

(12) 发明专利申请

(10) 申请公布号 CN 101976246 A

(43) 申请公布日 2011.02.16

(21) 申请号 201010501267.0

(22) 申请日 2010.09.30

(71) 申请人 互动在线(北京)科技有限公司
地址 100088 北京市海淀区知春路106号中
关村皇冠假日酒店写字楼8层

(72) 发明人 潘海东 梅春 陈岩

(74) 专利代理机构 北京正理专利代理有限公司
11257

代理人 张雪梅

(51) Int. Cl.

G06F 17/30(2006.01)

权利要求书 1 页 说明书 7 页 附图 2 页

(54) 发明名称

百科词条分类检索方法

(57) 摘要

一种百科词条分类检索方法,包括以下步骤:

(1) 按照百科词条分类将分类百科词条的多个关键特征设定为多个标准化字段,对所述标准化字段建立特征模板;(2) 用户利用特征模板,用词条中的字段信息编辑相应的标准化字段;(3) 数据处理步骤,用于将用户输入的相关字段信息转化为标准格式并生成包括各标准化字段信息的数据集;(4) 数据关联步骤,将各词条中对应于同一标准化字段的字段信息产生相互关联;(5) 词条检索步骤,将一个或数个与标准化字段相应的字段信息作为检索条件进行检索。本发明的百科词条的检索方法,技术投入成本小,用户操作便捷、自由,且搜索结果准确。



1. 一种百科词条的分类检索方法,其特征在于包括以下步骤:
 - (1) 按照百科词条分类将分类百科词条的多个关键特征设定为多个标准化字段,对所述标准化字段建立特征模板;
 - (2) 用户利用特征模板,用词条内容中的字段信息填充相应的标准化字段。
2. 如权利要求 1 所述的百科词条的分类检索方法,其特征在于所述对标准化字段建立特征模板的步骤包括:
 - (1) 设定标准化字段,该标准化字段至少包括:字段名称、字段描述、字段类型、字段长度、字段说明、该字段是否必填或是否显示;
 - (2) 编辑配置所述标准化字段,定义所述的字段名称、字段描述、字段类型、字段长度、字段说明、该字段是否必填或是否显示。
3. 如权利要求 2 所述的百科词条分类检索方法,其特征在于:该字段类型为文本、数字、选择或图片。
4. 如权利要求 2 所述的百科词条分类检索方法,其特征在于:该特征模板选用 XML 语言编写,并对每个标准化字段配置 XSL 文件。
5. 如权利要求 1 所述的百科词条分类检索方法,其特征在于该方法还包括数据处理步骤,用于将用户输入的相关字段信息转化为标准格式并生成包括各字段信息的标准化数据集。
6. 如权利要求 5 所述的百科词条分类检索方法,其特征在于:该方法还包括数据关联步骤,在生成的标准化的数据集中,针对不同百科词条的字段信息,在百科词条间产生相互关联。
7. 如权利要求 6 所述的百科词条分类检索方法,其特征在于:该方法还包括词条检索步骤,将一个或数个与标准化字段相应的字段信息作为检索条件利用字段信息间的相互关联进行检索。
8. 如权利要求 7 所述的百科词条分类检索方法,其特征在于:将检索步骤中得到的检索结果作关联的可视化显示。
9. 如权利要求 7 所述的百科词条分类检索方法,其特征在于:在数据关联步骤中,当用户检索某个特定百科词条时,通过检索结果中显示的字段进一步检索与该字段相关的内容。

百科词条分类检索方法

技术领域

[0001] 本发明涉及一种词条的分类检索方法,尤指一种百科词条的分类检索方法。

背景技术

[0002] 随着信息技术的不断发展,在海量的信息中如何迅速、准确、便捷的搜索到有用的信息已经成为亟待解决的问题,由此,不同的信息检索方法应运而生,目前市场上对于百科词条的分类与检索的方法主要有三种:

[0003] 1、关键字模糊搜索

[0004] 通过关键字模糊搜索,用户需找到适合的关键字进行搜索,系统按百科词条内容与该关键字的匹配程度计算并输出结果。该方法在实际使用过程中,因为用户对于百科词条中关键字的理解与系统不同,输出结果很可能出现偏差。在市场上常见的搜索引擎中,用户往往需要输入多个关键字或使用一些语法约定结果的范围,这在一定程度上提高了用户使用的成本。另一方面,搜索引擎技术门槛较高,互联网公司在搜索引擎技术上投入的成本包括大量技术人员、高级服务器、大量内容监管人员等,故此方法普适性不强。

[0005] 2、树状分类查询

[0006] 分类查询是指将百科词条内容置于预先设定好的分类中,分类之下可包含子分类,用户只需为百科词条选定分类即可。浏览分类下的百科词条,用户需要搜索或手动指定一个分类,此时系统会按照一定规则显示该分类下的全部百科词条。该方法操作简单,技术门槛低,但存在明显使用缺陷。首先,分类和百科词条为单向所属的层级关系,用户想要浏览某一百科词条就必须知道它的所属分类,如在《大英百科全书》中使用分类查询词条“Beijing”,需要依循“Geography > Aisa > East Aisa > China > Northeast ofChina > Beijing”的分类路径进行检索,对于不熟悉该分类体系的用户来说极为不便;其次,根据分类规则的不同,同样的百科词条可以同时属于不同分类,如词条“北京”即可属于分类“中国城市”,同时也属于分类“1000 万人口城市”,树状分类结构无法解决这种“一词多类”的现象;最后,对于网站

[0007] 而言,为了确保百科词条分类依循设定好的规则,势必限制用户的参与,打击用户积极性。

[0008] 3、标签归类

[0009] 标签归类是目前最为常用的一种内容分类形式,被广泛应用于博客等领域。标签归类的特点是灵活,用户可以为百科词条随意填写标签,具有相同标签的百科词条就被归为一类。而该方法的弊端是用户为百科词条添加标签的目的不同造成了标签的偏差。仍旧以词条“北京”为例,用户 A 从客观的角度为北京填写标签“中国”,用户 B 使用标签描述自身偏好,标注为“喜欢的城市”,用户 C 将标签用作行事历,添加标签“旅游目的地”。可见,不同用户对于标签的理解和使用不同,无法对同一内容进行近似的标注。

[0010] 上述三种百科词条的分类与检索的方法各有利弊,市场上还不曾出现一种技术投入成本既小,用户操作又便捷、自由,且搜索结果准确的检索方法。

[0011] 因此,需要一种既能使用户按照设定对百科词条分类,又能使用户自由设置检索词,让互联网公司掌握词条整体的信息架构的百科词条的分类和检索的方法。

发明内容

[0012] 本发明的目的在于,提供一种百科词条检索方法,由互联网提供者抽象出分类百科词条的多个关键特征,将该多个关键特征设定为多个标准化字段,构建百科词条信息的整体架构,且用户能够按照标准化字段的引导用词条的字段信息填充标准化字段,从而提高分类查询的灵活性及准确性,并使得用户操作更为便捷和自由。

[0013] 本发明的另一目的在于,提供一种百科词条检索方法,使得技术投入成本较小,降低维护成本。

[0014] 本发明的又一目的在于,提供一种百科词条检索方法,能够将关联的检索内容呈现给客户,满足不同客户的检索需求,提高网站的访问量。

[0015] 为达到上述目的,本发明采用了如下的技术手段:

[0016] 一种百科词条的分类检索方法,其特征在于包括以下步骤:

[0017] (1) 按照百科词条分类将分类百科词条的多个关键特征设定为多个标准化字段,对所述标准化字段建立特征模板;

[0018] (2) 用户利用特征模板,用词条内容中的字段信息填充相应的标准化字段。

[0019] 优选地,所述对标准化字段建立特征模板的步骤包括:

[0020] (3) 设定标准化字段,该标准化字段至少包括:字段名称、字段描述、字段类型、字段长度、字段说明、该字段是否必填或是否显示;

[0021] (4) 编辑配置所述标准化字段,定义所述的字段名称、字段描述、字段类型、字段长度、字段说明、该字段是否必填或是否显示。

[0022] 优选地,所述字段类型为文本、数字、选择或图片。

[0023] 优选地,所述特征模板选用 XML 语言编写,并对每个标准化字段配置 XSL 文件。

[0024] 优选地,该方法还包括数据处理步骤,用于将用户输入的相关字段信息转化为标准格式并生成包括各字段信息的二维数据集。

[0025] 优选地,该方法还包括数据关联步骤,在生成的标准化的数据集中,针对不同百科词条的字段信息,在百科词条间产生相互关联。

[0026] 优选地,该方法还包括词条检索步骤,将一个或数个与标准化字段相应的字段信息作为检索条件利用字段信息间的相互关联进行检索。

[0027] 优选地,将检索步骤中得到的检索结果作关联的可视化显示。

[0028] 优选地,在数据关联步骤中,当用户检索某个特定百科词条时,通过检索结果中显示的字段进一步检索与该字段信息相关的内容。

[0029] 本发明的有益效果在于:

[0030] 1、本百科词条分类检索方法区别于常见的树状分类和标签分类方法,本方法抽象出一类百科词条中的相同的关键特征,既保留了用户对词条内容填写的自由性,又让互联网提供者掌握整体内容的信息架构,不但提高了分类查询的灵活性及准确性,而且使得用户操作更为便捷和自由。

[0031] 2、使用本百科词条分类检索方法进行信息字段的编辑,可将出错率从 45%降低到

20%，站方后期维护的成本从每千词条 25 工时每人降低到每千词条 3 工时每人。

[0032] 3、使用根据本发明的百科词条分类检索方法能够显示关联性的内容，方便用户检索，且能够进行可视化显示，使网站人均浏览量提高 35%。

附图说明

[0033] 图 1 为本发明实施例的百科词条分类检索方法整体步骤流程图。

[0034] 图 2 为图 1 所示实施例中的步骤 S101 的子步骤流程图。

[0035] 图 3 为本发明一实施例搜索结果数据分布图。

具体实施方式

[0036] 百科词条的特点是信息量大，关联性强、分类繁多。互联网的百科产品需要遵循用户习惯和需求。在互联网上，用户接受信息的渠道众多，成本几乎为零，当用户一旦判断当前内容无法满足需求就会马上离开。另一方面，对于百科词条这类 UCG（用户创建内容）产品，用户个人对内容的编辑极易破坏整个产品的内容架构，为他人浏览造成障碍。

[0037] 本发明引导用户设置符合根据本发明的分类检索方法的相关性信息，并使用可视化的信息关联方法，为用户呈现关联内容。下面将结合附图对本发明的具体实施例做详细说明。

[0038] 参照图 1 所示，是本发明的百科词条分类检索方法整体步骤流程图，具体包括如下步骤：

[0039] 步骤 S101：按照百科词条分类，将分类百科词条的多个关键特征设定为多个标准化字段，对所述标准化字段建立特征模板。

[0040] 研究表明，用户对百科词条 80% 的关注度集中在 20% 的内容分类上，如：仅人物类百科词条，约占互动百科词条浏览量的 25%（2009 年数据）。相同类别的百科词条具有相同的关键特征，提取这些关键特征是本发明的第一步。如人物类别可提取特征为：中文名、英文名、职业、性别、国籍、籍贯、出生年月、去世年月等，针对该人物类百科词条的上述不同特征分别进行提取，生成为标准化的字段，以便将这些标准化的字段用作检索的标准化数据。如图 2 所示，上述步骤 S101 中，又可具体分为三个子步骤：步骤 S1011、步骤 S1012、步骤 S1013。

[0041] 步骤 S1011：设定标准化字段，该标准化字段至少包括：字段名称、字段描述、字段类别、字段长度、字段说明、该字段是否必填或是否显示。

[0042] 特征模板可选用 XML、HTML、JAVA 等语言进行编写，由于 XML 语言的特点是以描述数据为主，数据与结构分离，属性可以自定义、方便灵活、结构简单，考虑到该特征模板需要不懂编程语言的编辑进行维护，因此优选 XML 作为特征模板的编程语言。

[0043] 检索常用的字段类型有：文本、数字、选择、图片等，分别对其进行编程，代码结构如下：

[0044] `<property name = " 模块名称" describe = " 描述" type = " 类别" maxlenth = " 字段长度" description = " 说明" require = " 是否必填" available = " 是否显示" >`

[0045] `<rule> 规则 </rule>`

[0046] </property>

[0047] 字段名称 :该标准化字段的名称,不对普通用户显示,属性唯一 ;

[0048] 描述 :该字段的中文名称,即对外显示的名称 ;

[0049] 类别 :该字段的类型,如文本类型字段此处为“string”;

[0050] 字段长度 :用以限定该字段的最长字符数 ;

[0051] 说明 :用来显示针对该字段的用户提示 ;

[0052] 是否必填、是否显示 :是否为必填信息、是否显示该信息。

[0053] 步骤 S1012 :编辑配置所述标准化字段,定义所述的字段名称、字段描述、字段类别、字段长度、字段说明、该字段是否必填或是否显示。

[0054] 编辑配置特征模板的过程比较简单,仅需要将字段信息填入相应的 XML 文件即可。下面以人物信息中的“中文名”字段为例 :

```
[0055] <property name = " people_name_zh " describe = " 中文名 " type  
= " string " maxlenth = " 50 " description = " 请输入该人物的姓名,以常用名为  
准" require = " false " available = " true " >
```

```
[0056] <rule></rule>
```

```
[0057] </property>
```

[0058] 至此,网站按照上述步骤 S1011-S1012 为各个百科词条分类预先建立、配置、编辑好了特征模板,由此生成了大部分标准化字段。

[0059] 步骤 S1013 :为每个标准化字段配置 XSL 文件。

[0060] XML 语言可以定义信息的内容,却没有定义信息应该如何表达,这实际上就是 XML 的长处,它将信息的内容和形式分离开来。而 XSL (XML StyleLanguage) 语言能够将 XML 文档转换为 XHTML 文档,它具有很强的格式 XML 文档的能力。

[0061] 因此,可以利用 XML 语言存储内容,而利用 XSL 语言显示内容,XML 内容的表达通过 XSL 来实现。为不同模块单独配置 XSL 文件也增强了该产品的可扩展性,使得用户可按照网站给定的标准化字段添加相应的字段信息。

[0062] 回到图 1,步骤 S102 :用户利用特征模板,用词条内容中的字段信息填充相应的标准化字段。

[0063] 越来越多的 Web 界面应用技术出现在网站上,比如我们常见的日历控件,搜索下拉框等,这些 Web 界面应用技术大大的丰富了网页的表现形式,且增强了其功能性。Web 界面应用技术提供简单、直观和即时响应的用户界面,让用户花更少的精力和时间去完成事情。本发明所运用的 Web 界面编辑技术与目前网络上通用的 Web 界面技术相同,不再赘述。

[0064] 当网站为各个百科词条分类建立、配置、编辑好了特征模板,生成了标准化字段后,用户可以登录网站,按照站方设定好的标准化字段补充完善字段信息,或设置检索条件。需要注意的是,网站需要限定字段的填写格式,以降低后期数据处理的成本。如 :日期的表示方法有许多种 :dd/mm/yy、mm/dd/yy、yyyy-mm-dd 等,需要限定其中一种格式作为客户填写的格式。

[0065] 步骤 S103 :数据处理。

[0066] 用于将用户输入的相关字段信息转化为标准格式,并生成包括各标准化字段信息的标准化数据集。

[0067] 数据处理主要集中在用户填写的文本字段类型,当用户在网站前端完成字段信息的填写之后,后台即可进行数据处理,对所填的内容进行整理和索引。

[0068] 为实现高效的数据调取,需要对内容进行前期整理,该过程主要针对用户输入习惯差异造成的数据不规范,下面以日期字段为例:

[0069] 针对日期文本字段,虽然网站在建立、编辑、配置特征模板时已经预先输入提示信息,提示客户按照某种格式填写,但仍有用户会按照自己习惯的方式填写,如输入:2010年12月20日、2010年12月20号、2010-12-20、Dec 20/2010等,系统会将常见的填写格式转换为标准格式进行存储,便于统一结构,最后为每一个百科词条生成一个标准化的数据集。

[0070] 步骤 S104:数据关联。

[0071] 在本步骤中,将各词条中对应于同一标准化字段的字段信息产生相互关联。

[0072] 当完成数据处理步骤后,系统会将该分类百科词条的信息组合为一个二维数据集,用户可根据需要选择一个或多个字段对数个百科词条进行关联。

[0073] 步骤 S105:词条检索。

[0074] 在本步骤中,可将一个或数个与标准化字段相应的字段信息作为检索条件利用字段信息间的相互关联进行检索。

[0075] 通过数据关联步骤及词条检索步骤能够实现:

[0076] (1) 以数个字段作为选项的百科词条过滤器。

[0077] (2) 将一个或数个与标准化字段相应的字段信息作为检索条件进行检索,将各个百科词条进行关联,得到的检索结果作关联的可视化显示。

[0078] (3) 当用户检索某个特定百科词条时,可通过检索结果中显示的字段进一步检索与该字段相关内容。

[0079] 以下特举优选的实施例作更详细的说明:

[0080] 系统数据库中按照步骤 S101、S102 存储了数个国家词条信息:中国、美国、日本、英国、俄罗斯、法国、德国、澳大利亚、巴西、阿根廷,每个百科词条分别设置了数个字段:首都、所在大洲、面积、2008年人口数量、GDP,各个词条的每个字段的详细信息如下:

[0081]

| 国家 | 首都 | 所在大洲 | 面积 (km ²) | 人口 (2008 年) | GDP (兆美元) |
|------|---------|------|-----------------------|---------------|-----------|
| 中国 | 北京 | 亚洲 | 9,640,821 | 1,338,612,968 | 9.712 |
| 美国 | 华盛顿 | 北美洲 | 9,826,675 | 310,338,000 | 14.256 |
| 日本 | 东京 | 亚洲 | 377,944 | 127,420,000 | 4.267 |
| 俄罗斯 | 莫斯科 | 欧洲 | 17,075,400 | 141,927,297 | 2.109 |
| 英国 | 伦敦 | 欧洲 | 243,610 | 62,041,708 | 2.139 |
| 法国 | 巴黎 | 欧洲 | 674,843 | 65,447,374 | 2.108 |
| 德国 | 柏林 | 欧洲 | 357,021 | 81,757,600 | 2.806 |
| 澳大利亚 | 堪培拉 | 大洋洲 | 7,617,930 | 22,473,494 | 0.851 |
| 巴西 | 巴西利亚 | 南美洲 | 8,514,877 | 192,272,890 | 2.013 |
| 阿根廷 | 布宜诺斯艾利斯 | 南美洲 | 2,766,890 | 40,134,425 | 0.609 |

[0082] 客户在 Web 网页上可设置不同的检索条件,并实现如下检索功能:

[0083] (1) 以数个字段作为选项的百科词条过滤器。

[0084] 例如,以“2008 年人口数量”、“所在大洲”两个字段的结合作为检索条件,设置过滤器。当在“2008 年人口数量”处输入“一亿以上”、在“所在大洲”处选择“亚洲”时,即查询 2008 年人口数量在一亿以上的亚洲国家,得到的检索结果为:中国和日本。

[0085] 用户还可以在 Web 网页上设定显示搜索结果的其他相关字段信息,例如:用户还想了解在搜索结果中,各个国家的首都是哪里,网站前端提供输入选项,用户选择“首都”即可,得到的搜索结果显示如下:

[0086] 国家 所在大洲 人口(2008 年) 首都

[0087] 中国 亚洲 1,338,612,968 北京

[0088] 日本 亚洲 127,420,000 东京

[0089] 当然,在百科词条字段数量不多的情况下,也可以设置在搜索结果中,系统默认显示该百科词条的所有的字段信息。

[0090] (2) 用户给定字段内容的可视化显示。

[0091] 为了能够更加直观的显示各个百科词条中某个字段之间的关联,用户可以将搜索结果做可视化的显示,如:以柱形图、条形图、折线图、饼图等方式显示。

[0092] 举例来说,用户搜索上述 10 国的“人口”及“面积”,搜索结果中“人口”选择以柱形图显示、“面积”选择以折线图的方式显示,通过“人口”与“面积”字段的复合查询,得到可视化的人口面积分布图表,如图 3 所示。

[0093] (3) 当用户检索某个特定百科词条时,可通过检索结果中显示的字段进一步检索与该字段相关内容。

[0094] 如检索词条“中国”时,可得到如下结果:

| 中国 | |
|-------|---------------------|
| 所属大洲: | 亚洲 |
| 成立日: | 1949年10月1日 |
| 政体: | 人民代表大会制度 |
| 执政党: | 中国共产党 |
| 面积: | 960万km ² |
| 人口: | 13亿 |
| 常用语言: | 汉语 |
| 官方语言: | 汉语 |
| GDP: | 290000亿元人民币 |
| 基尼系数: | 0.47 |
| 首都: | 北京 |
| 主要城市: | 上海,北京,广州,天津,重庆,成都 |
| 国家代码: | P.R.C |
| 主要宗教: | 佛教,道教等 |

[0095]

[0096] 点击“所属大洲”字段,可以显示与中国同处亚洲的国家列表:中国,日本;

[0097] 点击“面积”字段,可以显示按照面积由大到小排列的国家列表:俄罗斯,美国,中国,巴西……;

[0098] 点击“人口”字段,可以显示按照人口数量由大到小排列的国家列表:中国,美国,巴西,俄罗斯……。

[0099] 上述检索结果也可以通过前述的可视化的方式呈现。

[0100] 在本实施例中,可以利用下拉菜单的选择框、单选框、复选框等形式实现数据的横向/纵向检索,如,点击“面积”字段,同时在选择框中选择“面积由大到小”或“面积由小到大”,使得检索结果得以进一步的扩展,用户获取信息更为全面、方便、快捷。

[0101] 综上所述,本发明的百科词条检索方法不但可以满足不同用户的需求,方便快捷灵活准确的得到检索结果,而且用户对内容的编辑不会破坏整个产品的内容架构,使得站方能够掌控整体内容的信息架构。

[0102] 应当理解,这里描述的实施例是示意性的而非限制性的。本领域技术人员通过阅读说明书,可以对本发明的技术方案有更好的了解,并可以在本发明的精神和宗旨下对本发明的实施例进行各种修改和变型。本发明的保护范围仅由随附权利要求书限定。

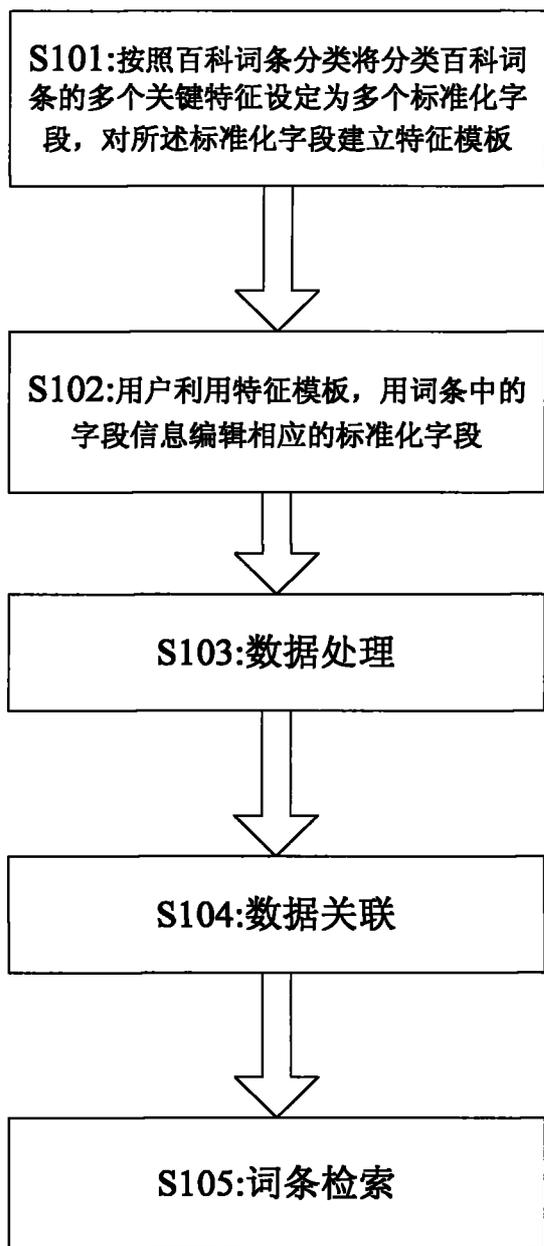


图 1

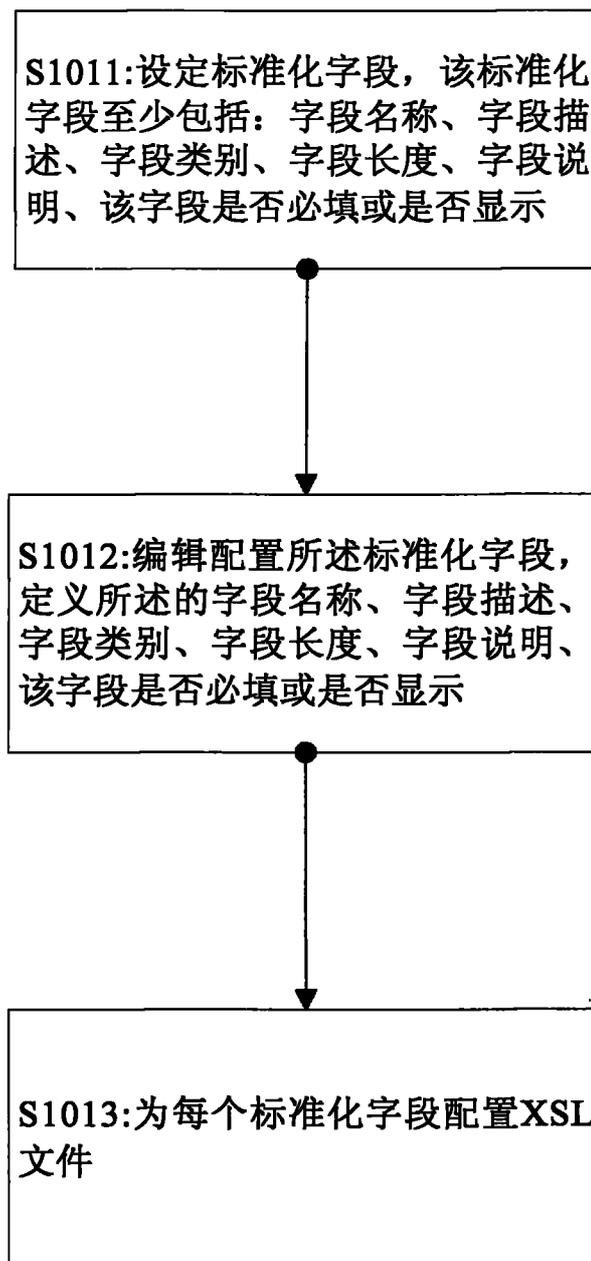


图 2

各国人口与面积分布图

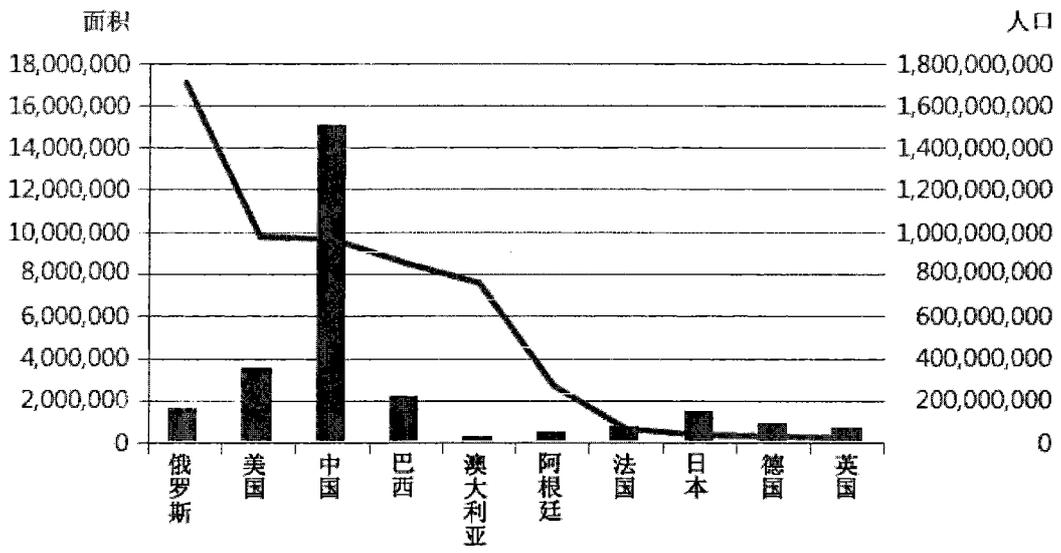


图 3