

(12) 发明专利申请

(10) 申请公布号 CN 102646099 A

(43) 申请公布日 2012. 08. 22

(21) 申请号 201110041757. 1

(22) 申请日 2011. 02. 21

(71) 申请人 株式会社理光

地址 日本东京都

(72) 发明人 姜珊珊 谢宣松 孙军 赵利军

郑继川

(74) 专利代理机构 北京市柳沈律师事务所

11105

代理人 黄小临

(51) Int. Cl.

G06F 17/30 (2006. 01)

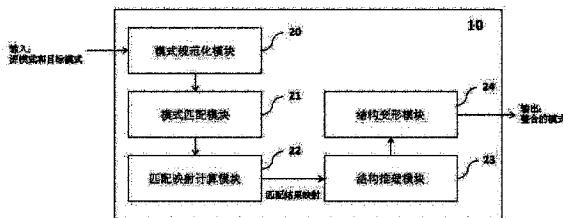
权利要求书 2 页 说明书 18 页 附图 9 页

(54) 发明名称

模式匹配系统、模式映射系统及方法

(57) 摘要

公开了基于混合属性 - 值匹配的模式匹配系统、模式映射系统、模式匹配方法和模式映射方法，用于匹配对象的源模式和目标模式中的对应项，模式代表对象的副本，并由具有层次结构的属性 - 值对组成。其中，对源模式和目标模式中的值进行规范化，以用于源模式和目标模式中的对应项的匹配，所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式，即为所述值添加元信息。通过上述模式匹配和模式映射系统及方法，可以使得源模式和目标模式的对应项的值更加可比较，减小了相似度计算的粒度，从而提高了模式匹配的精度。并且，由于无需引入领域相关的表单、词典以及本体知识，可以降低系统的成本，并便利用户的使用。



1. 一种基于混合属性 - 值匹配的模式匹配系统, 用于匹配对象的源模式和目标模式中的对应项, 模式代表对象的副本, 并由具有层次结构的属性 - 值对组成, 所述模式匹配系统包括 :

模式规范化模块, 对源模式和目标模式中的值进行规范化, 以用于源模式和目标模式中的对应项的匹配, 所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式, 即为所述值添加元信息。

2. 一种基于混合属性 - 值匹配的模式映射系统, 包括 :

模式匹配装置, 用于匹配对象的源模式和目标模式中的对应项以生成匹配结果映射, 模式代表对象的副本, 并由具有层次结构的属性 - 值对组成, 其中所述模式匹配装置对源模式和目标模式中的值进行规范化处理, 以匹配源模式和目标模式中的对应项, 所述规范化处理是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式, 即为所述值添加元信息 ;

模式整合装置, 与模式匹配装置相连接, 用于根据所述模式匹配装置生成的所述匹配结果映射来整合所述源模式和目标模式, 以生成整合的模式。

3. 根据权利要求 2 所述的模式映射系统, 其中, 所述模式匹配装置包括 :

模式规范化模块, 接收对象的源模式和目标模式作为输入, 对源模式和目标模式的属性和值进行规范化处理, 以使得所述属性和值更加可比较 ;

模式匹配模块, 与所述模式规范化模块相连接, 接收已由所述模式规范化模块进行了规范化的属性和值, 并计算源模式和目标模式之间的属性 - 属性匹配相似度、值 - 值匹配相似度和属性 - 值交叉匹配相似度 ;

匹配映射计算模块, 与所述模式匹配模块相连接, 接收由所述模式匹配模块计算出的源模式和目标模式之间的属性 - 属性匹配相似度、值 - 值匹配相似度和属性 - 值交叉匹配相似度, 从而计算所述源模式和目标模式的对应项之间的综合相似度并生成所述匹配结果映射。

4. 根据权利要求 3 所述的模式映射系统, 其中, 所述模式整合装置包括 :

结构推理模块, 与所述匹配映射计算模块相连接, 接收所述匹配映射计算模块所生成的匹配结构映射, 并根据所述匹配结果映射推理实际映射情况 ;

结构变形模块, 与所述结构推理模块相连接, 根据所述接收推理模块输出的所述实际映射情况对所述源模式或所述目标模式进行变形, 以生成所述整合的模式。

5. 根据权利要求 3 所述的模式映射系统, 其中, 所述值的规范化处理包括 :

值为复合的简单短语时, 分离处于并列关系的简短短语以成为简短短语集合的形式 ;

值为值表达式时, 借助于领域无关的度量单位字典来分离值表达式中的数值和度量单位以成为数值 + 度量单位的形式 ;

值为复合的值表达式时, 分离处于并列关系的值表达式, 并借助于领域无关的度量单位字典来分离值表达式中的数值和度量单位以成为数值 + 度量单位集合的形式 ;

值为表格和列表时, 分解表格和列表的项, 以成为简短短语或简短短语集合, 以及数值 + 度量单位或数值 + 度量单位集合的形式 ;

值为解释性段落时, 从解释性段落中抽取关键词语, 以成为简短短语或简短短语集合, 以及数值 + 度量单位或数值 + 度量单位集合的形式。

6. 根据权利要求 5 所述的模式映射系统, 其中, 所述值 - 值匹配相似度计算包括 :

在所述源模式和目标模式的值均为简短短语或简短短语集合时, 对于源模式和目标模式的两个简短短语集合中的每一个简短短语, 使用字符串相似度度量来计算相似度, 并取平均值作为值 - 值匹配相似度 ;

在所述源模式和目标模式的值均为数值 + 度量单位或数值 + 度量单位集合时, 对于源模式和目标模式的两个数值 + 度量单位集合中的每一个数值 + 度量单位, 借助于领域无关的度量单位字典来计算相似度, 并取平均值作为值 - 值匹配相似度 ;

在所述源模式和目标模式的值为简短短语集合和数值 + 度量单位集合的结合时, 对于源模式和目标模式的简短短语集合中的每一简短短语和数值 + 度量单位集合中的每一个数值 + 度量单位, 使用字符串相似度度量来计算相似度, 并取平均值作为值 - 值匹配相似度。

7. 根据权利要求 3 所述的模式映射系统, 其中, 所述源模式和目标模式的对应项之间的综合相似度为 :

$$\text{Score} = \alpha \cdot \text{Score}_{\text{attr}} + \beta \cdot \text{Score}_{\text{val}} + (1 - \alpha - \beta) \cdot \text{Score}_{\text{cross}}$$

其中,  $\text{Score}_{\text{attr}}$  为所述属性 - 属性匹配相似度,  $\text{Score}_{\text{val}}$  为所述值 - 值匹配相似度,  $\text{Score}_{\text{cross}}$  为所述属性 - 值交叉匹配相似度 ;  $\alpha$  和  $\beta$  为权重, 并满足如下关系 :  $0 \leq \beta \leq 1$ ,  $0 \leq \alpha \leq 1$ ,  $0 \leq \alpha + \beta \leq 1$ 。

8. 根据权利要求 3 所述的模式映射系统, 其中, 所述匹配映射结果的生成包括 :

生成所述源模式到所述目标模式的匹配映射 : 对源模式中的每个元素  $i$ , 取  $\text{Score}[i]$  中分数最高的  $\text{Score}[i][j]$ , 目标模式中的元素  $j$  即为元素  $i$  的对应项, 将  $\langle i, j \rangle$  添加到匹配映射中 ;

生成所述目标模式到所述源模式的匹配映射 : 对目标模式中的每个元素  $p$ , 取  $\text{Score}^T[p]$  中分数最高的  $\text{Score}^T[p][q]$ , 其中  $\text{Score}^T[\cdot][\cdot]$  为  $\text{Score}[\cdot][\cdot]$  的转置矩阵, 源模式中的元素  $q$  即为元素  $p$  的对应项, 将  $\langle p, q \rangle$  添加到匹配映射中。

9. 一种基于混合属性 - 值匹配的模式匹配方法, 用于匹配对象的源模式和目标模式中的对应项, 模式代表对象的副本, 并由具有层次结构的属性 - 值对组成, 所述模式匹配方法包括 :

对源模式和目标模式中的值进行规范化, 以用于源模式和目标模式中的对应项的匹配, 所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式, 即为所述值添加元信息。

10. 一种基于混合属性 - 值匹配的模式映射方法, 包括 :

模式匹配步骤, 用于匹配对象的源模式和目标模式中的对应项以生成匹配结果映射, 模式代表对象的副本, 并由具有层次结构的属性 - 值对组成, 其中所述模式匹配步骤对源模式和目标模式中的值进行规范化处理, 以匹配源模式和目标模式中的对应项, 所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式, 即为所述值添加元信息 ;

模式整合步骤, 用于根据所述模式匹配步骤生成的匹配结果映射来整合所述源模式和目标模式, 以生成整合的模式。

## 模式匹配系统、模式映射系统及方法

### 技术领域

[0001] 本发明总的来说涉及与信息处理和信息整合技术,且更具体地,涉及基于混合属性 - 值匹配的模式匹配系统和模式映射系统及其方法。

### 背景技术

[0002] 在信息处理和信息整合技术中,有时需要构建对象数据库,同时匹配不同对象副本中的对应项并整合异构的副本,这里,对象的副本通常被称为模式。

[0003] 在互联网上存在着大量含有对象属性 - 值信息的网页,比如产品的规范说明页面。这些属性 - 值的表格可以通过信息抽取获取,作为自动建立对象数据库的第一步工作。但是异构的数据源网页对产品信息的展示方式也不尽相同,涉及不同的措辞,不同的表格结构,针对特定用户的不完全信息。因此,需要从一个现实世界中的产品对象的多个模式副本识别出其中的对应项,并整合这些异构的副本为一个一致的模式。以上所涉及的具体任务可以被划分为模式匹配和模式整合。

[0004] 对于调和不同数据来源的模式,在 Reconciling schema of disparate datasources :a machine learning approach, Doan AH, 2001. In :Proc ACM SIGMODConf, pp. 509–520 中公开了一种机器学习方法。这种机器学习方法应用于数据集成系统,采用了基于元数据的学习方法。但是,当如上述情况,处理目标是网页中的表格而并非逻辑数据库中的表格或者 XML 文件时,由于所处理的数据缺少元数据和数据格式的约束,因此这种监督学习方法可能导致过度拟合且无法适应跨领域的数据。

[0005] 在 S-Match :an algorithm and an implementation of semantic matching 中公开了一种语义匹配的算法及实现,即, S-Match, 其是一种面向结构的模式匹配方法,通过使用 WordNet 计算词之间的距离,并使用 SAT 求解器推理映射。但是, WordNet 虽然可用于挖掘语义相关性,但是在产品信息的面向实例的模式匹配中,并不适用。这是因为对于例如上述产品规范说明页面中的值表达式和解释性段落来说,很难定义其语义相似度。

[0006] 在 US 2008/0021912 A1, Tools and methods for semi-automatic schemamatching 中,公开了一种半自动化模式匹配的工具和方法,这篇专利采用了多种外部词典,但是这种外部词典无法适应跨领域数据,并且其处理对象为富含元信息的 XML 数据。

[0007] 在网络数据库中模式匹配的方法和系统 (US 7249135 B2, Method and system for schema matching of web database., MICROSOFT CORP) 中,提供了一种方法实施在网络数据库中识别模式之间的匹配,这里的模式是网络数据库中表的模式;并且已知一个全局的模式,匹配主要依赖于模式与全局模式之间的匹配实现。但是,这里公开的方法和系统主要应用于网络数据库中的模式匹配,网络数据库为关系数据库,即输入的数据都是有完整元信息的数据库表格。但是对于数据源网页的表格,并没有元信息的约束,因此虽然实现了属性 - 属性匹配计算和值 - 值匹配计算,但是处理的数据主要为字符串类型,没有为数值数据提供特别的方法,因而在对于数值数据的匹配方面仍存在不足。此外,在上述方法和系统中

使用了全局模式,因此需要先验性的领域或本体知识。

[0008] 在一种从多网页中抽取和规范化产品属性的非监督方法 (AnUnsupervised Framework for Extracting and Normalizing Product Attributes fromMultiple Web Sites) 中,提供一种方法从多网页中同时抽取和规范化产品属性,这里属性的规范化即是发现其中的语义相似性,将产品属性通过某种距离度量聚类,聚类结果为一条属性的可能词表。但是,在上述方法中,产品属性没有区分属性和值,即将例如上述数据源网页的表格中涉及的产品的属性和值看作是一条属性,因此,在进行匹配时必然导致匹配精度降低。此外,上述方法中所采用的距离度量是使用监督的机器学习方法训练所得,即在一个特定领域内,要进行一次距离计算,而在另一个领域内,距离要重新计算,这显然提高了系统应用的成本并造成了用户的不便。

[0009] 因此,可以看到在以上提到的多篇现有技术文件中,大多数仅关注于特定领域,造成领域信息很难收集,需要大量的人力。并且,现有技术中的系统和方法大多数是处理关系数据库中的表格以及结构化的 XML 数据,这些数据富含元信息,如数据类型,取值范围和约束等。而对于非结构化的数据,比如无结构的 XML 数据或者网页中抽取出的表格,则不包含上述元信息。例如,网页中抽取出的表格只有表格结构和文本内容两类信息,因此并不适合于采取上述现有技术中的系统和方法来进行处理。

[0010] 因此,需要一种领域无关的模式匹配和模式映射系统及方法,能够对于对象的非结构化的模式副本进行处理,得到可以接受的结果精度,同时不需要先验性的领域或本体知识。

## 发明内容

[0011] 因此,本发明的目的是解决上述现有技术中的一个或多个问题和缺点。

[0012] 本发明的目的是提供模式匹配系统、模式映射系统、模式匹配方法和模式映射方法,其能够将对象的模式的无结构的纯文本形式的值规范化为有结构的形式,从而为所述值添加元信息以使其更加可比较。

[0013] 为实现上述目的,根据本发明的一方面,提供了一种基于混合属性 - 值匹配的模式匹配系统,用于匹配对象的源模式和目标模式中的对应项,模式代表对象的副本,并由具有层次结构的属性 - 值对组成,所述模式匹配系统包括 :模式规范化模块,对源模式和目标模式中的值进行规范化,以用于源模式和目标模式中的对应项的匹配,所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式,即为所述值添加元信息。

[0014] 根据本发明的另一方面,提供了一种基于混合属性 - 值匹配的模式映射系统,包括 :模式匹配装置,用于匹配对象的源模式和目标模式中的对应项以生成匹配结果映射,模式代表对象的副本,并由具有层次结构的属性 - 值对组成,其中所述模式匹配装置对源模式和目标模式中的值进行规范化处理,以匹配源模式和目标模式中的对应项,所述规范化处理是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式,即为其添加元信息 ;模式整合装置,与模式匹配装置相连接,用于根据所述模式匹配装置生成的所述匹配结果映射来整合所述源模式和目标模式,以生成整合的模式。

[0015] 在上述模式映射系统中,所述模式匹配装置包括 :模式规范化模块,接收对象的源

模式和目标模式作为输入,对源模式和目标模式的属性和值进行规范化处理,以使得所述属性和值更加可比较;模式匹配模块,与所述模式规范化模块相连接,接收已由所述模式规范化模块进行了规范化的属性和值,并计算源模式和目标模式之间的属性-属性匹配相似度、值-值匹配相似度和属性-值交叉匹配相似度;匹配映射计算模块,与所述模式匹配模块相连接,接收由所述模式匹配模块计算出的源模式和目标模式之间的属性-属性匹配相似度、值-值匹配相似度和属性-值交叉匹配相似度,从而计算所述源模式和目标模式的对应项之间的综合相似度并生成所述匹配结果映射。

[0016] 在上述模式映射系统中,所述模式整合装置包括:结构推理模块,与所述匹配映射计算模块相连接,接收所述匹配映射计算模块所生成的匹配结构映射,并根据所述匹配结果映射推理实际映射情况;结构变形模块,与所述结构推理模块相连接,根据所述接收推理模块输出的所述实际映射情况对所述源模式或所述目标模式进行变形,以生成所述整合的模式。

[0017] 在上述模式映射系统中,所述值的规范化处理包括:值为复合的简单短语时,分离处于并列关系的简短短语以成为简短短语集合的形式;值为值表达式时,借助于领域无关的度量单位字典来分离值表达式中的数值和度量单位以成为数值+度量单位的形式;值为复合的值表达式时,分离处于并列关系的值表达式,并借助于领域无关的度量单位字典来分离值表达式中的数值和度量单位以成为数值+度量单位集合的形式;值为表格和列表时,分解表格和列表的项,以成为简短短语或简短短语集合,以及数值+度量单位或数值+度量单位集合的形式;值为解释性段落时,从解释性段落中抽取关键词语,以成为简短短语或简短短语集合,以及数值+度量单位或数值+度量单位集合的形式。

[0018] 在上述模式映射系统中,所述值-值匹配相似度计算包括:在源模式和目标模式的值均为简短短语或简短短语集合时,对于源模式和目标模式的两个简短短语集合中的每一个简短短语,使用字符串相似度度量来计算相似度,并取平均值作为值-值匹配相似度;在源模式和目标模式的值均为数值+度量单位或数值+度量单位集合时,对于源模式和目标模式的两个数值+度量单位集合中的每一个数值+度量单位,借助于领域无关的度量单位字典来计算相似度,并取平均值作为值-值匹配相似度;在源模式和目标模式的值为简短短语集合和数值+度量单位集合的结合时,对于源模式和目标模式的简短短语集合中的每一简短短语和数值+度量单位集合中的每一数值+度量单位,使用字符串相似度度量来计算相似度,并取平均值作为值-值匹配相似度。

[0019] 在上述模式映射系统中,所述源模式和目标模式的对应项之间的综合相似度为:  

$$\text{Score} = \alpha \cdot \text{Score}_{\text{attr}} + \beta \cdot \text{Score}_{\text{val}} + (1 - \alpha - \beta) \cdot \text{Score}_{\text{cross}}$$

[0020] 其中,Score<sub>attr</sub>为所述属性-属性匹配相似度,Score<sub>val</sub>为所述值-值匹配相似度,Score<sub>cross</sub>为所述属性-值交叉匹配相似度; $\alpha$ 和 $\beta$ 为权重,并满足如下关系: $0 \leq \beta \leq 1$ , $0 \leq \alpha \leq 1$ , $0 \leq \alpha + \beta \leq 1$ 。

[0021] 在上述模式映射系统中,所述匹配映射结果的生成包括:生成源模式到目标模式的匹配映射:对源模式中的每个元素*i*,取Score[i]中分数最高的Score[i][j],目标模式中的元素*j*即为元素*i*的对应项,将*<i, j>*添加到匹配映射中;生成目标模式到源模式的匹配映射:对目标模式中的每个元素*p*,取Score<sup>T</sup>[p]中分数最高的Score<sup>T</sup>[p][q],其中Score<sup>T</sup>[ ] [ ] 为Score[ ] [ ] 的转置矩阵,源模式中的元素*q*即为元素*p*的对应项,将*<p, q>*添

加到匹配映射中。

[0022] 在上述模式映射系统中,所述属性的规范化处理包括:平滑层次关系:抽取从根到当前元素的绝对路径信息;和平滑模式中各元素的位置先后关系。

[0023] 在上述模式映射系统中,所述属性-属性匹配相似度的计算采用任意技术的字符串相似度度量。

[0024] 在上述模式映射系统中,所述属性-值交叉匹配相似度的计算包括:使用字符串相似度度量,计算源模式中属性和目标模式中值的匹配相似度;和使用字符串相似度度量,计算源模式中值和目标模式中属性的匹配相似度。

[0025] 在上述模式映射系统中,所述模式整合装置根据源模式到目标模式的匹配映射和目标模式到源模式的匹配映射来推理实际映射情况,并根据所述实际映射情况整合对应项和非对应项以对源模式或目标模式进行变形。

[0026] 在上述模式映射系统中,所述实际映射情况的推理包括:推理一对一映射:对源模式中的元素*i*,在目标模式中有元素*j*使得<*i*, *j*>和<*j*, *i*>成为匹配映射,并且在源模式中没有另一个元素*k*使得<*i*, *k*>或<*k*, *j*>成为匹配映射;推理一对多映射:对源模式中的元素*i*,在目标模式中有多于一个的元素{*j*, *k*}使得<*j*, *i*>和<*k*, *i*>成为匹配映射,并且<*i*, *j*>和<*i*, *k*>中至少有一个为匹配映射;推理多对一映射:对源模式中的多于一个的元素{*i*, *j*},在目标模式中有元素*k*使得<*i*, *k*>和<*j*, *k*>成为匹配映射,并且<*k*, *i*>和<*k*, *j*>中至少有一个为匹配映射;和推理无映射:对源模式中的元素*i*,在目标模式中没有元素*j*使得<*i*, *j*>或<*j*, *i*>成为匹配映射。

[0027] 在上述模式映射系统中,所述源模式的变形包括:一对一映射:不变形;一对多映射:将目标模式中的多个节点附加为源模式节点的子节点;多对一映射:将目标模式中的节点插入到源模式的多个节点和它们的父节点之间;和无映射:将目标模式中的节点附加为源模式根节点的子节点。

[0028] 在上述模式映射系统中,所述目标模式的变形包括:一对一映射:不变形;一对多映射:将源模式中的多个节点附加为目标模式节点的子节点;多对一映射:将源模式中的节点插入到目标模式的多个节点和它们的父节点之间;和无映射:将源模式中的节点附加为目标模式根节点的子节点。

[0029] 根据本发明的另一方面,提供了一种基于混合属性-值匹配的模式匹配方法,用于匹配对象的源模式和目标模式中的对应项,模式代表对象的副本,并由具有层次结构的属性-值对组成,所述模式匹配方法包括:对源模式和目标模式中的值进行规范化,以用于源模式和目标模式中的对应项的匹配,所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式,即为所述值添加元信息。

[0030] 根据本发明的再一方面,提供了一种基于混合属性-值匹配的模式映射方法,包括:模式匹配步骤,用于匹配对象的源模式和目标模式中的对应项以生成匹配结果映射,模式代表对象的副本,并由具有层次结构的属性-值对组成,其中所述模式匹配步骤对源模式和目标模式中的值进行规范化处理,以匹配源模式和目标模式中的对应项,所述规范化处理是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式,即为所述值添加元信息;模式整合步骤,用于根据所述模式匹配步骤生成的匹配结果映射来整合所述源模式和目标模式,以生成整合的模式。

[0031] 在上述模式匹配系统、模式映射系统及方法中，通过将对象的模式的无结构的纯文本形式的值规范化为有结构的形式，即为其添加元信息，可以使得源模式和目标模式的对应项的值更加可比较，同时也减小了相似度计算的粒度，从而提高了模式匹配的精度。

[0032] 并且，在上述模式匹配系统、模式映射系统及方法中，通过对对象的模式的属性和值进行交叉匹配计算，能够发现更多的匹配对应项，从而提高了模式匹配的精度。

[0033] 此外，在上述模式匹配系统、模式映射系统及方法中，通过借助于领域无关的字典将对象的模式的值规范化为简短短语或简短短语集合以及数值 + 度量单位或数值 + 度量单位集合，无需引入领域相关的表单、词典以及本体知识，可以降低系统的成本，并便利用户的使用。

[0034] 通过阅读结合附图考虑的以下本发明的优选实施例的详细描述，将更好地理解本发明的以上和其他目标、特征、优点和技术及工业重要性。

## 附图说明

- [0035] 图 1 是示出本发明实施例中的对象的示意图；
- [0036] 图 2 是示出如图 1 所示的对象的模式的树结构表示的图；
- [0037] 图 3 是示出如图 2 所示的模式以“\*.xml”格式存储在硬盘中的示意图；
- [0038] 图 4 是示出本发明实施例的模式匹配和模式映射系统的源模式和目标模式的匹配结果映射的示意图；
- [0039] 图 5 是示出源模式和目标模式的整合结果的示意图；
- [0040] 图 6 是示出了本发明实施例的模式映射系统的框图；
- [0041] 图 7 是示出本发明实施例的模式中的层次关系和位置顺序的示意图；
- [0042] 图 8 是示出了本发明实施例的模式规范化模块的值规范化的流程图；
- [0043] 图 9 是示出了本发明实施例的属性 - 属性匹配的流程图；
- [0044] 图 10 是示出了本发明实施例的值 - 值匹配的流程图；
- [0045] 图 11 是示出了本发明实施例的属性 - 值交叉匹配的流程图；
- [0046] 图 12 是示出了本发明实施例的一对多映射情况下源模式的结构变形的示意图；
- [0047] 图 13 是示出了本发明实施例的多对一映射情况下源模式的结构变形的示意图；
- [0048] 图 14 是示出了本发明实施例的模式映射方法的流程图。
- [0049] 图 15 是示出了以计算机实现本发明实施例的模式映射系统和模式映射方法的系统的硬件框图。

## 具体实施方式

[0050] 下面将结合附图来详细描述本发明的具体实施例。

[0051] 根据本发明的实施例，提供了一种基于混合属性 - 值匹配的模式匹配系统，用于匹配对象的源模式和目标模式中的对应项，模式代表对象的副本，并由具有层次结构的属性 - 值对组成，所述模式匹配系统包括：模式规范化模块，对源模式和目标模式中的值进行规范化，以用于源模式和目标模式中的对应项的匹配，所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式，即为所述值添加元信息。

[0052] 根据本发明的实施例，提供了一种基于混合属性 - 值匹配的模式映射系统，包括：

模式匹配装置，用于匹配对象的源模式和目标模式中的对应项以生成匹配结果映射，模式代表对象的副本，并由具有层次结构的属性 - 值对组成，其中所述模式匹配装置对源模式和目标模式中的值进行规范化处理，以匹配源模式和目标模式中的对应项，所述规范化处理是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式，即为所述值添加元信息；模式整合装置，与模式匹配装置相连接，用于根据所述模式匹配装置生成的所述匹配结果映射来整合所述源模式和目标模式，以生成整合的模式。

[0053] 首先，将描述本发明实施例的模式匹配和模式映射系统的原理。

[0054] 在本发明实施例的模式匹配和模式映射系统中，处理的对象通常是指现实世界中的一个产品，比如数码相机，并且模式是指这种现实产品的一个副本。由于应用等方面的差异，对于单一现实产品来说，可能存在多个异构的模式。因此，本发明实施例的模式匹配和模式映射系统意在识别出异构模式中的对应项并进行匹配，从而映射同一对象的不同模式，并整合这些异构的模式。

[0055] 例如，在对象为互联网上异构的数据源网页的情况下，各个不同模式中所包括的对象信息可以是通过信息抽取技术从网页中识别出来。图1是示出本发明实施例中的对象的示意图。例如，图1示出了网页中的表格，其是本发明实施例的模式匹配系统和模式映射系统中的模式的数据来源。这里，图1所示的对象是现实产品，具体地说，型号为“Canon EOS 7D”的数码相机。对于因特网上网页表格的抽取，通常包含表格识别和层次结构表格抽取两个步骤，本领域技术人员可以了解上述步骤的具体实现方式，因此在这里就不再赘述。

[0056] 这里，对象的内部表示即被称为模式，其通常由属性和值组成，也被称为模式的元素。模式的一个实例就是一个带有绝对路径信息的属性 - 值对，且属性可以有层次关系。图2是示出如图1所示的对象的模式的树结构表示的图。这里，图2示出的模式1和模式2即是本发明实施例的模式匹配和模式映射系统所要进行处理的源模式和目标模式的示例，即，本发明实施例的模式匹配和模式映射系统处理的是含有对象属性 - 值对信息的模式。

[0057] 以模式1为例，这个模式很好地代表了网页表格，很好地描述了对象“Canon EOS 7D”。可以看到，对象包含属性“General”和“Product Type”等，以及值“Digital camera-SLR”和“5.8in”等。属性的层次信息以树结构的表示是很清楚的：根元素为“top”，非叶子节点为属性，如“General”和“Product Type”等；叶子节点为值，如“Digital camera-SLR”和“5.8in”等。在硬盘存储中，模式被保存为“\*.xml”格式，如图3所示。

[0058] 在进行模式匹配和模式映射时，如果已知两个模式（源模式和目标模式）描述同一个对象，则首先要找出对应的元素。图4所示为本发明实施例的模式匹配和模式映射系统的源模式和目标模式的匹配结果映射的示意图。这里，匹配结果映射以TreeMap的数据结构存储在RAM中。比如，属性 - 值对 <“top->General->Product Type”, “Digital camera-SLR”> 和 <“Specification->Type->Type”, “Digital, AF/AE single-lens reflex camera”> 是语义上匹配的对应项。为记录对应项，定义了两个匹配结果映射以减少冲突，即源模式到目标模式的映射和目标模式到源模式的映射。在源模式到目标模式的映射中，<i, j> 表示源模式中的元素 i 和目标模式中的元素 j 为对应项。

[0059] 根据生成的匹配结果映射，通过源模式或目标模式的变形，将源模式和目标模式整合为一个结果模式。整合后的模式包含所有源模式和目标模式中的信息，并且没有冗余。图5所示为源模式和目标模式的整合结果。

[0060] 在本发明实施例的模式映射系统中，模式匹配装置包括：模式规范化模块，接收对象的源模式和目标模式作为输入，对源模式和目标模式的属性和值进行规范化处理，以使得所述属性和值更加可比较；模式匹配模块，与所述模式规范化模块相连接，接收已由所述模式规范化模块进行了规范化的属性和值，并计算源模式和目标模式之间的属性-属性匹配相似度、值-值匹配相似度和属性-值交叉匹配相似度；匹配映射计算模块，与所述模式匹配模块相连接，接收由所述模式匹配模块计算出的源模式和目标模式之间的属性-属性匹配相似度、值-值匹配相似度和属性-值交叉匹配相似度，从而计算所述源模式和目标模式的对应项之间的综合相似度并生成所述匹配结果映射。

[0061] 在本发明实施例的模式映射系统中，模式整合装置包括：结构推理模块，与所述匹配映射计算模块相连接，接收所述匹配映射计算模块所生成的匹配结构映射，并根据所述匹配结果映射推理实际映射情况；结构变形模块，与所述结构推理模块相连接，根据所述接收推理模块输出的所述实际映射情况对所述源模式或所述目标模式进行变形，以生成所述整合的模式。

[0062] 下面，将参考图6来详细描述本发明实施例的模式映射系统，图6是示出了本发明实施例的模式映射系统的框图。

[0063] 如图6所示，本发明实施例的模式映射系统10包括模式规范化模块20，模式匹配模块21，匹配映射计算模块22，结构推理模块23和结构变形模块24。其中，模式规范化模块20接收例如如图4所示的源模式和目标模式作为输入，从而对源模式和目标模式的属性和值进行规范化，以使得所述属性和值更加可比较。模式匹配模块21与模式规范化模块20相连接，接收已由模式规范化模块20进行了规范化的属性和值，并计算属性-属性匹配相似度，值-值匹配相似度和属性-值交叉匹配相似度。匹配映射计算模块22与模式匹配模块21相连接，接收由模式匹配模块计算出的源模式和目标模式之间的属性-属性匹配相似度，值-值匹配相似度和属性-值交叉匹配相似度，从而计算源模式和目标模式的对应项之间的综合相似度并生成匹配结果映射。结构推理模块23与匹配映射计算模块22相连接，从匹配映射计算模块22接收匹配结果映射，并根据匹配结果映射推理实际映射情况。结构变形模块24与结构推理模块23相连接，根据接收推理模块23输出的实际映射情况对源模式或目标模式进行变形，以生成整合的模式，例如如图5所示整合后的模式。本系统的输入是两个模式：源模式和目标模式，例如如图2所示的。系统的输出是一个整合的模式，例如如图5所示的。并且，中间结果为记录对应项的匹配结果映射，例如如图4所示的。

[0064] 下面，将对上述模式映射系统10的每个模块进行具体说明。

[0065] 首先说明模式规范化模块20。在实际引用中，虽然网页中的表格在视觉上是结构化的，但是实际上并没有设计为关系表格，并且描述风格和措词也是多样的。以数码相机产品为例，销售网站多倾向于列举用户感兴趣并易于理解的通用特征作为产品说明；而产品的官方网站往往给出详尽的偏向技术细节却不易理解的属性作为产品描述。由于无法给出确切地定义某一对象的哪一个属性是重要的，相似的模式结构并不说明内容也是相似的，也就是说模式中的结构信息对于匹配是无用的。因此，在本发明实施例的模式规范化模块20中，首先规范化模式中的属性，平滑掉对匹配无用的信息。

[0066] 在本发明实施例的模式映射系统中，属性的规范化包括：平滑层次关系：抽取从根到当前元素的绝对路径信息；和平滑模式中各元素的位置先后关系。

[0067] 图 7 示出了本发明实施例中的模式的层次关系和位置顺序信息。层次关系即是树中的父子关系,比如路径“Specification-> Type-> Recording Media”中的层次关系为：“Specification”为“Type”的上层(父节点);同时“Type”是“Recording Media”的上层(父节点)。位置顺序关系是节点在树中出现的顺序,比如各个属性的位置顺序为：“Type”,“Recording Media”,“ImageSensor Size”,“Lens Mount”,“Type”,“Pixels”,“Total Pixels”等。在本发明实施例的模式规范化模块中,规范化模式的属性的方法可以包括:

[0068] 1) 使用从根到当前元素的绝对路径作为属性,(路径;当前元素的属性),比如:

[0069] (Specification, Type ;Type)

[0070] (Specification, Type ;Recording Media)

[0071] (Specification, Type ;Image Sensor Size)

[0072] (Specification, Type ;Lens Mount)

[0073] (Specification, Image Sensor ;Type)

[0074] (Specification, Image Sensor ;Pixels)

[0075] (Specification, Image Sensor ;Total Pixels)

[0076] 2) 忽略路径信息,只考虑当前元素的属性,(当前元素的属性)。

[0077] 通过上述两种属性的规范化方法,属性都不再保有层次信息和位置顺序信息。当然,本领域技术人员可以理解,这里属性的规范化方法也可以采用现有技术当中的其它方法,本发明的实施例并不意在对此进行限制。

[0078] 上面对于模式规范化模块 20 的对于模式的属性的规范化进行了说明,下面将说明值规范化。

[0079] 在本发明实施例的模式映射系统中,值的规范化包括:值为复合的简单短语时,分离处于并列关系的简短短语以成为简短短语集合的形式;值为值表达式时,借助于领域无关的度量单位字典来分离值表达式中的数值和度量单位以成为数值+度量单位的形式;值为复合的值表达式时,分离处于并列关系的值表达式,并借助于领域无关的度量单位字典来分离值表达式中的数值和度量单位以成为数值+度量单位集合的形式;值为表格和列表时,分解表格和列表的项,以成为简短短语或简短短语集合,以及数值+度量单位或数值+度量单位集合的形式;值为解释性段落时,从解释性段落中抽取关键词语,以成为简短短语或简短短语集合,以及数值+度量单位或数值+度量单位集合的形式。

[0080] 相比于关系数据库中的表格和结构化的 XML 文档,网页中的表格没有元信息:其中的值只以无结构的字符串纯文本形式存在,没有任何类型,表约束,取值范围,命名空间等元信息;而元信息可以帮助建立结构化数据之间的联系。因此,本发明实施例的模式规范化模块 20 在进行值的规范化处理时,是将这些无结构的纯文本形式的值转化为结构化形式,即为所述值创建部分元信息,使得它们更加可比较。表 1 中列举了网页表格中值的各种形式的一个示例,而表 2 中列举了对应的规范化后的值的相应示例。

[0081] 表 1 :网页表格中值的形式

[0082]

值的形式	属性-值示例	
简短短语	Product type	Digital camera – SLR
复合的简短短语	Special effects	Neutral, Faithful, Portrait, Landscape, Monochrome
值表达式	Resolution	18 megapixels
复合的值表达式 (W*H*D)	Dimensions	Approx. 5.8*4.4*2.9 in.
表格或列表	Video Out Terminal	Video out terminal: NTSC/PAL selectable Mini-HDMI out terminal
解释性段落	AF-assist Beam	When an external EOS-dedicated Speedlite is attached to the camera, the AF-assist beam from the Speedlite will be emitted when necessary

[0083] 表 2 :规范化的结果

[0084]

值的形式	规范化示例	
简短短语	Digital camera – SLR	<Digital camera – SLR>
复合的简短 短语	Neutral, Faithful, Portrait, Landscape, Monochrome	<Neutral> <Faithful> <Portrait> <Landscape> <Monochrome>
值表达式	18 megapixels	<18(value) + megapixels(unit)>
复合的值表	Approx. 5.8*4.4*2.9 in.	<5.8 + in.> <4.4 + in.> <2.9 +

[0085]

达式		in.>
表格或列表	Video out terminal: NTSC/PAL selectable Mini-HDMI out terminal	<Video out terminal: NTSC/PAL selectable> <Mini-HDMI out terminal >
解释性段落	When an external EOS-dedicated Speedlite is attached to the camera, the AF-assist beam from the Speedlite will be emitted when necessary	<external EOS-dedicated Speedlite> <AF-assist beam from the Speedlite>

[0086] 图 8 是示出了本发明实施例的模式规范化模块的值规范化的流程图,如图 8 所示:

[0087] 在步骤 S21 中,判断值的形式:使用正则表达式来检测出数值;使用分隔符如逗号和分号来分隔并列关系的项;使用索引标号找出隐藏的表格或列表。

[0088] 在步骤 S22 中,使用逗号或者乘号等分隔符,分隔处于并列关系的项(简短短语,值表达式),比如“Neutral, Faithful, Portrait, Landscape, Monochrome”和“5.8\*4.4\*2.9 in.”。规范化后的结构为(<简短短语>)\*或(<值表达式>)\*。

[0089] 在步骤 S23 中,分隔值表达式中的数值和度量单位,比如将“18megapixels”规范化为数值“18”和度量单位“megapixels”。数值可以使用正则表达式匹配,度量单位可以借助于一个领域无关的词典。规范化后的结果为<数值+度量单位>。

[0090] 在步骤 S24 中,根据索引标号分解表格和列表,规范化后的结果为(<表格列表项>)\*。

[0091] 在步骤 S25 中,为了使解释性段落更好比较,抽取其中的关键词语或名词短语来代表整段文本,借助于关键词抽取工具或者词性标注工具。规范化后的结果为(<关键词语>)\*或(<名词短语>)\*。

[0092] 这样,在本发明实施例的模式规范化模块 20 的对于模式的值进行规范化之后,所述模式的值被由无结构的字符串纯文本形式转化为结构化的数据,即,(<简短短语>)\*和(<数值+度量单位>)\*两种形式。这里,(<简短短语>)\*表示简短短语或简短短语的集合,同样,(<数值+度量单位>)\*表示值表达式或者值表达式的集合。这里,上述步骤 S24 中获得的(<表格列表项>)\*和步骤 S25 中获得的(<关键词语>)\*或(<名词短语>)\*均可认为是以(<简短短语>)\*和(<数值+度量单位>)\*形式的。

[0093] 当然,本领域技术人员可以理解,在上述实施例中,将模式的值的形式划分为“简短短语”、“复合的简短短语”、“值表达式”、“复合的值表达式”、“表格或列表”和“解释性段落”六种形式,并根据所述值的这六种形式规范化为(<简短短语>)\*和(<数值+度量单位>)\*两种形式。但是,根据所采用的模式的值的具体形式,也可以将值划分为其它的多种形式,并相应地规范化为其它的多种形式。

[0094] 例如,根据本发明实施例的模式的值的规范化处理的另一示例中,并不将模式的值的形式划分为上述的六种形式,而是仅将模式的值看作是单一的字符串纯文本。与此相应的,该示例性的规范化处理可以包括:分离处于并列关系的项;抽取文本中的值表达式,这是因为通常在含有值表达式的文本中,值表达式为其中的重要信息;和抽取文本中的关键词语作为代表性信息。

[0095] 这里,本领域技术人员可以理解,本发明实施例中值的规范化处理可以根据具体问题的数据和目的选择规范化的粒度,比如在上述示例中,处于并列关系的项在分离后仍可进一步的抽取关键词语,或者对于纯文本中的值表达式是否重要可以自行判断。因此,对于本发明实施例的模式的值的规范化处理,本申请的说明书文本并不意在进行任何限制。

[0096] 并且,在上述描述中,模式规范化模块 20 对于模式的属性和值进行规范化,本领域技术人员可以理解,这里模式规范化模块 20 可以包括属性规范化单元和值规范化单元来分别对于模式的属性和值进行规范化处理,或者上述属性和规范化处理和值的规范化处理也可以由单一组件进行,本发明的实施例并不意在对此进行限制。

[0097] 在由模式规范化模块 20 进行了模式的属性和值的规范化之后,模式匹配模块 21 从模式规范化模块 20 接收经过规范化之后的属性和值,并进行匹配。所述模式匹配模块 21 可以包括三个单元,以分别进行属性 - 属性匹配、值 - 值匹配和属性 - 值匹配。

[0098] 在本发明实施例的模式映射系统中,属性 - 属性匹配相似度的计算采用任意技术的字符串相似度度量。

[0099] 具体地说,在属性 - 属性匹配单元中,对于模式的属性进行匹配计算的相似度分数被存储在一个二维矩阵中,对源模式中的每个元素和目标模式的每个元素都有一个相似度值,这个值是一个 [0, 1] 区间上的实数。图 9 是示出了本发明实施例的属性 - 属性匹配的流程图。如图 9 所示,步骤 S31 和步骤 S32 执行了一个双层“for”循环以计算属性 - 属性匹配分数矩阵 Score<sub>attr</sub>[], 其中 Score<sub>attr</sub>[i][j] 为源模式中的元素 i 和目标模式中的元素 j 的属性匹配相似度分数。经过上述的属性规范化,属性的层次结构被平滑为文本形式的绝对路径及属性本身,因此属性的匹配可以采用字符串相似度度量来计算(步骤 S33),比如 Smith-Waterman 距离, LSC 等。

[0100] 在本发明实施例的模式映射系统中,值 - 值匹配相似度计算包括:在源模式和目标模式的值均为简短短语或简短短语集合时,对于源模式和目标模式的两个简短短语集合中的每一个简短短语,使用字符串相似度度量来计算相似度,并取平均值作为值 - 值匹配相似度;在源模式和目标模式的值均为数值 + 度量单位或数值 + 度量单位集合时,对于源模式和目标模式的两个数值 + 度量单位集合中的每一个数值 + 度量单位,借助于领域无关的度量单位字典来计算相似度,并取平均值作为值 - 值匹配相似度;在源模式和目标模式的值为简短短语集合和数值 + 度量单位集合的结合时,对于源模式和目标模式的简短短语集合中的每一简短短语和数值 + 度量单位集合中的每一数值 + 度量单位,使用字符串相似度度量来计算相似度,并取平均值作为值 - 值匹配相似度。

[0101] 具体地说,在值 - 值匹配单元中,经过上述模式规范化模块 20 所进行的值的规范化之后,如上述实施例中所述的那样,值的无结构的字符串纯文本被转换为如下两种形式:1) 简短短语或简短短语的集合:(<简短短语>)\*,其中的简短短语可以是普通的简短短语,表格或列表中的项,或解释性段落中抽出的关键词语或名词短语;2) 值表达式或值表达式

的集合 :(< 数值 + 度量单位 >)\*, 其中度量单位可能缺失。这里, 本领域技术人员可以看到, 显然同一形式下的值对比更加有意义 : 比较简短短语与简短短语, 以及比较值表达式与值表达式, 比单纯使用字符串相似性度量比较所有的值更加合理。

[0102] 图 10 是示出了本发明实施例的值 - 值匹配的流程图。如图 10 所示, 步骤 S41 和步骤 S42 执行了一个双层 “for” 循环以计算值 - 值匹配分数矩阵 Score<sub>val</sub>[][] , 其中 Score<sub>val</sub>[i][j] 为源模式中的元素 i 和目标模式中的元素 j 的属性匹配相似度分数。步骤 S43 计算元素 i 的值和元素 j 的值之间的相似度, 具体可分解如下步骤 : 首先, 步骤 S61 判断两个值的形式是否相同, 以确定两个值是否可比较, 判定的结果可能为 :

[0103] 1) 元素 i 和元素 j 的值都为 (< 简短短语 >)\*。

[0104] 步骤 S62 中, 对每个短语, 可以使用任意字符串相似度度量来计算其相似度, 取每次匹配的平均值赋给 Score<sub>val</sub>[i][j] : a) 如果两个值都是单个的短语, 计算这两个短语的匹配相似度 ; b) 如果两个值都是简短短语的集合 (复合简短短语, 解释性段落, 表格), 对两个简短短语集合中的每一对简短短语计算匹配相似度, 最后取各次相似度计算的平均值作为结果 ; c) 如果一个值为单个短语而另一个值为简短短语的集合, 计算单个短语和简短短语集合中的每一个简短短语的匹配相似度, 并取各次相似度计算的平均值作为结果。

[0105] 2) 元素 i 和元素 j 的值形式不同, 即元素 i 的值为 (< 短语 >)\* 而元素 j 的值为 (< 数值 + 度量单位 >)\*; 或者元素 i 的值为 (< 数值 + 度量单位 >)\* 而元素 j 的值都为 (< 短语 >)\*。

[0106] 这里对于 (< 短语 >)\* 和 (< 数值 + 度量单位 >)\* 进行相似度计算。但在一些复杂的情况下, 解释性段落或表格中可能含有可以表示为 (< 数值 + 度量单位 >)\* 的值表达式, 而这些值表达式在规范化中不会被发现, 因为其他的文本信息可能更为重要。因此在步骤 S62 中使用字符串相似度度量计算 Score<sub>val</sub>[i][j]。

[0107] 3) 元素 i 和元素 j 的值都为 (< 数值 + 度量单位 >)\*。

[0108] 对两个集合中的每一对值表达式计算相似度, 取每次匹配的平均值赋给 Score<sub>val</sub>[i][j]。在步骤 S63 中, 判断度量单位是否可比 : 如果度量单位相同, 比较两个值表达式中的数值 ; 如果度量单位缺失, 默认为数值可比较, 比较两个值表达式中的数值 ; 如果度量单位不同, 在步骤 S64 中进行单位换算, 这里可以借助于一个领域无关的度量单位换算词典。在步骤 S65 中, 比较两个数值是否相等, 结果精度只能为 0.0 和 1.0。比如, “18 megapixels” 和 “1800000 pixels”的相似度为 1.0 : 将 “megapixels” 换算为 “pixels” 导致 “18” 变为 “1800000”, 即 18 megapixels 等于 1800000 pixels。

[0109] 上述值 - 值匹配处理是基于将值的无结构的纯文本形式规范化为 (< 简短短语 >)\* 和 (< 数值 + 度量单位 >)\* 所进行的匹配处理。本领域技术人员可以理解, 如上文所述, 通过根据具体问题的数据和目的, 可以选择值的规范化处理后的值的不同形式, 及选择规范化处理的粒度。在这种情况下, 本发明实施例的值 - 值匹配处理可以根据相应的规范化处理后值的不同形式来进行值 - 值匹配相似度计算, 其原理与以上所述的相同, 本发明的实施例并不意在对此进行任何限制。

[0110] 在本发明实施例的模式映射系统中, 属性 - 值交叉匹配相似度的计算包括 : 使用字符串相似度度量, 计算源模式中属性和目标模式中值的匹配相似度 ; 和使用字符串相似度度量, 计算源模式中值和目标模式中属性的匹配相似度。

[0111] 这里,属性 - 值交叉匹配单元针对可能存在的以下情况,比如,源模式中的元素 i 为〈Resolution-18 megapixels〉,目标模式中的元素 j 为〈Pixels-18,000,000〉,属性 - 属性匹配计算和值 - 值匹配计算都不会判定其为对应项。首先,属性“Resolution”和属性“Pixels”通过字符串匹配无法判定相似,使用 WordNet 也无法发现它们语义相似,即它们之间的语义关系非常弱,尽管它们在数码相机领域里频繁地共现。如果参考属性的绝对路径,“top, Mainfeatures ;Resolution”和“Specification, Image sensor ;Pixels”,字符串匹配也无法发现匹配。其次,值“18 megapixels”和值“Approx. 18,000,000”直观上看起来很相似,但是“18,000,000”只是一个数值缺失度量单位,使得两个值表达式中的数值无法直接进行比较。这里,需要注意的是数值的比较必须非常谨慎,缺失度量单位意味着缺失约束,比较的结果会不可靠。而如果比较元素 i 的值“18 megapixels”和元素 j 的属性“Pixels”,很容易产生匹配,使用简单的字符串相似度度量即可。

[0112] 图 11 是示出了本发明实施例的属性 - 值交叉匹配的流程图。如图 11 所示,步骤 S51 和步骤 S52 执行了一个双层的“for”循环来计算属性 - 值交叉匹配分数矩阵 Score<sub>cross</sub>[], 其中 Score<sub>cross</sub>[i][j] 为源模式中的元素 i 和目标模式中的元素 j 的交叉匹配相似度分数。匹配分为两个步骤:步骤 S53 中计算元素 i 的属性和元素 j 的值之间的字符串相似度 s<sub>ij</sub>;步骤 S54 中计算元素 i 的值和元素 j 的属性之间的字符串相似度 s<sub>ji</sub>。最后在步骤 S55 中,取 s<sub>ij</sub> 和 s<sub>ji</sub> 的平均值赋给 Score<sub>cross</sub>[i][j]。

[0113] 这里,本领域技术人员可以理解,以上所述的属性 - 属性匹配、值 - 值匹配和属性 - 值匹配计算的流程仅为本发明实施例的模式匹配模块 21 所执行的计算的特定示例,根据模式规范化模块 20 进行的属性和值的规范化结果,模式匹配模块 21 可以进行相应的匹配计算,本发明的实施例并不意在对此进行任何限制。

[0114] 在本发明实施例的模式映射系统中,所述源模式和目标模式的对应项之间的综合相似度为:Score = α • Score<sub>attr</sub> + β • Score<sub>val</sub> + (1 - α - β) • Score<sub>cross</sub>

[0115] 其中,Score<sub>attr</sub> 为所述属性 - 属性匹配相似度,Score<sub>val</sub> 为所述值 - 值匹配相似度,Score<sub>cross</sub> 为所述属性 - 值交叉匹配相似度;α 和 β 为权重,并满足如下关系:0 ≤ β ≤ 1,0 ≤ α ≤ 1,0 ≤ α + β ≤ 1。

[0116] 具体地说,匹配映射计算模块 22 接收模式匹配模块 21 计算出的属性 - 属性匹配、值 - 值匹配和属性 - 值交叉匹配的分数,如上述实施例所述,Score<sub>attr</sub>[] [] 是属性 - 属性匹配计算的分数,Score<sub>val</sub>[] [] 是值 - 值匹配计算的分数,且 Score<sub>cross</sub>[] [] 是属性 - 值交叉匹配计算的分数。这里,匹配映射计算模块 22 将上述三个计算分数分别乘以相应的权重,从而计算出对应项的相似度分数为:

[0117] Score[i][j] = α • Score<sub>attr</sub>[i][j] + β • Score<sub>val</sub>[i][j] + (1 - α - β) • Score<sub>cross</sub>[i][j]

[0118] 其中 0 ≤ β ≤ 1,0 ≤ α ≤ 1,0 ≤ α + β ≤ 1;优选地,α 取 0.7,β 取 0.2。

[0119] 在计算出相应项的相似度分数之后,匹配映射计算模块 22 进一步根据相似度分数生成匹配结果映射。

[0120] 这里,匹配结果映射的生成有两种:

[0121] 1) 生成源模式到目标模式的匹配映射:对源模式中的每个元素 i,取 Score[i] 中分数最高的 Score[i][j],目标模式中的元素 j 即为元素 i 的对应项,将 <i, j> 添加到匹配

映射中。

[0122] 2) 生成目标模式到源模式的匹配映射 : 对目标模式中的每个元素 p, 取  $\text{Score}^T[p]$  中分数最高的  $\text{Score}^T[p][q]$ , 其中  $\text{Score}^T[]$  为  $\text{Score}[]$  的转置矩阵 ; 源模式中的元素 q 即为元素 p 的对应项, 将  $\langle p, q \rangle$  添加到匹配映射中。

[0123] 注意到, 对于每个元素, 只有一个匹配被记录, 即相似度分数的最大值, 尽管有时会有多个相似度的匹配发生。同时, 每个实际的匹配情况都不会被错过, 它将在后面步骤的结构推理中被发现。举例说明, 源模式中的元素 k “shutter speed” 和目标模式中的元素 i “max shutter speed” 以及元素 j “minshutter speed”, 显然是对应项。对元素 k 来说, 由于最大值只有一个, 只有一个匹配映射被记录, 可能是  $\langle k, i \rangle$  或者是  $\langle k, j \rangle$  保存在源模式到目标模式的映射中。同时,  $\langle i, k \rangle$  和  $\langle j, k \rangle$  都会记录到目标模式到源模式的映射中。通过检查两个映射中的连通路径, 就能发现元素 k 和元素 i 及元素 j 的关系。

[0124] 在本发明实施例的模式映射系统中, 模式整合装置根据源模式到目标模式的匹配映射和目标模式到源模式的匹配映射来推理实际映射情况, 并根据所述实际映射情况整合对应项和非对应项以对源模式或目标模式进行变形。

[0125] 在本发明实施例的模式映射系统中, 实际映射情况的推理包括 : 推理一对一映射 : 对源模式中的元素 i, 在目标模式中有元素 j 使得  $\langle i, j \rangle$  和  $\langle j, i \rangle$  成为匹配映射, 并且在源模式中没有另一个元素 k 使得  $\langle i, k \rangle$  或  $\langle k, j \rangle$  成为匹配映射 ; 推理一对多映射 : 对源模式中的元素 i, 在目标模式中有多个元素 {j, k} 使得  $\langle j, i \rangle$  和  $\langle k, i \rangle$  成为匹配映射, 并且  $\langle i, j \rangle$  和  $\langle i, k \rangle$  中至少有一个为匹配映射 ; 推理多对一映射 : 对源模式中的多个元素 {i, j}, 在目标模式中有元素 k 使得  $\langle i, k \rangle$  和  $\langle j, k \rangle$  成为匹配映射, 并且  $\langle k, i \rangle$  和  $\langle k, j \rangle$  中至少有一个为匹配映射 ; 和推理无映射 : 对源模式中的元素 i, 在目标模式中没有元素 j 使得  $\langle i, j \rangle$  或  $\langle j, i \rangle$  成为匹配映射。

[0126] 具体地说, 结构推理模块 23 接收匹配映射计算模块 22 生成的匹配结果映射, 以进行结构推理。其中, 在已经得到源模式到目标模式的匹配结果映射和目标模式到源模式的匹配结构映射后, 通过推理得到实际映射情况。如表 3 所示, 实际映射类型包括 :

[0127] 1) 一对一映射 : 匹配发生在一个源模式的元素和一个目标模式的元素之间。

[0128] 2) 一对多映射 : 匹配发生在同一个源模式的元素和多个目标模式的元素之间。

[0129] 3) 多对一映射 : 匹配发生在多个源模式的元素和同一个目标模式的元素之间。

[0130] 4) 无映射 : 一个源模式的元素, 和任意目标模式的元素之间, 没有匹配发生。

[0131] 表 3 : 实际映射类型

[0132]

映射类型	示例
一对一	Source. element[i] = Target. element[j] 源模式的元素 i = 目标模式的元素 j
一对多	Source. element[i] = Target. element[j], and Source. element[i] = Target. element[k]

[0133]

	源模式的元素 i = 目标模式的元素 j, 且 源模式的元素 i = 目标模式的元素 k
多对一	Source. element[i] = Target. element[k], and Source. element[j] = Target. element[k] 源模式的元素 i = 目标模式的元素 k, 且 源模式的元素 j = 目标模式的元素 k
无	Source. element[i] ≠ Target. element[j] 源模式的元素 i ≠ 目标模式的元素 j

[0134] 这里假设, 网页表格中的模式结构都是合理的, 其层次结构是遵循现实世界规律的; 并且在一个模式中没有冗余项。则具体推理出各种实际映射的方法为:

[0135] 1) 推理一对一映射: 对源模式中的元素 i, 在目标模式中有元素 j 使得  $\langle i, j \rangle$  和  $\langle j, i \rangle$  成为匹配映射, 并且在源模式中没有另一个元素 k 使得  $\langle i, k \rangle$  或  $\langle k, i \rangle$  成为匹配映射。

[0136] 2) 推理一对多映射: 对源模式中的元素 i, 在目标模式中有多个元素 {j, k} 使得  $\langle j, i \rangle$  和  $\langle k, i \rangle$  成为匹配映射, 并且  $\langle i, j \rangle$  和  $\langle i, k \rangle$  中至少有一个为匹配映射。

[0137] 3) 推理多对一映射: 对源模式中的元素 i 和元素 j, 在目标模式中有元素 k 使得  $\langle i, k \rangle$  和  $\langle j, k \rangle$  成为匹配映射, 并且  $\langle k, i \rangle$  和  $\langle k, j \rangle$  中至少有一个为匹配映射。

[0138] 4) 推理无映射: 对源模式中的元素 i, 在目标模式中没有元素 j 使得  $\langle i, j \rangle$  或  $\langle j, i \rangle$  成为匹配映射。

[0139] 在本发明实施例的模式映射系统中, 所述源模式的变形包括: 一对一映射: 不变形; 一对多映射: 将目标模式中的多个节点附加为源模式节点的子节点; 多对一映射: 将目标模式中的节点插入到源模式的多个节点和它们的父节点之间; 和无映射: 将目标模式中的节点附加为源模式根节点的子节点。

[0140] 在本发明实施例的模式映射系统中, 所述目标模式的变形包括: 一对一映射: 不变形; 一对多映射: 将源模式中的多个节点附加为目标模式节点的子节点; 多对一映射: 将源模式中的节点插入到目标模式的多个节点和它们的父节点之间; 和无映射: 将源模式中的节点附加为目标模式根节点的子节点。

[0141] 具体地说, 结构变形模块 24 基于结构推理模块 23 做出的结构推理进行结构变形。如表 4 所示, 各种映射类型均会导致源模式的结构变形, 本质上是将目标模式中的对应项和非对应项整合到源模式中。

[0142] 表 4: 各种映射类型下的变形

[0143]

映射类型	示例	变形
一对一	Source. element[i] = Target. Element[j]	无变形。
一对多	Source. element[i] = Target. element[j]	将节点 j 和节点 k 附加为节点 i 的子节点。
	Source. element[i] = Target. Element[k]	
多对一	Source. element[i] = Target. element[k]	将节点 i 和节点 j 附加为节点 k 的子节点；
	Source. element[j] = Target. element[k]	将节点 k 附加为节点 i 的父节点的子节点。
无	Source. element[i] ≠ Target. element[j]	将节点 j 附加为源模式根节点的子节点。

[0144] 针对不同的映射类型,进行源模式的结构变形如下:

[0145] 1) 一对一映射:不发生变形。

[0146] 2) 一对多映射:将目标模式中的各个节点附加为源模式节点的子节点,如图 12 所示。

[0147] 3) 多对一映射:将目标模式的节点插入到源模式的各个节点和它们的父节点之间,如图 13 所示。

[0148] 4) 无映射:将目标模式中的节点附加为源模式根节点的子节点。

[0149] 这样,通过结构变形模块对于源模式的结构变形,生成了整合后的模式,作为整个模式映射系统的输出。

[0150] 当然,本领域技术人员可以理解也可以对目标模式进行结构变形,从而将源模式中的对应项和非对应项整合到目标模式中,生成整合后的模式,作为整个模式映射系统的输出。

[0151] 这里,关于图 6 所示的模式映射系统的框图对于本发明实施例的模式映射系统进行了解释。本领域技术人员可以理解,对于本发明实施例的模式匹配系统,例如,可以仅包括图 6 的系统框图中的模式规范化模块、模式匹配模块和匹配映射计算模块,从而接收对象的源模式和目标模式作为输入,并输出匹配结果映射。所述匹配结果映射除用于模式整合之外,还可用于数据库中的重复记录挖掘和数据清理等,以及用于帮助建立索引和检索。

[0152] 因此,本发明实施例的模式匹配系统既可以作为单独的系统应用,也可以作为模式匹配装置应用于如上所述的模式映射系统中,并且,在单独应用或与模式整合装置结合应用于模式映射系统的情况下,其均可以包括如图 6 的系统框图所示的模式规范化模块、模式匹配模块和匹配映射计算模块,本发明的实施例并不意在对此进行任何限制。

[0153] 根据本发明的实施例,提供了一种基于混合属性 - 值匹配的模式匹配方法,用于

匹配对象的源模式和目标模式中的对应项，模式代表对象的副本，并由具有层次结构的属性 - 值对组成，所述模式匹配方法包括：对源模式和目标模式中的值进行规范化，以用于源模式和目标模式中的对应项的匹配，所述规范化是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式，即为所述值添加元信息。

[0154] 根据本发明的实施例，提供了一种基于混合属性 - 值匹配的模式映射方法，包括：模式匹配步骤，用于匹配对象的源模式和目标模式中的对应项以生成匹配结果映射，模式代表对象的副本，并由具有层次结构的属性 - 值对组成，其中所述模式匹配步骤对源模式和目标模式中的值进行规范化处理，以匹配源模式和目标模式中的对应项，所述规范化处理是指将源模式和目标模式中的值的无结构的纯文本形式转化为结构化形式，即为所述值添加元信息；模式整合步骤，用于根据所述模式匹配步骤生成的匹配结果映射来整合所述源模式和目标模式，以生成整合的模式。

[0155] 图 14 是示出了本发明实施例的模式映射系统的流程图。如图 14 所示，本发明实施例的模式映射方法包括如下步骤：

[0156] 在步骤 S11(规范化属性) 中，模式实例中的属性被规范化，该步骤例如由上述实施例中的模式规范化模块 20 执行。该步骤的输入是源模式和目标模式，输出是属性被规范化后的源模式和目标模式。

[0157] 在步骤 S12(规范化值) 中，模式实例中的值被规范化，该步骤例如由上述实施例中的模式规范化模块 20 执行。该步骤的输入是属性被规范化后的源模式和目标模式，输出是属性和值都被规范化后的源模式和目标模式。

[0158] 在步骤 S13(属性 - 属性匹配) 中，计算模式中属性的相似度，该步骤例如由上述实施例中的模式匹配模块 21 执行。该步骤的输入是规范化后的源模式和目标模式，输出是属性匹配相似度矩阵。

[0159] 在步骤 S14(值 - 值匹配) 中，计算模式中值的相似度，该步骤例如由上述实施例中的模式匹配模块 21 执行。该步骤的输入是规范化后的源模式和目标模式，输出是值匹配相似度矩阵。

[0160] 在步骤 S15(属性 - 值交叉匹配) 中，交叉计算模式中属性 - 值的相似度，该步骤例如由上述实施例中的模式匹配模块 21 执行。该步骤的输入是规范化后的源模式和目标模式，输出是属性 - 值交叉匹配相似度矩阵。

[0161] 在步骤 S16(计算相似度分数) 中，计算模式中对应项的相似度，该步骤例如由上述实施例中的匹配映射计算模块 22 执行。该步骤的输入是属性匹配相似度矩阵，值匹配相似度矩阵，属性 - 值交叉匹配相似度矩阵；输出是综合相似度矩阵。

[0162] 在步骤 S17(生成匹配映射) 中，生成两个匹配结果映射分别记录源模式到目标模式的映射和目标模式到源模式的映射，该步骤例如由上述实施例中的匹配映射计算模块 22 执行。该步骤的输入是综合相似度矩阵，输出是两个映射。

[0163] 在步骤 S18(推理映射) 中，根据两个匹配结果映射，除去冗余和冲突，推理实际映射情况，该步骤例如由上述实施例中的结构推理模块 23 执行。该步骤的输入是两个映射，输出是一个整合后的源模式到目标模式的映射或目标模式到源模式的映射。

[0164] 在步骤 S19(结构变形) 中，根据整合后的映射变形源模式或目标模式，该步骤例如由上述实施例中的结构变形模块 24 执行。该步骤的输入是源模式或目标模式和整合后

的映射,输出是整合后的模式。

[0165] 图 15 是示出了以计算机实现本发明实施例的模式匹配系统和模式映射方法的系统的硬件框图。如图 15 所示,本发明实施例的模式匹配系统和模式映射系统可以 PC 系统实现:输入和输出存储在如硬盘之类的存储设备(13)中,功能模块和中间结果都存储于 RAM(11) 中,功能模块由中央处理单元 CPU(10) 执行。

[0166] 本发明实施例提供了一种领域无关的模式匹配和模式映射系统及其方法,其通过采用值规范化方法,增加了值表达式的可比较性,把纯文本形式的无结构的值表达式转化为结构化的各种形式,创建了数值 - 度量单位的约束,并用抽取出的关键信息来代表解释性段落;由于现有技术对于值表达式通常不做特殊处理,忽略了它们对于匹配计算的价值,而只把它们当作字符串文本来处理,这使得处理效率很低,而本发明实施例的模式匹配和模式映射系统及其方法通过对值的规范化,显著提高了处理效率和匹配精度。并且,通过采用属性 - 值交叉匹配方法,可以发现更多的匹配对应项,从而改进了匹配的准确度。此外,在现有方法中,仅采用了属性之间的匹配和值之间的匹配,并需要借助于外部资源,而本发明实施例的模式匹配和模式映射系统及其方法通过借助于领域无关的字典来进行值的规范化处理,能够避免引入领域相关的表单,词典以及本体知识等,从而节省了系统的成本,并便利用户的使用。

[0167] 在说明书中说明的一系列操作能够通过硬件、软件、或者硬件与软件的组合来执行。当由软件执行该一系列操作时,可以把其中的计算机程序安装到内置于专用硬件的计算机中的存储器中,使得计算机执行该计算机程序。或者,可以把计算机程序安装到能够执行各种类型的处理的通用计算机中,使得计算机执行该计算机程序。

[0168] 例如,可以把计算机程序预先存储到作为记录介质的硬盘或者 ROM(只读存储器)中。或者,可以临时或者永久地存储(记录)计算机程序到可移动记录介质中,诸如软盘、CD-ROM(光盘只读存储器)、MO(磁光)盘、DVD(数字多功能盘)、磁盘、或半导体存储器。可以把这样的可移动记录介质作为封装软件提供。

[0169] 本发明已经参考具体实施例进行了详细说明。然而,很明显,在不背离本发明的精神的情况下,本领域技术人员能够对实施例执行更改和替换。换句话说,本发明用说明的形式公开,而不是被限制地解释。要判断本发明的要旨,应该考虑所附的权利要求。

对象 1: (reviews.cnet.com)

Manufacturer:	Canon
Part Number:	3814B004
<b>General</b>	
Product Type	Digital camera-SLR with Live View mode, with Movie recording
Width	5.8in
Depth	2.9in
Height	4.4in
Weight	1.8lbs
Body Material	Magnesium alloy
<b>Main Features</b>	
Resolution	18 megapixels
Color Support	Color
Optical Sensor Type	CMOS
Total Pixels	19,000,000 pixels
Effective Sensor Resolution	18,000,000 pixels

对象 2: (usa.canon.com)

Specifications	
Type	Digital,AF/AE single-lens reflex camera with built-in flash
Recording Media	
	CF Card Type I and II,UDMA-compliant CF cards,via external media (USB v.2.0 hard drive,via optional Wireless File Transmitter WFT-E5A)
Image Format	22.3×14.9mm(APS-C size)
Compatible Lenses	
	Canon EF lenses Including EF-S lenses (35mm-equivalent focal length is approx. 1.6x the lens focal length)
Lens Mount	Canon EF mount
Type	
	Hign-sensitivity,high-resolution,large single-plate CMOS sensor
Pixels	
	Effective pixels:Approx.18.0 megapixels
Total Pixels	Approx.19.0 megapixels

图 1

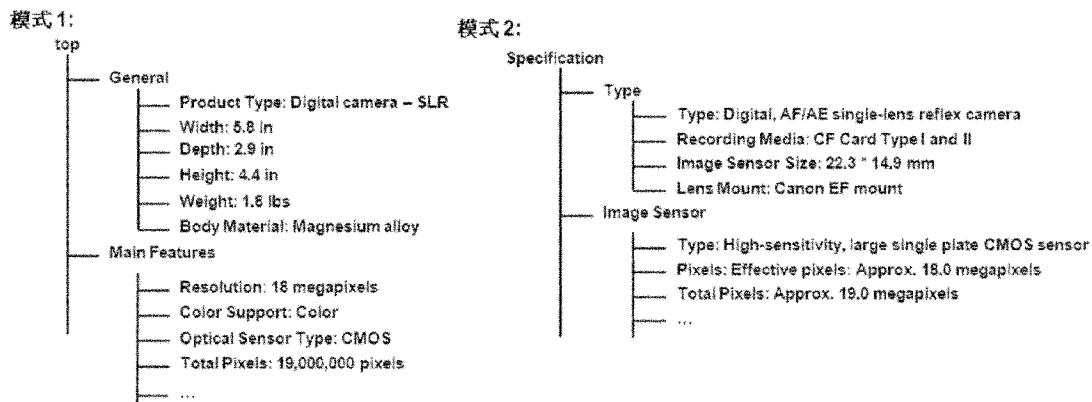


图 2

```

<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
- <attribute text="(top)">
- <attribute text="Manufacturer">
  <value text="Canon USA" />
</attribute>
- <attribute text="Part Number">
  <value text="3614B004" />
</attribute>
- <attribute text="General">
- <attribute text="Product Type">
  <value text="Digital camera - SLR with Live View mode , with Movie recording" />
</attribute>
- <attribute text="Width">
  <value text="5.8 in" />
</attribute>
- <attribute text="Depth">
  <value text="2.9 in" />
</attribute>
- <attribute text="Height">
  <value text="4.4 in" />
</attribute>
- <attribute text="Weight">
  <value text="1.8 lbs" />
</attribute>
- <attribute text="Body Material">
  <value text="Magnesium alloy" />
</attribute>
</attribute>
- <attribute text="Main Features">
- <attribute text="Resolution">
  <value text="18 megapixels" />
</attribute>
- <attribute text="Color Support">
  <value text="Color" />
</attribute>
- <attribute text="Optical Sensor Type">
  <value text="CMOS" />
</attribute>
- <attribute text="Total Pixels">
  <value text="19,000,000 pixels" />
</attribute>
- <attribute text="Effective Sensor Resolution">
  
```

图 3

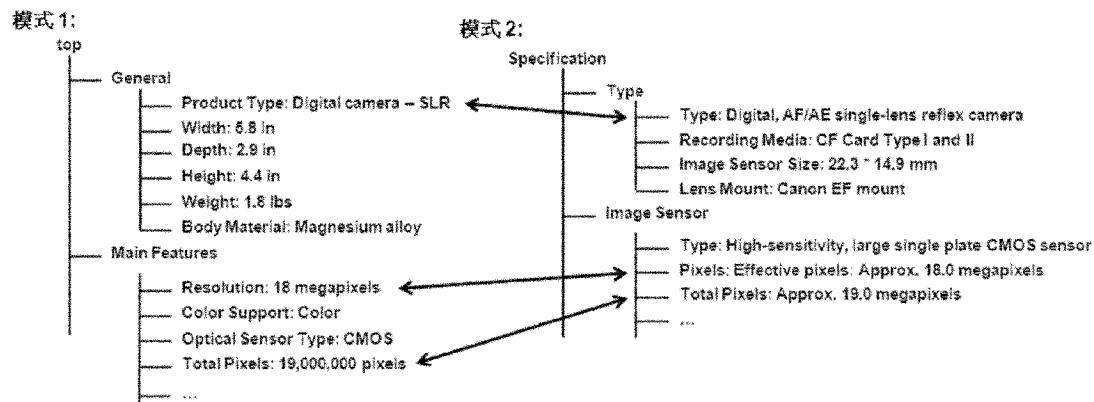


图 4

**整合后的模式:**

General		Image Sensor		Specification	
Product Type: Digital camera – SLR	Resolution: 18 megapixels	Type: Digital, AF/AE single-lens reflex camera	Type: High-sensitivity, large single plate CMOS sensor		
Dimensions	Color Support: Color	Recording Media: CF Card Type I and II	Pixels: Effective pixels: Approx. 18.0 megapixels		
Width: 5.8 in	Optical Sensor Type: CMOS	Image Sensor Size: 22.3 × 14.9 mm	Total Pixels: Approx. 19.0 megapixels		
Depth: 2.9 in	Total Pixels: 19,000,000 pixels	Lens Mount: Canon EF mount	...		
Height: 4.4 in	...	...	...		
Weight: 1.8 lbs	...	...	...		
Body Material: Magnesium alloy	...	...	...		

图 5

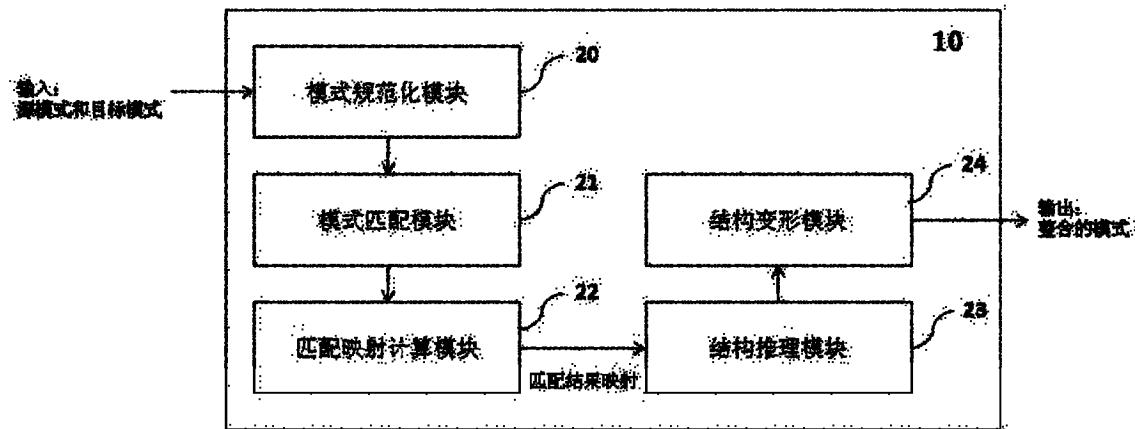


图 6

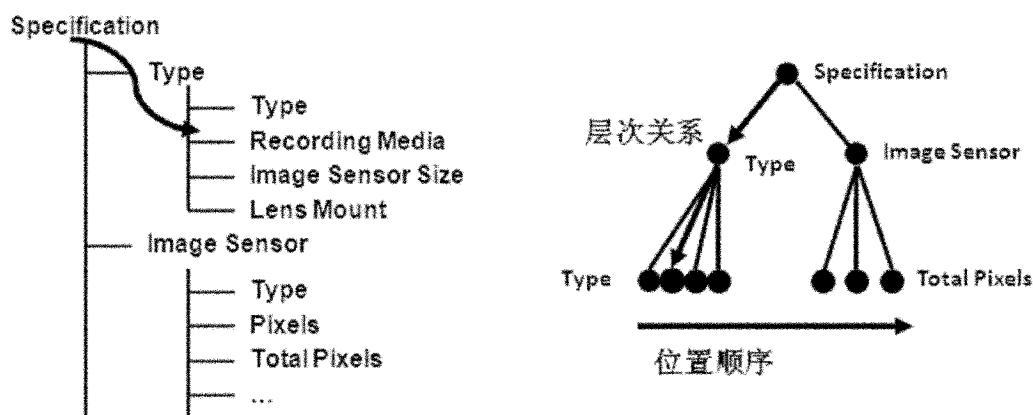


图 7

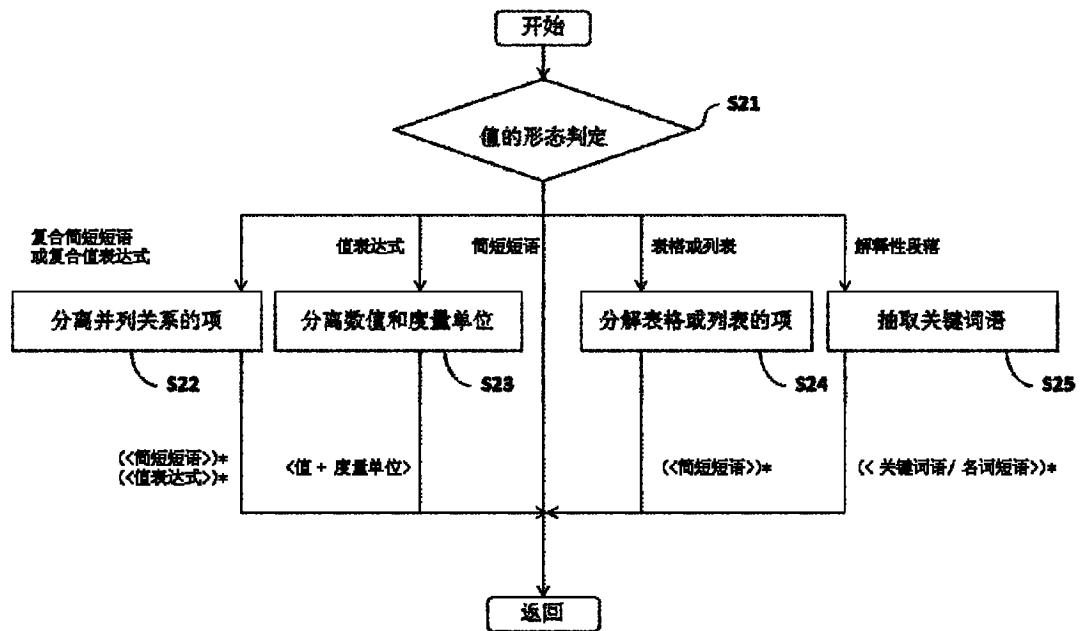


图 8

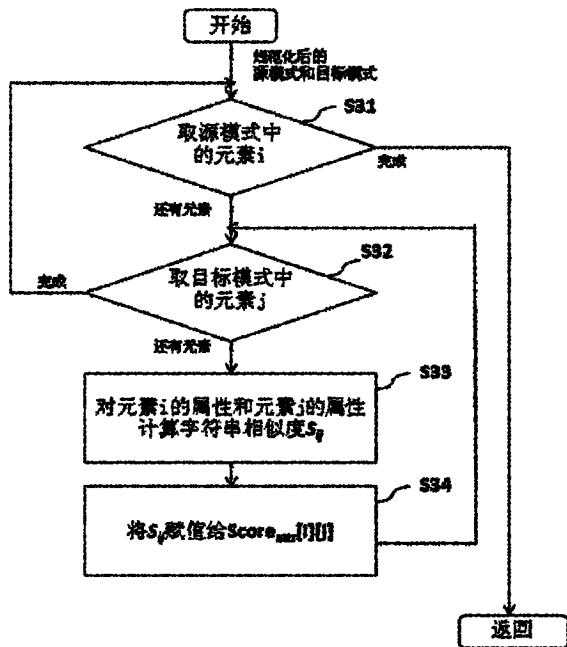


图 9

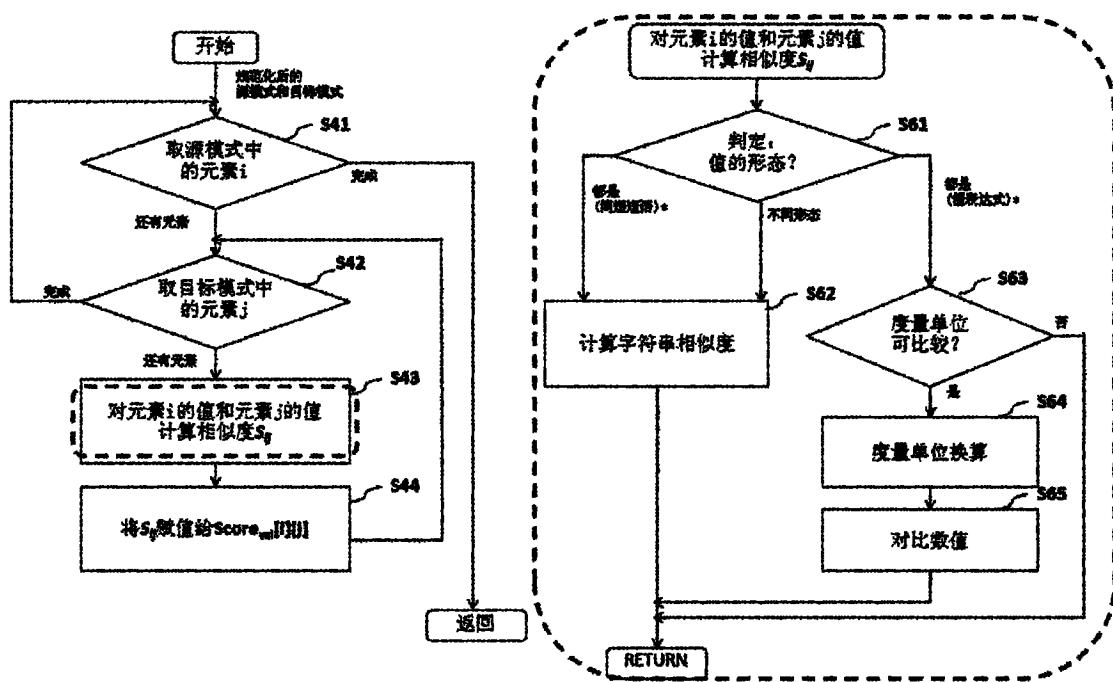


图 10

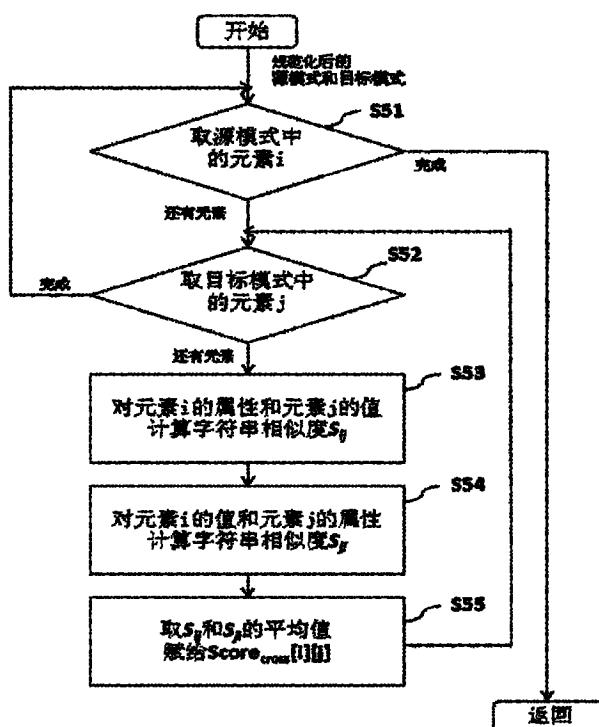


图 11

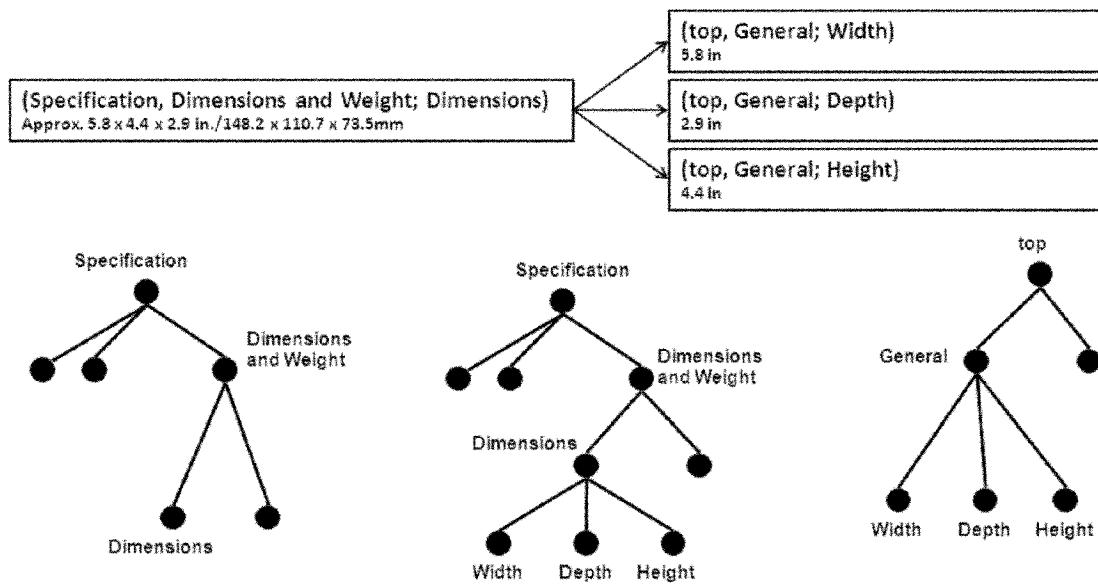


图 12

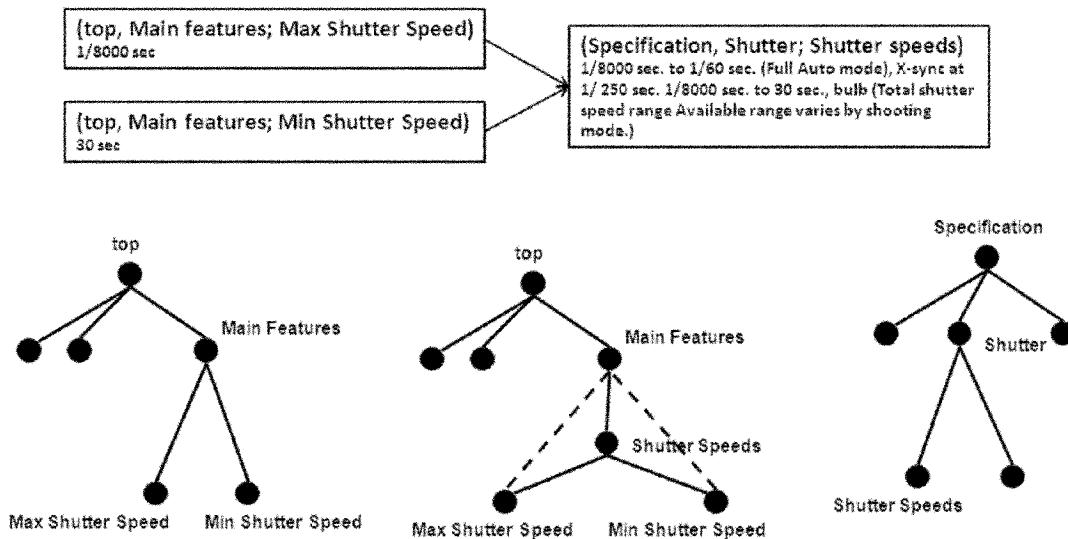


图 13

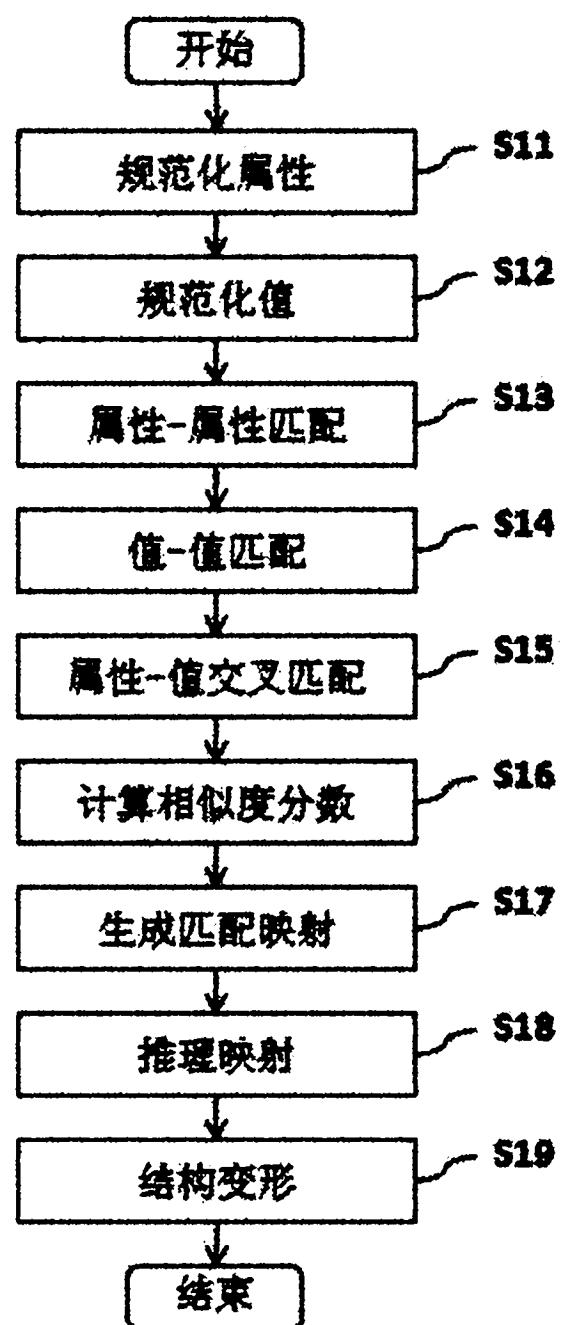


图 14

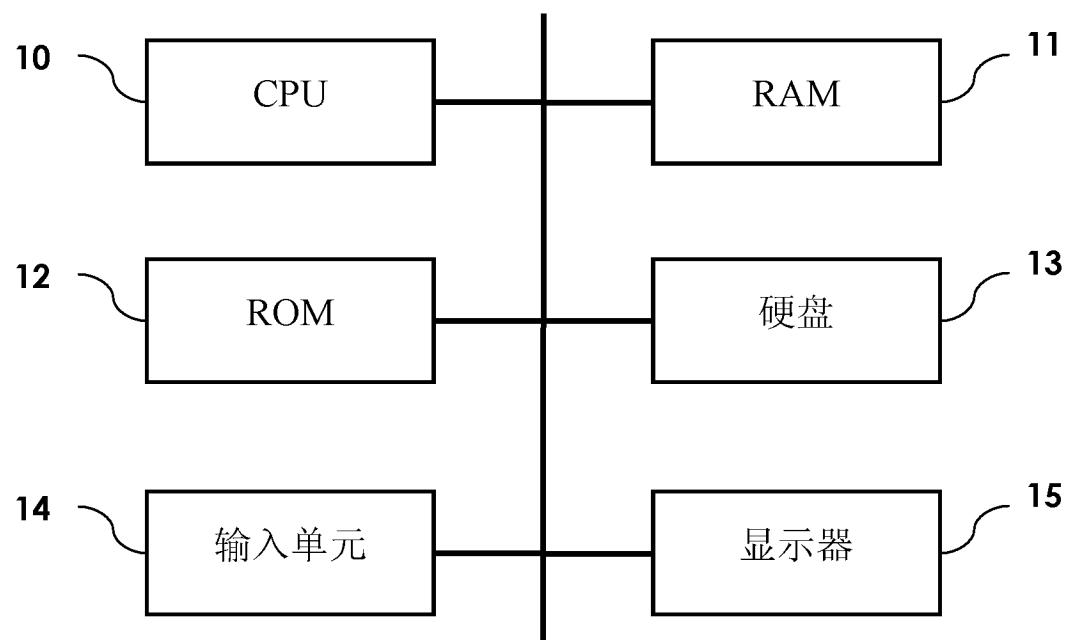


图 15