(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
2007 年12 月21 日 (21.12.2007)

PCT

(10) 国际公布号
WO 2007/143899 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2007/001662

(22) 国际申请日: 2007 年5 月22 日 (22.05.2007)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
200610081367.6
2006 年5 月22 日 (22.05.2006) CN

(71) 申请人及
(72) 发明人: 赵开灏(ZHAO, Kaihao) [CN/CN]; 中国北京市朝阳区团结湖北五条1507号, Beijing 100026 (CN)。

(74) 代理人: 北京中海智圣知识产权代理有限公司 (ZHONGHAI WISDOM INTELLECTUAL PROPERTY AGENT CO., LTD.); 中国北京市海淀区知春路1号学院国际大厦602室, Beijing 100083 (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, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

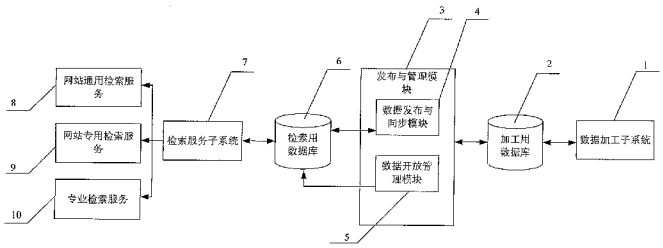
(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告。

[见续页]

(54) Title: SYSTEM AND METHOD FOR INTELLIGENT RETRIEVAL AND TREATING OF INFORMATION

(54) 发明名称: 信息智能检索加工的系统和方法



1 DATA TREATING SUBSYSTEM
2 TREATING DATABASE
3 ISSUING AND MANAGING MODULE
4 DATA ISSUING AND SYNCHRONIZING MODULE
5 DATA DEVELOP MANAGING MODULE
6 SEARCHING DATABASE
7 SEARCHING SERVING SUBSYSTEM
8 WEB SEARCHING SERVICE FOR GENERAL PURPOSE
9 WEB SEARCHING SERVICE FOR SPECIAL PURPOSE
10 PROFESSIONAL SEARCHING SERVICE

(57) Abstract: A system and a method for intelligently searching and treating information are disclosed. The system comprises a data intelligent treating subsystem, a treating database, an issuing and managing module, a searching database, and an intelligent searching serving subsystem, while the issuing and managing module comprises a data issuing and synchronizing module, a data develop managing module; the system can treat the data to divide the data into 12 kinds. Said method for intelligently searching and treating information comprises the following steps: inputting a search condition; pretreating; dividing the search request into simple direct search, advanced integrated search, classified browse search, full-text search, and intelligent logic search, while the first three modes directly use the relationship search engine to search; the full-text search uses the full search engine; the intelligent logic search uses the logic relationship calculation to recombine the inquire condition to search with the relationship search engine; after obtaining result, feeds back the search result.

[见续页]

WO 2007/143899 A1

所引用双字母代码及其它缩写符号，请参考刊登在每期PCT公报期刊起始的“代码及缩写符号简要说明”。

(57) 摘要：

本发明公开了一种信息智能检索加工的系统和方法，该系统包括数据智能加工子系统、加工用数据库、发布与管理模块、检索用数据库、智能检索服务子系统，其中发布与管理模块包括数据发布与同步模块、数据开放管理模块；该系统对数据进行加工，将数据分为 12 大类。本发明的信息智能检索加工的方法，包括：输入检索条件；预处理；搜索请求细分为简单直接搜索、高级组合搜索、分类浏览搜索、全文检索以及智能逻辑搜索，前三种搜索方式将直接通过关系搜索引擎进行搜索，全文检索将通过全文检索引擎进行检索，智能逻辑搜索将通过逻辑关系演算进行查询条件重组后再由关系搜索引擎进行搜索；获得搜索结果后，返回搜索结果。

信息智能检索加工的系统和方法

技术领域

本发明涉及一种信息智能检索加工的系统和方法，尤其涉及一种文本、图像、音频、视频的智能检索加工的系统和方法。

背景技术

数据信息和文档的有效检索和加工，是数据库应用领域中的核心和重要内容，广泛存在于各种电子数据、文献、商业数据库资源和互联网内容搜索的应用当中。

目前这一领域中的数据信息检索技术，一般是基于关键词的统计方法，用关键词的布尔表达式作为查询语句。对于文件数据库，使用关键词加关键词出现在文件中位置的字典，通过比较查询语句的关键词与文件数据库字典中的关键词，找到相应文件。另外，有些改进采用了模糊逻辑模型、向量空间模型和概率检索模型等。

但是这些方式只能实现以整篇文档为单位的检索，而且对于检索关键词与整篇文档之间的相似性程度，目前均处于努力提高和改善阶段，尚无满意的解决方案，导致无法实现信息检索结果的精准。如：无法找到同义不同形的关键词，或找到的是同形不同义的关键词。对于整篇文档内部包含的各类数据和信息，这些方式无法进行单独识别、区分，并进行基于知识属性关系的加工和利用，更无法实现不同文档内容之间的交叉分析和比较，以及在不同文档之间实现信息内容的多次加工利用。

目前各种数据库中的知识处理和检索结果均以整篇文档为最小单位，由于整篇文档中所含的知识属性非常丰富，所以这种方式在知识处理和检索结果提供环节均存在问题。

在知识处理环节，目前操作都是通过主题词标引、个别关键词标注、文档摘要方式对整篇文档进行属性标识，并作为检索过程中的检索关键词，这种方式远不能完全反映整篇文档中的全部知识信息，最终结果表现为检索结果中的文档缺失。

在检索结果提供环节，整篇文档中所携带的大量非相关信息会产生信息冗余和噪音，影响结果的精准，最终结果表现为检索结果中的文档泛滥、有效性降低。

发明内容

为了解决上述存在的问题，本发明提供一种新型的信息智能检索加工系统和方法，可以解决数据信息和文档中所包含的各类的检索问题，更可以满足文档中不同关键词之间、不同文档间的关键词之间，进行信息、知识的比较、分析、重新建立关系等信息数据的智能加工要求，能够支持诸如 “隐含指代”等较为复杂的搜索请求；同时通过系统支持的多格式位置表达技术，包含文本、图像、音频、视频等多种介质格式在内的内容都可以得到检索和加工。

本发明通过以下方案实现：一种信息智能检索加工的系统，包括数据智能加工子系统、加工用数据库、发布与管理模块、检索用数据库、智能检索服务子系统，其中发布与管理模块包括数据发布与同步模块、数据开放管理模块；

其中所述数据智能加工子系统对数据进行智能加工，将数据处理成深度拆解和标引的单元内容以及灵活准确的智能索引信息，存入加工用数据库中，加工用数据库中还存储着大量标志信息以及为了加快加工而生成的中间结果；

所述发布与管理模块完成将经过审批的内容和索引信息与智能检索服务子系统显示的数据进行同步；数据同步由数据发布与同步模块执行，将加工用数据库的内容同步到检索用数据库，将检索过程中的反馈信息从检索用数据库同步到加工用数据库中；数据开放管理模块负责对数据访问进行权限设置；

所述智能检索服务子系统提供智能检索服务平台，对来自用户的搜索请求进行统一处理，查询检索用数据库，智能检索相关内容。

一种信息智能检索加工的方法，其步骤为：

1、输入检索条件；除了目前大多数的检索服务系统提供的关键词输入和索引浏览两种输入方式之外，本系统中还可通过字根、笔顺输入法输入 Unicode 字库中包含或未包含的大量偏稀汉字；

2、对检索条件进行预处理，这其中包含了代码转换和索引复杂度评估；

3、搜索请求细分为常规的简单直接搜索、高级组合搜索、分类浏览搜索、全文检索以及智能逻辑搜索，前三种搜索方式将直接通过关系搜索引擎进行搜索，全文检索将通过全文检索引擎进行检索，而智能逻辑搜索将会通过逻辑关系演算进行查询条件重组后再由关系搜索引擎进行搜索；

4、通过关系搜索引擎或全文检索引擎获得搜索结果后，返回搜索结果。

本发明建立以深度拆解和标引文本内容以及建立高度灵活智能索引机制为基础的多介质检索方法和索引系统。通过设计和实现三元关系模型以及对汉字字词之间关系的多角度描述；建立了灵活高效的交叉索引体系；并且在这个交叉索引系统基础上，实现了具有语义分析功能的智能检索技术；同时通过对内容标引方法的规范化，使得字词和内容的关联比对更加智能化，能够支持诸如 “隐含指代”等较为复杂的搜索请求；同时通过智能系统支持的多格式位置表达技术，包含文字以及图像、表格、音频、视频等多种介质格式在内的内容都可以得到检索。

本发明具有如下显著优点：

1、能够实现信息内容检索的精准化，真正符合检索意愿，最大限度减少检索结果的冗余。

2、能够满足检索过程中的使用者随机需求。

3、能够通过系统丰富的知识背景和准确的知识发散路径，提供基于知识而非信息的检索结果。

4、能够实现任意知识源间基于知识元层次的全新的信息内容组合和知识组合，实现任意信息内容基于人、事、时、地、物等人类基本生产、生活、活动通用属性之间的交叉比对，实现对文本、图像、音频、视频等多种介质格式在内的内容进行二次加工，可自动生成二次、三次或多次文献。

5、能够实现对海量知识的活化和二次加工，实现信息向知识的快速有序转化。

6、涵盖人类生产、生活、活动中的各方面及不同的知识点，解决了海量信息检索中最佳知识路径问题，体现出较好的完备性。

7、完全对应于人类对知识的主观需求，具有很好的通用性和适用性，可正向、逆向进行检索，便于查询与记忆，操作方便，不需培训。

附图说明

图 1 是本发明的索引三元关系模型的典型情况；

图 2 是本发明的实施例中人物索引关键词之间的关系；

图 3 是本发明的实施例中关系关键词之间的关系；

图 4 是本发明的实施例中“逆关系”的推演路径；

图 5 是本发明的实施例中“二次传递”的推演路径；

图 6 是本发明的实施例中“相同主词”的推演路径；

图 7 是本发明的实施例中“对称”的推演路径；

图 8 是本发明的系统框图；

图 9 是本发明的流程图。

具体实施方式

下面结合附图和具体实施例对本发明进行更详细的描述。

本发明的出发点是拆解被搜索或处理的信息内容的内在含义和结构，在此基础上构建搜索和加工体系。因此，本发明将完全不受文本比对的限制，能够一方面做到含义的精确，即不包含无关或仅仅是字面相同的信息；一方面做到信息的完备，即能够包含字面不同但含义相同或具有用户指定关联特征的信息。

另一方面，本发明建立了高度灵活的智能索引机制，并在此基础上一方面充分保证各种分类信息的科学性，另一方面具备符合人们各种习惯和约定的易用性。

本发明并不排斥现有的搜索引擎和搜索服务系统，相反地，本发明可以和现有搜索引擎和搜索服务系统很好的集成，在不同的搜索需求下发挥相应的功能，并组合出更强大的搜索服务能力。

本发明中，准确的内容检索的实现，是对以“知识”形式出现的检索结果进行拆解。这种拆解包括两个层次，一是对检索结果本身进行拆分，形成了以完整、独立含义为特征的“知识单元”或“知识片”；二是对内容中包含的关键词进行了提取，增加关键词的含义相关度信息及加强了具有“隐含指代”等关系的有效关键词，丰富了主要信息的知识属性。去除了相关度不大的无效关键词，以此减少 所有数据源中相对检索来说次要的信息对主要的信息的干扰。

本发明的智能检索方式，结合了按索引分类浏览和按文本关键词匹配两种常用的检索方式。和常用的搜索引擎不同的是，本发明中按索引分类浏览除了能按照常见的学科、习惯等分类方法按照成员隶属关系逐级细分以外，还提供了基于等价别名关系和背景参考关系两种横向拓展的检索途径。和常见系统中的相关链接跳转不同，这种横向检索途径依旧是按照索引分类进行的，具有明确的指向性。和常用的搜索引擎不同的另一方面在于，本发明中按文本关键词匹配进行检索可能会检索到同名不同义的关键词，这是检索用户可以通过系统提示，清楚地了解同名不同义的关键词的相关信息，直接进行快捷的二

次检索，定位自己所需要的结果集。

本发明提出按照自然语义，以表达完整自然语义的最小单元作为知识单元的划分标准，这既可以在知识处理时穷尽每一个知识单元的属性，又可以在检索结果呈现时反应精准，减少信息噪音。

本发明的信息智能检索加工系统是完全依照人性自然需求本能的思维逻辑进行分类的系统。本发明根据人类检索和使用知识的思维方式，将信息数据分为 12 大类，即人物、事件、时间、地点、物品、生物、衣物、食物、住物、行物、育物、乐物（简称人、事、时、地、物、生、衣、食、住、行、育、乐），每个大类再细分为若干子类，如人物的子类有人之姓名、人之性别、人之籍贯等等；每个子类又有若干子类，如人之姓名又分为姓赵、姓张、姓李等等。这样形成树状多层次结构，作为索引结构 30 层就足以表达各种细分数据。每个大类和其子类的索引都用相应代码表示，在此基础上再进行索引的二次加工，穷尽索引结构的背景信息进行标引，重新排序和聚类，形成高度灵活、准确、多维指向、互有交叉的智能化索引。

将各种信息数据，包括各种文献、电子数据，按照其内容长度或容量划分为若干知识单元，文本知识单元容量为 600 个字符以内，对该知识单元编号。然后，对每个知识单元的内容进行分析、拆解，把各关键词按照上述分类方法，编号后对应到上述关系树的子类上。

本发明的分类方法与过去传统的专业分类逻辑有本质上的区别，完全打破传统的分类概念。目前其他各种专业的分类方法，主要基于符合专业的结构层次，并非首要考虑满足人的自然知识需求，通

用性程度不高。对于用户基于自然知识需求的检索要求，实现过程中转换的复杂程度较高，并不适用。

本发明的另一大特点是可以包含其他各种专业分类法，由于本发明的分类方法着眼于符合人类基本知识需求，其分类角度的普适性使其可以涵盖和包容其他各种基于专业的分类方法，这样就可以将各种不同的分类方法进行统一和整合，为知识处理和使用的整合创造技术条件。

本发明中，构建高度灵活智能索引机制的实现，是建立了一种自包含、自组织的三元关系模型。各种常见语言都具有主要语法结构：（主语、谓语、宾语），本发明对这种三元关系进行了模拟，实现了基于三元关系模型的数据表达、存储和检索。

如图 1 所示，本发明的三元关系模型采用三元组 Ka,Kr,Kb 形式，其中 Ka 代表关键词 a，Kb 代表关键词 b，Kr 代表关键词 a 和关键词 b 之间的关系。该三元组形式表示和实现关键词之间的三种类型的关联关系，包括成员隶属关系、等价别名关系以及背景参考关系。

每种类型中可以不断细分，同时各种关系之间仍能实现三种类型的关联。在这种三元关系模型的基础上进行演算，可以进行包含逻辑含义的检索，有别于单纯进行关键词组合的查询方式。

Kr_r 代表关系关键词之间的关系，如逆关系、二次传递、相同主词、对称等，Kr'代表 Kr 根据 Kr_r 推导出的关系，由此 Ka'关键词和 Kb'关键词具有了新的关系 Kr'。

图 2 为人物索引关键词之间的关系的一个例子：如果系统中的

人物关键词包含了以下三个三元组：

（张老三，儿子，张三）（张三，儿子，张小三）（张三，儿子，张小四）。

同时，图 3 所示，系统中定义了以下针对关系关键词的三元组：

（儿子，逆关系，父亲）（儿子，二次传递，孙子）（儿子，相同主词，兄弟）（兄弟，对称，兄弟）。

那么系统可以在不增加其它信息的情况下，自动推演出以下结论：

如图 4 所示，根据“逆关系”可以推演出：（张三，父亲，张老三）（张小三，父亲，张三）（张小四，父亲，张三）。

如图 5 所示，根据“二次传递”关系可以推演出：（张老三，孙子，张小三）（张老三，孙子，张小四）。

如图 6、图 7 所示，根据“相同主词”关系可以推演出：（张小三，兄弟，张小四）并在此基础上根据“对称”关系推演出（张小四，兄弟，张小三）。

注意：推演的先后次序根据实际情况可能不同。

以上结果只是应用一次关系关键词三元组的结论，如果多次、组合应用，可以产生出更多的逻辑结果。

和目前已有检索系统相比，上述推演具有以下特点：

1、基础数据量大幅度减少：上例中，基础数据仅有 3 个人物三元组和 4 个关系三元组，而目前已有检索系统为了满足不同的检索要求，需要完备的基础数据，上例中所有推演的结论都需要作为基础数

据进入系统。

2、可检索数据的大幅度增加：从上例的推演可以看出，用户可以检索的数据，不再是仅仅依赖于基础数据量，同时也和关系三元组的数量相关。由于关系三元组具有很强的通用性，因此，当增加一个关系三元组，带来的可检索数据的增加将会是成倍甚至是几何级数的。

3、数据关系一致性更强：由于大量结论是系统经过逻辑推演得到的，因此具有严密的逻辑性。而目前已有检索系统由于基础数据都是独立进入数据库的，因此可能会有同时出现（张老三，儿子，张三）（张三，兄弟，张老三）的情况，数据一致性得不到保障。

4、关系的扩展性：从上例的推演可以看出，只要是符合逻辑的关系三元组就可以在系统中进行定义，从这个意义上说，一方面根据生活经验以及现有的科技发展状况总结出来的关系可以通过这个系统实现，同时随着社会、科技的不断进步，新的关系将不断出现，而这些新的关系同样也可以在系统中实现；并且对于了新定义的关系三元组，所有以前的数据将马上得到相应的组织以备查询。

本发明采用了知识单元标引方法，类似关键词的三元模型，知识单元的标引采用(C,R,K)组和(Ca,R,Cb)三元组进行表示和实现，其中 C 表示知识单元的内容，K 表示关键词，R 代表知识单元和关键词之间的关系；Ca 代表知识单元 a 的内容，Cb 代表知识单元 b 的内容，R 代表知识单元 a 和知识单元 b 之间的关系。该方法记录知识单元中关键词的位置、长度、相关度等以及知识单元之间的相互引用等关联知

识。通过这种标引，一方面知识单元能够以结构化的方式呈现，满足用户对关联信息的需要，同时另一方面，也可以按照知识来源的最初模式进行呈现。

另外，通过(C,R,K)三元组，标引方法很好的解决了知识单元中的“指代”关系，例如，对于一个知识单元中出现的代词“他”，通过在三元组中确定实际的指代目标，系统就可以向用户提供针对指代目标的检索，而不仅限于文字上的相同或近似。

图 8 说明了本发明中信息智能检索加工系统的整体结构。本系统包括数据智能加工子系统 1、加工用数据库 2、发布与管理模块 3、检索用数据库 6、智能检索服务子系统 7，其中发布与管理模块 3 包括数据发布与同步模块 4、数据开放管理模块 5。

数据加工由数据智能加工子系统 1 完成。数据将在这里从不同介质的各种来源经处理后成为经深度拆解和标引的文本或其他介质格式内容以及灵活准确的智能索引信息。这个阶段主要对加工用数据库 2 进行操作，除了最终用于检索的各项信息外，加工用数据库 2 中还存储着大量标志信息以及为了加快加工而生成的中间结果。

在数据加工阶段，整个处理过程被分为三个步骤：

（1）首先，基础数据加工，这是针对文本内容正确性的加工步骤。在这一步中系统将对进入数据库的数据进行校对，校对的内容包括文字、目录和段落层次、注释的引用等。本发明还能够支持标准 Unicode 字库中包含或未包含的大量偏稀汉字，即所谓异体字或图像字的查询和显示，这是通过对异体字或图像字进行编号实现。

（2）其次，在基础数据的正确性得到保证的前提下，进行知识单元智能加工。在这一步骤中，系统将对原来以自然段为基础单元的数据进行拆解，形成具有独立完整含义的“知识单元”。同时在这个步骤中，系统还将建立起“知识单元”和索引关键词之间的关联关系。

（3）数据加工的第三个步骤为智能索引加工，和上一个步骤知识单元智能加工在实际操作中是并行进行的，智能索引加工需要对来自知识单元智能加工中提取的关键词进行索引化处理，再将经索引化处理过的结果进行二次加工，编制出灵活准确、多维指向、互为交叉的智能化索引。

（4）智能化索引反作用于知识单元过程，可根据使用者的随机需求形成新的分类、排序与聚类，生成二次、三次或多次文献、表单、图像、音频、视频。

数据智能加工子系统 1 还包括工序管理和控制模块，对这些步骤中的中间结果、数据状态进行管理。这个模块本身不对数据产生直接的影响，而是对数据的流向进行监控和管理。

发布与管理模块 3 承担发布与管理的任务。该模块主要完成将后台经过审批的内容与索引信息和前台数据进行同步。实现过程是双向的，主要的数据流是从加工用数据库 2 流向检索用数据库 6，但与此同时，一些检索过程中的反馈信息也将从检索用数据库 6 同步到加工用数据库 7 中，这些数据同步的过程将由数据发布与同步模块 4 执行。发布与管理模块 3 还有一个重要的任务，就是对数据访问进行权限设置，由数据开放管理模块 5 承担的功能。

由网络用户发起的检索操作，由智能检索服务子系统 7 完成。用户发起的检索请求，包括横向的通用检索以及纵向的专用检索（通用检索请求指使用常用的关键词或关键词的组合的检索请求，而专用检索请求指通过本系统提供的分类法的检索请求），将被转换为相应的内部检索请求，对内容和索引信息进行智能检索。另外，在这个阶段，系统还提供公共的访问接口，为一些专业检索请求服务，如其它网站可以通过链接到本系统提供专业检索服务。

本系统提供了一个公共的智能检索平台——智能搜索服务平台，对来自不同用户的各种搜索请求进行统一处理。在此基础上，系统自身提供了以获取丰富关联内容为目的的横向的网站通用检索服务 8 功能和以获取深层次知识为目的的纵向的网站专用检索服务 9。另外，前面提到的公共的访问接口，是以专业检索服务 10 的形式提供的。

图 9 说明了使用本发明的信息智能检索加工的方法，对于用户 11 的检索请求，本发明是如何进行处理的。图中的方框表示了各种处理操作，圆柱图表示了包含索引数据 61 和内容数据 62 的检索用数据库 6。图中的实线箭头表示了操作的流程，而虚线箭头则代表主要的数据流向。

在实际的操作中，用户 11 主要通过系统提供的网站或者通过开放接口接入本系统的其它系统的用户界面输入检索条件 12。除了提供的关键词输入和索引浏览两种输入方式之外，本系统中还提供了使用拼音或笔顺的输入方式来输入Unicode字库中包含或未包含的大量偏稀汉字。

在系统获得了用户的检索请求后，本系统将对检索条件进行预处理 13。这其中既包含了常用的代码转换 14 技术，同时也包含了索引复杂度评估 15 技术。在经过了预处理条件 13 后，搜索请求会被细分为常规的简单直接搜索 16、高级组合搜索 17、分类浏览搜索 18、全文检索 19 以及智能逻辑搜索 20，前三种常规搜索方式将直接通过关系搜索引擎 22 进行搜索，全文检索 19 将直接通过全文检索引擎 23 进行检索，而智能逻辑搜索 20 将会通过逻辑关系演算进行查询条件重组后再由关系搜索引擎 22 进行搜索，该逻辑关系演算是基于上述三元关系模型、分类索引库和知识单元标引的方法。在最终通过关系搜索引擎 22 和全文检索引擎 23 获得搜索结果后，系统会使用能充分体现搜索条件和搜索结果的内在逻辑联系的界面返回搜索结果 24。

本发明涉及的系统和方法，可以应用于计算机单机、局域网络、企业内部网络（Intranet）、互联网（Internet）等多种环境下，系统使用者可以扩展为有信息内容检索需求的任意人群。

本发明能够实现信息内容的智能化检索和加工，真正符合检索意愿，最大限度减少检索结果的冗余，实现任意知识源间基于知识元层次的全新的信息内容和知识的智能化组合，实现任意信息内容基于人、事、时、地、物等人类基本生产、生活、活动通用属性之间的智能化分类、排序、聚类过程。

本发明的特定实施例已对发明内容做了详尽说明。对本领域一般技术人员而言，在不背离本发明原理的前提下对它所做的任何显而易见的改动，都构成对本发明专利的侵犯，将承担相应的法律责任。

1．一种信息智能检索加工的系统，其特征包括数据智能加工子系统、加工用数据库、发布与管理模块、检索用数据库、智能检索服务子系统，其中发布与管理模块包括数据发布与同步模块、数据开放管理模块；

其中所述数据智能加工子系统对文本、图像、音频、视频数据进行加工，将数据处理成深度拆解和标引的知识单元内容以及灵活准确的索引信息，存入加工用数据库中，加工用数据库中还存储着大量标志信息以及为了加快加工而生成的中间结果；

所述发布与管理模块完成将经过审批的内容和索引信息与智能检索服务子系统显示的数据进行同步；数据同步由数据发布与同步模块执行，将加工用数据库的内容同步到检索用数据库，将检索过程中的反馈信息从检索用数据库同步到加工用数据库中；数据开放管理模块负责对数据访问进行权限设置；

所述智能检索服务子系统提供智能检索服务平台，对来自用户的搜索请求进行统一处理，查询检索用数据库，智能检索相关内容。

2．根据权利要求 1 所述的信息智能检索加工的系统，其特征在于：所述数据智能加工子系统对数据进行加工，将数据分为 12 大类，即人物、事件、时间、地点、物品、生物、衣物、食物、住物、行物、育物、乐物。

3．根据权利要求 2 所述的信息智能检索加工的系统，其特征在于：每个大类再细分为若干子类，每个子类又有若干子类，将形成的树状

多层次结构，作为索引结构。树状结构中的知识条目节点具有多种交叉隶属关系；每个大类和其子类的索引都用代码表示。

4. 根据权利要求 3 所述的信息智能检索加工的系统，其特征在于：所述子类小于等于 30 层。

5．根据权利要求 1 所述的信息智能检索加工的系统，其特征在于：所述数据智能加工子系统对数据进行加工，将信息数据，按照其内容长度或容量划分为若干个知识单元。

6．根据权利要求 5 所述的信息智能检索加工的系统，其特征在于：一个文本知识单元容量为 600 个字符以内。

7．根据权利要求 1 所述的信息智能检索加工的系统，其特征在于：所述的数据智能加工子系统采用三元关系模型，三元组 Ka,Kr,Kb 形式，其中 Ka 代表关键词 a，Kb 代表关键词 b，Kr 代表关键词 a 和关键词 b 之间的关系，该三元组形式表示和实现关键词之间的三种类型的关联关系，包括成员隶属关系、等价别名关系以及背景参考关系。

8．一种信息智能检索加工的方法，其步骤为：

（1）输入检索条件；

（2）对检索条件进行预处理，这其中包含了代码转换和索引复杂度评估；

（3）搜索请求细分为常规的简单直接搜索、高级组合搜索、分类浏览搜索、全文检索以及智能逻辑搜索，前三种搜索方式将直接通过关系搜索引擎进行搜索，全文检索将通过全文检索引擎进行检索，而智能逻辑搜索将会通过逻辑关系演算进行查询条件重组后再由关

系搜索引擎进行搜索；

（4）通过关系搜索引擎或全文检索引擎获得搜索结果后，返回搜索结果。

9．一种数据智能加工处理方法，其步骤为：

（1）基础数据智能加工，系统将对进入数据库的数据进行智能校对，校对的内容包括文字、目录和段落层次、注释的引用。

（2）进行知识单元智能加工，系统将对原来以自然段为基础单元的数据进行智能拆解，形成具有独立完整含义的知识单元；同时在这个步骤中，系统还将建立起知识单元和索引关键词之间的关联关系。

（3）索引智能加工，和上一个步骤知识单元智能加工在实际操作中是并行进行的；索引智能加工需要对来自知识单元智能加工中提取的关键词进行索引化处理，再将经索引化处理过的结果进行二次加工，穷尽索引结构的背景信息进行标引，重新排序和聚类，形成高度灵活、准确、多维指向、互有交叉的智能化索引。

（4）智能化索引反作用于知识单元过程，可根据使用者的随机需求形成新的分类、排序与聚类，生成二次、三次或多次文献、表单、图像、音频、视频。

10．根据权利要求 8 所述的信息智能检索加工的方法，其特征在于：通过对异体字或图像字进行拆分、编排、编号，支持标准 Unicode 字库中包含或未包含的大量偏稀汉字，实现对异体字或图像字的笔顺、字根的调用、查询和显示。

11 ．根据权利要求 9 所述的数据智能加工处理方法，其特征在于：通过对异体字或图像字进行拆分、编排、编号，支持标准 Unicode 字库中包含或未包含的大量偏稀汉字，实现对异体字或图像字的笔顺、字根的调用、查询和显示。

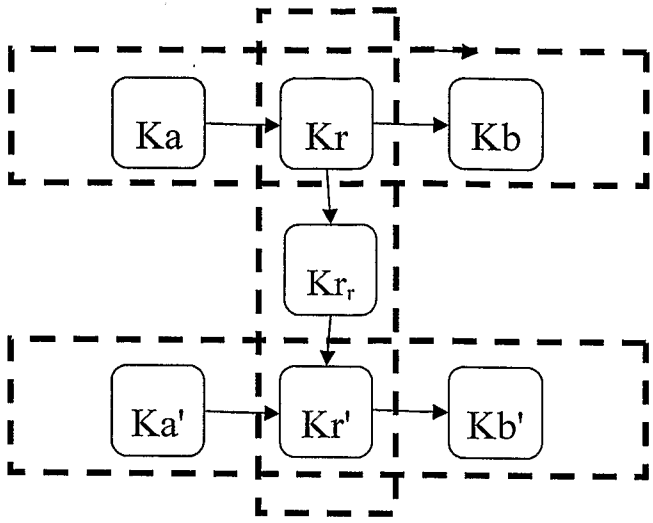

Ka
Kr
Kb
Kr_r
Ka'
Kr'
Kb'

图 1

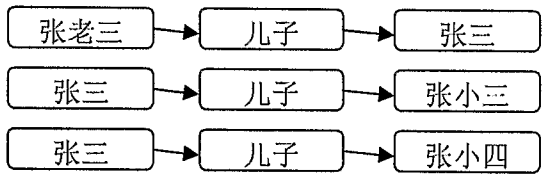

张老三
儿子
张三
张三
儿子
张小三
张三
儿子
张小四

图 2

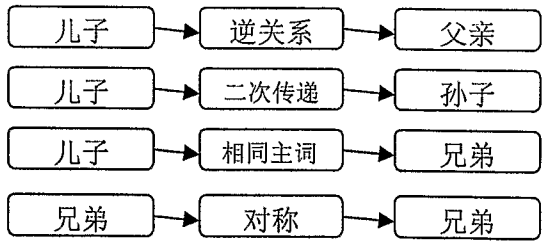

儿子
逆关系
父亲
儿子
二次传递
孙子
儿子
相同主词
兄弟
兄弟
对称
兄弟

图 3

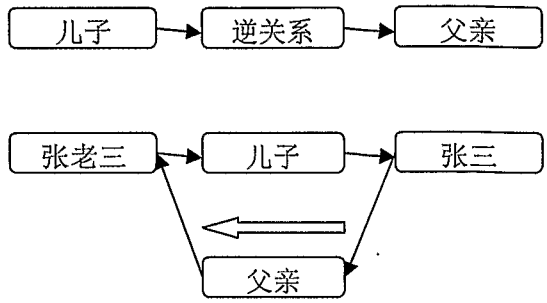

儿子
逆关系
父亲
张老三
儿子
张三
父亲

图 4

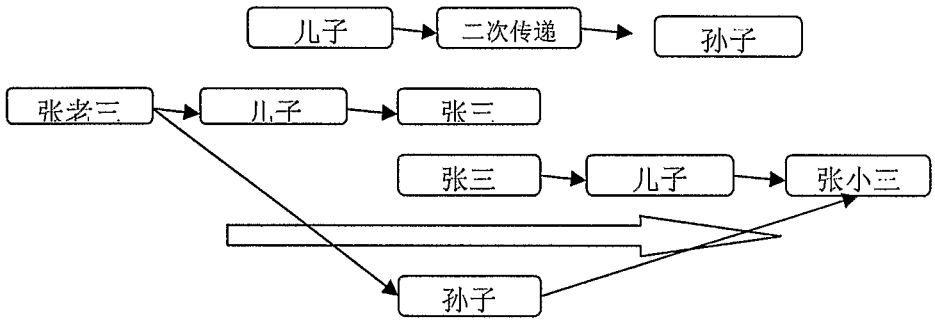

儿子
二次传递
孙子
张老三
儿子
张三
张三
儿子
张小三
孙子

图 5

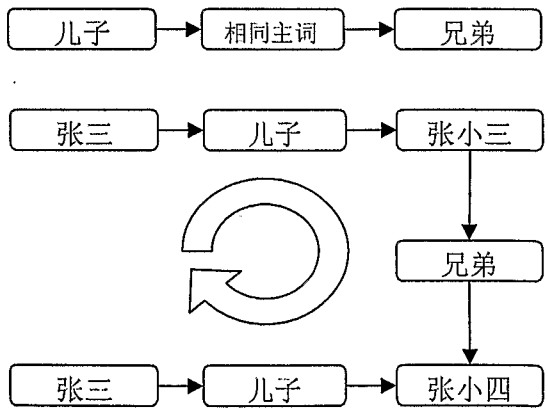

儿子
相同主词
兄弟
张三
儿子
张小三
兄弟
张三
儿子
张小四

图 6

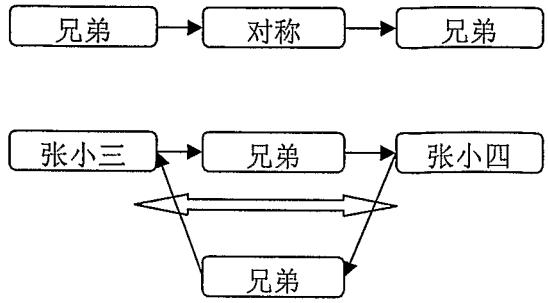


图 7

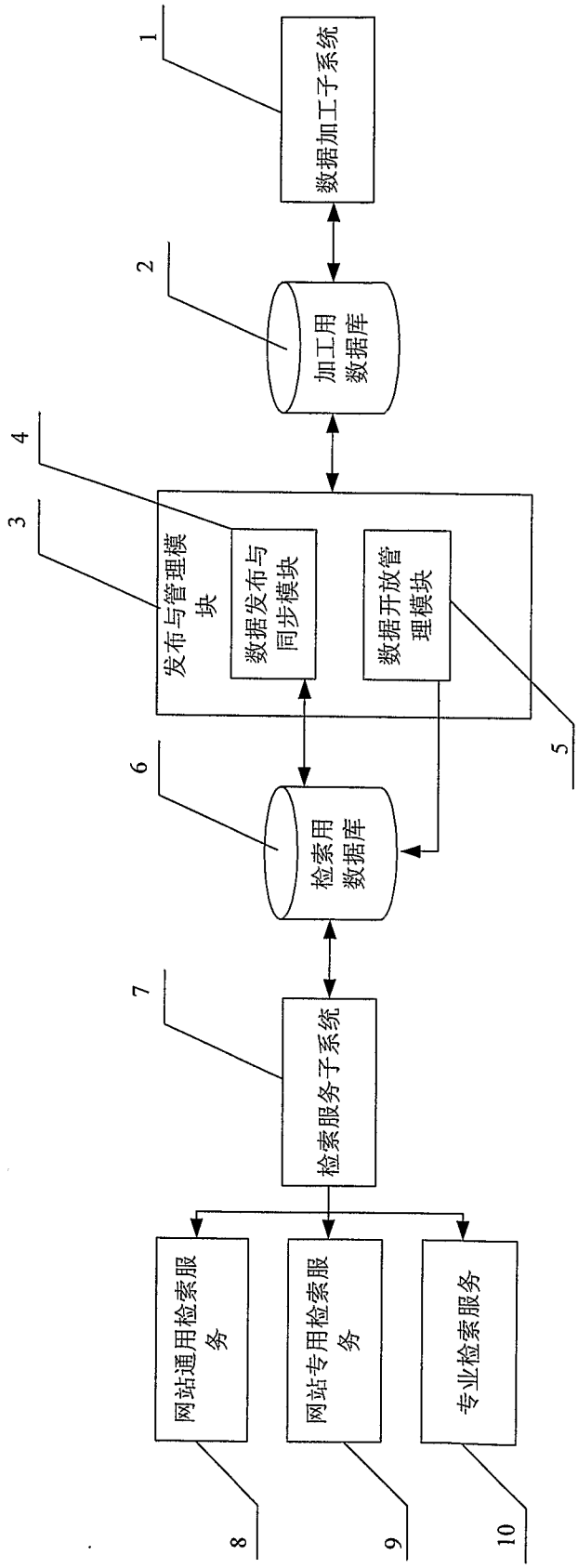

1
数据加工子系统
2
加工用数据库
4
3
发布与管理模块
数据发布与同步模块
数据开放管理模块
5
6
检索用数据库
7
检索服务子系统
网站通用检索服务
网站专用检索服务
专业检索服务
8
9
10

图 8

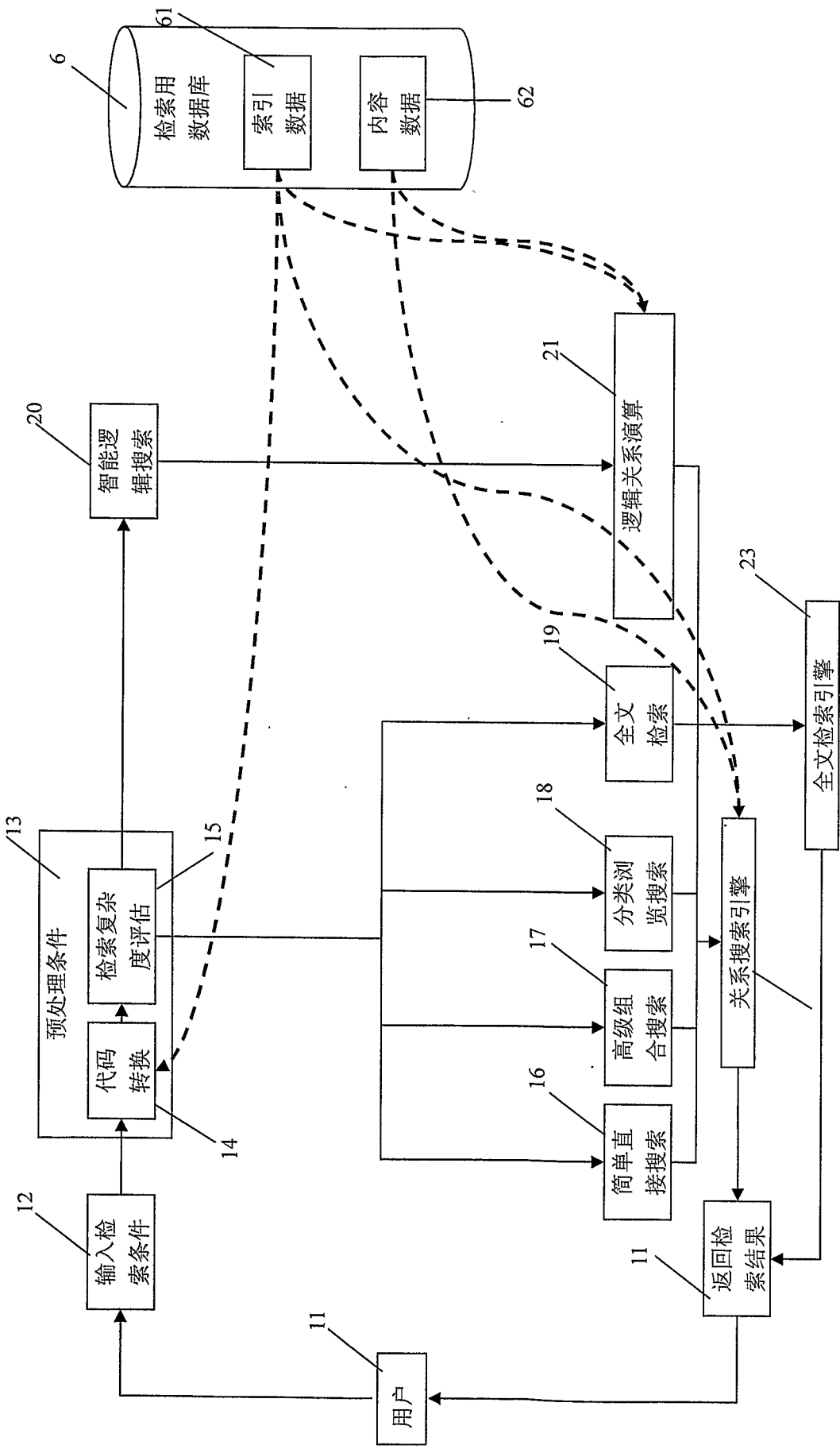

6
检索用数据库
61
索引数据
内容数据
62
20
智能逻辑搜索
21
逻辑关系演算
23
全文检索引擎
19
全文检索
18
分类浏览搜索
17
高级组合搜索
16
简单直接搜索
关系搜索引擎
22
13
预处理条件
检索复杂度评估
15
代码转换
14
12
输入检索条件
11
返回检索结果
11
用户

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2007/001662

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/30(2006.01)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)

IPC: G06F 17/30

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)

CNPAT, CNKI, WPI, PAJ, EPODOC: search, retrieval, database, index, issue, process, treat, serve, intelligent

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP2004206629A (HITACHI LTD) 22 Jul.2004(22.07.2004) the whole document	1-11
A	WO9905614A1(DATOPS SA) 04 Feb.1999(04.02.1999) the whole document	1-11
A	CN1713176A(HUAXIA SCI & TECHNOLOGY CO LTD)28 Dec.2005(28.12.2005) the whole document	1-11
A	CN1410915A (HUIXUN INFORMATION TECHNOLOGY CO LTD BEI) 16 Apr.2003(16.04.2003) the whole document	1-11

☐ 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	Date of mailing of the international search report
22 Agu.2007(22.08.2007)	**06 Sep. 2007 (06.09.2007)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer WANG,Ning Telephone No. (86-10)62084970

Form PCT/ISA/210 (second sheet) (April 2007)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2007/001662

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
JP2004206629A	22.07.2004	NONE	
WO9905614A1	04.02.1999	NONE	
CN1713176A	28.12.2005	NONE	
CN1410915A	16.04.2003	NONE	

国际检索报告

国际申请号
PCT/CN2007/001662

A. 主题的分类

G06F17/30(2006.01)i

按照国际专利分类表(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)

IPC: G06F 17/30

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）)

CNPAT, CNKI, WPI, PAJ, EPODOC: 检索、搜索、查询、数据库、加工、标引、同步、search, retrieval, database, index, issue, process, treat, serve, intelligent

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
A	JP2004206629A（HITACHI LTD）22.7 月 2004（22.07.2004） 全文	1-11
A	WO9905614A1（DATOPS SA）04.2 月 1999（04.02.1999）全文	1-11
A	CN1713176A(华夏科技股份有限公司)28.12 月 2005（28.12.2005）全文	1-11
A	CN1410915A（北京慧讯信息技术有限公司）16.4 月 2003（16.04.2003）全文	1-11

☐ 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

"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	国际检索报告邮寄日期
22.8 月 2007（22.08.2007）	**06.9 月 2007 (06.09.2007)**
中华人民共和国国家知识产权局(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 汪宁 电话号码：(86-10) **62084970**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2007/001662

检索报告中引用的专利文件	公布日期	同族专利	公布日期
JP2004206629A	22.07.2004	无	
WO9905614A1	04.02.1999	无	
CN1713176A	28.12.2005	无	
CN1410915A	16.04.2003	无	