成果发布表单下载模板

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **科技成果项目** | | | | | | |
| 成果标题\* | 一种高效倒排索引组织方法 | | | | | |
| 行业领域\* | 人工智能——自然语言处理 | | | | | |
| 技术领域\* | 绿色化工技术□ **电子信息技术**■ 航空航天技术□ 先进制造技术□ 生物、医药和医疗器械技术□ 新材料及其应用□ 新能源与高效节能□ 环境保护和资源综合利用技术□ 核应用技术□ 农业技术□ 现代交通□ 城市建设和社会发展□ 现代纺织□ 其他□ | | | | | |
| 成熟度\* | 报告级□ 方案级□ 功能级□ 仿真级别□ 初样级□ 正样级□ 环境级□ 产品级□ 系统级□ **销售级■** | | | | | |
| 合作方式\* | 技术转让□ 技术许可□ 作价入股□ 合作开发□ 技术咨询□ **技术服务**■ 创业融资□ 股权融资□ 委托开发□ | | | | | |
| 成果类型（多选）\* | **发明专利■** 实用新型专利□ 软件著作权□ 著作权□ 商标权□ 新品种□ 外观设计□ 新技术□ | | | | | |
| 交易金额\* | —— 万元 双方协商■ | | | | | |
| 成果介绍\*（500-1000字） | | | | | | |
| **成果类型**：发明专利，专利号ZL 2012 1 0511478.1，专利授权公告日：2017年2月8日  **技术领域：**涉及全文检索系统中使用的高效倒排索引组织方法。  **背景技术：**  1、大数据时代的网络环境中，信息量和用户量呈爆炸式增长，给大规模信息检索系统准确高效的服务带来了压力和挑战。倒排索引是信息检索的核心，其存储结构及组织方式对信息检索性能的提升有很大影响。在改进检索算法的同时，人们也在努力优化倒排索引结构及其组织方式。  2、目前国内外针对倒排索引优化的研究，主要从以下三方面展开：(1)通过压缩技术减少索引在外存上的体积；(2)对倒排表内容的组织方式进行优化，减少需要访问的倒排表内容；(3)对倒排表的磁盘存储结构进行管理，尽量减少磁盘的IO次数。但现有的成果包括压缩技术、倒排索引组织方式及存储结构还不成熟，很难适应大数据时代信息检索系统信息量大、快速响应精而准的需求，因此本发明提出了一种融合压缩技术的高效的倒排索引结构及组织方法，以解决当前倒排索引优化问题，实现海量数据的有效管理。  **发明内容：**  发明提供了一种高效的倒排索引结构，该结构包括：  S101倒排索引由倒排项组成，倒排项包括记录倒排项和属性倒排项两种类型，它们分开存储以减少信息冗余，同时在逻辑上倒排索引也就被分为记录倒排索引和属性倒排索引；  .....  发明还提供一种高效的倒排索引组织方法，该方法包括：  201倒排索引在文件中采用分块的方式进行组织存储，倒排块综合采用链表和集中定位表两种方式进行管理，使得倒排索引既便于动态维护又有利于SKIP技术下的随机访问；为提高数据加载效率，同时有利于数据随机访问，以数据库索引的优化作为临界点来选择倒排块的管理方式：优化后的倒排索引采用集中定位表进行管理，而优化后新加载的数据所对应的倒排索引仍然用链表管理，直到下一次数据库索引的优化；  .....  发明技术方案提供的高效倒排索引结构及组织方法，将倒排索引中倒排项分为记录倒排项和属性倒排项进行存储，以减少信息冗余；倒排索引以分块的方式组织存储，综合链表和集中定位表的优势，对倒排块进行管理，使得倒排索引既便于动态维护又有利于SKIP技术下的随机访问，同时在数据存储时采用了一定的压缩技术，以实现在尽量少的空间存储尽量多的信息；减少了数据存储空间，提高了数据索引效率。 | | | | | | |
| 成果亮点\*（500-1000字） | | | | | | |
| 1. **减少信息冗余**：本发明公开了一种高效倒排索引组织方法。将记录倒排项和属性倒排项分开存储，减少信息冗余； 2. **可实现数据的高效访问**：采用分块方式对倒排索引进行存储，以数据库索引优化为切入点，综合利用链表和集中定位表的优势对倒排块进行管理，既有利于索引动态维护又便于SKIP技术下的随即访问； 3. **信息储存效率大大提高**：根据域间关联关系，采用了压缩技术，以达到尽量少的空间存储尽量多的信息量。 4. **实现海量数据有效管理**：所提供的倒排索引结构及组织方法，减少了存储空间、优化了访问时间，在一定程度解决了倒排索引优化问题，能够实现海量数据的有效管理。 | | | | | | |
| 应用前景\*（500-1000字） | | | | | | |
| **应用“一种高效倒排索引组织方法”发明的“TRS海贝搜索引擎数据库”可实现全文检索数据库国产替代**  当今的全文检索数据库市场，基本上处于开源软件ElasticSearch一家独大的局面。Elasticsearch是一款基于Lucene的搜索服务器，可以提供分布式多用户能力的全文搜索引擎。ElasticSearch自发布以来，凭借高效的性能，开放的架构，以及开源免费的商业模式迅速占领了全文检索的市场。目前国内外很多企业已放弃自主研发全文检索数据库，逐渐转投ElasticSearch阵营，而还在坚持全文检索数据库自主研发并保持技术领先的企业凤毛麟角。  随着ElasticSearch的市场占有率越来越高，暴露出的安全问题也逐渐增多。2019年底发生的用户隐私泄漏事件，致使超过27亿个电邮地址（有10亿个密码是简单的明文）被盗，其中不乏我国的互联网大牌供应商；2020年初，又爆出黑客入侵超过1.5W台ElasticSearch服务器事件，以上问题对用户信息安全造成了极大的威胁。  数字时代，数据检索在数据管理中的重要性大大提高，没有高效的检索，大数据几乎无法可用；没有准确的检索，大数据的分析挖掘手段也难以施展；特别是互联网搜索引擎普及应用以后，在一个基于大数据的应用系统里，没有便捷的搜索引擎更是无法想象的。从原理上讲，作为一款大数据管理系统，数据存储并不是目的，存储数据的目的是通过数据分析来挖掘数据价值。  不可否认，ElasticSearch是一款非常优秀的搜索引擎，它拥有开放的架构和灵活的接口，但因其主要侧重大众市场，所以产品功能在某些方面存在一定的局限。其一，在数据规模要求较高的应用中，会随着集群规模的不断变大而集群效率大大降低；其二，由于定位于通用检索，因此在特定领域缺乏专业功能，如中文检索中常用到的段、句、位检索等；其三，缺乏在企业级应用中常用的多租户、读写分离、冷热分区等功能。  TRS海贝搜索引擎数据库就是一款从内核到系统完全国产自研的全文检索数据库，其定位为大数据应用支撑软件，为大数据应用提供高效的数据存储、全文检索、分类统计等数据管理服务，支持几乎所有类型的数据，包括文本、数字、地理空间、图片等各种结构化、半结构化和非结构化数据。  目前，我国已有很多企事业单位启动了全文检索数据库的国产替代工作。这里就以提供全国企业（亿级）信用信息、信用代码、以及相关站内文章公开查询服务的“信用中国“网站为例，2018年底“信用中国”的搜索服务就由ElasticSearch迁移到TRS海贝大数据管理平台，当下“信用中国”已实现单集群日检索量突破10亿次，单台服务器的每秒并发检索突破1K。 | | | | | | |
| 团队介绍\*（500-1000字） | | | | | | |
|  | | | | | | |
| 产生的效益\*（500-1000字） | | | | | | |
| 拓尔思信创方面拳头产品——“TRS海贝搜索引擎数据库”便应用了“一种高效倒排索引组织方法”发明，使得该产品成为一款“高性能搜索引擎”，不仅高效、安全，还集成了众多开源检索系统不具备的企业级功能，能够有效降低用户的系统建设成本、开发成本和运维成本。  应用该项发明的“TRS海贝搜索引擎数据库”已经广泛应用于安全大数据、媒体大数据、政府大数据等众多细分行业，公安部、新华社、市场监督管理局、海关总署、专利局、商标局等一系列重量级客户都选择了海贝产品；国家重大基础数据库，如企业信用信息公示系统、专利检索系统等都构建在TRS海贝搜索引擎数据库之上。  **其典型行业应用如下：**  **某网安项目：数据量巨大、数据类型丰富，单集群每日新增数据超过60亿**  该项目采用海贝搜索引擎数据库用于网络数据安全管理，特点主要为集群规模大，单集群从数十个节点到数百个节点不等；数据量巨大，单集群每日新增数据超过60亿，每天处理超过20T数据，总记录数超过万亿，存储总数据量PB级别；项目分布范围广，已经在全国多个省市展开。项目经过一段时间的运行，在多个同类产品对比中，用户对海贝搜索引擎数据库赞不绝口。  **某项目：超大规模集群，大大降低资源消耗**  在某项目系统建设中，单集群的服务器数量达到了600台。众所周知，在分布式系统里，单集群机器的数量越多，系统在机器之间协调通讯的工作也就会成倍增长。我们通过架构的调整和系统的优化，大大降低了这部分的资源消耗，让超大规模集群成为了可能。  **信用中国项目：并发访问量大，日检索量突破10亿**  “信用中国”网站由国家发展改革委、人民银行指导，国家信息中心主办，是政府褒扬诚信、惩戒失信的总窗口。网站提供全国企业（一亿多家）信用信息、信用代码、以及相关站内文章的公开查询服务。2018年底“信用中国”的搜索服务迁移到TRS海贝大数据管理平台。在信用中国项目中，海贝单集群日检索量突破10亿。其中，单台服务器的每秒并发检索突破1K。 | | | | | | |
| 转化方式\*（500-1000字） | | | | | | |
| 全文检索数据库是以倒排索引为核心构建的新一代数据库。  而“TRS海贝搜索引擎数据库”正运用了“一种高效倒排索引组织方法”发明 | | | | | | |
| 成果资料\* |  | | | | | |
| 成果视频 |  | | | | | |
| 联系人\* | 周建辉 | | 联系电话\* | | 18201118114 | |
| 单位名称\* | 拓尔思信息技术股份有限公司 | | | | | |
| 所在地区\* | 北京市海淀区 | | | | | |
| 详细地址\* | 建枫路(南延)6号院3号楼 | | | | | |
| 拟转化落地试点城市（园区） |  | | | | | |
| 成果商业计划书 | 可选择相关附件，支持格式：pdf、ppt。 | | | | | |
| 是否为成果代理人 | 是□ 否□ | | | | | |
| 是否已有技术评定 | 是□ 否□ | | | | | |
| 是否参与路演 | 是□ 否□ | | | | | |
| 为此成果贡献服务的相关助力方 | 试点城市（园区） | 科技服务团 | | 企业技术问题征集活动 | | 技术经理人 |
|  |  | |  | |  |
| 是否提交产业化落地方案\* | 是□ 否■ | | | | | |

\*为必填项

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **科技成果综合评价报告** | | | | |
| 成果名称\* |  | | | |
| 所属单位\* |  | | | |
| 成果简介\* |  | | | |
| 创新水平\* | 关键共性技术○ 前沿引领技术○ 现代工程技术○ 颠覆性技术○ 其他○ | | | |
| 技术进度\* | 新设备或新装置 | 原理样机○ 工程样机○ 中试原型机○ 产业化○ | | |
| 新材料或新技术 | 实验室阶段○ 工程化阶段○ 产业化阶段○ | | |
| 技术成果 | 国际专利□ 国家专利□（多选） | | 专利号 |  |
| 国际奖项□ 国家奖项□（多选） | | 奖项名称 |  |
| 产品方向\* | 有多个应用方向○ 有一个应用方向○ 没有应用方向○ 无法判断○ | | | |
| 市场空间\* | 需求前景巨大○ 需求前景较大○ 需求前景一般○ 无法判断○ | | | |
| 成本竞争\* | 优势明显○ 优势一般○ 没有优势○ 无法判断○ | | | |
| 政策影响\* | 政策鼓励○ 政策限制○ 政策淘汰○ 无法判断○ | | | |
| 市场周期\* | 进入期○ 成长期○ 饱和期○ 衰退期○ 无法判断○ | | | |
| 转化周期\* | 近期可控(1年内)○ 周期较长(2年内)○ 很难转化(3年起)○ 无法判断○ | | | |
| 科技成果的创新基因评价（不少于150字）\* | | | | |
| 主要从技术研发团队和技术创新水平两个方面进行评价：可从技术带头人的资历背景、领域地位、研发业绩和领军经验方面，以及研发团队规模、团队背景、团队经验、团队分工明确程度和团队配合状态等方面评价。再从科技成果创新水平、创新进度、创新优势等方面客观评价。 | | | | |
| 科技成果的技术亮点评价（不少于150字）\* | | | | |
| 主要从科技成果的技术方案、技术优势、产品化、产业化建议等方面进行评价。 | | | | |
| 科技成果的应用市场评价（不少于150字）\* | | | | |
| 主要从应用前景、推广优势、转化目标市场、资本化建议等方面进行评价。 | | | | |
| 评价专家组综合意见\* | | | | |
| 主要从科技成果创新水平，市场前景，产业化路径、投资风险与回报等角度对成果转化提出综合性评价意见和建议。 | | | | |
| 评价专家姓名\* |  | | | |
| 评价专家职务\* |  | | | |
| 评价专家所在单位\* |  | | | |

\*为必填项