词一致的字符串，判断抽出的字符串是否包含未知词区域，当判断为不包含未知词区域时，进行调整，从而使抽出的字符串的相似度下降，以与该调整后的相似度对应的顺序，将字符串作为检索结果输出。



1．一种文档检索装置，其特征在于，包括：

对照装置，基于检索查询词和与该检索查询词相似的扩展查询词，对作为检索对象的文档数据进行检索，抽取与所述检索查询词和所述扩展查询词一致的字符串；

分析装置，分析作为所述检索对象的文档数据，识别未知词区域；

判断装置，判断由所述对照装置抽出的字符串的至少一部分是否包含在所述未知词区域中；以及

检索结果输出装置，根据所述对照装置和所述判断装置的处理结果，在判断为由所述对照装置抽出的字符串未包含在所述未知词区域中的情况下，降低所述字符串的相似度，输出所述字符串作为检索结果。

2．根据权利要求1所述的文档检索装置，其特征在于：

所述检索结果输出装置具有分数校正装置，该分数校正装置在由所述判断装置判断为不包含所述未知词区域时，使由所述对照装置抽出的字符串的相似度下降；以与所述分数校正装置取得的相似度对应的顺序，将所述字符串作为检索结果输出。

3．根据权利要求1所述的文档检索装置，其特征在于：

所述扩展查询词，是将对构成所述检索查询词的字符进行字符识别时误识别为另一字符的概率高的字符，替换成所述另一字符，这样构成的字符串。

4．根据权利要求2所述的文档检索装置，其特征在于：

所述对照装置，根据所述一致的字符串中包含的与所述检索查询词的字符一致的字符数、和所述一致的字符串中包含的与所述扩展查询词的字符一致的字符数，求出所述一致的字符串的相似度。

5．根据权利要求1所述的文档检索装置，其特征在于：

所述分析装置进行所述文档数据的词素分析，根据词素分析取

得的短语中包含的单词是否包含在单词字典中，识别所述未知词区域。

6. 根据权利要求 2 所述的文档检索装置，其特征在于：

所述分析装置分析作为所述检索对象的文档数据，以短语为单位分割该文档数据，

所述分数校正装置，从所述抽出的字符串的相似度，减去包含由所述对照装置抽出的字符串的至少一部分的短语的最大字符数，来降低相似度。

7. 一种文档检索方法，包括：

对照步骤，基于检索查询词和该检索查询词的扩展查询词，对作为检索对象的文档数据进行检索，抽取与所述检索查询词和所述扩展查询词一致的字符串；

分析步骤，分析所述检索对象的文档数据，识别未知词区域；

判断步骤，判断由所述对照步骤抽出的字符串的至少一部分是否包含在所述的未知词区域中；以及

检索结果输出步骤，根据所述对照步骤和判断步骤的处理结果，在判断为由所述对照步骤所抽取的字符串未包含在所述未知词区域中的情况下，降低所述字符串的相似度，输出在所述字符串作为检索结果。

8. 根据权利要求 7 所述的文档检索方法，其特征在于：

在所述检索结果输出步骤中具有分数校正步骤，该分数校正步骤在由所述判断步骤判断为不包含所述未知词区域时，使由所述对照步骤抽出的字符串的相似度下降；

以与所述分数校正步骤取得的相似度对应的顺序，将所述字符串作为检索结果输出。

9. 根据权利要求 7 所述的文档检索方法，其特征在于：

所述扩展查询词，是将对构成所述检索查询词的字符进行字符识别时误识别为另一字符的概率高的字符，替换成所述另一字符，这样构成的字符串。

10. 根据权利要求 8 所述的文档检索方法，其特征在于：

所述对照步骤，根据所述一致的字符串中包含的与所述检索查询词的字符一致的字符数、和所述一致的字符串中包含的与所述扩展查询词的字符一致的字符数，求出所述一致的字符串的相似度。

11. 根据权利要求 7 所述的文档检索方法，其特征在于：

所述分析步骤，进行所述文档数据的词素分析，根据词素分析取得的短语中包含的单词是否包含在单词字典中，识别所述未知词区域。

12. 根据权利要求 8 所述的文档检索方法，其特征在于：

所述分析步骤分析作为所述检索对象的文档数据，以短语为单位分割该文档数据，

所述分数校正步骤，从所述抽出的字符串的相似度，减去包含所述对照步骤抽出的字符串的至少一部分的短语的最大字符数，来降低相似度。

文档检索装置及方法

技术领域

本发明涉及按照检索查询词（query），检索文档数据的文档检索装置、检索方法以及存储介质。

背景技术

伴随着个人计算机（PC）的普及，文档的生成一般使用文档生成软件等PC上的应用软件来进行。具体而言，广泛进行在PC的画面上生成、编辑各种文件，对文件进行复制、检索这样的作业。此外，伴随着网络的发展和普及，通常，这样在PC上生成的电子的文档数据（电子文档数据），不使用打印机等作为纸文档打印，而由其它PC等访问，用电子邮件发送、分发，无纸的文档生成环境正在扩大。

构筑文档管理系统，以由计算机系统地管理实现这样的无纸化的电子文档数据。该电子文档数据，在基于文档共享的高效的信息量的削减、文档间建立关联等方面，极其便利。伴随着电子文档数据的普及，文档数据的全文检索、关键字检索等检索操作普及，检索的有效性逐渐广为人知。

另一方面，在纸上打印文档数据的纸文档，与电子文档数据相比，具有容易阅读、处理通用、容易搬运、容易把握全貌等优点。例如当需要分发资料时，依然以打印装置打印电子数据后形成的纸文档这种形式进行分配。可是，纸文档无法以原样的形式检索，所以不容易检索打印有所需信息的纸文档。因此，以往使用扫描纸文档并进行OCR（Optical Character Recognition）处理电子文本化的电子文档数据，进行检索。可是，在OCR处理中，如果发生误识别，就无法正确检索用户所需的文档数据。

为了解决这样的问题，以往提出了各种方案。作为一个例子，有假定文档数据的字符串的字符遗漏、字符混合、字符走样，对照检索查询词（query）和文档数据的字符串来进行检索的方法。此外，还提出将检索查询词的各字符扩展成假定的误识别字符，并与该扩展的检索查询词对照，从而检索文档数据的方法。这里，将这些对照方法统称为模糊对照。

通过这样的模糊对照，能命中（hit）由于误识别而漏取的字符串，但是也有很多弊端。例如作为检索查询词输入“イラク”，连由于字符识别“イラク”时的误识别而产生的误识别字符（例如“イテク”）也要作为检索对象使其命中。这时，也命中例如“ハイテク”（未误识别）中的“イテク”。据此，在文档数据中每当使用“ハイテク”这一单词，就将产生无关的命中。这样的不想要的命中多发时，需要一种用于选择有意义的命中的作业，成为对于用户来说作业负荷增大、难以使用的检索装置。

作为与此相关的技术，有日本特开2004-334334号公报。

可是，在日本特开2004-334334号公报的技术中，依然会产生无关的命中。例如想检索“人間”这一字符串（或者误识别“人間”后产生的字符串）时，作为查询词，指定“人間”。汉字“間”和“関”相似，所以作为用于误识别的字符的检索查询词，也设定“人関”。于是，当文档数据中存在“被告人関係者”这一字符串时，尽管该字符串未被误识别，但是该字符串中的“人関”也命中了。这时，“人関”、“告人関”、“人関係”等并不形成字典单词的一部分，在以往技术中无法抑制该命中。

发明内容

本发明在于解决所述以往技术的缺点。

此外，本发明的特征在于，提供能高效检索被误识别的字符串的文档检索装置及其方法。

本发明提供一种文档检索装置，包括：对照装置，基于检索查

询词和与该检索查询词相似的扩展查询词，对作为检索对象的文档数据进行检索，抽取与所述检索查询词和所述扩展查询词一致的字符串；分析装置，分析作为所述检索对象的文档数据，识别未知词区域；判断装置，判断由所述对照装置抽出的字符串的至少一部分是否包含在所述未知词区域中；以及检索结果输出装置，根据所述对照装置和所述判断装置的处理结果，在判断为由所述对照装置抽出的字符串未包含在所述未知词区域中的情况下，降低由所述对照装置抽出的字符串的相似度，输出所述字符串作为检索结果。

并且，本发明提供一种文档检索方法，包括：对照步骤，基于检索查询词和该检索查询词的扩展查询词，对作为检索对象的文档数据进行检索，抽取与所述检索查询词和所述扩展查询词一致的字符串；分析步骤，分析所述检索对象的文档数据，识别未知词区域；判断步骤，判断由所述对照步骤抽出的字符串的至少一部分是否包含在所述的未知词区域中；以及检索结果输出步骤，根据所述对照步骤和判断步骤的处理结果，在判断为由所述对照步骤所抽取的字符串未包含在所述未知词区域中的情况下，降低所述字符串的相似度，输出在所述字符串作为检索结果。

本发明的概要并非完全列举必要的特征，因此这些特征群的子组合（sub-combination）也能成为发明。

在以下的参照附图的说明中，本发明其它特征、目的和优势将变得明显，在附图中相似的参照符号表示相同或相似的部分。

附图说明

附图构成说明书的一部分，描述本发明的实施例，与说明一起用来解释发明的原理。

图 1 是表示本发明实施例的文档检索装置的结构的框图。

图 2A、图 2B 是说明本实施例的检索的操作例的图。

图 3 是说明本实施例的未知词分析处理的一例的图。

图 4 是说明本实施例的存储未知词区域的未知词区域表的数据

结构的图。

图 5 是说明本实施例的用于将检索查询词的各字符（字符串）扩展为相似字符串的查询词扩展表的数据结构的图。

图 6 是说明将本实施例的检索查询词扩展为有可能通过误识别扩展的字符串的扩展查询词格网的图。

图 7 是表示本实施例的用于认定词素分析结果中哪部分应该成为未知词区域的规则即未知词区域认定规则的存储形式的图。

图 8 是表示本实施例的存储检索结果的候选的检索结果表的结构的图。

图 9 是说明本实施例的规定检索结果的输出顺序的分数的计算式的图。

图 10 是说明本实施例的文档检索装置的处理的流程图。

图 11 是说明作为图 10 的步骤 S4 中的事件对应处理的一部分的检索处理的流程图。

图 12 是说明图 11 的步骤 S11 的未知词分析处理的流程图。

图 13 是说明图 12 的步骤 S23 的未知词区域的抽取处理的流程图。

图 14 是说明图 11 的步骤 S13 的模糊对照处理的流程图。

图 15 是说明图 11 的步骤 S14 的分数调整处理的流程图。

具体实施方式

下面参照附图详细说明本发明的优选实施例。以下的实施例并不限定关于权利要求书的发明，此外本实施例中说明的特征的组合的全部并不一定是发明的解决手段所必须的。

图 1 是表示本发明实施例的文档检索装置的结构的框图。

在图中，CPU101 是微处理器，按照 ROM102 或 RAM103 中存储的程序，进行用于图象处理（image processing）、字符处理（character processing）、字符识别处理（character recognition processing）、检索处理（search processing）的运算、逻辑判断等，控制通过总线 120

连接的各构成要素。总线 120 是系统总线，传送指示作为 CPU101 的控制对象的各构成要素的地址信号、数据以及控制信号。ROM102 是读出专用的非易失性存储器，存储由 CPU101 执行的引导程序和各种数据。该引导程序在系统的起动时，将硬盘（HD）108 中存储的控制程序加载到 RAM103 中，使 CPU101 执行。关于该控制程序，以后参照流程图详细说明。RAM103 是可读写的随机存储器，存储从 HD108 加载并由 CPU101 执行的各种程序，并且在 CPU101 的动作时作为工作区使用，用于暂时存储来自各构成要素的各种数据。

输入部（input unit）104 包含键盘、鼠标、触摸板等，通过用户的操作，进行菜单项目的选择、各种数据的输入等。显示部（display unit）105 具有液晶、CRT、等离子体等显示器，用于将各种菜单、处理结果、错误、警告、检索结果等显示从而呈现给用户。扫描仪（scanner）106 进行光学地读取作为原稿的纸文档并数字化等处理。打印机（printer）107 在打印文档和图象时使用。在该文档检索装置中，也能打印由通信部（communication unit）110 接收的 PDL（打印控制语言）格网式的电子文档数据。

HD（hard disk）108 中存储有由 CPU101 执行的控制程序 111、用于进行自然语言分析的词素分析字典（morpheme analyzing dictionary）112、记述了用于认定未知词区域（unknown-word area）的规则的未知词区域认定规则（unknown-word area recognition rule）113 等。并且，根据需要，也存储用于管理未知词区域的未知词区域表（unknown-word area table）114、保持检索结果的检索结果表（search result table）115、将检索查询词扩展（develop）并且保持的查询词扩展表（query development table）116 等作业用数据。这些各种数据，根据需要加载到 RAM 中被参照，并根据需要变更后写回到 HD108。词素分析字典 112 中存储在一般的自然语言分析中所提出的必要信息，例如单词书写、词性信息、变化（conjugation）信息、单词搭配信息等。

可移动的外部存储装置 109 是 USB 存储设备、IC 卡等可插拔的

存储设备（storage device）。它们与普通的 PC 同样，也可以是用于访问软盘、CD、DVD 等外部存储的驱动等。该外部存储装置 109 能与 HD108 同样使用，能通过这些存储介质与其它装置进行数据交换。硬盘 108 中存储的控制程序 111，可根据需要从外部存储装置 109 将全部或一部分复制（安装）到 HD108。通信部 110 是网络控制器，能通过通信线路与外部进行数据交换。

具有以上的结构的本实施例的文档检索装置，根据来自输入部 104 等的各种事件进行工作。当有来自输入部 104 的中断时，将该中断信号发送给 CPU101，与此相伴随地产生事件。根据该事件，CPU101 读出 ROM102 或 RAM103 中存储的各种命令，通过执行命令，按照该控制程序进行各种控制。

图 2A、图 2B 是说明本实施例的检索的操作例的图。

在图 2A 所示的例 1 中，本来要检索的文本是“イラクへのハイテク兵器輸出の是非を問う”（201）。它在 OCR 处理时，被识别为“イテクへのハイテク兵器輸出の是非を問う”，并被登录（202）。这里，本来是“イラク”的字符串因为误识别，所以变为“イテク”。

接着，操作员为了寻找该误识别的字符串“イラク”，发出检索查询词“イラク”（203）。在本实施例的检索处理中，通过相似词扩展处理，将相似的字符（扩展查询词“イテク”）视为一致地进行检索。据此，作为检索结果找到被误识别的字符串“イテク”（210）。这里，在字符串“ハイテク”中也存在相同的字符串（“イテク”）（211），但是它们是在分析“ハイテク”的短语时命中的，所以降低命中位次地作为检索结果输出。

在图 2B 所示的例 2 中，检索的文本是“法律に詳しい人間が被告人関係者に必要”（205）。与例 1 同样字符识别该原文时，将“間”误识别为“関”，并被登录（206）。这里，为了检索“人間”，发出检索查询词“人間”的命令（207）。据此，包含相似字符（“人関”）地进行检索。命中字符串“人関”212 的“人関”。另外，此时“被告人関係者”中的“人関”213 成为可分析为“被告人” “関

係者”的短语，所以降低命中位次地作为检索结果输出。

图 3 是说明本实施例的未知词分析处理的一例的图。

页面图像 301 表示通过纸文档的扫描或电子文档的光栅化所生成的文档数据，它与原文一致。用文本 302 表示对它进行 OCR 处理后的结果。在该文本 302 中，因为误识别，所以“人間”变为“人関”（310），“望まれる”变为“望申れる”（311），“イラク”变为“イテク”（312）(误识别的字符带有下划线地显示)。

文本 303 表示将进行了字符识别的文本 302 进行词素分析后的分析结果，将文本分割为短语单位。“/”表示短语的划分。这里，没能进行词素分析的地方（无法分析字符串 311），带有框 313 地显示。这时，被误识别的“人関”（310）、“イテク”（312），如果分割为词素就能分析，所以不判断为无法分析字符串。文本 304 表示对 303 所示的分析结果进一步进行后述的未知词区域的抽取处理后的结果（带有框 314~315 的字符串表示未知词区域）。据此，除了 303 中的无法分析字符串(“望申れる”（311）)以外，按照后述的未知词区域认定规则 113，也将被误识别的“人関”（310）、“イテク”（312）的部分也设定为未知词区域。

图 4 是说明本实施例的存储未知词区域的未知词区域表 114 的数据结构的图。

对于各未知词区域，存储开始位置（start position）401 和末尾位置（end position）402。这些开始位置和末尾位置，存储表示文本上的未知词区域的开始位置和末尾位置的值（表示页、行数、该行的第几个字符的信息）。字符代码 403，是在图 3 的文本 304 中与被认定为未知词区域的字符串 316 对应的字符代码。图 4 中，存储指定为未知词区域的“イテク”316 的开始位置和末尾位置。

图 5 是说明本实施例的用于将检索查询词的各字符（字符串）扩展为相似字符串的查询词扩展表 116 的数据结构的图。

在这里存储像是被误识别的具有相似性的字符（串）的对。例如，片假名“ン”和“ソ”相似，所以在原字符串（original character

string）501 中存储“ン”，在与它对应的扩展字符串（developed character string）502 中存储“ソ”。此外，片假名“デ”有可能被误识别为片假名 2 字符的“テリ”，所以作为与原字符串 501 的“デ”对应的扩展字符串 502 存储“テリ”。此外，片假名“ク”有可能误识别为“ワ”从而被登录。此外，片假名 2 字符的“イン”有可能被误识别为 1 字符的汉字“仁”，所以作为与原字符串 501 的“イン”对应的扩展字符串 502 存储“仁”。此外，还登录有被误识别可能性高的字符和字符串，但是这里省略它们。

图 6 是说明将本实施例的检索查询词扩展为有可能通过误识别扩展的字符串的扩展查询词的格网（lattice）的图。

这里，各扩展字符串的连接状况形成格网。选择从开始节点到末尾节点 602 的路径时，原来的检索查询词表现为扩展成有可能被误识别的相似字符串的一个查询词。例如作为检索查询词，“インデクス”按照 610 所示的规则，就变为“イソテリワス”，按照 611 所示的规则，就扩展为“仁デワス”。按照其它规则，就扩展为“インテリクス”、“イソデクス”、“仁デワス”等。这样，检索查询词，按照是其检索查询词的原文不变（这里“インテクス”）还是该扩展查询词，区分是否为被扩展成上述相似字符串的字符串地进行存储。在图 6 中，椭圆内的字符表示与原来的检索查询词的字符一致的字符，矩形内的字符表示其被误识别时的字符。

图 7 是表示用于认定词素分析结果中哪部分应该成为未知词区域的规则即未知词区域认定规则的存储形式的图。

关于各规则，记述第一短语 701 和第二短语 702 各短语所满足的条件。例如，在规则 1 中，第一短语 701 的短语长度（字符数）为“1”，并且第二短语 702 的独立词长度（即除去附属词的字符数）为“1”时，满足规则 1（即认定为未知词区域）。据此，在图 3 的 304 所示的例子中，由“人”和“関ガ”构成的 2 个短语 314 被认定为未知词区域。

同样，规则 2 记述着的这样规则，当第一短语 701 的短语长度

小于等于2，书写用片假名，并且第二短语702的独立词长度小于等于2，书写用片假名时，认定为未知词。据此，在图3的304所示的例子中，由“イ”和“テク”构成的2个短语316被认定为未知词区域。

图8是表示本实施例的存储检索结果的候选的检索结果表115的结构的图。

这里，对于各检索结果候选，在开始位置801中存储检索结果（命中字符串）的开始字符在文本上的位置。在末尾位置802中存储检索结果（命中字符串）的末尾字符在文本上的位置。在分数803中存储规定该检索结果的显示位次的值（后面描述）。将该检索结果最终按分数803的值的顺序排序，并作为检索结果输出。

图9是说明本实施例的规定检索结果的输出位次的分数的计算式的图。

首先，相似度由（完全一致字符数）×2+（模糊一致字符数）的表达式计算。这里，“完全一致字符数”是检索查询词和命中字符串正确（不进行相似字符串扩展）一致的字符数。此外，“模糊一致字符数”是与将检索查询词扩展为相似字符串后的结果即命中字符串一致的字符数。这里，当命中字符串一部分在未知词区域时，所述求出的“相似度”原封不动成为“分数”。此外，命中字符串不在未知词区域时，即命中字符串的整个区域都已被分析成短语时，“分数”为从“相似度”中减去“最大分析短语长度”。这里，“最大分析短语长度”是关于命中字符串的词素分析结果的短语中最长的短语长度（字符数）。以下详细说明。

例1表示命中字符串在未知词区域中时的例子。这时，定义为“分数”=“相似度”。换言之，“最大分析短语长度”=0。

例2~例4表示命中字符串不在未知词区域中时的例子。在例2中，分析短语1和分析短语2在命中字符串中，其中字符数多的短语的字符数n成为“最大分析短语长度”。

在例3中，命中字符串仅覆盖一个短语，因此该短语的字符数n

为“最大分析短语长度”。

例 4 中，命中字符串覆盖 3 个分析短语，而其中最长短语的字符数 k 为“最大分析短语长度”。

使用图 3 的 304 所示的例子说明以上说明的分数的求解方法。当检索查询词为“人間”时，作为与扩展查询词“人関”一致的字符串，检索“人関が”314 和“被告人関係者”317。此外，检索查询词为“イラク”时，作为与扩展查询词“イテク”一致的字符串，检索“イテク”316 和“ハイテク”318。这时，首先“人関が”314 的相似度与(例 1)对应。这时，根据图 9 的表达式，完全一致的字符(“人”)的数（1）×2+（模糊一致的字符（“関”）的数（1）=3，此外，字符串“人関”的一部分在未知词区域，所以分数也变为“3”。

而“被告人関係者に”317 的情况，与（例 2）对应。这时，相似度也是与上述的计算相同的“3”。可是，这时任何部分都不在未知词区域中，所以分数是相似度“3”－最大分析短语长度(4)(3-4=)-1。同样，检索查询词为“イラク”时，与（例 1）对应。因此，“イテク”316 的相似度，根据图 9 的表达式，成为完全一致的字符(“イ”、“ク”的数（2）×2+（模糊一致的字符“テ”的数（1）=“5”，此外，由于在未知词区域中，所以分数也为“5”。而在“ハイテク”318 的情况下，“ハイテク”318 收容在一个分析短语内，所以与（例 3）对应。通过上述的计算，相似度为“5”，但由于不在未知词区域，所以分数为相似度“5”－最大分析短语长度（4）（5-4=）1。

按照流程图，说明上述的动作。

图 10 是说明本实施例的文档检索装置的处理的流程图，执行该处理的程序在执行时预先存储在 RAM103 中，在 CPU101 的控制下执行。

首先在步骤 S1 中，执行系统的初始化处理，这里进行各种参数的初始化和初始画面的显示等。接着在步骤 S2 中，等待来自输入部 104 或经由网络等连接的设备的请求等产生的任意事件。这里，事件发生后进入步骤 S3，判别该发生的事件，根据该判别出的事件的种

类，分支为各种处理。这里，用步骤 S4 综合表现与各种事件对应的分支目标的多个处理。作为与各种事件对应的分支目标的处理的一个例子，有图 11 所示的检索处理。作为其它处理未记述细节，有指定检索条件的处理、扫描原稿并且生成文档图象的处理、指定文档的处理等通常的检索装置的处理。然后，进入步骤 S5，显示步骤 S4 的各处理的处理结果。这里的处理，是检索结果的显示处理、存在错误时的错误显示、正常结束时的显示处理等通常广泛进行的处理。

图 11 是说明作为图 10 的步骤 S4 的事件对应处理的一部分的检索处理的流程图。

首先，在步骤 S11 中，执行图 12 的流程图中详细描述的未知词分析处理。这里，根据所指定的文档的图象，进行字符识别，生成 OCR 文本，进而通过词素分析、未知词分析，生成未知词区域表 114（图 4）。接着在步骤 S12 中，将输入的检索查询词扩展为有可能被误识别的相似字符串，生成扩展查询词格网（参照图 6）。接着在步骤 S13 中，根据该生成的扩展查询词格网，参照图 14 的流程图，执行后述的模糊对照处理，生成检索结果表 115。接着在步骤 S14 中，参照图 15 的流程图，如后所述，根据未知词区域表 114 和图 9 的表达式，求出检索结果表 115 的分数。接着，在步骤 S15 中，按照求出的分数的顺序，将检索结果按分数顺序排序。然后在步骤 S16 中，显示输出按照分数顺序排序后的检索结果。

图 12 是说明图 11 的步骤 S11 的未知词分析处理的流程图。

首先，在步骤 S21 中，对所指定的文档图象进行字符识别，取得文本信息（图 3 的 302）。接着在步骤 S22 中，对文本信息进行词素分析，分割为短语（图 3 的 303）。接着在步骤 S23 中，参照图 13 的流程图，如后所述，抽取未知词区域（图 3 的 304）。接着在步骤 S24 中，将抽出的未知词区域作为未知词区域表 114 输出。

图 13 是说明图 12 的步骤 S23 的未知词区域的抽取处理的流程图。

首先，在步骤 S31 中初始设定变量等，使指示短语的指针指示

文本的开始地进行初始化。接着在步骤 S32 中，取得由该指针所指示的短语的信息。接着在步骤 S33 中，参照词素分析字典 112，判断步骤 S32 中取得的短语是否为不能分析的短语。当判断为是不能分析的短语时，视为未知词区域，分支到步骤 S35，而判断为不是不能分析的短语时，进入步骤 S34，参照未知词区域认定规则 113，判断该短语是否属于未知词区域。这里，如果判断为不属于未知词区域，就分支到步骤 S36，而如果判断为属于未知词区域，就进入步骤 S35，关于抽出的未知词区域收集必要的信息，设定为未知词区域。然后，进入步骤 S36，更新表示短语的指针，以指示下一短语。接着在步骤 S37 中，判断是否存在下一短语，如果判断为存在，就回到步骤 S32，执行所述的处理。而如果判断为下一短语不存在，就结束未知词区域的抽取处理。

图 14 是说明图 11 的步骤 S13 的模糊对照处理的流程图。

首先，在步骤 S41 中，进行初始化设定，以将指示字符位置的指针指向文本的开始。接着在步骤 S42 中，对照扩展查询词格网和由该指针指示的文本上的字符。然后在步骤 S43 中，判断扩展查询词格网与字符是否一致，当不一致时，跳到步骤 S46，移动到下一字符。在步骤 S43 中，当判断为一致时，进入步骤 S44，将一致的程度作为相似度计算。该相似度的计算处理按照所述图 9 所示的表达式进行。接着进入步骤 S45，将该一致的字符位置登录到检索结果表 115（图 8）。这里，将在步骤 S44 中求出的相似度原封不动地设定为分数。接着进入步骤 S46，更新指示字符位置的指针，将字符位置向下一个前进。然后，在步骤 S47 中，判断字符位置是否到达文本的末尾，当未到达末尾时，回到步骤 S42，执行所述的处理，当到达末尾时，结束该模糊对照处理。

图 15 是说明图 11 的步骤 S14 的分数调整处理的流程图。

首先，在步骤 S51 中，进行初始化设定，使得指示检索结果的指针指示检索结果表 115（图 8）的开始。接着在步骤 S52 中，取得指针指示的检索结果的信息（位置和分数）。接着在步骤 S53 中，

根据未知词区域表 114（图 4）检查检索结果表示的命中字符串在文本上是否包含未知词区域。然后在步骤 S54 中，如果判断为包含未知词区域，就分支到步骤 S58，将分数确定为与相似度相等的值，而如果判断为不包含，就进入步骤 S55，如图 9 所示，求出命中字符串所涉及的最长分析短语长度。接着在步骤 S56 中，从分数减去该求出的最长分析短语长度，来校正分数。接着在步骤 S57 中，将该校正后的分数反映到检索结果表 115（图 8）中。接着在步骤 S58 中，将指示检索结果的指针更新为指示下一检索结果。然后在步骤 S59 中，判断是否为检索结果的最后，当不是最后时，回到步骤 S52，执行所述的处理，当判断为结束时，结束分数调整处理。

（其它实施例）

本发明并不局限于上述的实施例，只要不脱离本发明的宗旨，可以进行适当变更。在上述的实施例中，作为语言分析的方法，使用词素分析，但是也考虑此外的实现方式。例如，也考虑基于只分割为单词的手法的方式。这时，附属词的部分完全不分析，作为未知词区域处理。据此，存在分析的精度下降这样的缺点，但是与词素分析相比，分析处理轻松完成，能构筑负荷更轻的系统。

此外，在上述的实施例中，作为模糊对照的手法，将查询词的字符扩展为相似字符串地进行检索，但是也考虑不进行扩展，而将相似的字符组汇总，标准化为要代表的代表字符进行对照的手法。通过这样构成，能减轻处理负载，能应用于更小规模的装置。

此外，也考虑完全不同的模糊对照的实施方式。例如，还可采用如通配符检索那样，即使存在不一致的部分也判断为对照成功的手法。这时，相似度的计算方法有若干改变，但是如果除去模糊对照的部分，就能完全同样地构成，能取得完全同样的效果。并且，在所述以外，只要不脱离本发明的宗旨，就能适当变更该实施例的结构。

如上所述，根据本实施例，对于存在误识别的文本，能进行允许误识别的字符串检索。并且即使存在允许了误识别的命中字符串

时，如果包含在可分析的字符串中则分数评价低，因此能控制不想要的误命中。据此，相对优先地显示实际被误识别的字符串的命中，能提供操作性高的文档检索装置。

本发明并不局限于上面的实施例，在本发明的精神和范围中能进行各种变更和修改。因此，为了向大众通知本发明的范围，产生了以下的权利要求书。

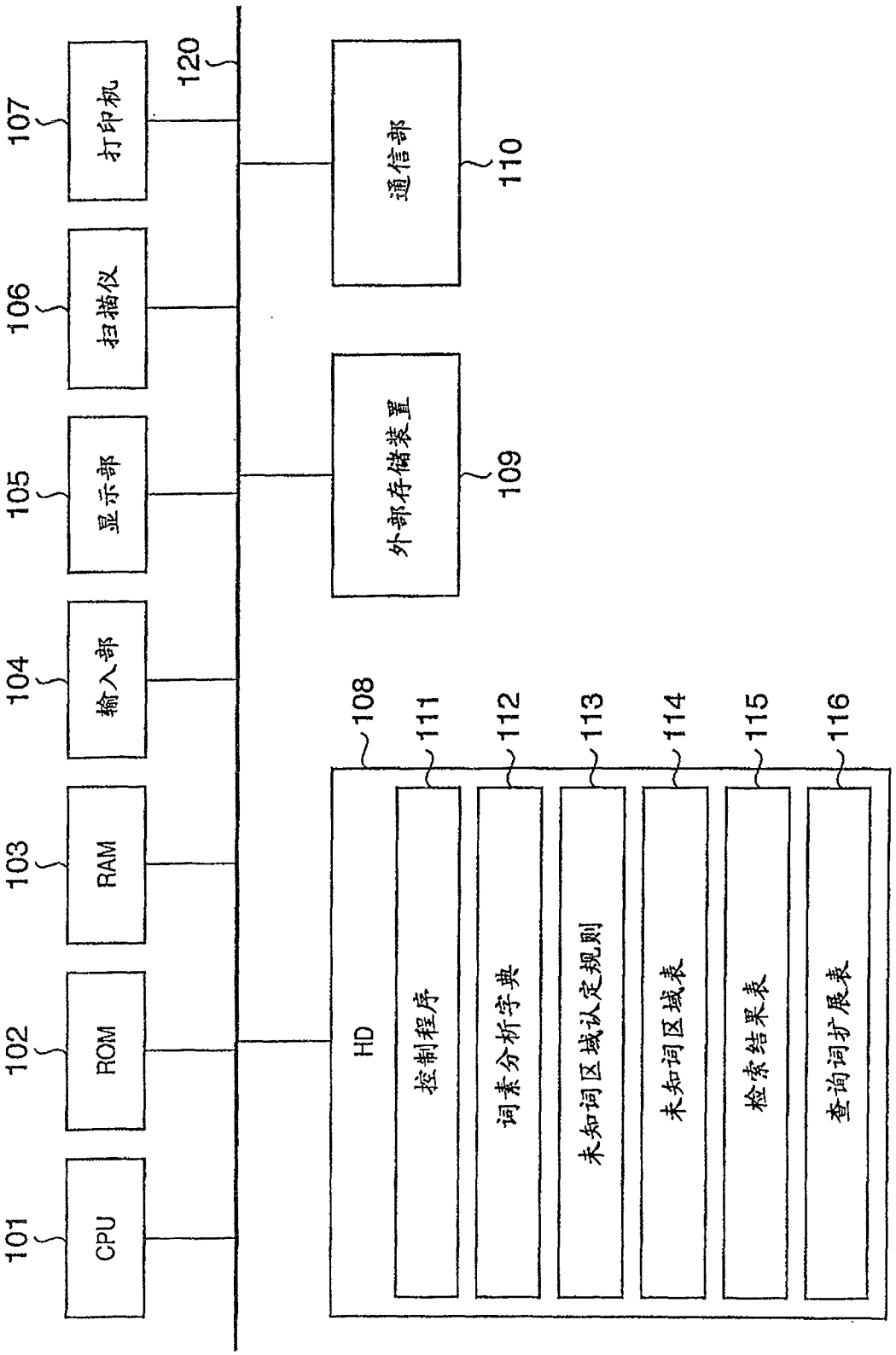

101
CPU
102
ROM
103
RAM
104
输入部
105
显示部
106
扫描仪
107
打印机
120
通信部
110
外部存储装置
109
HD
108
控制程序
111
词素分析字典
112
未知词区域认定规则
113
未知词区域表
114
检索结果表
115
查询词扩展表
116

图 1

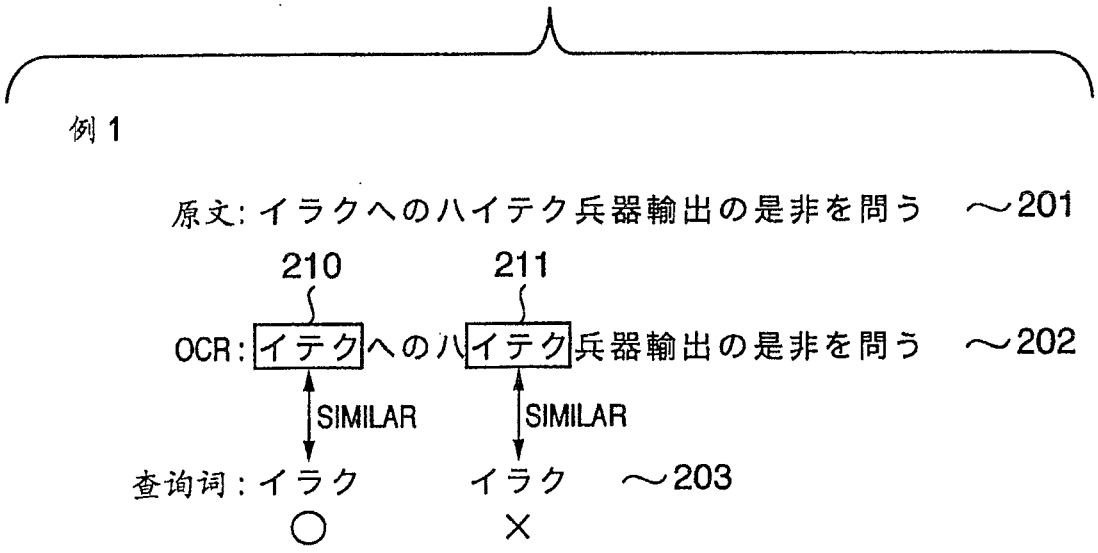

例 1
原文: イラクへのハイテク兵器輸出の是非を問う ～201
210
211
OCR: イテクへのハイテク兵器輸出の是非を問う ～202
SIMILAR
SIMILAR
查询词: イラク イラク ～203
○
×

图 2A

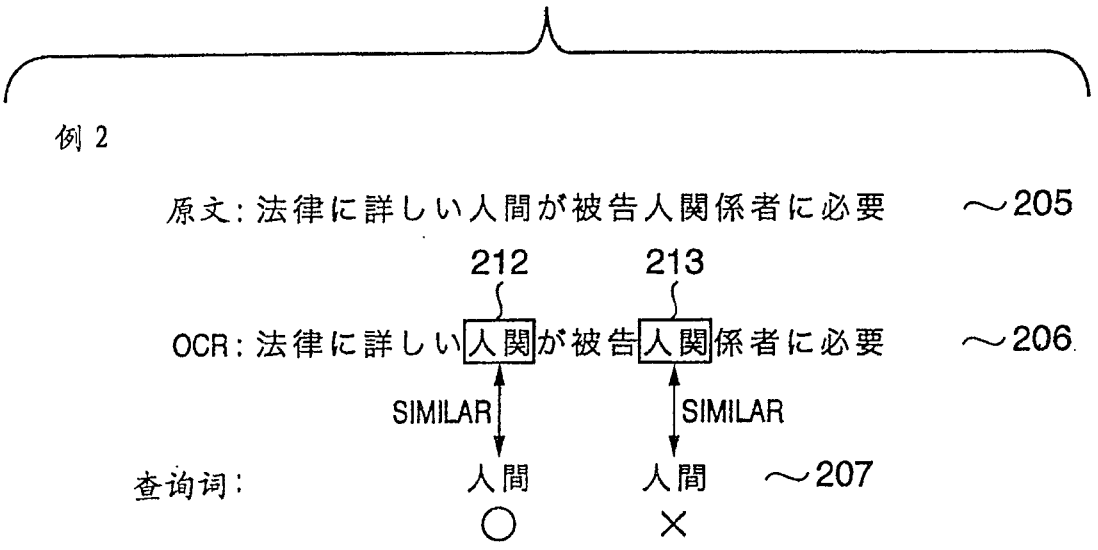

例 2
原文: 法律に詳しい人間が被告人関係者に必要 ～205
212
213
OCR: 法律に詳しい人関が被告人関係者に必要 ～206
SIMILAR
SIMILAR
查询词:
人間 人間 ～207
○
×

图 2B

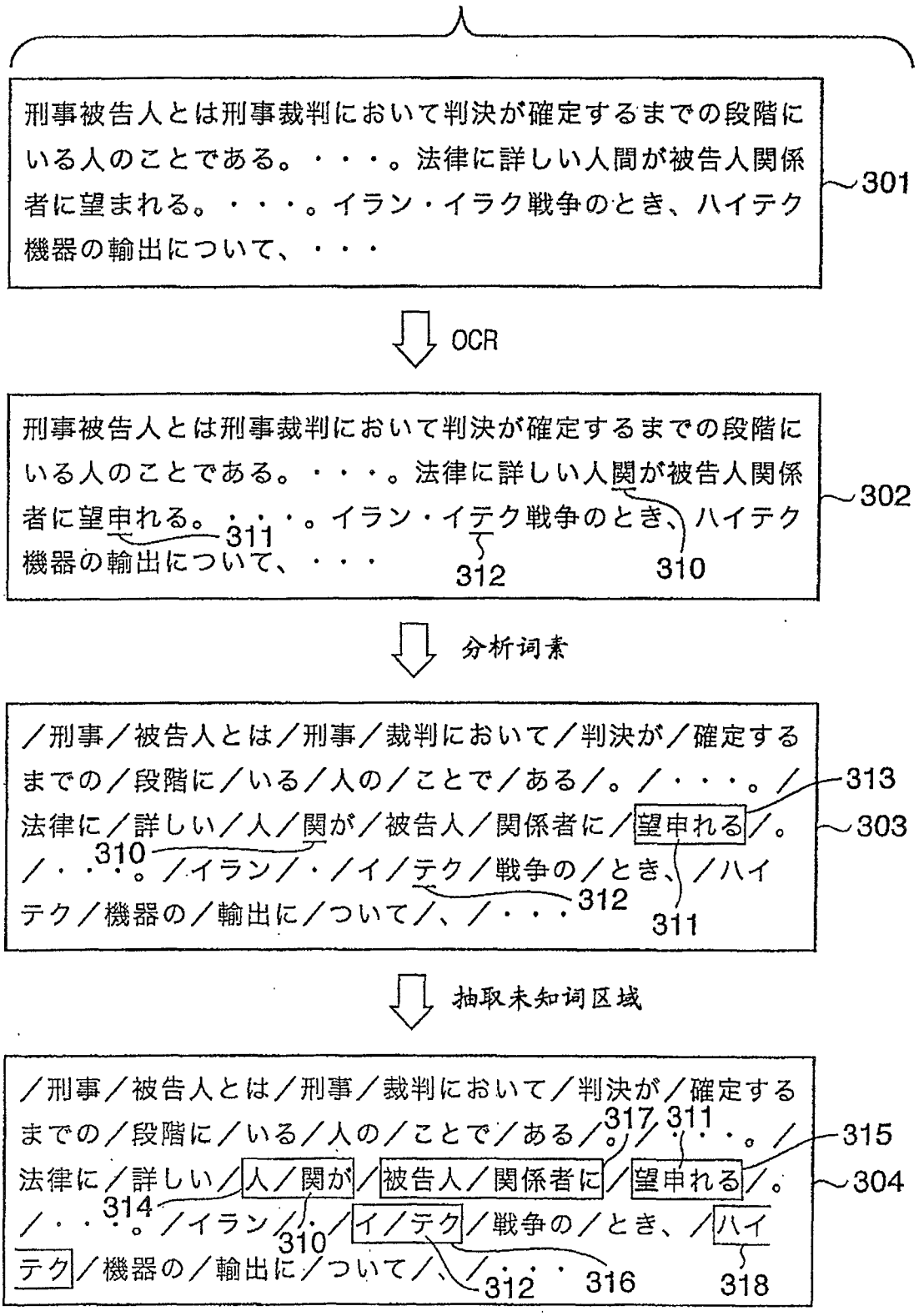

刑事被告人とは刑事裁判において判決が確定するまでの段階にいる人のことである。・・・。法律に詳しい人間が被告人関係者に望まれる。・・・。イラン・イラク戦争のとき、ハイテク機器の輸出について、・・・
301
OCR
刑事被告人とは刑事裁判において判決が確定するまでの段階にいる人のことである。・・・。法律に詳しい人関が被告人関係者に望申れる。・・・。イラン・イテク戦争のとき、ハイテク機器の輸出について、・・・
302
311
312
310
分析词素
／刑事／被告人とは／刑事／裁判において／判決が／確定するまでの／段階に／いる／人の／ことで／ある／。／・・・。／法律に／詳しい／人／関が／被告人／関係者に／望申れる／。／・・・。／イラン／・／イ／テク／戦争の／とき、／ハイテク／機器の／輸出に／ついて／、／・・・
313
303
310
312
311
抽取未知词区域
／刑事／被告人とは／刑事／裁判において／判決が／確定するまでの／段階に／いる／人の／ことで／ある／。／・・・。／法律に／詳しい／人／関が／被告人／関係者に／望申れる／。／・・・。／イラン／・／イ／テク／戦争の／とき、／ハイテク／機器の／輸出に／ついて／、／・・・
317
311
315
304
314
310
312
316
318

图 3

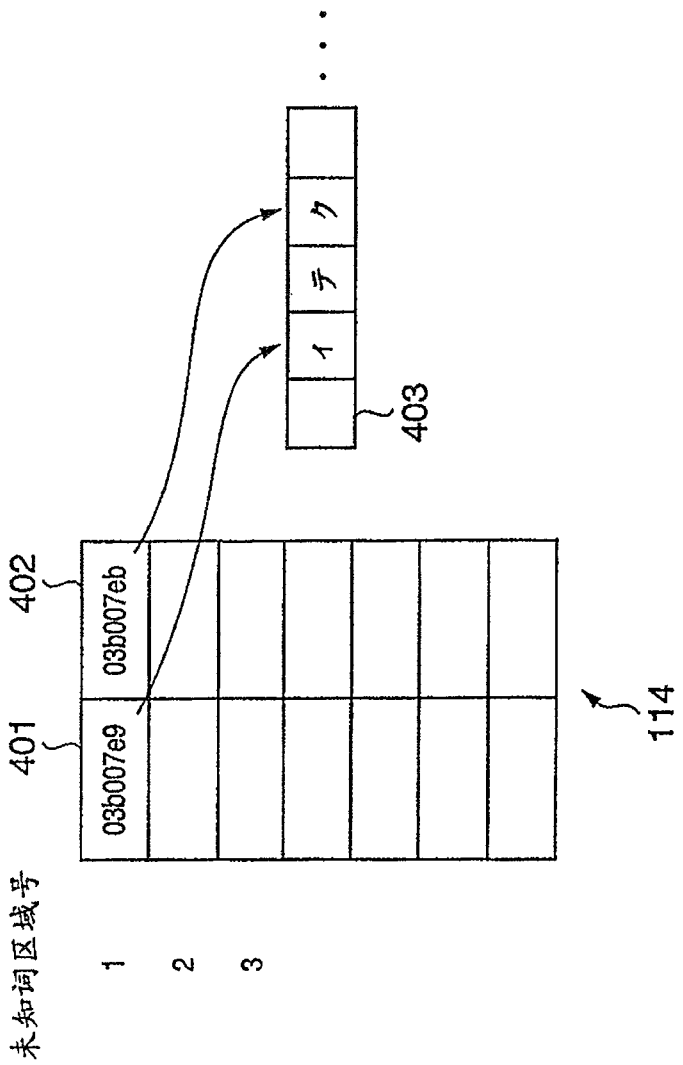


图 4

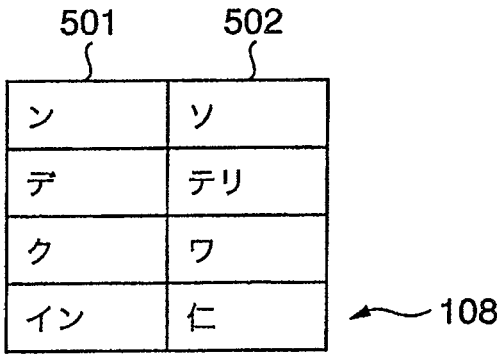


图 5

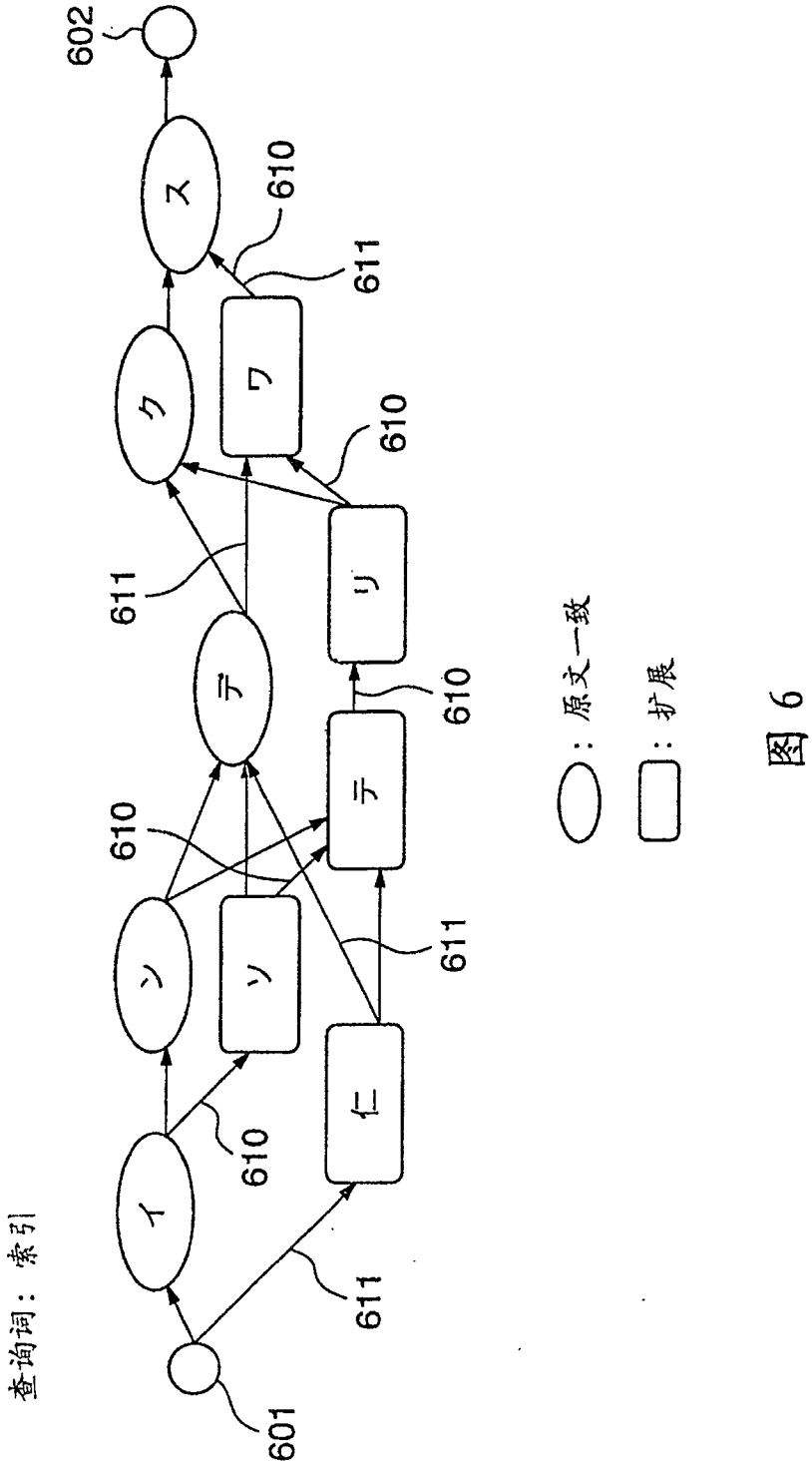

查询词：索引
601
602
610
611
イ
ロ
ハ
ニ
ホ
ヘ
ト
チ
リ
ヌ
ル
ヲ
ワ
カ
ヨ
タ
レ
ソ
ツ
ネ
ナ
ラ
ム
ウ
ヰ
ノ
オ
ク
ヤ
マ
ケ
フ
コ
エ
テ
ア
サ
キ
ユ
メ
ミ
シ
ヱ
ヒ
モ
セ
ス
：原文一致
：扩展

图 6

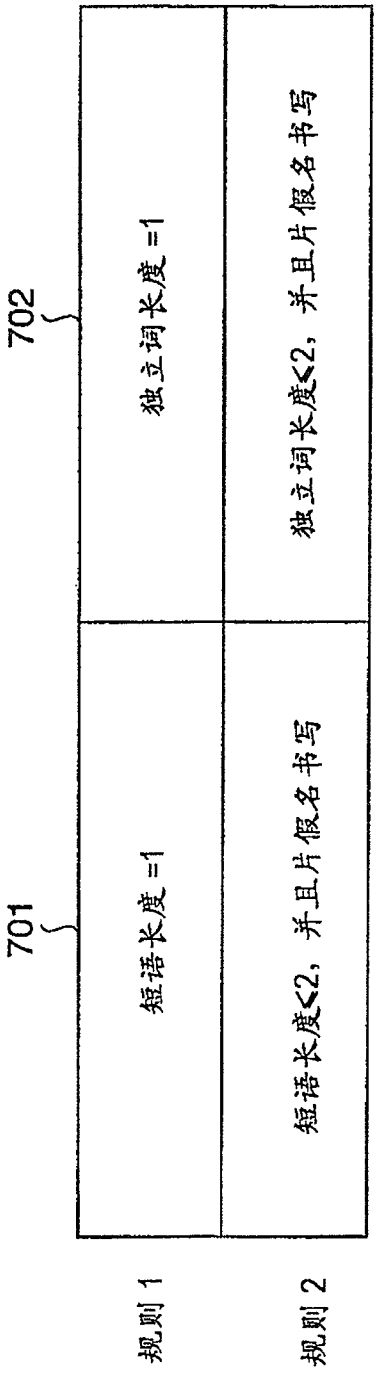

701
702
短语长度 =1
独立词长度 =1
短语长度<2，并且片假名书写
独立词长度<2，并且片假名书写
规则 1
规则 2
113

图 7

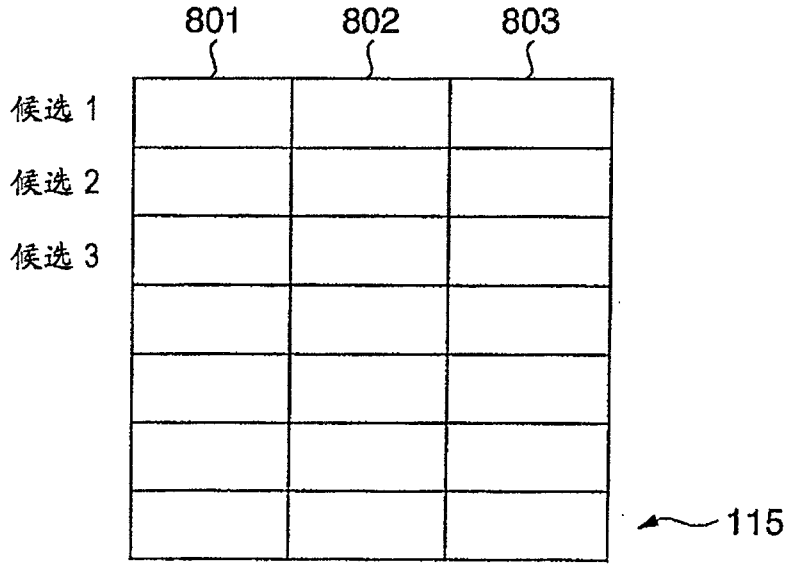
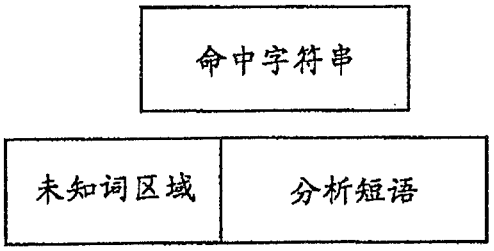


图 8

相似度=（完全一致字符数）×2+（模糊一致字符数）

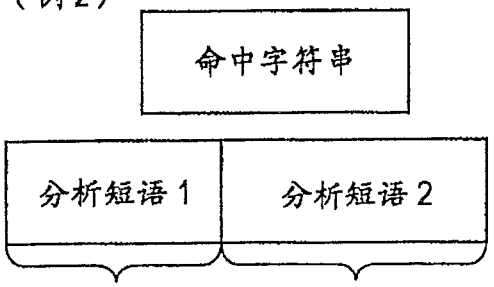
分数=相似度－最大的分析短语长度

（例1）



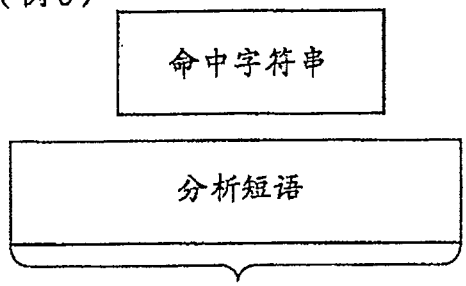
最大的分析短语长度=0

（例2）



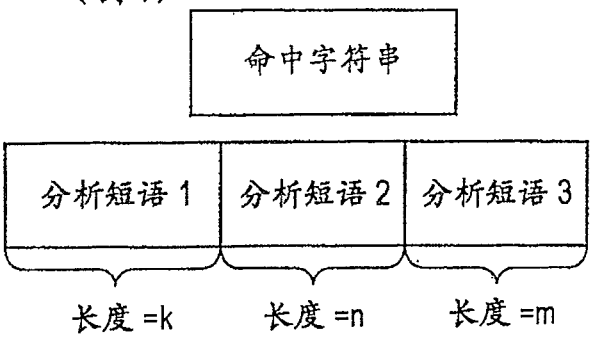
n>m→ 最大的分析短语长度=n

（例3）



最大的分析短语长度=n

（例4）



k>n>m→最大的分析短语长度=k

图 9

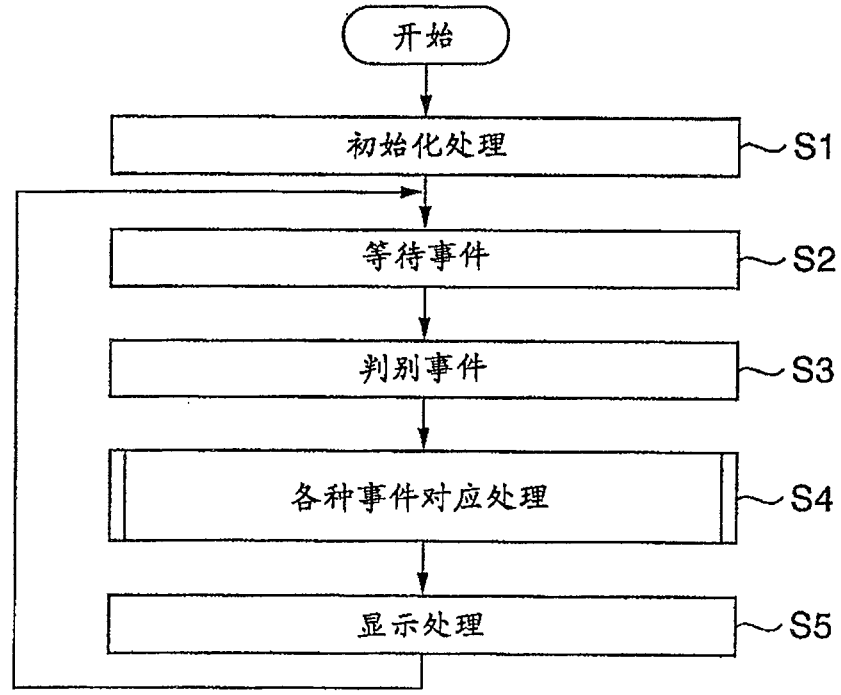


图 10

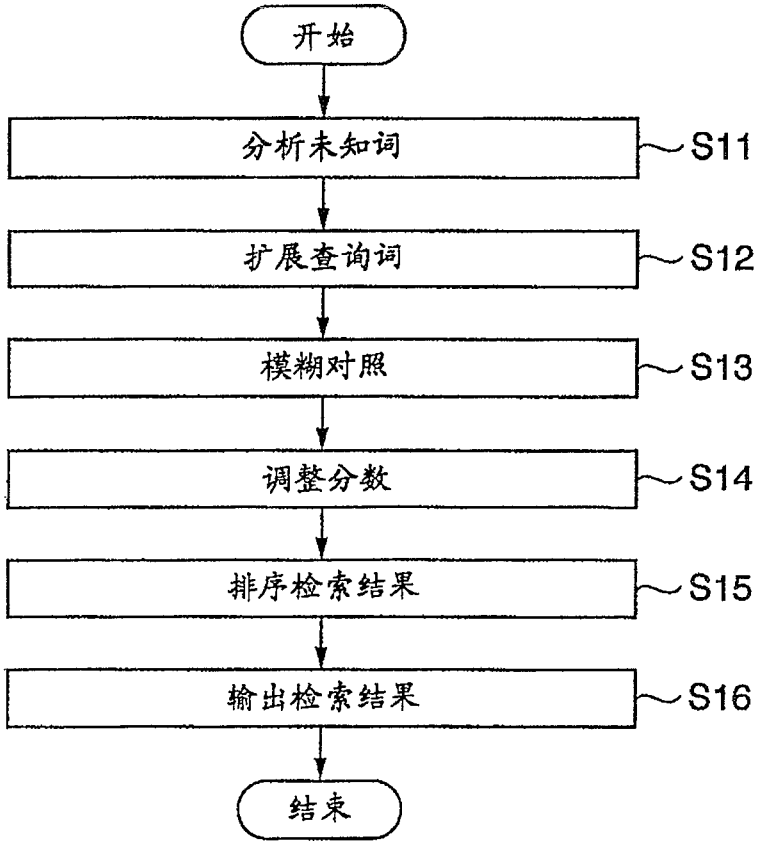

开始
分析未知词
S11
扩展查询词
S12
模糊对照
S13
调整分数
S14
排序检索结果
S15
输出检索结果
S16
结束

图 11

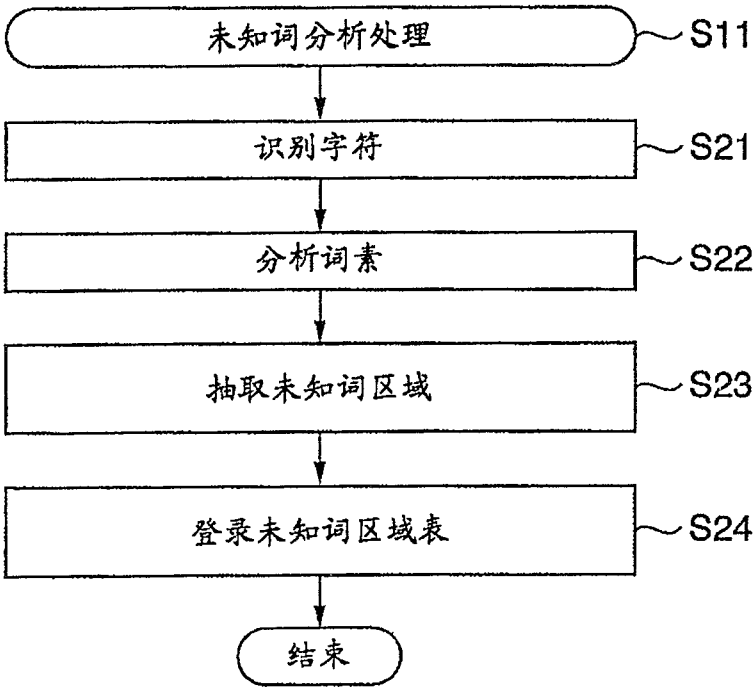


图 12

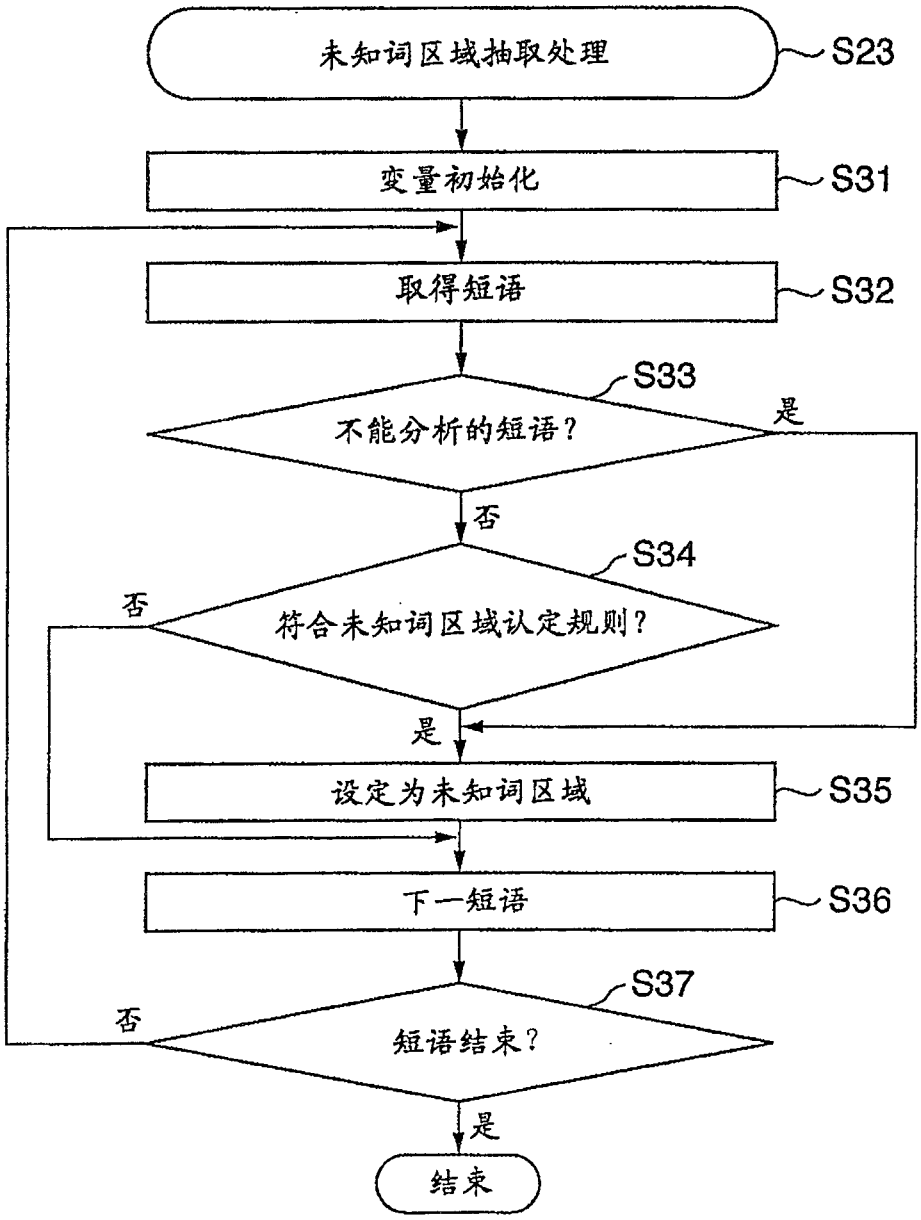

未知词区域抽取处理
S23
变量初始化
S31
取得短语
S32
S33
不能分析的短语？
是
否
S34
否
符合未知词区域认定规则？
是
设定为未知词区域
S35
下一短语
S36
S37
否
短语结束？
是
结束

图 13

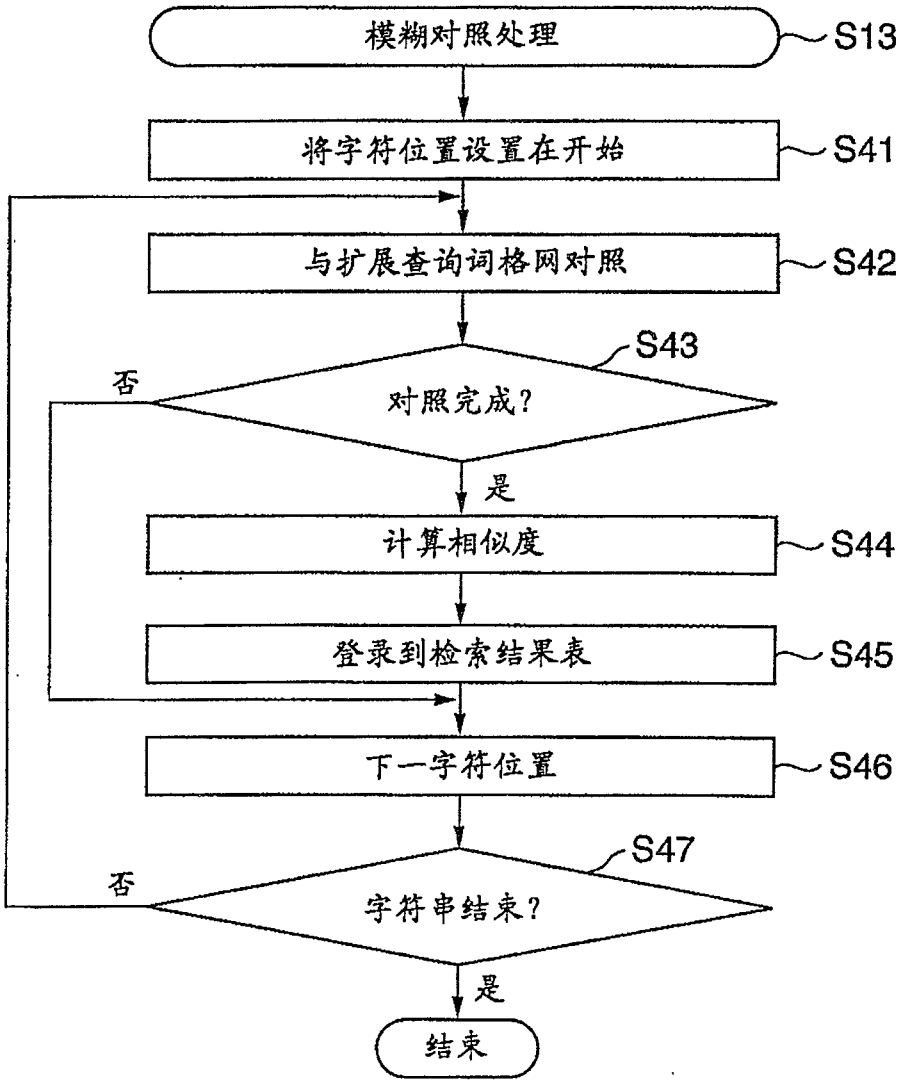

模糊对照处理
S13
将字符位置设置在开始
S41
与扩展查询词格网对照
S42
S43
对照完成?
否
是
计算相似度
S44
登录到检索结果表
S45
下一字符位置
S46
S47
字符串结束?
否
是
结束

图 14

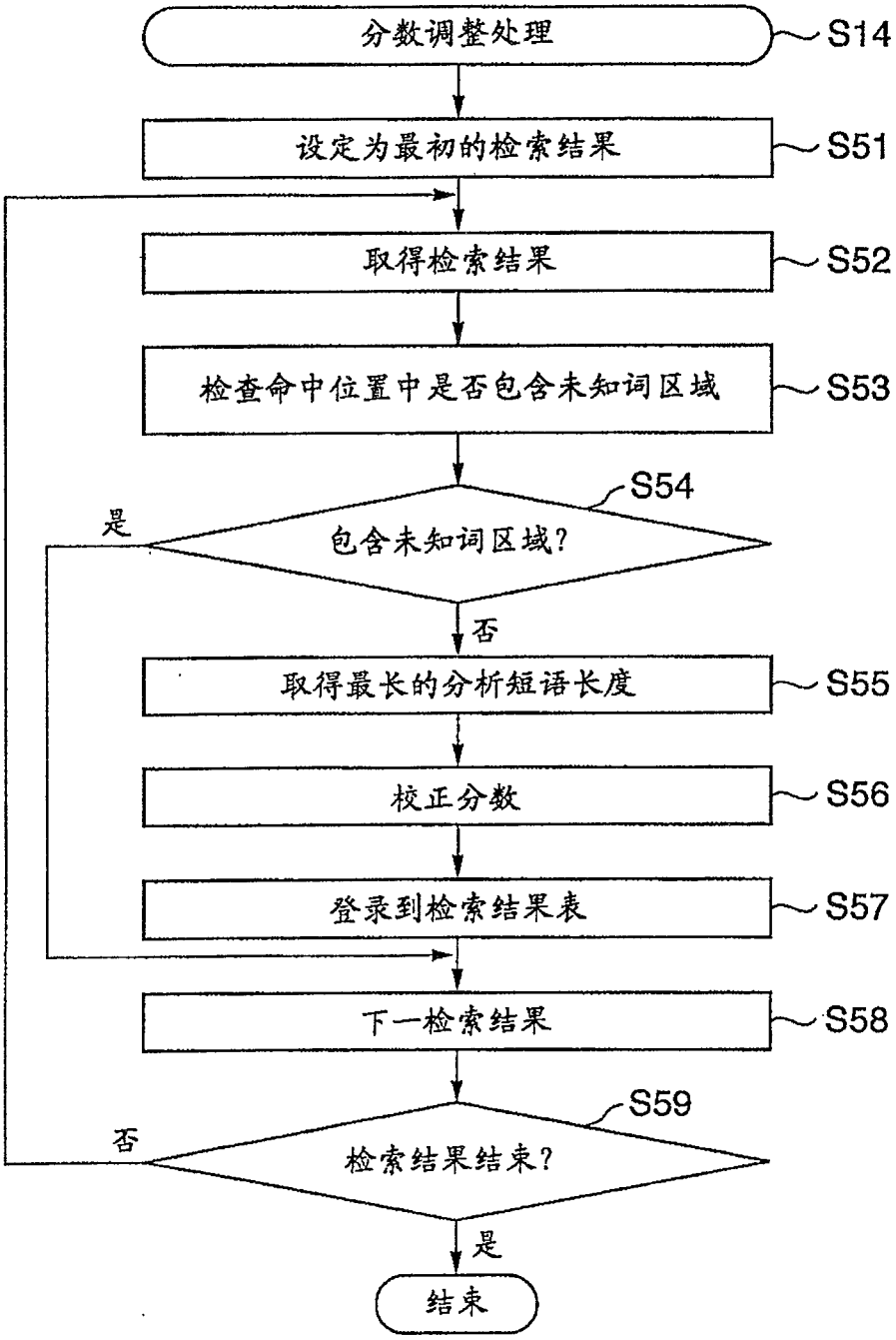

分数调整处理
S14
设定为最初的检索结果
S51
取得检索结果
S52
检查命中位置中是否包含未知词区域
S53
S54
包含未知词区域?
是
否
取得最长的分析短语长度
S55
校正分数
S56
登录到检索结果表
S57
下一检索结果
S58
S59
检索结果结束?
否
是
结束

图 15