



(12) 发明专利

(10) 授权公告号 CN 1716258 B

(45) 授权公告日 2012.05.23

(21) 申请号 200510083733.7

(22) 申请日 2005.05.14

(30) 优先权数据

10/846,396 2004.05.14 US

(73) 专利权人 微软公司

地址 美国华盛顿州

(72) 发明人 J-R·文 马维英

(74) 专利代理机构 上海专利商标事务所有限公司 31100

代理人 钱慰民

(51) Int. Cl.

G06F 17/30 (2006.01)

(56) 对比文件

Jiying Wang et al. Data Extraction and Label Assignment for Web Databases. proceedings of the 12th international conference on world wide web. 2003, 187-196.

Jacob Berlin et al. Data Schema

Matching Using Machine Learning with FeatureSelection. proceedings of the 14th international conference on advanced information systems engineering. 2002, 452-466.

审查员 王雪莲

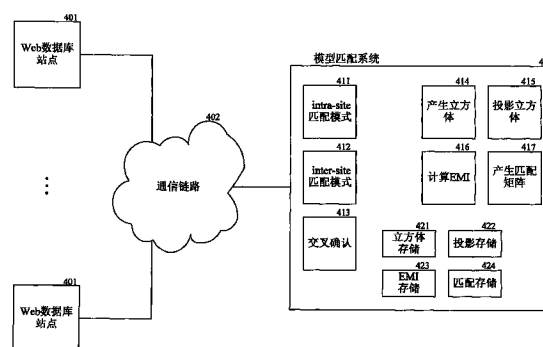
权利要求书 2 页 说明书 11 页 附图 12 页

(54) 发明名称

用于 web 数据库模式匹配的方法和系统

(57) 摘要

提供一种识别 web 数据库模式的的方法和装置。模式匹配系统生成一个 web 数据库的界面模式和结果模式之间的映射,用于表示底层数据库模式。该模式匹配系统也生成一个 web 数据库的界面属性和结果属性到全局模式的全局属性的映射,所述全局模式的语义是已知的。利用所述映射,搜索引擎服务可以用全局属性把查询公式化,映射那些查询至对应的界面属性,提交所述查询,并从对应期望的全局属性的结果属性中检索数值。



1. 一种在计算机系统中生成出现次数立方体的方法,该方法包括:

对于数据库的域每个全局属性:

对于所述数据库的每个界面属性,向所述数据库提交查询,每一查询具有所述数据库的所述界面属性的值,该值被设置为所述数据库的所述域的所述全局属性的全局属性值;
和

对于每一个提交的查询的结果,计算所述全局属性的所述全局属性值在所述结果的每一结果属性中出现的次数;以及

对于每个全局属性、界面属性和结果属性的组合,将所述次数的计数存储为所述出现次数立方体的元素,

其中,存储的元素构成了所述出现次数立方体。

2. 如权利要求 1 所述的方法,包括从出现次数立方体生成一个与全局属性和界面属性相关联的出现次数矩阵。

3. 如权利要求 1 所述的方法,包括从出现次数立方体生成一个与全局属性和结果属性相关联的出现次数矩阵。

4. 如权利要求 1 所述的方法,包括从出现次数立方体生成一个与界面属性和结果属性相关联的出现次数矩阵。

5. 一种在计算机系统中识别一个域内的数据库的属性的方法,该方法包括:

提供与所述域的全局模型的全局属性、所述数据库的界面模型的界面属性和结果模型的结果属性相关联的出现次数的计数,其中,对于每个全局属性、界面属性和结果属性的组合,每个计数表示在将该界面属性设置为该全局属性的全局属性值时,该全局属性值作为提交给所述数据库的查询的结果的结果属性的值而出现的出现次数;

基于提供的计数来估计模型对间的交互信息;和

从估计的交互信息中识别哪些属性相匹配。

6. 如权利要求 5 所述的方法,其中提供计数包括将出现次数立方体投影到与模型对相关联的出现次数矩阵中,所述与全局属性、界面属性和结果属性相关联的出现次数的计数被存储为所述出现次数立方体的元素,其中所存储的元素构成了所述出现次数立方体。

7. 如权利要求 6 所述的方法,包括通过提交查询至数据库来生成出现次数立方体,其中界面属性的值被设置为全局属性的全局属性值。

8. 如权利要求 5 所述的方法,其中基于 HTML 输入相关元素识别界面属性。

9. 如权利要求 5 所述的方法,其中利用规则表达式包装来识别结果属性。

10. 如权利要求 6 所述的方法,其中交互信息用下面的方程式估计:

$$EMI(S_{1i}, S_{2j}) = \frac{m_{ij}}{M} \log \frac{\frac{m_{ij}}{M}}{\frac{m_{i+}}{M} * \frac{m_{+j}}{M}}$$

其中,EMI 是模型 S_{1i} 的第 i 个属性和模型 S_{2j} 的第 j 个属性之间的估计的交互信息, M 是 $\sum_{i,j} m_{ij}$, m_{i+} 是 $\sum_i m_{ij}$, m_{+j} 是 $\sum_j m_{ij}$, m_{ij} 是出现次数矩阵中的元素。

11. 如权利要求 5 所述的方法,其中从估计的交互信息中识别哪些属性相匹配包括:当具有对另一模型的一个属性最高的估计的交互信息的一个模型的一个属性没有对该另一

个模型的另一属性更高的估计的交互信息时,标识出该对模型中属性间的匹配。

用于 web 数据库模式匹配的方法和系统

技术领域

[0001] 所述技术一般涉及 web 数据库模式的确定。

背景技术

[0002] 万维网 (“web”) 提供大量可经网页访问的信息。网页可以包含静态的内容或动态的内容。静态内容一般涉及停留在许多访问网页的相同交叉上的信息。动态内容一般涉及存储在一个 web 数据库中并响应搜索请求增加到网页的信息。动态内容表示被称为深层 web 或隐藏 web 的内容。

[0003] 许多搜索引擎服务允许用户搜索 web 的静态内容。当一个用户提交一个搜索请求或包括检索词的查询后, 搜索引擎服务识别可能涉及那些检索词的网页。该网页就是搜索结果。为了快速识别涉及的网页, 搜索引擎服务可以保存一个关键词到网页的映射。可以通过“爬行 (crawl)”web 生成映射以识别每一网页的关键词。为了爬行 web, 搜索引擎服务可以采用主页列表来识别通过主页接入的所有网页。任何特定网页的关键词可以用各种众所周知的信息检索技术来识别, 例如识别标题文字、网页元数据中提供的文字、突出显示的文字等等。

[0004] 然而, 这些搜索引擎一般不提供动态内容的搜索, 所述动态内容也被认为是不可爬行的内容。搜索动态内容伴随的问题是, 如果没有提供 web 数据库的网站的配合, 要直接得到相应 web 数据库的模式, 是很困难或是不可能的。一个模式定义存储在数据库中的信息或属性。例如, 一个书商的 web 数据库可以具有一个书籍目录 (即 web 数据库) 的模式, 包括每一本书的标题和作者。在不知道模式的情况下, 对搜索引擎服务来说, 通过爬行 web 数据库的内容来确定什么信息对于搜索是可用的, 是很困难的。即使 web 数据库的模式是已知的, 搜索引擎服务仍需确定怎样爬行 web 数据库来检索它的内容。假设搜索引擎可以检索 web 数据库的内容, 该搜索引擎仍需识别什么时候不同模式的属性在语法上表示相同属性。例如, 售书网站可以有一个规定该书是平装本、硬装本或光盘的目录。一个售书网站也许把该属性叫做“类型”, 另一售书网站也许会把同样的属性叫做“格式”。为了允许通过多个网站对动态内容进行有效的搜索, 搜索引擎服务需要知道 web 数据库属性的含义或语义。

[0005] 需要这样一项技术, 可以自动识别相应 web 数据库的模式, 并识别表示相同语义内容的不同模式的属性。

发明内容

[0006] 提供一种识别 web 数据库模式的方法和系统。模式匹配系统生成一个 web 数据库的界面模式和结果模式之间的映射, 用来表示底层数据库模式。所述模式匹配系统也生成一个 web 数据库的界面属性和结果属性到全局模式的全局属性的映射, 所述全局模式的语义是已知的。利用这些映射, 搜索引擎服务可以用全局属性把查询公式化, 映射那些查询至相应的界面属性, 提交所述查询, 并从对应期望全局属性的结果属性中检索数值。

[0007] 附图简介

[0008] 图 1 表示一个书商的 web 数据库的各种模式的框图。

[0009] 图 2 表示一个实施例中的内部站 (intra-site) 匹配和相互站 (inter-site) 匹配。

[0010] 图 3 表示一个实施例中模式匹配系统的分区的一个通道。

[0011] 图 4 表示一个实施例中模式匹配系统部件方框图。

[0012] 图 5 表示一个实施例中 intra-site 匹配部件处理过程的流程图。

[0013] 图 6 表示一个实施例中立方体生成部件处理过程的流程图。

[0014] 图 7 表示一个实施例中立方体更新部件处理过程的流程图。

[0015] 图 8 表示一个实施例中立方体投影部件处理过程的流程图。

[0016] 图 9 表示一个实施例中 EMI 计算部件处理过程的流程图。

[0017] 图 10 表示一个实施例中匹配矩阵生成部件处理过程的流程图。

[0018] 图 11 表示一个实施例中 inter-site 匹配部件的处理过程的流程图。

[0019] 图 12 表示一个实施例中计算估计向量相似性的部件处理过程的流程图。

[0020] 图 13 表示一个实施例中交叉确认 (cross-validate) 部件处理过程的流程图。

具体实施方式

[0021] 提供一种识别 web 数据库模式的方法和系统。在一个实施例中,模式匹配系统生成一个 web 数据库的界面模式和结果模式之间的映射,用来表示基础的数据库模式。Web 数据库的界面模式表示可用于搜索的数据库属性。Web 数据库的结果模式表示作为搜索结果部分显示的数据库属性。所述映射指示哪些界面属性与哪些结果属性具有相同的(也涉及相当的或相匹配的)含义。模式匹配系统也生成 web 数据库中的界面属性和结果属性到全局模式中的全局属性的映射,所述全局属性的语义是已知的。利用这些映射,搜索引擎服务可以用全局属性把查询公式化,映射所述查询至相应的界面属性,提交所述查询,并从对期望全局属性的结果属性中检索数值。这样,该模式匹配系统识别 web 数据库的模式,该模式可用于搜索 web 数据库。

[0022] 图 1 表示一个书商的 web 数据库的各种模式的框图。该 Web 数据库包括一个数据库模式 101,一个界面模式 102 和一个结果模式 103。该数据库模式表示 web 数据库的底层模式,在本例中包括属性标题、作者、出版商、ISBN 号、格式和出版日期。网站提供一个搜索网页,以便用户可以访问来查找书籍。该 web 数据库的界面模式包括属性标题、作者、格式和 ISBN 号。用户可以为任何界面属性的组合指定搜索字符串,来搜索书籍数据库。网页中“你的搜索”字段允许用户在 web 数据库的所有属性中进行搜索。搜索的结果显示在一个结果网页上。该 web 数据库的结果模式包括标题、作者、出版商、格式和出版日期。搜索结果典型地提供多个多个条目,这些条目对应于与搜索请求匹配的数据库的每个条目。结果的每一条目一般包含每个结果属性的一个值。在本例中,所述界面模式有一个未包括在结果模式中的属性(即 ISBN 号),结果模式有一个未包括在界面模式中的属性(即出版日期)。

[0023] 除了使用 web 数据库的界面模式和结果模式外,模式匹配系统还使用一个特殊区域(domain-specific)的全局模式。一个区域的全局模式表示在该区域内 web 数据库通常使用的属性设置。例如,在书籍区域内的 web 数据库典型地具有包括标题、作者和出版商的

属性,并且在汽车区域内的 web 数据库典型地具有包括构造、型号和年代的属性。一个全局模式也可以具有与其对应的样本全局属性值。例如,书籍区域的出版商属性可以具有包括“Random House”和“MITPress”的全局属性值。

[0024] 为了生成映射,该模式匹配系统一开始就识别 web 数据库的区域全局模式,和 web 数据库的界面模式和结果模式。(识别所述模式的技术在下面进行描述)。模式匹配系统从全局属性的全局属性值(例如,从一组样本值)生成查询,并通过界面网页向 web 数据库提交查询(例如,通过搜索网页发送一个对应提交查询的 HTTP 请求)。模式匹配系统分析结果网页给出的结果,来确定那些界面属性对应哪些结果属性(“界面-结果对应”),哪些全局属性对应哪些界面属性(“全局-界面对应”),哪些全局属性对应哪些结果属性(“全局-结果对应”)。因为界面和结果模式对应单个网站的模式,因此该对应关系称为“内部站(intra-site)”匹配。当搜索时使用的界面属性的值与结果属性的值相匹配时,模式匹配系统识别一个界面属性可以对应一个结果属性。例如,当给出标题界面属性的值为“Harry Potter”时,许多结果条目都可能在标题界面属性里具有“Harry Potter”值。相反,当给出作者界面属性的值为“Harry Potter”来搜索时,只有少数的结果条目可能在标题界面属性里具有“Harry Potter”值。同样的,所述标题界面属性可能对应标题结果属性,但是作者界面属性可能不对应标题结果属性。

[0025] 在一个实施例中,该模式匹配系统也可以产生不同网站的界面模式和结果模式之间的映射。模式匹配系统分析如上述提交的查询结果,并识别某一网站模式的哪个界面属性对应另一网站模式的哪个界面属性(“界面-界面对应”),和某一网站模式的哪个结果属性对应另一网站模式的哪个结果属性(“结果-结果对应”)。例如,该模式匹配系统可以识别,某一网站的类型(type)界面属性可以对应另一网站的格式(format)界面属性。因为是在不同网站之间进行模式的匹配,所以该对应关系称为“相互站(inter-site)”匹配。该 inter-site 匹配信息可以在一个区域中搜索多个 web 数据库时使用。所述 inter-site 匹配信息也可以用来帮助确认 intra-site 匹配是否正确。

[0026] 图 2 表示一个实施例中的 intra-site 匹配和 inter-site 匹配。椭圆 202 表示关于在一个书籍区域内的 web 数据库的模式。每个站点 1...N 具有一个界面模式(“IS”)和一个结果模式(“RS”),并且所述区域具有一个全局模式(“GS”)。模式间的连线表示“intra-site”匹配和“inter-site”匹配。例如,站点 1 的 IS 和 GS 之间的线表示 intra-site 的全局-界面对应,站点 1 的 IS 和站点 1 的 RS 之间的线表示 intra-site 的界面-结果对应,站点 1 的 IS 和站点 2 的 IS 之间的线表示站点 1 和站点 2 间 inter-site 的界面-界面对应。

[0027] 在一个实施例中,模式匹配系统生成一个事件立方体(occurrence cube),对应 web 数据库的全局属性、界面属性和结果属性的每一组合,当搜索时全局属性值被用作该界面属性值时,该立方体识别该全局属性的全局属性值在该结果属性中出现的次数。模式匹配系统为每一界面属性提交多个查询。每个查询具有设为不同全局属性值的该界面属性值。例如,如果全局属性包括一个附带平装本、硬装本和光盘值的格式属性,和一个附带 Rowling 值的作者属性,那么模式匹配系统提交一个附带设为平装本的标题属性的查询,一个附带设为硬装本的标题属性的查询,一个附带设为光盘的标题属性的查询,和一个附带设为 Rowling 的标题属性的查询。该模式匹配系统为每个界面属性提交平装本、硬装本、光

盘和 Rowling 的全局属性值的查询。该模式匹配系统为每一查询结果计数查询的全局属性值作为每一结果属性值出现的次数。例如,当提交一个带有设为平装本标题界面属性的查询时,可能只有很少或没有匹配,这表示该标题界面属性很可能与格式全局属性不匹配。相反,当提交一个带有设为平装本格式界面属性的查询时,可能找到许多匹配,且可以在格式结果属性内许多结果条目中找到搜索项“平装本”,这表示格式全局属性、格式界面属性和格式结果属性是相互对应的。特定的全局属性、界面属性和结果属性组合尤其是相对于其它的组合的高次数,表示所述属性很可能是对应的,即它们表示相同的语义内容。

[0028] 生成事件立方体后,模式匹配系统创建用于全局-界面对应、全局-结果对应和界面-结果对应的事件矩阵。在一个实施例中,模式匹配系统通过把事件立方体的维数投影到一个平面上,来创建一个事件矩阵。为了生成全局-界面对应的事件矩阵,模式匹配系统把每一全局属性和界面属性的组合的所有结果属性出现的次数相加。该模式匹配系统以类似的方式生成全局-结果对应和界面-结果对应的生成事件矩阵。表 1 是一个全局-界面对应的事件矩阵的例子。

[0029] 表 1

[0030]		标题 _{GS}	作者 _{GS}	出版商 _{GS}	ISBN _{GS}
[0031]	作者 _{IS}	93	498	534	0
[0032]	标题 _{IS}	451	345	501	0
[0033]	出版商 _{IS}	62	184	468	2
[0034]	关键词 _{IS}	120	248	143	275
[0035]	ISBN _{IS}	0	0	0	258

[0036] 虽然计算的值是成对属性之间对应关系的一个指示,但是相对值比绝对值更能指示匹配关系。尤其是,一个高出现次数不一定表示对应的属性。例如,作者_{IS}和出版商_{GS}的矩阵元素(534)在矩阵中是最大值,但是作者_{IS}和出版商_{GS}在语义上不是相互对应的。通常,给出一个特定矩阵元素 m_{ij} ,在界面属性 i 和全局属性 j 的所有元素中,它的相对值比它的绝对值更加重要。例如,可能包括“你的搜索”字段,并且不是书籍区域的真正属性的关键词_{IS},具有一个与全局属性相似的特性,表示其不可以与任何一个全局属性很好得匹配。出版商_{IS}和出版商_{GS}(468)的元素不是出版商_{GS}元素中最高的。然而,与其它出版商_{IS}的元素相比,它较大。

[0037] 为了识别哪一对属性是对应的,所述模式匹配系统估计属性对的一个交互信息内容。交互信息内容也可以被称为交叉熵(cross-entropy)和信息增益(information gain)。所述模式匹配系统根据模式的属性,假设每一模式代表一个 web 数据库的分区。来自分区重叠最大的不同模式的成对属性可能对应。在一个实施例中,所述模式匹配系统根据下列方程式估计一对属性间的交互信息:

$$[0038] \quad \text{EMI}(S_{1i}, S_{2j}) = \frac{m_{ij}}{M} \log \frac{\frac{m_{ij}}{M}}{\frac{m_{i+}}{M} * \frac{m_{+j}}{M}} \quad (1)$$

[0039] 其中 EMI 是估计的模式 S_{1i} 的第 i 属性和模式 S_{2j} 的第 j 属性之间的交互信息, M 是 $\sum_{i,j} m_{ij}$, m_{i+} 是 $\sum_i m_{ij}$, m_{+j} 是 $\sum_j m_{ij}$ 。

[0040] 表 1 事件矩阵的 EMI 矩阵如表 2 所示。

[0041] 表 2

[0042]	标题 GS	作者 GS	出版商 GS	ISBN GS
[0043] 作者 IS	-0.04	0.11	0.06	0.00
[0044] 标题 IS	0.19	-0.03	-0.01	0.00
[0045] 出版商 IS	-0.03	-0.02	0.14	-0.01
[0046] 关键词 IS	-0.01	0.01	-0.07	0.17
[0047] ISBN IS	0.00	0.00	0.00	0.32

[0048] 当一个 EMI 矩阵元素大于其它相同界面属性的元素 (即在同一行中), 并且也大于其它相同全局属性的元素 (即在同一列中) 时, 所述模式匹配系统检测属性间的匹配关系。如长方形所示的, 相应的属性彼此之间在信息内容上, 比与相反模式的其它属性具有更大重叠。例如, 对应作者界面属性和作者全局属性, 作者 IS 和作者 GS 的 EMI 矩阵元素 (即 0.11) 是最大的, 并是正确的匹配。属性的匹配由下列方程表示:

$$[0049] \quad \text{MAP}(S_{1i}, S_{2j}) = \text{match} \text{ between } i, j \geq e_{ik}/k \neq j \text{ and } e_{ij} \geq e_{ik} \text{ } k \neq i \quad (2)$$

[0050] 其中, MAP 表示模式 S_1 的第 i 属性和模式 S_2 的第 j 属性是否相匹配, e_{ij} 是对应模式 S_1 的第 i 属性和模式 S_2 的第 j 属性的 EMI 矩阵元素。

[0051] 在一个实施例中, 所述模式匹配系统识别不同 web 数据库的属性间的匹配关系。该模式匹配系统基于 web 数据库的对应事件矩阵的向量之间的相似性, 来识别匹配关系。例如, 表 3 表示模式 S_1 的全局 - 界面事件矩阵, 表 4 表示模式 S_2 的全局 - 界面事件矩阵。全局模式 GS 是 {标题, 作者, 出版商, ISBN}, 站点 1 的界面模式 IS_1 是 {作者₁, 题目₁, 出版商₁, 关键词₁, ISBN₁}, 站点 2 的界面模式 IS_2 是 {题目₂, 作者₂, ISBN₂}。

[0052] 表 3

[0053]	T_G	A_G	P_G	I_G
[0054] A_1	93	498	534	0
[0055] T_1	451	345	501	0
[0056] P_1	62	184	468	2
[0057] K_1	120	248	143	275
[0058] I_1	0	0	0	258

[0059] 表 4

[0060]	T_G	A_G	P_G	I_G
[0061] T_2	166	177	118	0
[0062] $A_2(P)$	39	331	406	0
[0063] I_2	0	0	0	18

[0064] 属性 $A1$ 由表 3 中第一行的向量表示, 属性 $A2$ 由表 4 中第二行的向量表示。该模式匹配系统用下列方程来计算两属性间的相似性:

$$[0065] \quad EVS(S_{1i}, S_{2j}) = \frac{\sum_k a_{ik} b_{jk}}{\sqrt{\sum_k a_{ik}^2} * \sqrt{\sum_k b_{jk}^2}}$$

(3)

[0066] 其中 EVS 是模式 S_1 的第 i 属性和模式 S_2 的第 j 属性间的估计向量相似性，

[0067] a_{ik} 表示模式 S_1 的事件矩阵的值， b_{jk} 表示模式 S_2 的事件矩阵的值。

[0068] 表 5 表示从表 3 和表 4 中导出的估计向量相似性。

[0069] 表 5

[0070]		T2	A2 (P)	I2
[0071]	A_1	0.84	0.99	0
[0072]	T_1	0.96	0.84	0
[0073]	P_1	0.71	0.95	0.01
[0074]	K_1	0.72	0.67	0.66
[0075]	I_1	0	0	1.00

[0076] 当一个 EVS 矩阵元素大于同一网站的相同界面属性的其它元素，并也大于其它网站的相同界面属性的其它元素时，所述模式匹配系统检测属性间的匹配关系。表 5 中的长方形表示行和列中的最大相似性值，也表示正确的匹配。虽然 IS_2 ，作者₂的第二属性不能与 GS 的出版商₂正确匹配，但是模式匹配系统运用 inter-site 匹配来校正所述匹配。

[0077] 在一个实施例中，所述模式匹配系统交叉确认 (cross-validates) 全局 - 界面对应、全局 - 结果对应、界面 - 结果对应、界面 - 界面对应和结果 - 结果对应，以便识别和校正可能错误的匹配。该模式匹配系统基于匹配的全局属性将界面属性（和类似的结果属性）集群为多个群集 (cluster)。例如，与某一个全局属性匹配的各种 web 数据库的属性表示一个群集。该群集是基于 intra-site 匹配的。inter-site 匹配也可以用于交叉确认群集。如果 intra-site 和 inter-site 匹配完全正确，那么 web 数据库的每一属性将只映射到在相同群集中其它 web 数据库的属性。换句话说，web 数据库的属性一向是彼此映射，且被映射到全局属性。在一个实施例中，模式匹配系统表示 web 数据库模式的属性为顶点，inter-site 匹配为顶点之间的边缘。模式匹配系统划分顶点，这样使边缘切割 (edge-cut) 最小化。所述边缘切割 (edge-cut) 是分区间所有边缘加权和（例如，每一边缘具有相同的权）。通过最小化边缘切割 (edge-cut)，模式匹配系统使不同群集的顶点之间的边缘数目最小化。

[0078] 在一个实施例中，使用初始群集作为初始分区，且只要切割的数目降低就把顶点从一个群集移动到另一群集，模式匹配系统将边缘切割 (edge-cut) 近似于最小化。通常，一个顶点被移动至其中有它大多数相邻顶点的群集。相邻的顶点之间具有一个边缘。因为如果多个临近顶点发生移动，这个顶点就需要移动，那么模式匹配系统可以使用多通道以便边缘切割 (edge-cut) 集中到局部最佳。当边缘切割 (edge-cut) 集中时，模式匹配系统分解交叉的群集，该群集通过放弃交叉群集匹配在站点 S_1 的 A_i 属性和站点 S_2 的 B_i 属性之间匹配，所述站点 S_2 包含在两个群集 C_1 和 C_2 中，并再匹配 A_i 到站点 S_2 的 B_k 属性，所述站点

S_2 群集到 C_1 , 反之亦然。

[0079] 图 3 表示一个实施例中模式匹配系统的分区的一个通道。在该例中, 全局模式包含两个属性 { 作者, 出版商 }, 5 个 web 数据库包含 IS 属性 $IS_1 = \{Aa\}$, $IS_2 = \{Ba, Bp\}$, $IS_3 = \{Ca, Cp\}$, $IS_4 = \{Da, Dp\}$ 和 $IS_5 = \{Ea, Ep\}$ 。群集 301 和 302 基于其匹配 (intra-site 匹配) 的全局属性来指示属性 (用顶点表示) 的初始群集, 成对属性之间的边缘指示属性匹配 (inter-site 匹配)。在初始状态, Aa 不能与出版商全局属性正确匹配, 也不能与 Bp 正确匹配, 但是它在作者范畴里已经与其它三个属性正确匹配。因此, 该模式匹配系统将交叉群集边缘数目从 3 减少到 1 来移动 Aa。该移动把 Aa 的属性从出版商校正到作者全局属性。该移动之后, 模式匹配系统消除 Aa 和 Bp 之间的边缘, 并增加 Aa 和 Ba 之间新的边缘 (站点 2 的属性匹配到作者全局属性)。群集 311 和 312 表示校正后的对应。

[0080] 全局模式、界面模式和结果模式可以用各种技术来识别。一些识别全局模式的技术依靠属性的名字和元素的结构。(参见 S. Castano, V. Antonellis, and S. Vimercati. Global Viewing of Heterogeneous Data Sources. IEEE Trans. Data and Knowledge Eng., vol. 13, no. 2, 2001; 和 B. He, and C. C. Chang. Statistical Schema Matching across Web Query Interfaces. Proc. ACM SIGMOD conf., 2003, 这里引入作为参考) 其它技术依靠形式本体论 (参见 B. He, and C. C. Chang. Statistical Schema Matching across Web Query Interfaces. Proc. ACM SIGMOD Conf., 2003; 和 F. Hakimpour, and A. Geppert. Global Schema Generation Using Formal Ontologies. Proc. 21st Conf. on Conceptual Modeling, 2002, 这里引入作为参考) 全局属性值样本可以从各种 web 数据库样本中收集或手动产生。Web 数据库的界面模式可以从作为 HTML 说明的查询网页的输入关联 (input-related) 标签中识别。(参见 A. Arasu, and H. Garcia-Molina. Extracting Structured Data from Web Pages. Proc. ACM SIGMOD Conf., 2003; C. H. Chang, and S. C. Lui. IEPAD: Information Extraction based on Pattern Discovery. Proc. 10th World Wide Web Conf., 681-688, 2001; V. Crescenzi, G. Mecca and P. Merialdo. ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites. Proc. 27th VLDB. Conf., 109-118, 2001; 和 J. Wang and F. Lochovsky. Data Extraction and Label Assignment for Web Databases. Proc. 12th World Wide Web Conf., 187-196, 2003, 这里引入作为参考) 一项技术基于嵌套在 HTML 页面中的重复图案 (repeated-pattern) 来生成一个规则表达式 (regular-expression) (参见 J. Wang and F. Lochovsky. Data Extraction and Label Assignment for Web Databases. Proc. 12th World Wide Web Conf., 187-196, 2003, 这里引入作为参考) 本领域普通技术人员可以理解, 每一模式也可以手动或结合手动来识别, 也可以用自动化设备识别。

[0081] 图 4 表示一个实施例中模式匹配系统的部件的方框图。模式匹配系统 410 通过通信链路 402 连接到多个 web 数据库站点 401。该模式匹配系统包括一个内部站 (intra-site) 部件 411, 一个相互站 (inter-site) 匹配部件 412, 一个交叉确认 (cross-validate) 部件 413, 一个立方体生成部件 414, 一个立方体投影部件 415, 一个 EMI 计算部件 416, 和一个匹配矩阵生成部件 417。该模式匹配系统还包括一个立方体存储器 421, 一个投影存储器 422, 一个 EMI 存储器 423, 和一个匹配存储器 424。intra-site 匹配部件调用立方体生成部件来生成一个事件立方体, 并调用立方体投影部件来生成全局 - 界面、全局 - 结果和界面 - 结果

事件矩阵。intra-site 匹配部件也调用 EMI 计算部件基于事件矩阵计算估计交互信息,并调用匹配矩阵生成部件来识别哪个属性对是匹配的。cross-validate 部件改变没有正确匹配的属性间的匹配。立方体存储器包含事件立方体,投影存储器包含事件矩阵,EMI 存储器包含 EMI 矩阵,匹配存储器包含匹配矩阵。

[0082] 模式匹配系统上的计算装置包括一个中央处理单元、内存、输入装置(例如,键盘和点击设备)、输出设备(例如,显示设备)和存储设备(例如磁盘驱动器)。所述内存和存储设备是计算机可读介质,可以包含实现所述模式匹配系统的指令。另外,数据结构和消息结构可被存储或通过数据传输媒体被发送,例如作为通信链路上的信号。可以采用各种通信链路,如因特网、局域网、广域网或点对点的拨号连接。

[0083] 该模式匹配系统可以在各种操作环境下实现,包括个人计算机、服务器计算机、手提或膝上计算机、多处理器系统、微处理器系统、可编程电子消费装置、网络 PC、小型计算机、大型计算机、包括任何上述系统或设备分布式计算环境,等等。

[0084] 该模式匹配系统在上下文中可以概括地描述为计算机可执行指令,例如由一个或更多计算机或其它装置执行程序模块。通常,程序模块包括例程、程序、对象、分量、数据结构等等,可执行特定的任务或实现特定的抽象数据类型。典型地,在各种实施例中程序模块的功能可以按期望的进行结合或分配。

[0085] 图 5 表示一个实施例中 intra-site 匹配部件处理过程的流程图。该部件识别 web 数据库的全局-界面、全局-结果和界面-结果的对应关系。在方框 501 中,该部件调用立方体生成部件来生成事件立方体。在方框 502-506 中,该部件循环选择成对的模式(即全局和界面,全局和结果,界面和结果),并生成表示每对对应关系的匹配矩阵。在方框 502 中,该部件选择下一对模式。在判别方框 503 中,如果所有的成对模式都已经选定,那么该部件工作完成,否则该部件继续到方框 504。在方框 504 中,该部件调用立方体投影部件来生成选定成对模式的事件矩阵。在方框 505 中,该部件调用 EMI 计算部件,来估计选定成对模式的成对属性之间的交互信息。在方框 506 中,该部件调用匹配矩阵生成部件,来生成指示选定成对模式的属性对应关系的匹配矩阵。然后该部件循环至方框 502,选择下一对模式。

[0086] 图 6 表示一个实施例中立方体生成部件处理过程的流程图。该部件基于全局模式、界面模式和结果模式,生成一个 web 数据库的事件立方体。一个事件立方体是一个三维矩阵,它把每一全局属性、界面属性和结果属性的组合映射为计数。该计数是查询的结果条目在该结果属性中具有该全局属性值的次数,其中查询具有被设置为该全局属性的全局属性值的该结果属性。在方框 601 中,该部件选择下一全局属性。在判定方框 602 中,如果所有的全局属性都已经选定,该部件返回,否则该部件继续到方框 603。在方框 603 中,该部件选择已选定全局属性的下一全局属性值。在判定方框 604 中,如果所有的已选定全局属性的全局属性值都已经选定,该部件循环至方框 601 来选择下一全局属性。否则该部件继续到方框 605。在方框 605-609 中,该部件循环选择每一界面属性并提交一个带有设为选定全局属性值的那个界面属性的查询。本领域普通技术人员可以理解,一些界面属性的值域是可以有限制的。例如,如果一个界面属性是用 HTML SELECT 元素来表示的,那么它的值域可以限制为与 OPTION 元素相关的值。在这种情况下,该部件可以只提交用于与选择(option)值“相似”的全局属性值的查询。如果一个全局属性值包含一个 option 值,可以认为它是相似的。本领域普通技术人员可以理解,也可以采用其它衡量相似性的方法。CHECKBOX 和

RADIOBOX 元素的查询可以按类似的方式处理。因为 TEXTBOX 的值域可能未知,该部件可以完全提交所有使用 TEXTBOX 表示界面属性的全局属性值的查询。在一个实施例中,该部件只为每一查询的一个界面属性设置数值。其它界面属性的值可以具有一个网站定义的默认值。在方框 605 中,该部件选择下一界面属性。在判定方框 606 中,如果所有界面属性已经选定,该部件循环至方框 603 来为选定的全局属性选择下一全局属性值。在方框 607 中,该部件利用选定的界面属性和选定的全局属性值,来公式化查询。在方框 608 中,该部件将公式化的查询提交至网站。在方框 609 中,该部件基于查询结果更新事件立方体,并循环至方框 605 来选择下一界面属性。

[0087] 图 7 表示一个实施例中立方体更新部件处理过程的流程图。该部件传送一个全局属性、一个全局属性值、一个界面属性和一个查询结果的指示。在方框 701 中,该部件选择结果的下一条目或行。在判定方框 702 中,如果所有结果条目都已经选定,该部件返回,否则该部件继续到方框 703。在方框 703 中,该部件选择下一结果属性或列。在判定方框 704 中,如果所有的结果属性已经选定,该部件循环至方框 701 来选择下一结果条目,否则该部件继续到方框 705。在方框 705 中,如果全局属性值等于选定条目的选定结果属性值,那么该部件继续到方框 706,否则该部件循环至方框 703 来选择下一选定条目的结果属性。在方框 706 中,该部件在通过全局属性、通过界面属性和选定结果属性的事件立方体内增加计数。然后该部件循环至方框 703 来选择下一选定条目的结果属性。

[0088] 图 8 表示一个实施例中立方体投影部件处理过程的流程图。在该实施例中,该部件生成全局-界面对应的事件矩阵。所述模式匹配系统可以以相似的方式生成全局-结果对应和界面-结构对应的事件矩阵。在该实施例中,该部件对全局属性的结果属性和界面属性对的结果求和,来将事件立方体的三维投影为相应矩阵的二维。本领域普通技术人员可以理解,除直接求和外还可以采用其它投影技术。例如,该部件可以采用加权求和,其中加权是基于在结果模式的自动识别期间得到的可信度。在方框 801 中,该部件选择下一全局属性。在判定方框 802 中,如果所有的全局属性都已经选定,该部件返回,否则该部件继续到方框 803。在方框 803 中,该部件选择下一界面属性。在判定方框 804 中,如果所有的界面属性已经选定,该部件循环至方框 801 来选择下一全局属性,否则该部件继续到方框 805。在方框 805 中,该部件选择下一结果属性。在判定方框 806 中,如果所有的结果属性已经选定,该部件循环至方框 803 来选择下一界面属性,否则该部件继续到方框 807。在方框 807 中,该部件将选定界面属性和结果属性的事件立方体中的计数加上选定全局属性,界面属性和全局属性的事件矩阵中的计数。然后该部件循环至方框 805 来选择下一结果属性。

[0089] 图 9 表示一个实施例中 EMI 计算部件处理过程的流程图。该部件采用方程式 1 估计在一个事件矩阵中成对属性的交互信息。本领域普通技术人员可以理解,可以采用各种技术来估计成对属性匹配的似然性。该部件传送一个事件矩阵并返回 EMI 矩阵。在方框 901 中,该部件计算事件矩阵内所有计数的和。在方框 902 中,该部件计算事件矩阵的每一行内计数的和。在方框 903 中,该部件计算事件矩阵的每一列内计数的和。在方框 904-908 中,该部件循环选择事件矩阵每一对事件矩阵属性,并确定属性匹配的似然性。在方框 904 中,该部件选择事件矩阵的下一行。在判定模块 905 中,如果事件矩阵的所有行都已经选定,那么该部件返回,否则该部件继续到方框 906。在方框 906 中,该部件选择事件矩阵的下一列。在判定方框 907 中,如果事件矩阵的所有列都已经选定,那么该部件循环至方框 904 来选择

事件矩阵的下一行,否则该部件继续到方框 908。在方框 908 中,该部件计算由选定的行和列来表示的属性的估计交互信息。然后该部件循环至方框 906 来选择下一列。

[0090] 图 10 表示一个实施例中匹配矩阵生成部件处理过程的流程图。该部件传送一个矩阵,例如 EMI 矩阵,该矩阵指示成对属性匹配的似然性。如果一对属性的似然性是双属性中最高的(例如,在表示一个属性的行中最高的和表示其它属性的列中最高的),该部件找出所述的属性匹配。在方框 1001 中,该部件选择传送矩阵(passed matrix)的下一行。在判定方框 1002 中,如果传送矩阵(passed matrix)的所有行都已经选定,那么该部件返回,否则该部件继续到方框 1003。在方框 1003,该部件选择传送矩阵(passed matrix)的下一列。在判定方框 1004 中,如果传送矩阵(passed matrix)的所有列都已经选定,那么该部件循环至方框 1101 来选择传送矩阵(passed matrix)的下一行,否则该部件继续到方框 1005。在判定方框 1005 中,如果选定行和列的值是该行内最高的,那么该部件继续到方框 1006,否则该部件循环至方框 1003 来选择下一列。在判定方框 1006 中,如果选定行和列的值是该列内最高的,那么该部件继续到方框 1007,否则该部件循环至方框 1003 来选择下一列。在方框 1007 中,该部件设定选定行和列的匹配矩阵的值,来指示匹配关系,然后循环至方框 1003 来选择选定行的下一列。

[0091] 图 11 表示一个实施例中 inter-site 匹配部件处理过程的流程图。该部件识别一个网站的哪一属性(界面和结果)与另一网站的哪一属性相匹配。该部件使用网站的全局-界面对应的事件矩阵,来识别界面模式的匹配,使用网站的全局-结果对应的事件矩阵,来识别结果模式的匹配。在方框 1101 中,该部件调用立方体生成部件来生成站点 A 的事件立方体。在方框 1102 中,该部件调用立方体投影部件来生成站点 A 的事件矩阵。在方框 1103 中,该部件调用立方体生产部件来生成站点 B 的事件立方体。在方框 1104 中,该部件调用立方体投影部件来生成站点 B 的事件矩阵。在方框 1105 中,该部件调用一个计算界面属性的估计向量相似性的部件,来生成来自站点 A 和站点 B 的界面属性的匹配似然性。本领域普通技术人员可以理解,可以采用许多不同技术来估计这个似然性,向量相似性只是一个例子。在方框 1106 中,该部件调用匹配矩阵生成部件,来生成一个指示哪些成对界面属性是相匹配的矩阵,该匹配矩阵生成部件传送估计界面属性的矢量相似性矩阵。在方框 1107 中,该部件调用一个计算估计向量相似性的部件,来生成结果属性的估计向量相似性矩阵。在方框 1108 中,该部件调用一个匹配矩阵生成部件,来生成一个指示哪些成对结果属性相匹配的矩阵。然后该部件完成处理。

[0092] 图 12 表示一个实施例中计算估计向量相似性的部件处理过程的流程图。该部件传送一个界面-界面对应的或一个结果-结果对应的事件矩阵,并确定每一对属性匹配的似然性。在方框 1201 中,该部件选择站点 A 的下一属性。在判定方框 1202 中,如果站点 A 的所有属性已经选定,那么该部件返回,否则该部件继续到方框 1203。在方框 1203 中,该部件选择站点 B 的下一属性。在判定方框 1204 中,如果站点 B 的所有属性已经选定,那么该部件循环至方框 1201 来选择站点 A 的下一属性,否则继续到方框 1205。在方框 1205 中,该部件按照方程式 3 计算选定属性的估计向量的相似性,然后循环至方框 1203 来选择站点 B 的下一属性。

[0093] 图 13 表示一个实施例中 cross-validate 部件处理过程的流程图。当 inter-site 匹配指示一个 intra-site 匹配不正确时,该部件改变属性的匹配。在方框 1301 中,该部

件选择下一全局属性。在判定方框 1302 中,如果所有的全局属性已经选定,那么该部件完成处理,否则该部件继续到方框 1303。在方框 1303 中,该部件选择下一网站。在判定方框 1304 中,如果所有网站已经选定,那么该部件循环至方框 1301 中来选择下一全局属性,否则该部件继续到方框 1305。在方框 1305 中,如果选定网站具有一个属性与选定的全局属性相匹配,那么该部件继续到方框 1306,否则该部件循环至方框 1303 来选择下一网站。在判定方框 1306 中,如果该选定属性移动至另一全局属性,那么该部件继续到方框 1307,否则该部件循环至方框 1303 来选择下一网站。在方框 1307 中,该部件改变选定属性来匹配不同的全局属性。在方框 1308 中,该部件改变选定属性的 intra-site 匹配。然后该部件循环至方框 1303 来选择下一网站。

[0094] 本领域普通技术人员可以理解,虽然在这里结合附图描述了模式匹配系统的特定实施例,但是可以在不背离本发明精神和范围的情况下做出各种改变。因此,本发明不仅限于所附的权利要求。

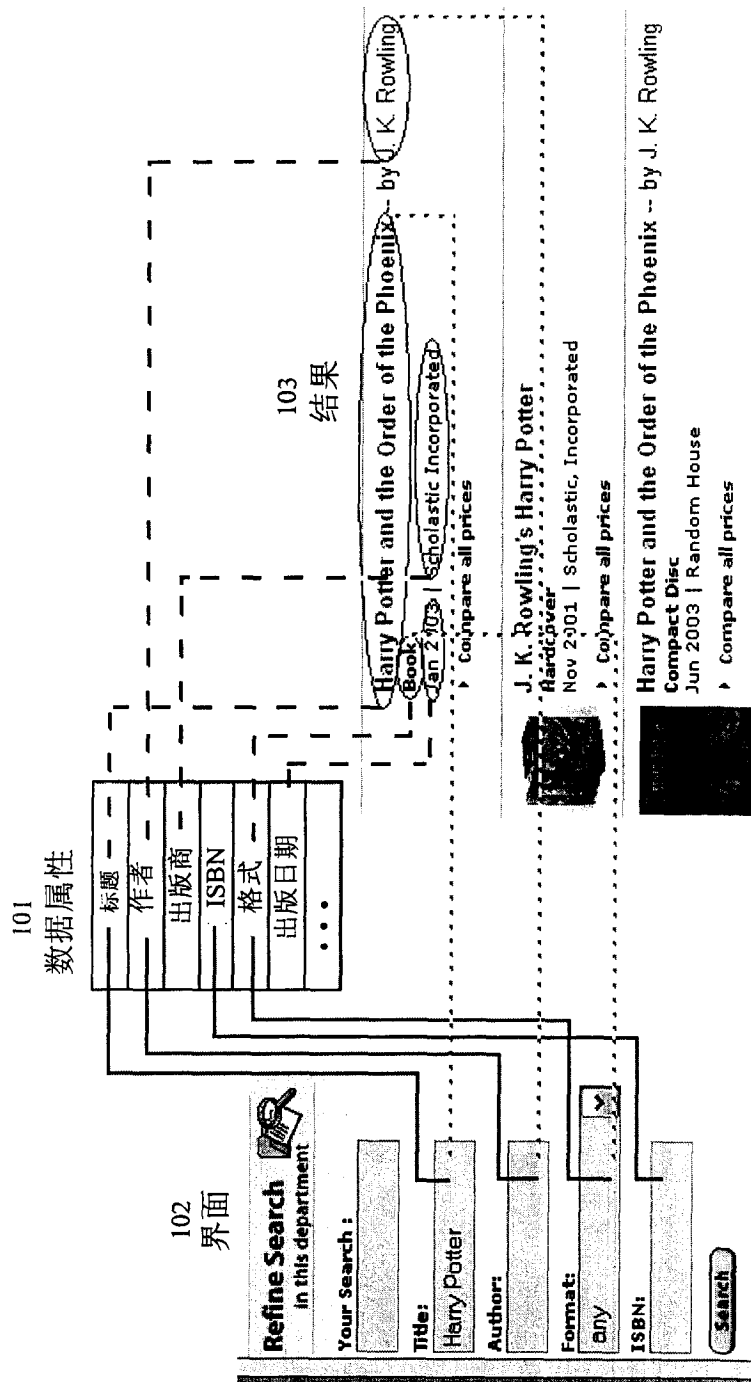
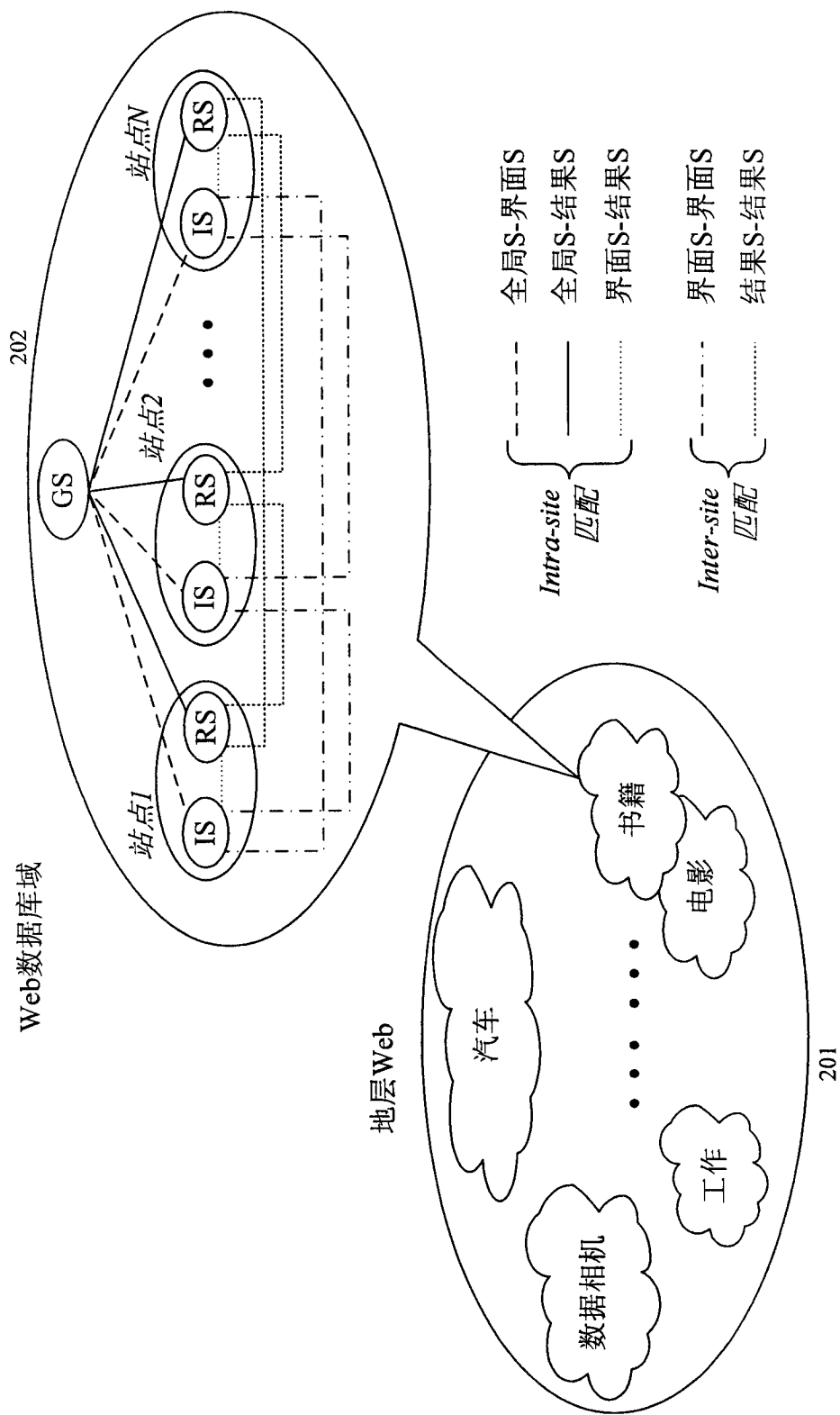


图 1



2
四

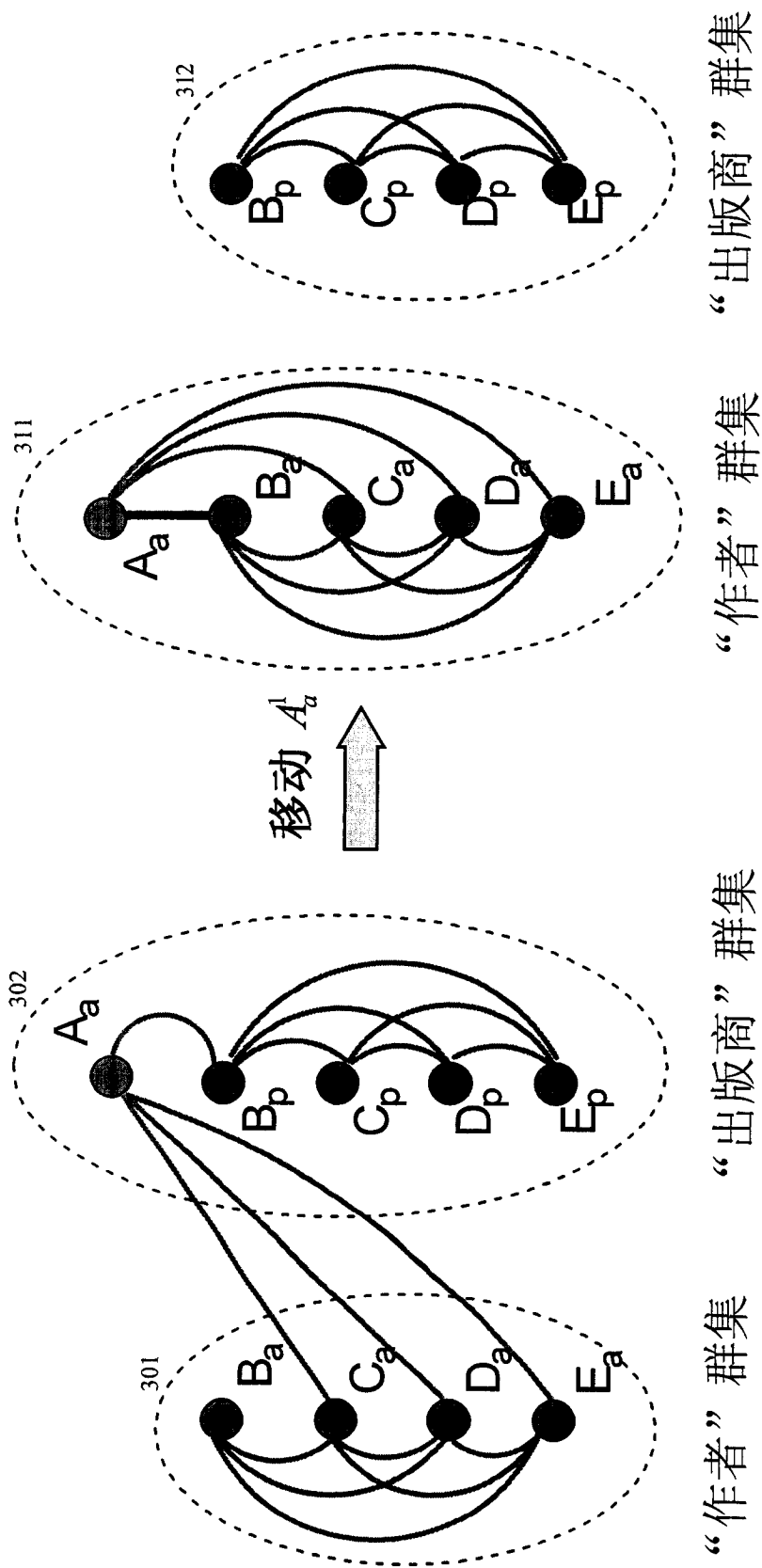


图 3

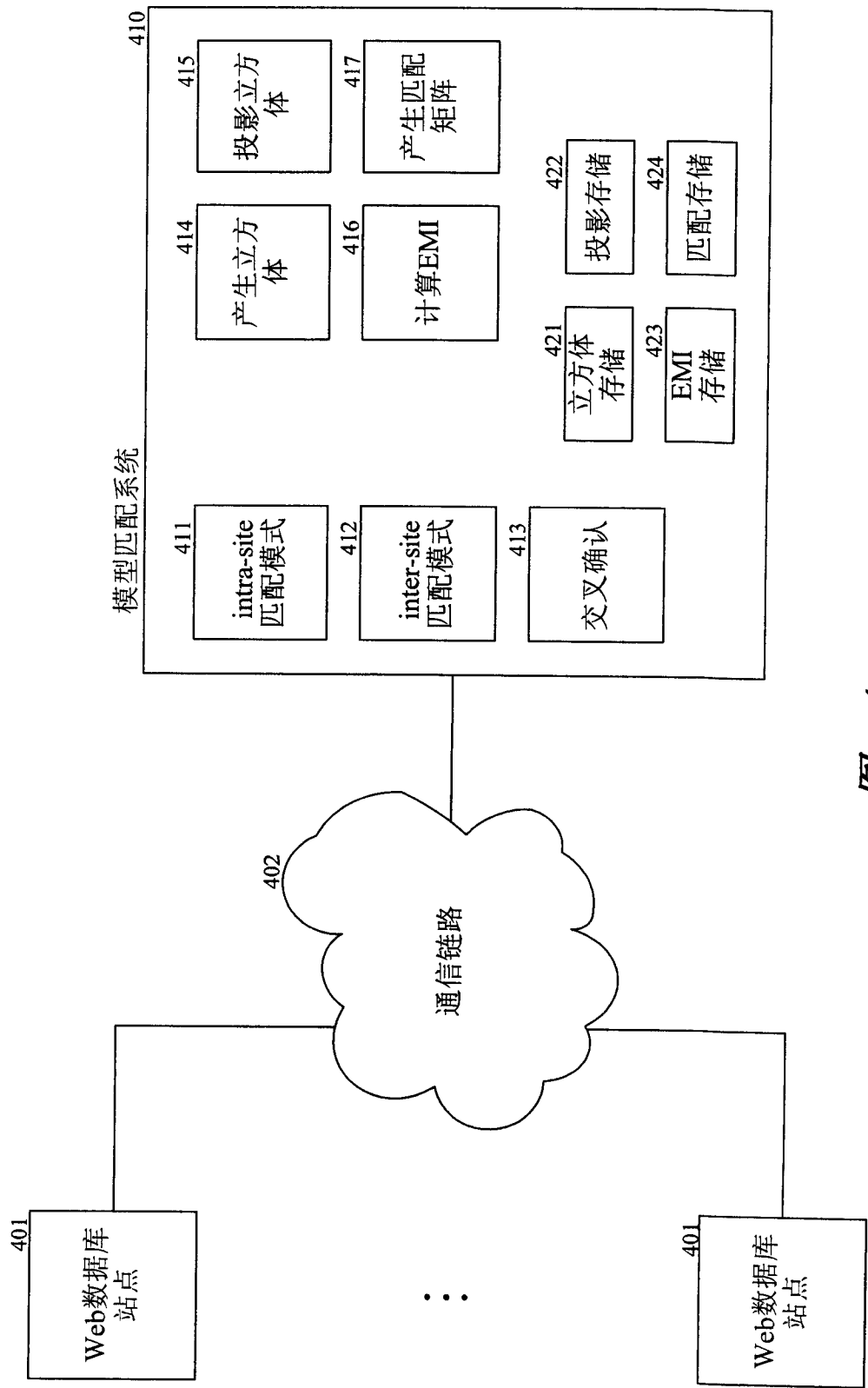


图 4

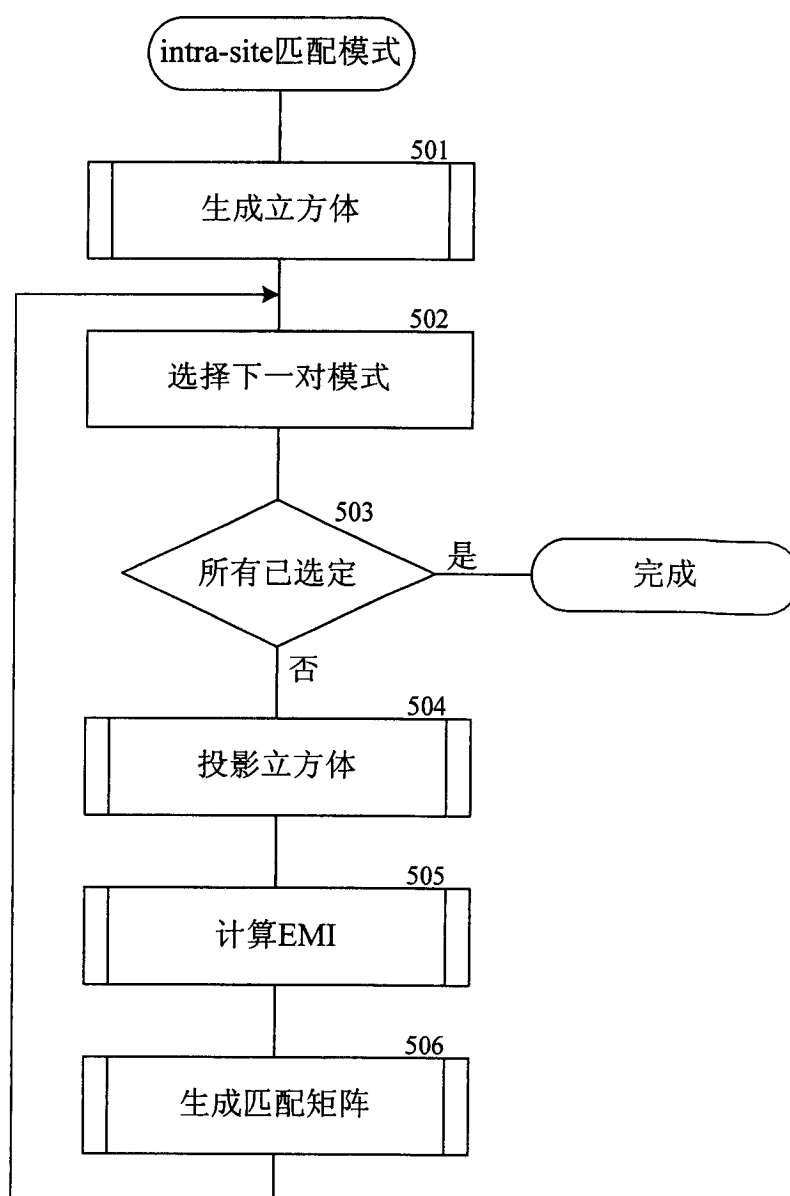


图 5

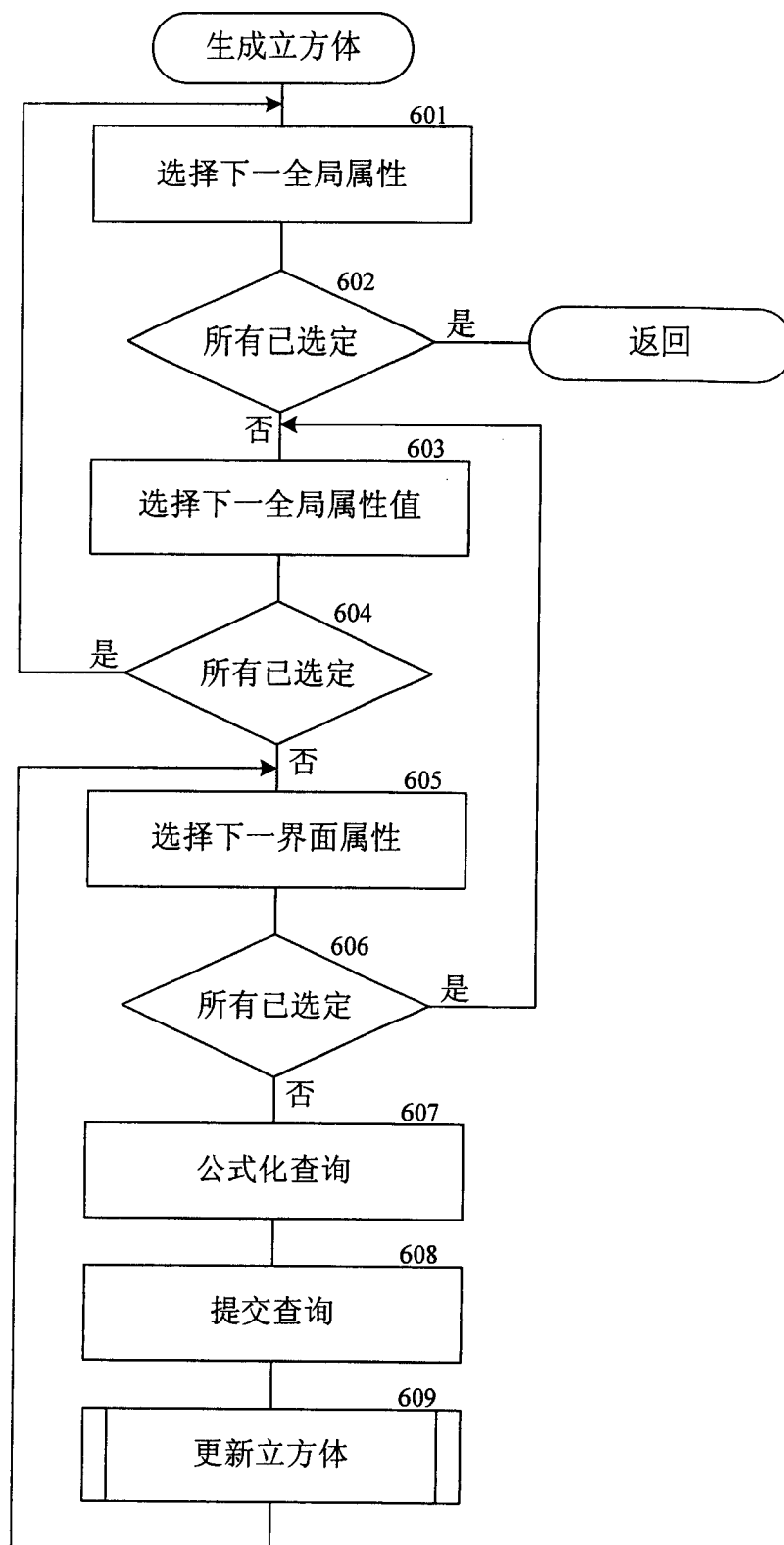


图 6

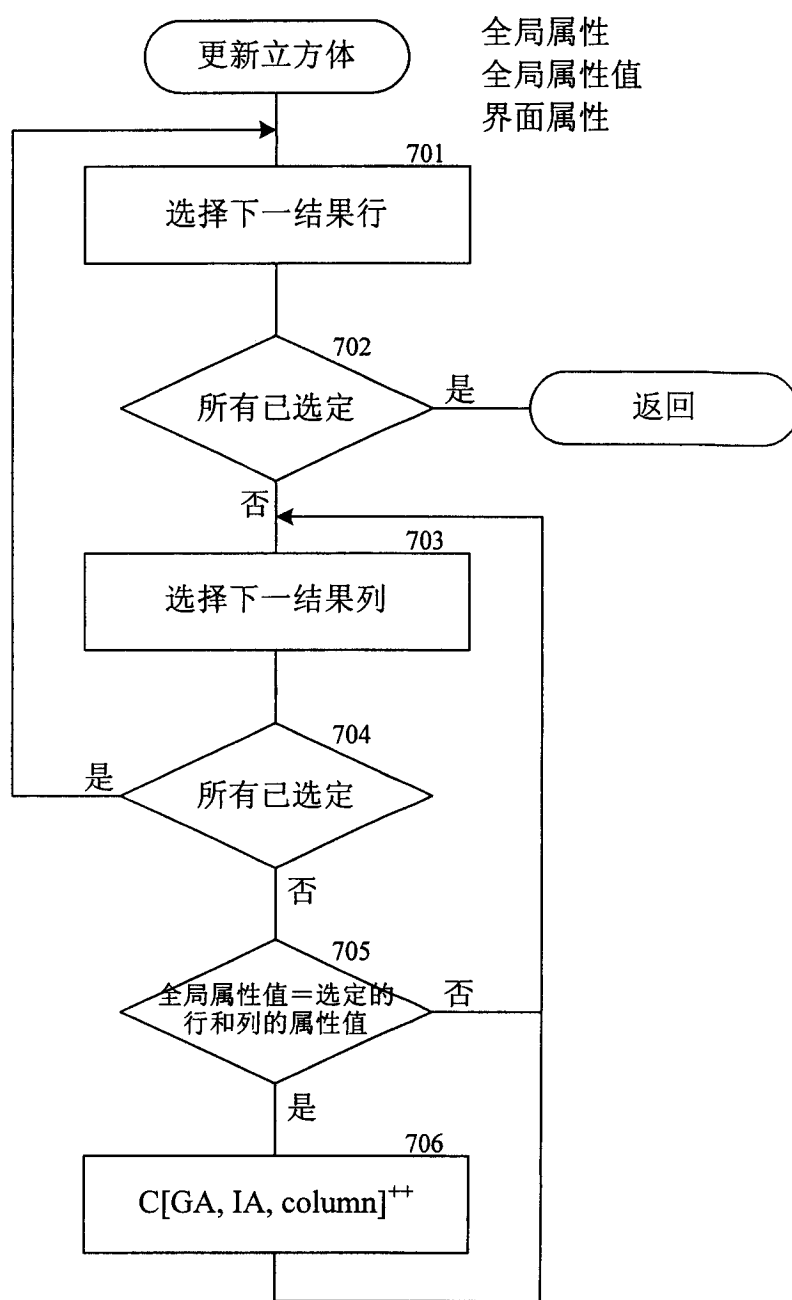


图 7

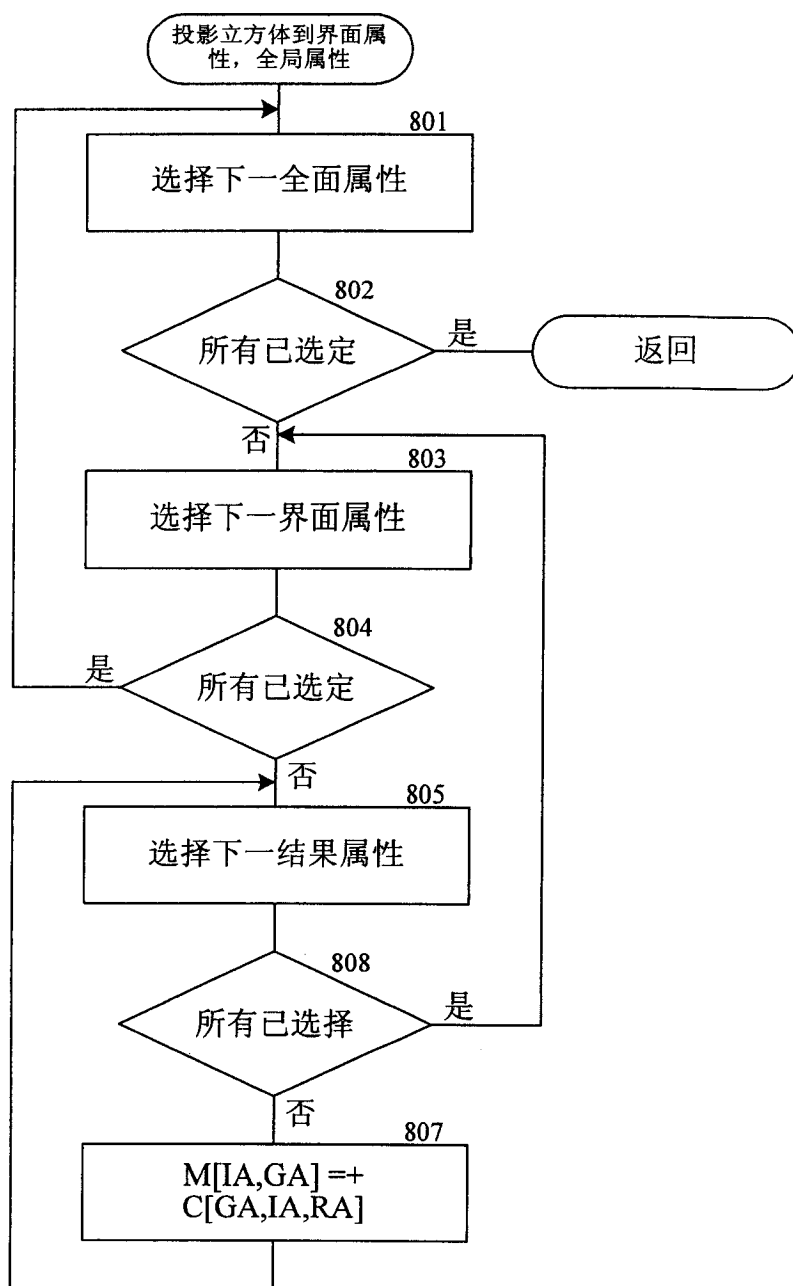


图 8

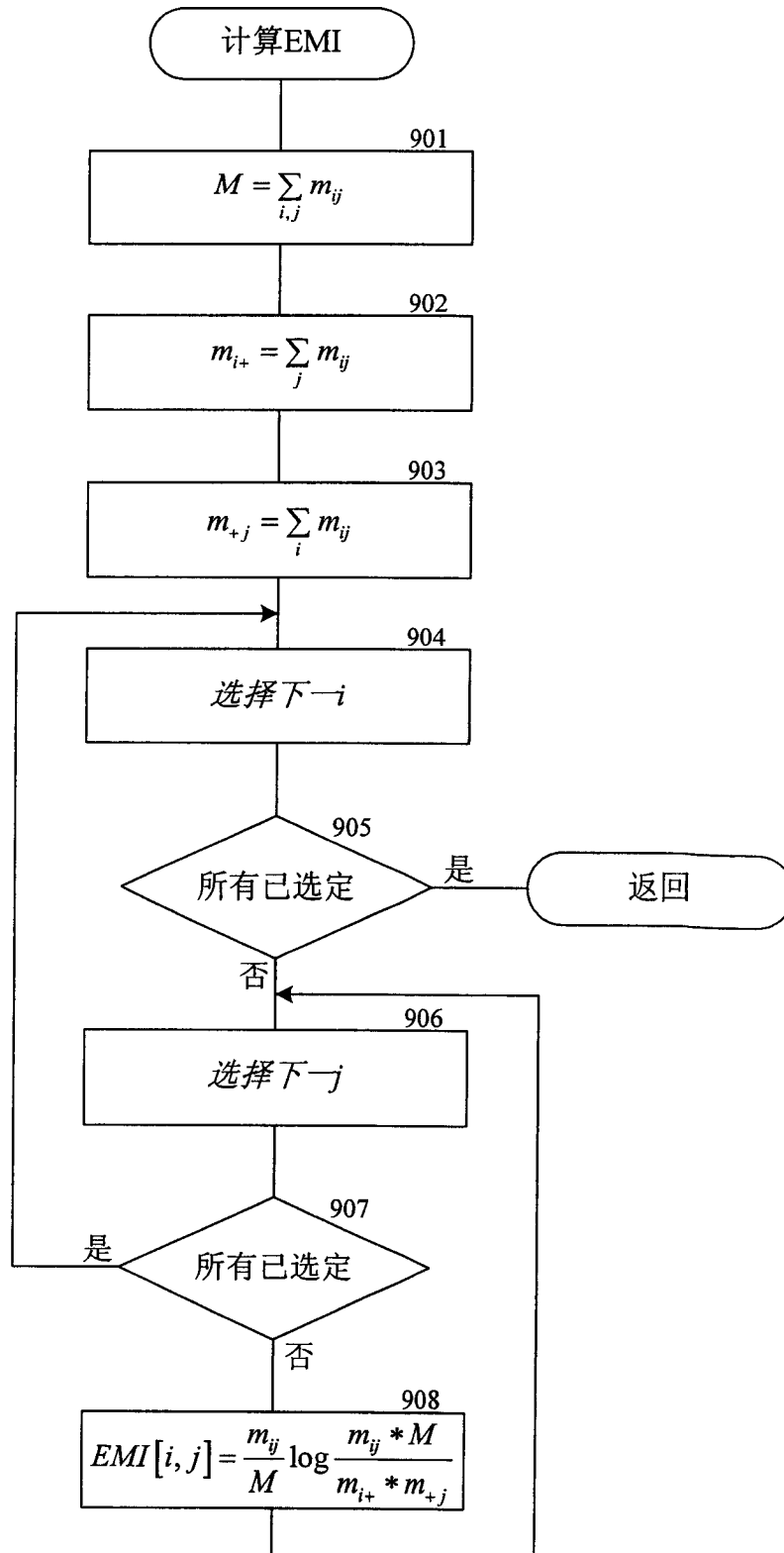


图 9

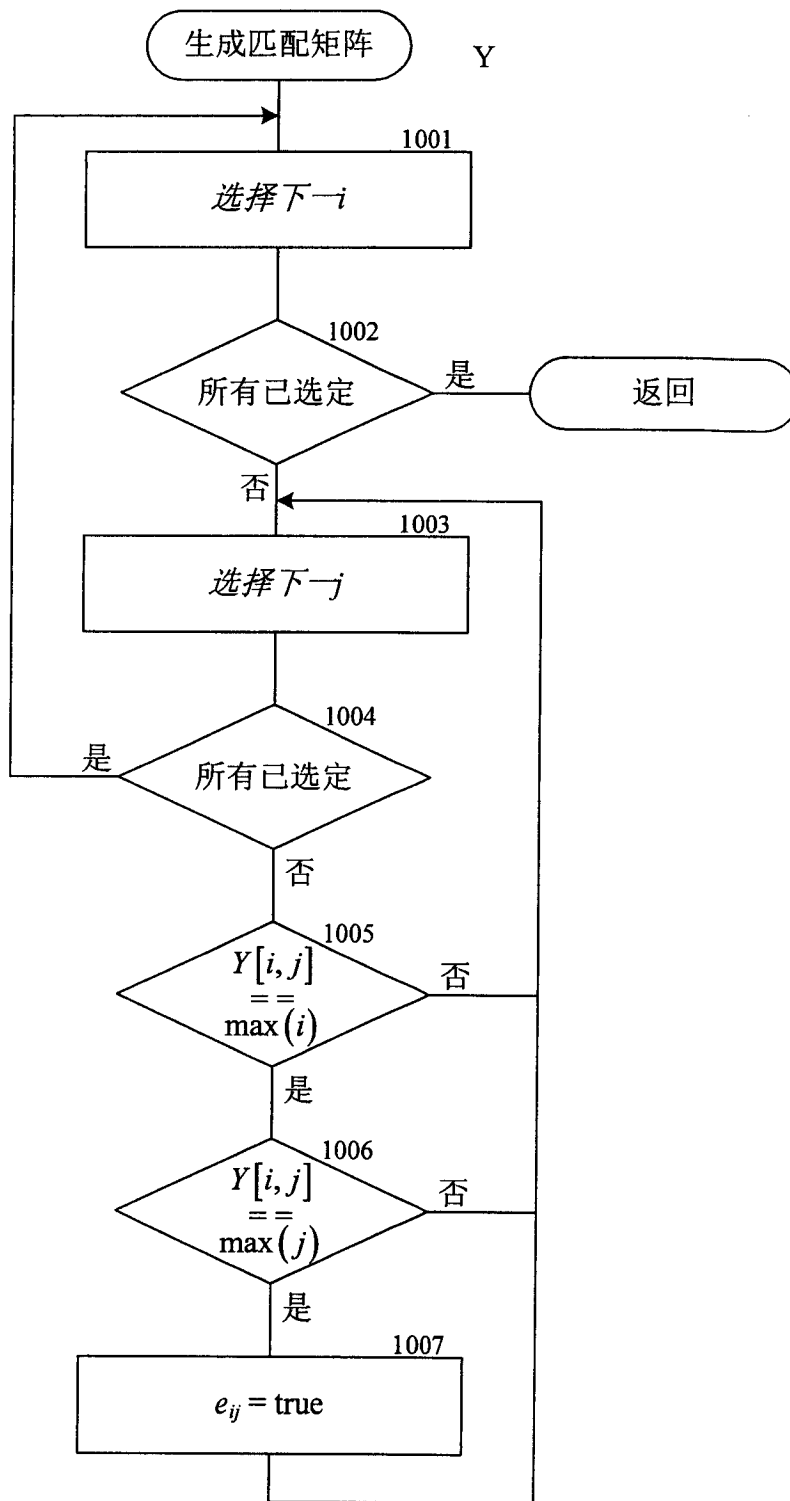


图 10

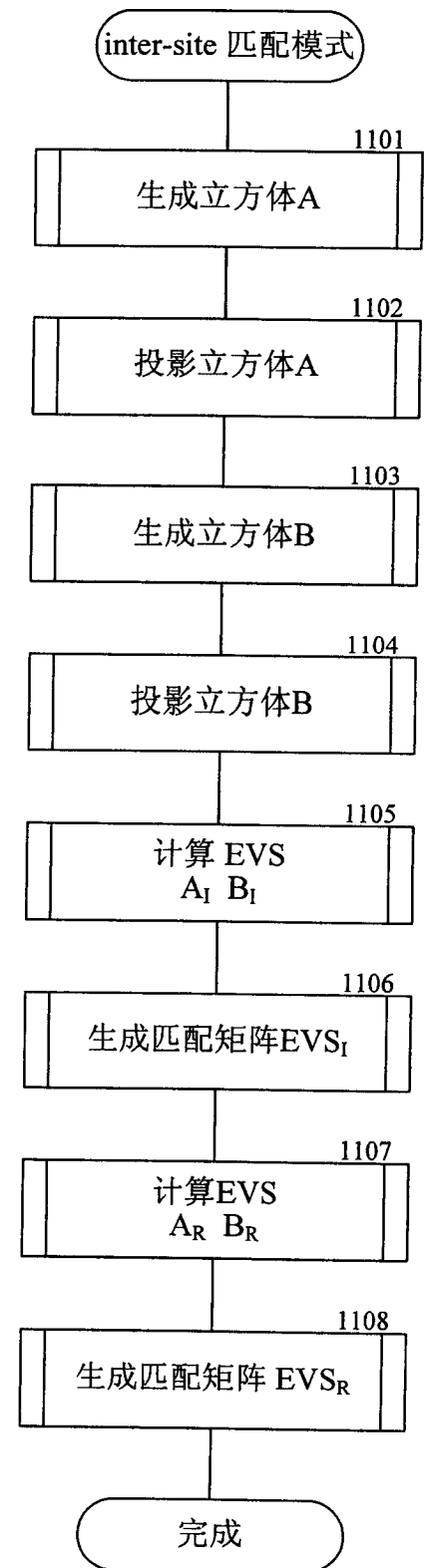


图 11

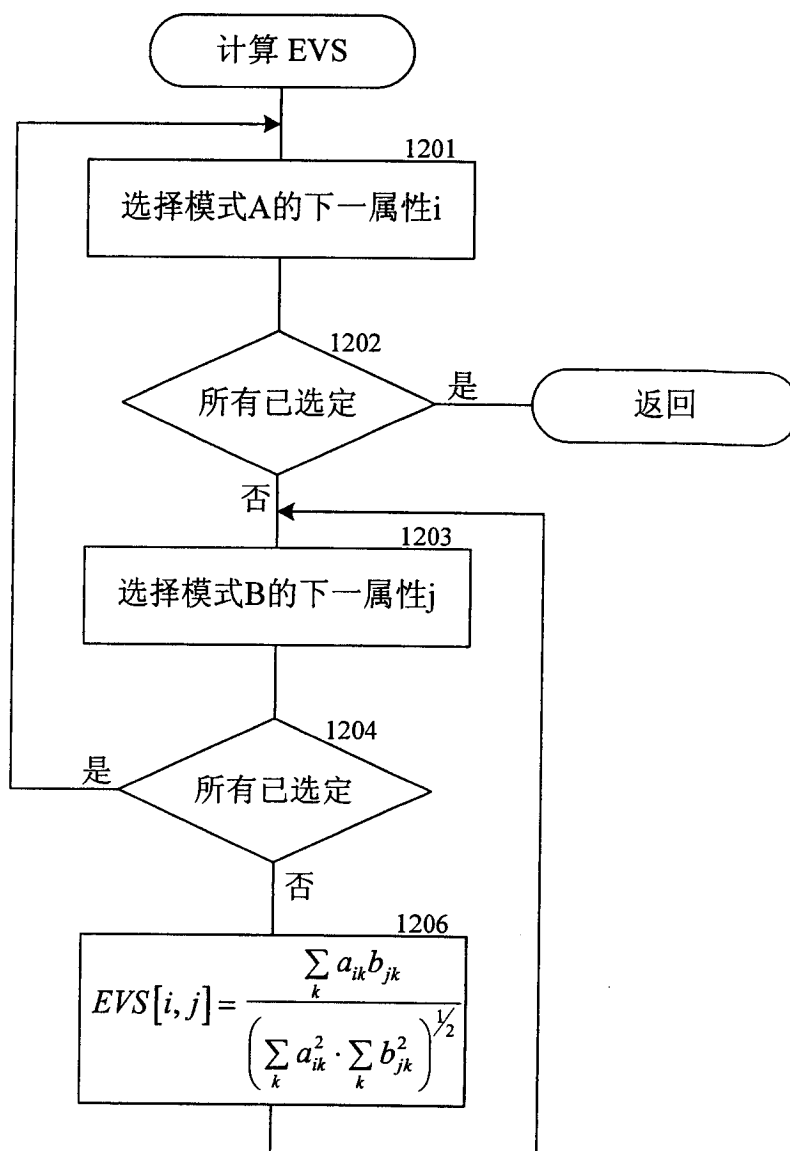


图 12

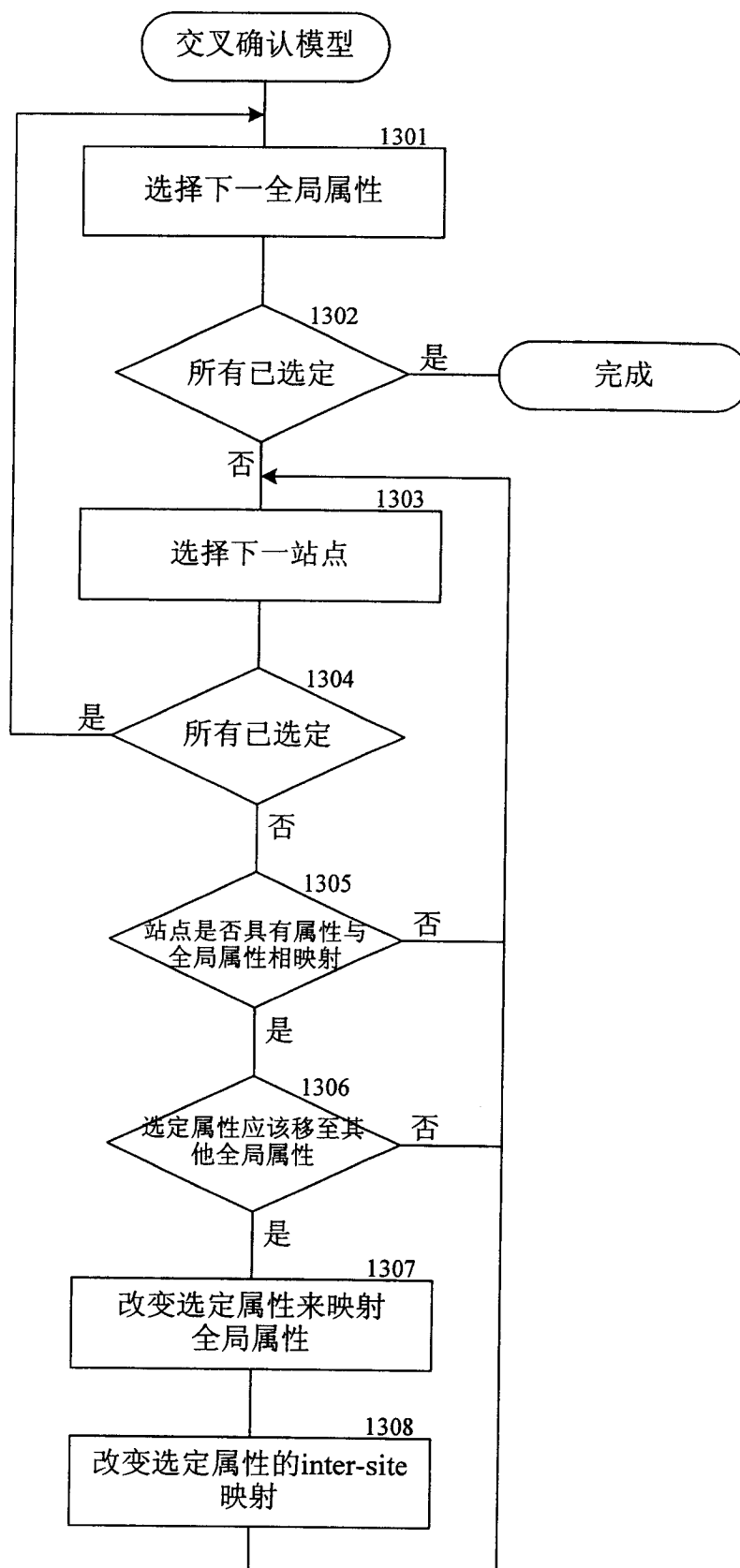


图 13