[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/28 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200710030770.0

[43] 公开日 2009 年 4 月 15 日

[11] 公开号 CN 101408873A

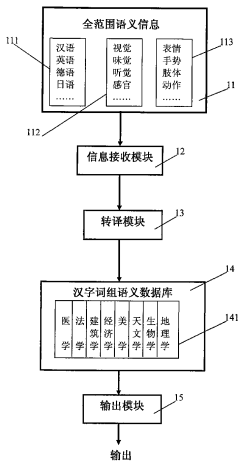
[22] 申请日 2007.10.9
[21] 申请号 200710030770.0
[71] 申请人 劳英杰
地址 200031 上海市徐汇区武康路 103 号 2 楼
[72] 发明人 劳英杰
[74] 专利代理机构 广州新诺专利商标事务所有限公司
代理人 刘菁菁
权利要求书 2 页 说明书 12 页 附图 6 页

[54] 发明名称

全范围语义信息综合认知系统及其应用

[57] 摘要

本发明公开了一种全范围语义信息认知系统，包括：一信息接收模块，用于接收任何一种可被自然语言或文字所表达的信息源；以及一转译模块，将上述信息源根据语义转译至语义信息数据库；以及一语义数据库，由汉字词组构成，汉字具有按照部首属性编码规则编码成可应用至计算机系统的数字编码；以及一输出模块，将上述数字编码转换并输出。 本发明可对任何一种可用语言或文字表达的信息源进行综合认知，可通过电子系统撷取各种信息的数字数据，对应汉字词组语义，能进行综合理解及认知，然后以综合数据、模拟方式作出回应。本系统应用于语言及文字的翻译及检索等领域，速度和效率均可大幅度提高。



1、一种全范围语义信息综合认知系统，其特征在于包括：

一信息接收模块，用于接收任何一种可被自然语言或文字所表达的信息源；以及

一转译模块，将上述信息源根据语义转译至语义信息数据库；以及

一语义数据库，由汉字词组构成，汉字按照部首属性编码规则编码成可应用至计算机系统的数字编码；以及

一输出模块，将上述数字编码转换并输出；

所述部首属性编码规则是指汉字按照预定笔画集合和笔画顺序拆分成至少一个笔画、与数字构成的编码一一对应，每数字为1字节，每字节最多为3位元(bit) 编码表示。

2、根据权利要求1所述的系统，其特征在于：所述预定笔画集合由点“、”——代表点类笔画、短撇“丿” ——代表短撇及短捺类笔画、 长撇“丿”——代表长撇及长捺类笔画、 短划“－”——代表短横及短竖类笔画及长划“一” ——代表长横及长竖类笔画组成。

3、根据权利要求2所述的系统，其特征在于：所述数字构成的编码为1、2、3、4、5，分别对应点“、”、短撇“丿”、 长撇“丿”、 短划“－”及长划“一”，字型笔画不足部分以数字“0”表示。

4、根据权利要求1或2或3所述的系统，其特征在于：所述汉字根据字型结构以两组共6个数字字节，每字节最多为3位元(bit) 编码表示。

5、根据权利要求1所述的系统，其特征在于：所述语义数据库内根据汉字部首分类功能设有知识分类丛集词库，以实现汉字词组按照部首义项属性对同一应用领域汉字词组的丛集及分类，应用所述丛集词库对多义词进行部首义项属性关系匹配比较，判断出符合匹配关系的词组。

6、根据权利要求1所述的系统，其特征在于：所述接收模块接收感官信息数据转换为汉字词组的文字信息，并表达成可被计算机读取的数字编码。

7、根据权利要求1所述的系统，其特征在于：所述接收模块接收动作信息数据转换为汉字词组的文字信息，并表达成可被计算机读取的数字编码。

8、应用权利要求 1 所述的系统进行任何语言及文字系统信息数据的结构化处理。

9、应用权利要求 1 所述的系统进行任何自然语言及文字系统的互译。

10、一种应用权利要求 1 所述的系统对任何自然语言系统进行语音操控的电子机器。

全范围语义信息综合认知系统及其应用

技术领域

本发明涉及计算机技术领域，尤其涉及应用于计算机系统的人工智能的综合数据编码处理技术领域。

背景技术

以机器认知人类全范围语义信息，一直是个极难解决的问题。机器要被人类利用，必需能以自动方式对于人类全范围语义信息，有准确的理解及认知，才能进行正确的沟通及回应。任何语义信息都存在大量歧义，机器难以排除歧义、判断正确语义信息。人类之间沟通的目的是传达信息，信息内含有特定语义，人类赖以利用的主要是语言及文字，目前已出现了数以千计的语言及文字系统。

但事实上，世界不断的发展，人类所要传达及表示的信息及语义内容也更丰富多彩，这些信息及语义内容最终以各种语言及文字系统反映出来。所以每种语言及文字系统都出现相同情况，即存在大量的同音及近音词，及同义及近义词，产生语义上的混乱及错误；这是机器难于进行认知的原因所在。语义编码的目的，是机器能夠以自动方式认知人类全范围语义信息，信息必需要以一种标准语义符号作为标准来进行综合编码。汉字是人类社会其中一种自然语言的文字表示系统，亦是一种唯一的语义符号表示系统，能对应现时人类任何自然语言及文字系统内的语义；同时，汉字语义符号的独特结构，使机器能夠以固定及极少的数据量，达成高效率的语义搜索、判断及认知。

汉字以外的文字都是拼音文字，拼音文字的特色主要是由数十个字母符号，组合成一个或多个语音，代表某个特定语义。拼音文字的出现，源自语音，语音由字母串组成，表示特定语义信息；但字母符号本身並没有任何语义。汉字是目前仍在使用的最古老的文字，世界上的使用率仅次于英语。汉语是自然语言的一种，汉字发展至现在，拥有丰富的词组体系及简约的表达力。

现代汉字由数千个单一的汉字有机性地复合成两字、三字及四字词语，表达不同语义；单字词的例子是书、樹及光等，两字词组例子有衣服、飞机

及教師等，三字词组例子有电视机、飞行员及旅游社等。东方及西方经过三百多年文明的交接及融合，在全球化影响下，汉字词语的语义表述结构基本上能对应任何一种自然语言及文本语义信息。

过往关于文字的编码方法，目的是为了以电子方式记录及贮存文字，所以都是以每个唯一的字母符号进行编码，如 ASCII 内的 256 个组合能容纳英语及西欧文字，汉字的中文字型编码有大五码繁体字形、国标码 2312 简体字形、国标码 18030 简体字形及现时已能夠涵盖绝大部份世界文字的统一码等。汉字的数量繁多，不同字库有不同字量，国标码 2312 简体字形是 6,700 个，大五码繁体字形是 13,500 个及国标码 18030 简体字形的 18,030 个等。这些编码方法都是以记录唯一的字型为原则，以字型数量编码，目前是以多字节的数据量满足编码所需。

最早的文字编码方法，主要是以每个字母或字型编码，方法是分别将字型符号编入 128、256 及 65,536 个组合内，以不同长度的字符串表示不同语义。电脑发明于西方世界，应用的是拼音文字。普遍应用的 ASCII 和 ANSI 符号编码规则，每个字母或符号为 1 字节，每字节以 8 位元的数据长度表示。

由于 ASCII 只规定了 128 个最常用的字母符号，随着计算机字符集的增长，逐渐出现了很多种在 ASCII 上扩充的编码方式。信息领域的急速发展，累积了極大量以记錄为目的的文字数据，分别由不同的字母、数字或文字符号组成，但越大量的数据出现，就越需要强大的硬件运算能力，才能滿足在不断扩大的数据内搜索的需要。在任何计算机或电子系统内，字符组合的数量直接影响到文字的检索效率，在浩如煙海的信息世界或庞大的数据库内，数量大的字符组合的排序及比较等效率绝对比数量小的字符组合慢很多倍。

人类应用的文字及语言系统种类繁多，而任何的文字及语言系统都有一相同特性，都存在为数不少的同词异义(Homonyms, Polysemy or Homophomes)及异词同义 (Synonym or Hyponyms)。同词异义的定义是，同一单词或词组，或同音词组，在不同的语境中，具有完全不同的语义。这些都是任何语言及文字发展过程中所出现的必然现象。以机器自动认知方式区分这些特性，往往会产生难以解决的歧义问题，特别是要结合语境判断正确的语义，此亦是自动翻译系统难于解决的难题。人类在应用已熟悉的语言及文字系统时，会根据歧义词的上下文语境，判断正确语义。所以，目前的技

术只能在有限语言或文字范围内认知，在局部范围内的语言或文字，出现一词多义时不能以自动判断方式来确定符合上下文语境的正确语义。

任何拼音文字都是由不同长度的字符串组成，组成结构中没有类似于汉字部首的分类特性，当需要自动判断同名异义词组的语义时，就会出现模棱两可的情况。与任何拼音文字完全不同的是，汉字系统从古代到现在，都存在一特点，即汉字本身内存在着固定的部首系统，部首解释及表示该汉字的属性，包含有基本语义项；例如部首“疒”的语义项是“病理的”，部首“水”的语义项是“与水有关的”及部首“金”的语义项是“与金属有关”等。汉字部首的类别发展至目前，数量有214个。

汉字由部首及部件组成，只有汉字部首的结构具备语义分类功能，特别是在语义的排歧方面。在绝大部份的语境内，内容上互有关联的，其用于表述的汉字的部首，也会互有关联。例如部首“疒”是有关病理的，“医”是关于医学科等；这些汉字及词组通常会在同一语境范围内出现。若汉字内容需要判断歧义词的含义时，就能以部首的分类原则，排除同音同形但非关联部首的汉字或词组。任何自然语言及文字系统，都能以汉字及词组对应其语义。但目前的汉字编码方法，都没有对汉字的部首及语义编码。

另一方面，任何拼音文字及语言系统，都会出现極多的异名同义词，即是语义相同，而拼写不同的词。例如英语 Britian 就有 8 个相同语义的字母串，分别为 England, UK, U.K., United Kingdom, GB, G.B., Britian and Great Britian 等；其汉语的相同语义分别是英国、英格兰、大不列颠及大英帝国等，亦可概括为语义“英国”。到目前为止，尚未有高效率的对同义词进行准确自动获取的方法。若用户需搜索异名同义词时，都必需以多个不同词组提出搜索请求，才能获取最大范围内的搜索结果。

过往的语言及文字搜索模式，都是在相同的文字系统内匹配相同语音或文字词组，再进一步通过不同语种的字典，以相同语义进行互换从而得到不同自然语言之间的语言表达。另外，一般的同义词搜索方法，用户都需要分别输入源语言中所有语义相同的词组，才能匹配出目标语言中语义相同的词组。事实上，用户真正需要搜索的是该单一语义本身，但单一语义会存在多个表达词组，这些表达词组存在于海量的文字数据库内，要以不同的关键词逐个进行搜索。任何拼音文字的困难都在于，需要在海量的非结构化文字数

据内，进行上述多个相同语义的关键词搜索。若能以单一词组进行同义词的检索，将会大大缩小检索的范围，提高检索的效率。

现时的全文搜索，一般都是按照相同文字进行匹配，但事实上，用户需要搜索的是某个特定语义概念，或相关语义；以越少的汉字词组对应相同语义的同义词，对数据进行自动认知的过程就越高效率。以往少量的数据，可以用手工方式进行结构化分类建立目录进行查找；但以手工分类，会由于操作个体对语义认知的偏差而导致分类歧义。目前人类的文明已累积了極大量的信息数据，需要以综合及标准的运算原则进行自动分类及排序。任何数据都不是独立存在的，而是互有关联的，所以难于以手工方式进行绝对一致的分类，需以自动方式对随时更新的数据，以最高效率建立最有关联关系的数据结构。

过往的文字编码方法，是以记录最大范围的文本信息为目的，但这种编码方法只能满足以往对文字处理及贮存的需求。大量的信息组织成为数据，具有综合结构化的数据，才是有用的数据，才能最宽广及最深度地进行挖掘。现时的技术，是以人手方式对相同语义数据加入标签，标签後的数据自动进行文本分类及丛集，才能进行文字挖掘；丛集结构化或文本数据化的功能是建立语义目录，但拼音文字组成的词组，词组与词组混合使用时容易产生多义性，自动认知难于排除歧义。语义数据以部首标签方法，能正确表示及区分语义数据与数据之间的关系及属性。

发明内容

本发明的目的在于提供一种可对任何可用语言或文字表达的信息源进行综合认知的系统，以及应用该系统实现检索，翻译等功能。

本发明还提供了一种应用上述系统对任何自然语言系统进行语音认知，可以操控的电子机器。

为了综合达到上述发明目的，本发明采用了以下技术方案：一种全范围语义信息认知系统，其特征在于包括：

一信息接收模块，用于接收任何一种可被自然语言或文字所表达的信息源；以及

一转译模块，将上述信息源根据语义转译至语义信息数据库；以及

一语义数据库，由汉字词组构成，汉字具有按照部首属性编码规则编码

成可应用至计算机系统的数字编码；以及

一输出模块，将上述数字编码转换并输出；

所述部首属性编码规则是指汉字按照预定笔画集合和笔画顺序拆分成至少一个笔画、与数字构成的编码一一对应，每个数字表示 1 字节，每字节最多只以 3 位元(bit)表示。

所述预定笔画集合由点“、”——代表点类笔画、短撇“丿” ——代表短撇及短捺类笔画、 长撇“丿” ——代表长撇及长捺类笔画、 短划“－”——代表短横及短竖类笔画及长划“一” ——代表长横及长竖类笔画组成。

为提高系统运作效率，限定上述数字构成的编码为 1、2、3、4、5，分别对应点“、”、短撇“丿”、 长撇“丿”、 短划“－”及长划“一”，字型笔画不足部分以数字“0”表示。

为进一步地简化及明确汉字编码以提高效率，限定上述汉字根据字型结构以两组共 6 个数字，每个数字表示 1 字节，每字节最多只以 3 位元(bit)表示。 以下为 6 个数字对应二进制数字系统的表示方式：

数字	3 位元数字编码
0	000
1	001
2	010
3	011
4	100
5	101

为了能对同音、近音歧义词或同名多义词进行有效排歧及筛选，所述语义数据库内设有若干丛集词库分类，以实现汉字词组按照部首义项属性对同一应用领域汉字词组的丛集及分类，应用所述丛集词库对多义词进行部首义项关系匹配比较，筛选出符合匹配关系的词组。

进一步地，上述接收模块可接收感官信息或动作信息数据转换为汉字词组的文字信息，并表达成可被计算机读取的数字编码。

最有效率的数据搜索，是需要数据本身先以字母数字或字符组合的顺序排列，然後进行搜索及匹配；新发明以汉字词组对任何信息语义进行认知，

即是对应任何语义数据，每个汉字符号分别以不同部首或部件组成，每个部件以不同笔划组成。新发明以最少的笔划型态对应不同部首或部件的分组编码，以笔划对应不同数字，每个数字为 1 字节，每种笔划型态最多只有 3 位元(bit)的数据长度，每个汉字最少只有 6 个字节组成，且是固定长度数据编码组合，与拼音文字的非固定长度数据进行排序比较，效率肯定是最快。

现在每天都湧现大量的电子数据信息，在数据库内有任何新的数据出现，都需要进行更新、插入及排序，永远是需要重复这些运算过程，所以高效率的综合编码排序方法是必需的。新发明以汉字词组对应任何自然语言及文本的语义信息，任何语义都能以此最少综合数据组合的分组编码进行高速排序。

新发明以汉字词组对应任何自然语言及文本信息，汉语是自然语言的一种，汉字系统内具备部首系统，任何汉字词组都能以部首属性进行自动分类及丛集，任何自然语言及文本信息数据都能对应汉字词组进行自动认知，自动排除歧义完成正确的语义认知过程。以往的语言及文字翻译系统，被翻译的原文内容在语义上出现多重歧义，自动方式难于判断歧义词组与上下文语境的关联关系；新发明对于任何自然语言及文本信息，自动翻译为任何自然语言及文本信息，在内容上出现多重语义的情况，都能对应汉字词组，以部首的分类属性，正确的自动判断语境中出现歧义的语义。

人类的认知方式，除了通过语言和文字以外，还会以视觉、听觉、味觉和感官实现，例如视觉上看见红色，心理上浮现的语义有热情、危险和停止等；通过听觉能分辨悠闲、悦耳、轻快或嘈吵等；味觉上亦会理解到甜、酸、苦、辣等；身体的感官知觉受压亦能分辨出是轻压还是痛打。以上这些感官通过不同的电子系统撷取後，一般都会以数字作为语义数据贮存，新发明能够以不同的数字数据所表示的感官信息以适当的汉字词组与之对应。例如目前颜色的数字化，都以三原色(R,G,B)表示；“255,0,0”表示为红色，可对应的汉字词组编码为“红色”，“0, 255 ,0”表示为绿色，可对应的汉字词组编码为“绿色”等。人类还会以其他途径进行沟通，例如表情、手势及肢体动作等，自动认知系统撷取表情需要对应语义表示；例如：面部的唇形向上露齿等的表情语义，是对应汉字词组“笑”，人类点头的动作语义对应汉字词组“允许”或“赞成”，肢体方面，左右两手掌轻力互拍，表示的语义对应

为汉字词组“拍掌”、“欣赏”或“欢迎”等。新发明通过电子系统撷取各种信息的数字数据，对应汉字词组语义，能进行综合理解及认知，然後以综合数据；模擬方式作出回应。

本发明的汉字符号编码系统及方法以分组数字编码表示，单一汉字符号的其中一组数字对应不同部首属性，系统就能以不同部首属性进行语义认知。

任何自然语言及文字等语义信息要成为高效率的搜索数据，需要信息高度结构化，以最少的数据量达至最准确的分类。新发明利用汉字的部首属性对全范围语义信息进行分类，人类的知识本身是以不同的类别呈现，而呈现的方式都是以文字固定下来。不同的知识领域包含特定语义，在汉字系统内，特定语义有特定部首表示，如关于医学科的部首有“疒”，“医”及“月”等。所对应的汉字有“病”，“医”及“肿”等。所述语义数据库会以部首属性对不同知识领域进行有效丛集及分类。

本发明能以汉字词组对应不同词组搜索请求，集中搜索语义本身，就能以相同关联语义方式得出相同语义结果。

机械及电子机器的出现，已体现在各种各样的生活应用需求上，但到目前为止，只能以局部范围的语音信息能表示为少数指令集，进行认知及操控。不能进行全范围语义信息认知的原因是任何自然语言语音的重复性，即同音字词数量太多，出现太多歧义，不能转换为唯一指令进行准确操控。人类一直以来都希望能实现全范围自然语言操控机器运作，但侷限于认知全范围语音因同音及近音词组，容易出现认知上的错误。目前的技术，只能进行局部范围自然语言的认知运作上，例如通过语音查询天气、票务或银行账户等；转换为正确指令，进行数据的存取过程，或进一步以指令转换为已予设的电子机械动作。本发明能对人类全范围语义信息，包括任何自然语言及文字语义信息，进行准确认知，并表示及对应为指令操控机械及电子机器。实现全范围语音指令的可能，并能以部首属性编码，组织及丛集相关语义，作出相关回应，此亦是机器人能以相关范围思考学习的实现方法。

附图说明

图 1 是全范围语义认知系统结构示意图。

图 2a 是汉字笔划形态与数字编码对应关系图。

图 2b 是汉字笔划的数字编码示例图。

图 3 是语义排歧工作流程图。

图 4a 是实施例中自然语言的输入内容。

图 4b 是对图 4a 文字输入内容中的关键词进行部首义项分析。

图 4c 是关键词的部首编码与词组的对应关系。

图 5 是实施例 3 中汉字词组与英语同义词的对应关系示意图。

图 6 是关键词以笔划对应分组数字编码示意图。

具体实施方式

现结合附图进一步对本发明的实施例进行说明及解释，本发明的特点、目的和优点将变得更加明显。本处所描述的实施例仅用于说明和解释本发明，并不因此而限定本发明。

如图 1 所示为本认知系统结构，包括信息接收模块 12，转译模块 13，语义数据库 14，输出模块 15。

全范围语义信息 11，包括任一种自然语言及文字信息 111，如汉语、英语、德语、西班牙语、日语等语种的语音及文字；或者可用任一种自然语言及文字表达的信息，如视觉、听觉、味觉等感官信息 112；以及表情、手势、肢体动作等动作信息 113；通过信息接收模块 12 输入计算机系统中。接收模块可包括多类别的接收及数据输入装置，可将声音、动作、感官等信息接收并最终以文字方式表达。接收及数据输入装置可采用现有的装置，在此不作赘述。

语言或文字信息通过转译模块 13，根据语义转译至语义信息数据库 14。语义数据库 14 由汉字词组构成。语义数据库内的汉字按照部首属性编码规则编码成可应用至计算机系统的数字编码。部首属性编码规则是指汉字按照预定笔画集合和笔画顺序拆分成至少一个笔画、与数字构成的编码一一对应。

编码后通过输出模块 15 进行转换及输出模拟数据，以实现检索或翻译等功能。

该预定笔画集合由点“、”——代表点类笔画、短撇“丿” ——代表短

撇及短捺类笔画、 长撇“丿” ——代表长撇及长捺类笔画、 短划“-”——代表短横及短竖类笔画及长划“一” ——代表长横及长竖类笔画组成。

具体地来说，是以1、2、3、4、5作为数字编码，分别对应点“、”、短撇“丶”、 长撇“丿”、 短划“-”及长划“一”五种笔划型态。当汉字笔画不足时，不足部分以数字“0”表示。

汉字字型在形式分类上，分为横排和竖排两种；而在字形结构上分为单体字及合体字两种，每个汉字皆以两组数字组合进行编码。因此，每个汉字根据字型结构以两组共6个数字字节组成表示。笔划型态组合编码只有6个，转为二进制数字表示，每笔划数据长度为最多3位元，每个汉字数据长度为18位元。

现以实例解释上述汉字编码规则。

实施例1

如图2a所示，为五种汉字笔划形态“、”、“丶”、“丿”、“-”、“一”，分别以1、2、3、4、5编码，笔划不足的编以数字0，一共为6个数字。如图2b所示，以汉字“我”为例，“我”字为单体字，首部件笔划顺序编码为255，“我”字没有次部件，因此编码为000，完整分组编码即为255·000。又以“统”为例，首部件笔划顺序编码为222，次部件编码为142，整字分组编码即为222·142。

为简化输入及提高操作效率，本发明制定的规则中，五种汉字笔划形态分别是以1、2、3、4、5作为编码的，笔划不足的编以数字0。但若以另外6个数字，甚至以字母字符来编码各汉字笔划形态，亦不违背本发明的精神，应视为在本发明的保护范围之内。

目前被广泛应用的自然语言及文字系统，都存在歧义问题，分别存在于同音词组及同义词组内。以任何一种自然语言及文字系统的同音词组，对应不同的汉字词组，不同的汉字词组具备不同的部首义项属性，即：

同音词组A → 汉字词组A → 部首义项集1

同音词组B → 汉字词组B → 部首义项集2

⋮ → ⋮ → ⋮

同音词组 n → 汉字词组 n → 部首义项集 n

在语义数据库 14 内设有若干丛集词库 141，汉字词组按照部首义项对同一应用领域的汉字词组进行丛集及分类，如医学、法学、建筑学、经济学、美学及天文学等等。这相当是应用了汉字部首特有的标签分类功能，能对同音、近音歧义词及同名异义词进行排歧及筛选，从而确定符合匹配关系的词组。

该排歧筛选过程可见图 3 所示的流程。

步骤 301 表示，任何一种自然语言或文字在文字输入时，语义内容出现了歧义，即一词多义，如同音、近音歧义词或同名异义词。

步骤 302 表示，对上述多义词的各个语义通过转译模块对应为汉字词组认知信息数据库 14 内的不同语义的汉字词组。

步骤 303 表示，各不同语义的汉字词组存在着不同的部首义项属性，可以数字编码的形式进行提取。

步骤 304 表示，对于歧义的各语义词组需与上下文的语义关系进行匹配比较，实际上即是以部首义项与上下文的部首义项进行语义匹配。

步骤 305 表示，先进行上文部首义项关系属性的匹配比较。

步骤 306 表示， 然後进行下文部首义项关系属性的匹配比较。

步骤 307 表示， 歧义词组的多个语义部首义项匹配规则是优先选择上下文语义的部首义项最大关联语义者作为匹配语义。

现以具体实例解释上述流程。

实施例 2

任何自然语言系统内都存在同名异义，同音、近音歧义的情况，即具有着相同或相近的字母拼写的词语有着完全不同的语义，当转换为电子数据进行语义识别时，就会出现歧义问题。如图 4a 所示，输入一段英语文字内容。如图 4b 所示，对这段文字内容的多个关键词进行部首义项分析。在这段文字内容中含有同名多义词 “cancer”。英语单词 “Cancer” 在不同的语境内，具有完全不同的语义；语境与医学有关的，其语义为癌病、癌症及肿瘤等；当语境与星相学有关时，其语义为巨蟹座。语音内容对应为汉字语义词组时，例如名词“Cancer”就会出现两种不同语义。“Cancer” 有多个语义，如“癌症”，对应部首为“疒”；肿瘤，对应部首为“月疒”；“巨蟹座”，对应部首为“匚虫广”，见图 4b 的 402。上文“hospital”语义为“医院”。 “医” 的部首是“医”，见 401。下文 “patient” 语义为“病人”，“病”

的部首义项是“疒”。如图4c所示，上述部首义项的编码分别为555及153，在部首丛集词库内，“医”部与“疒”属于与医学有关的，丛集于同一词库内，因此“cancer”在此语境内会自动判断为与病理有关的语义，排除另一语义“巨蟹座”。

同理，“treatment”对应的汉字词组是“疗法”或“处理”，“疗法”的部首分别是“疒”与“氵”；“处理”的部首分别是“夂”与“王”。通过上下文匹配关系自动判断为“疗法”。

一般的关键词搜索过程，都是以关键词的拼写形式或书写方式在数据库内进行搜索及匹配。当同一语义有多个表达方式时，要搜索出该语义的相关文献，就必须要把所有的拼写表达方式都分别输入，过程变得复杂、缓慢、低效。新发明以汉字语义词组对应任何自然语言的语义，根据唯一的语义进行搜索，大大减小搜索数据量，有效地提高操作效率。

现以具体例子加以说明。

实施例 3

如图 5 所示，501 列出与 Britian 具有相同语义的字母串组合，包括 England, UK, U.K., United Kingdom, GB, G.B., Britian and Great Britian 等。

当需要搜索含有“英国”含义的英文相关文献时，由于不确切该文献中“英语”的拼写表达方式，可能是 England, UK, U.K., United Kingdom, GB, G.B., Britian and Great Britian 的任何一种，因此可能需要分别输入以上所有的表达方式才能找到所需文献。

502 表示上述各种拼写所表达的语义是唯一的，对应为汉字词组即为“英国”。如图 6 所示，“英国”所对应的数字编为 554.454 和 555.545。每个汉字以 6 个数字字节表示，每个字节为 3 位元，所以 6 字节的位元数量为 18 位元。503 表示以汉字语义词组数据库综合对语义信息进行搜索。因此，应用本法进行关键词搜索时，只需要搜索“英国”的数字编码 555.531，相关语义的词组都能一并出现，减少关键词冗余列表数量，检索过程大为简化，数据量也大大减小。

实施例 4

人类一直以人手、完整逻辑指令集及希望以语音操控电子机器。本发明

对人类全范围语义信息，包括任何自然语言及文字语义信息，进行准确认知，并表示及对应为指令操控机械及电子机器。实现全范围语音指令的可能，并能以部首属性编码，组织及丛集相关语义，作出相关回应，此亦是机器人能以相关范围思考学习的实现方法。

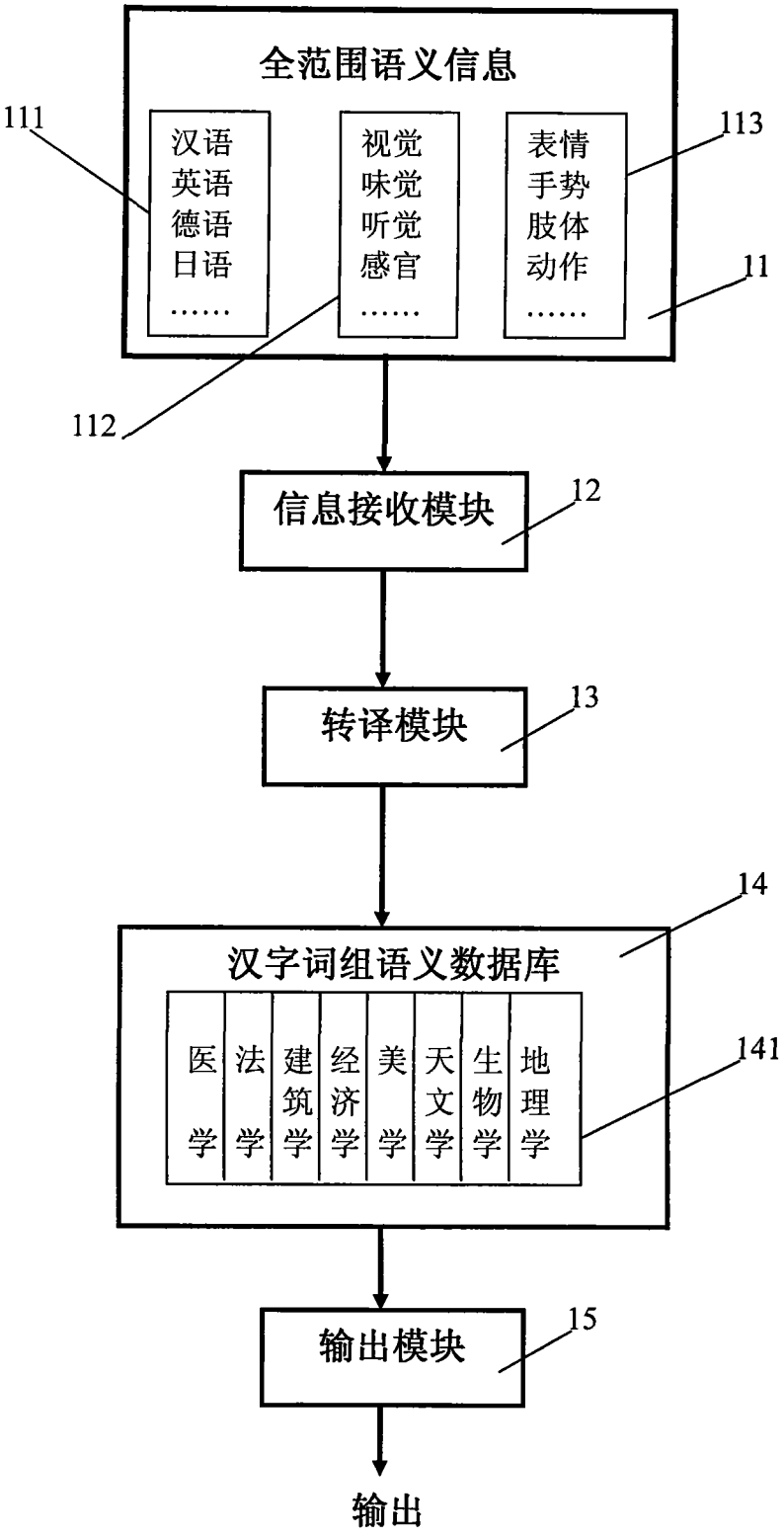

全范围语义信息
111
汉语
英语
德语
日语
……
视觉
味觉
听觉
感官
……
表情
手势
肢体
动作
……
113
11
112
信息接收模块
12
转译模块
13
14
汉字词组语义数据库
医学
法学
建筑学
经济学
美学
天文学
生物学
地理学
141
输出模块
15
输出

图 1

	点	短撇	长撇	短划	长划	不足笔划
笔划形态	丶	丿	丿	一	一	
编码	1	2	3	4	5	0

图 2a

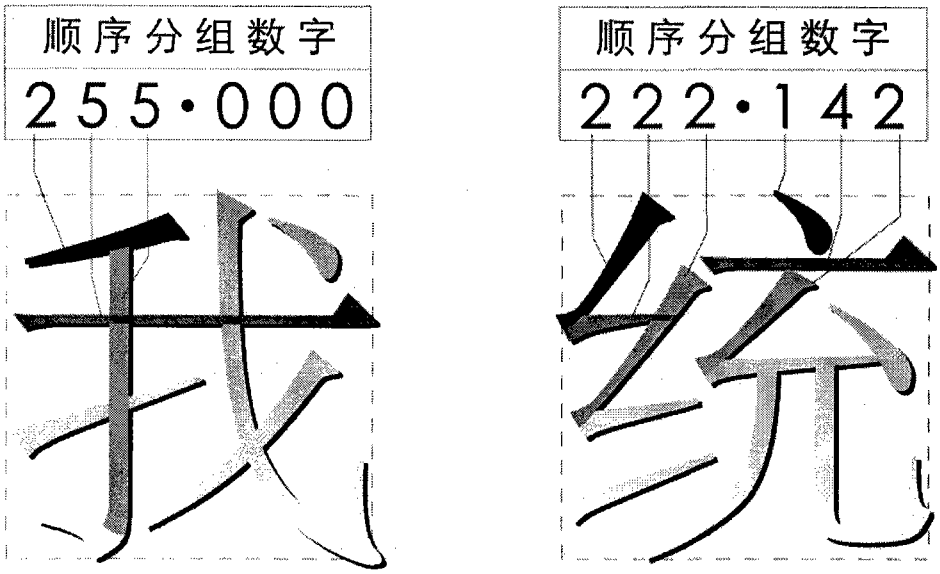


图 2b

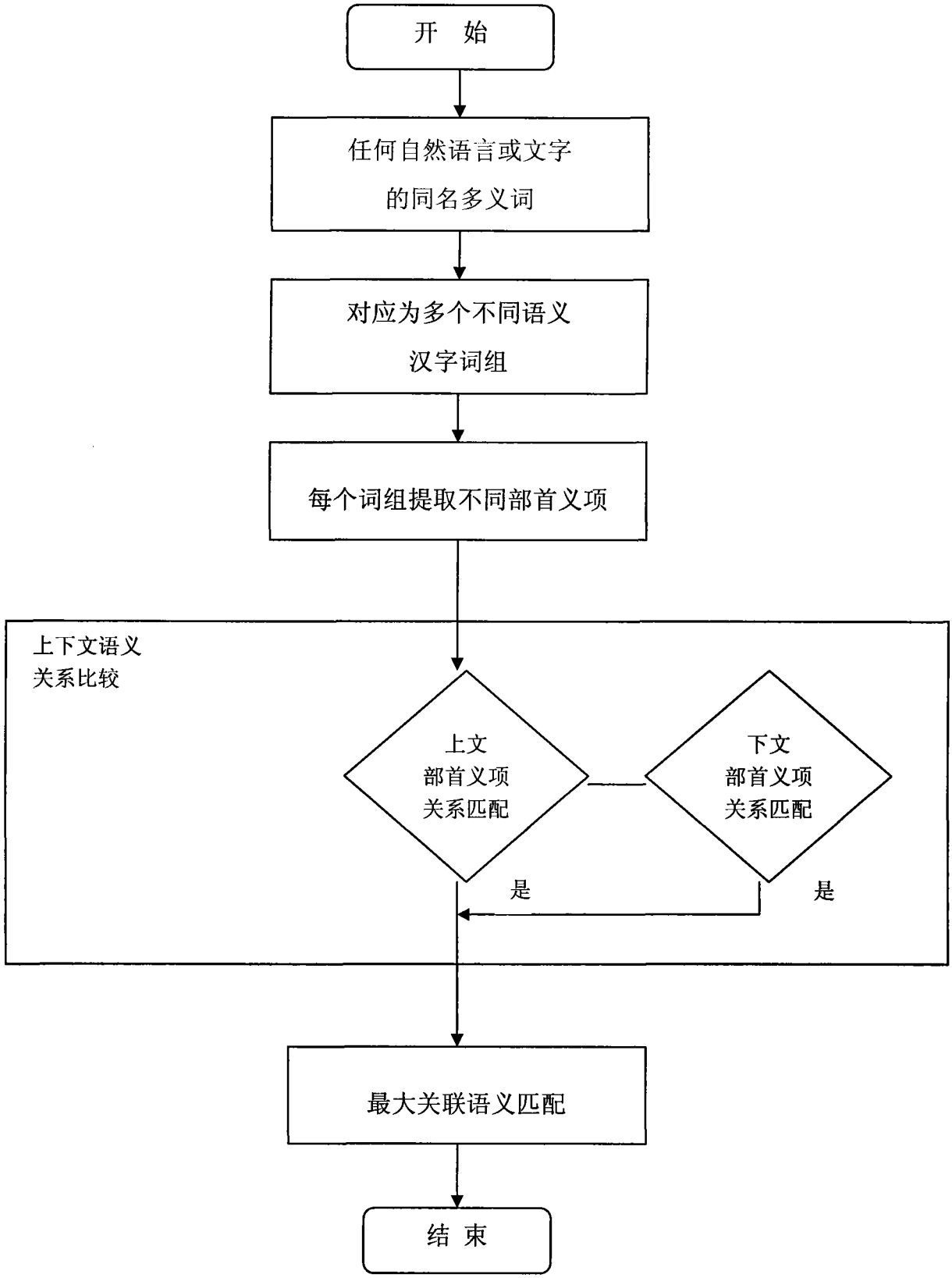

开 始
任何自然语言或文字
的同名多义词
对应为多个不同语义
汉字词组
每个词组提取不同部首义项
上下文语义
关系比较
上文
部首义项
关系匹配
下文
部首义项
关系匹配
是
是
最大关联语义匹配
结 束

图 3

自然语言口语输入内容
Oasis of Hope Hospital was created with the specific purpose of providing alternative cancer treatment using a multi disciplinary approach that meets the physical, emotional, and spiritual needs of the patient.

图 4a

属性分类	Oasis	hope	hospital		purpose	cancer		treatment		patient	
医学的	绿洲	希望	医 阝	医院	目的	疒疒	癌症	疒氵	疗法	疒人	病人
医学的						月疒	肿瘤	夕王	处理		
星座学						厂虫疒	巨蟹座				

图 4b

关系词组	部首	部首数字编码	汉字词组
Hospital, doctor	医	555	医院、医生
Cancer, patient	疒	153	癌症、肿瘤、病人
……	…	……	……

图 4c

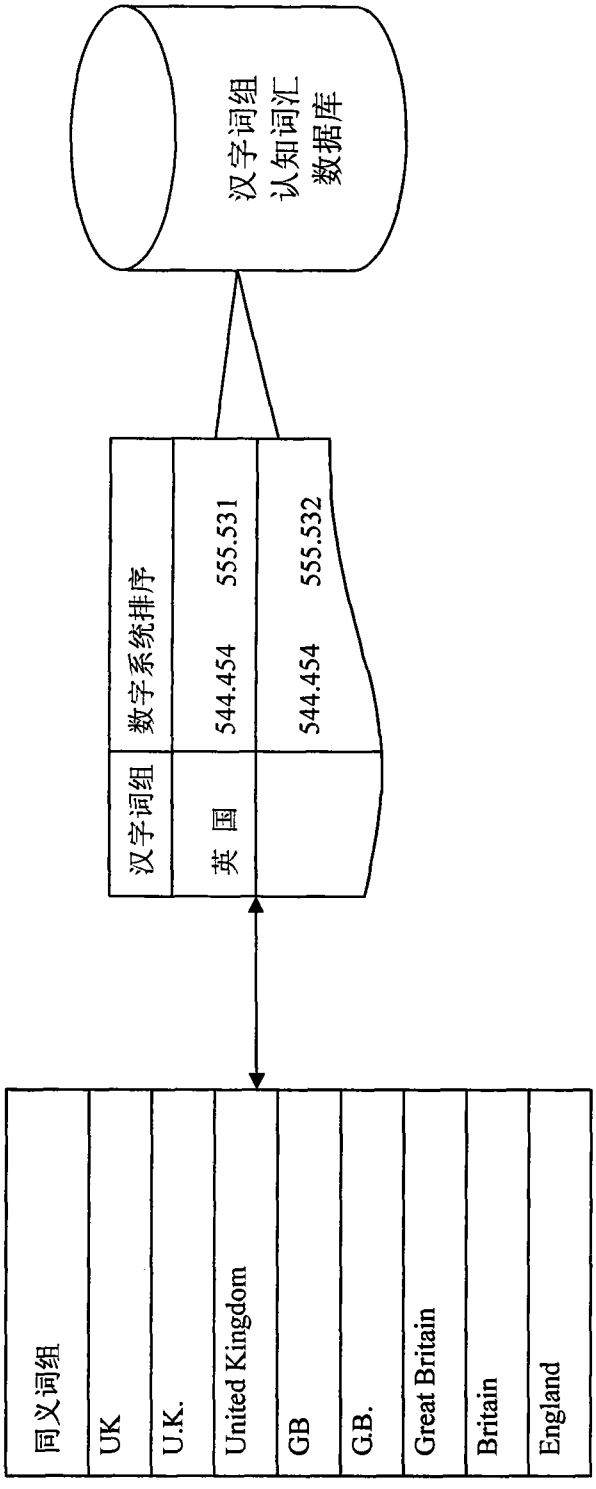


图 5

英语词组	对应汉字	首部件	首部件的首三划	次部件	次部件的首三划	3 字节分组 顺序数字编码	位元数量
Britain	英	艹	一 一 一	央	一 一 一	554.454	18 位元
	国	口	一 一 一	玉	一 一 一	555.545	18 位元

图 6