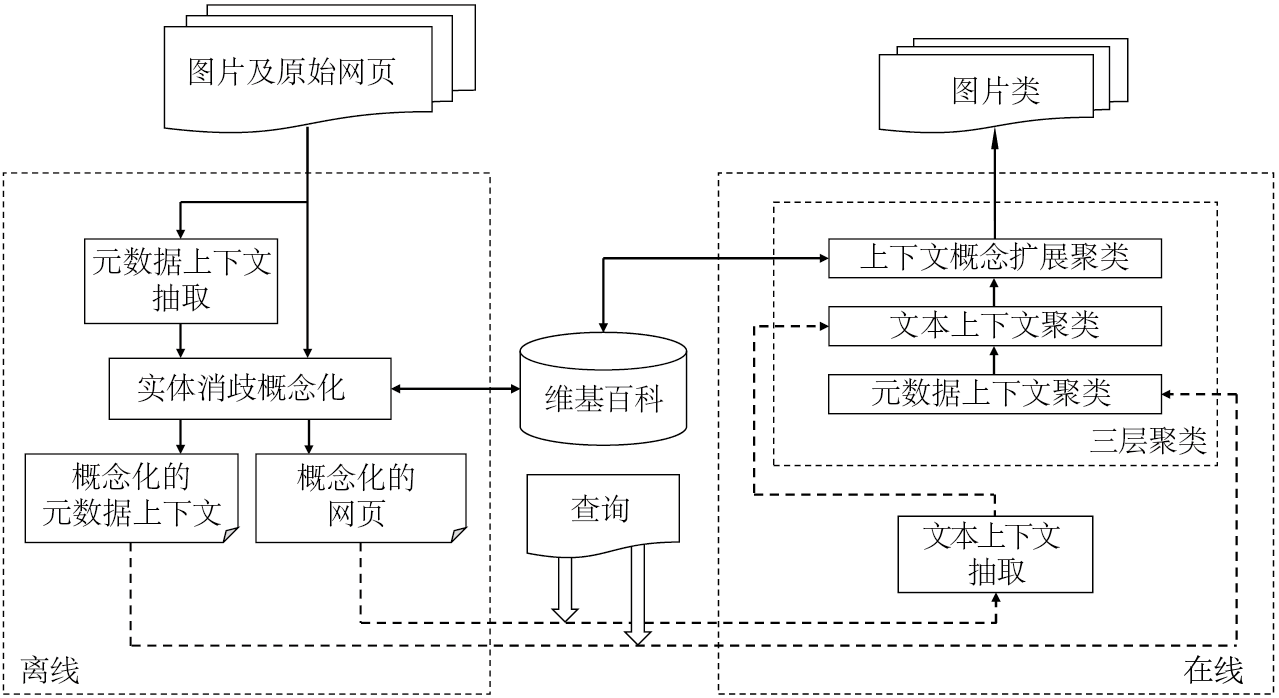
**说 明 书 摘 要**

本发明是用于对搜索引擎返回的网页图片进行实体聚类的系统，系统为每一个类别自动生成一些描述该类别的关键词（概念）。该系统分为离线与在线两个组件。离线组件主要抽取网页元数据，并对图片所在的网页的文本和元数据进行概念化，即把文本抽象为一组带权概念的集合。在线组件首先抽取搜索关键词在图片所在网页的上下文；其次，基于抽取的概念进行聚类；最后，为每个类归纳出可以表述该类特征的一组描述概念。其中，在线聚类算法采用了三层聚类的方式，即对抽取的特征按照置信度分为三类，从置信度最高的特征开始，每一层不断合并已有的图片类，以减少噪声数据对聚类效果的影响。每一层的聚类采用基于概念化的层次聚类算法，在聚类的同时可以给每一个类生成描述关键词。本系统的三层聚类算法与一般的层次聚类算法具有相同的时间复杂度，而对于特征的细分使得每一层的输入即前一层的输出更加精确，能有效提升聚类效果，并且给出准确的描述概念。

**摘 要 附 图**



**权 利 要 求 书**

1、一种对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，离线系统和在线系统，其中：

离线系统，用于对所有图片所在的源网页进行预处理，包括抽取网页元数据，把原网页文本和元数据概念化成一组带权概念的集合（概念向量）。概念化后的元数据和网页内容供在线系统查询使用。

在线系统，用于接收用户查询，提交到搜索引擎并接收返回的多页图片结果。对于每一个页的返回结果，找到源网页的概念化元数据和文本，并在概念化的文本中抽取用户查询关键词的上下文（查询上下文）以及图片上下文。在线系统分别利用元数据，上下文，以及通过维基百科对上下文进行概念扩展后的扩展上下文进行三层聚类，并为每一个类别自动标注相关的描述性概念，以供用户了解每一个类别的实体。

2、根据权利要求1所述的对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，离线系统进行元数据抽取，包括对URL中有效词条的抽取，图片ALT属性。对URL有效词条的抽取，利用二类分类器对有效和无效词条进行分类，并返回有效词条。

3、根据权利要求1所述的对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，离线系统包括概念化模块，用于将文本中的词条映射到维基百科中的概念。文本通过概念化模块，转换成带权概念的集合，以供聚类算法使用。每个概念的权值为该概念对图片的重要性，其定义如下：

其中， 为概念对图片的重要性，包括两部分的乘积：概念在图片上下文出现的频率，以及反向上下文频率。其中反向上下文频率反比于概念出现过的上下文的数量。

4、根据权利要求1所述的对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，在线系统包括文本上下文抽取模块，用于对用户所输入的查询关键词，抽取其概念化查询上下文和图片上下文。

5、根据权利要求1所述的对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，包含三层聚类算法模块。该模块根据抽取的元数据，上下文，以及扩展的上下文三类特征从置信度最高的元数据，到上下文，到扩展上下文进行三个层次的聚类， 其中：

第一层聚类，通过元数据概念化后的概念向量进行聚合层次聚类，获得类内精度高的聚类结果，并且合并每个类里所有图片的概念向量作为类的概念向量。

第二层聚类，向每个图片的概念向量中加入概念化上下文的概念向量，更新所有第一层聚类后得到的类的概念向量，并进一步对这些得到的类进行聚合层次聚类。

第三层聚类，把每个图片的向量替换成扩展的概念向量，更新所有第二层聚类后得到的类的概念向量，并进一步对这些概念向量进行聚合层次聚类。

6、根据权利要求5所述的对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，所使用的聚合层次聚类算法利用类的概念化进行类的相似度计算。类的概念化通过把类中的图片的概念向量进行相加，并且去除向量中值比较低的概念，得到高精度的类概念。类的概念化用如下公式定义：

其中，为概念，为类，为类中图片，为概念对图片的重要性。

7、根据权利要求5所述的对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，第三层聚类通过维基百科进行上下文的扩展，把图片的概念向量替换成扩展的概念向量，并且更新每个类的概念向量。其更新定义为如下公式：

其中， 为概念对概念的维基百科描述页面的重要性，为当前类概念向量中的概念。此上下文扩展过程通过选取值最大的前个概念对噪声数据进行过滤。

8、根据权利要求1所述的对搜索引擎返回的网页图片进行实体聚类的系统，其特征在于，利用三层聚类后得出的类概念向量给每个图片类标注相关的描述概念：选取每个类的概念向量中值最高的前几个概念用于描述该类所代表的实体。

**说 明 书**

对搜索引擎返回的网页图片进行实体聚类的系统

**技术领域**

本发明涉及计算机技术领域的自然语言处理，文本挖掘，具体地，涉及上下文抽取，文本聚类算法。

**背景技术**

随着互联网的普及以及网页图片日益增长，网页图片搜索逐渐成为互联网用户的一大日常应用。目前的图片搜索引擎主要返回跟查询关键词相关的图片。而这些图片往往包含多个同名的实体。用户需要从搜索结果中找到所要的图片，需要浏览查看每张返回的图片。为了提高搜索结果的可读性，按照不同实体区分搜索结果成为了图像搜索引擎的一个改良反向。

图像聚类是自动区分不同实体的方法。在过去的研究中，D. Cai（参见Cai, D., He, X., Ma, W.Y., Wen, J.R., Zhang, H.: Organizing www images based on the analysis of page layout and web link structure. ICME 2004）利用基于视觉的分块的方式抽取网页图片的上下文，并且利用该上下文和网页链接信息进行聚类。然而由于视觉分块的不稳定，以及上下文中的噪声数据，聚类的精度有很大的限制；Z.Fu（参见Fu, Z., Ip, H.H.S., Lu, H., Lu, Z.: Multi-modal constraint propagation for heterogeneous image clustering. MultiMedia 2011）提供了一种结合照图像的标签和图像的视觉特征等多个模块的框架，在多个图上通过传递类的约束来实现图像聚类。目前视觉特征的抽取精度的不足，该框架会传播视觉特征所包含的错误。而且，该方法需要在多个图中进行约束传递，导致聚类效率低下，不适合于对在线图片搜索结果的聚类。目前的图像聚类方法并不能提供描述性的概念去给每一个类进行标注。

经检索还发现如下相关专利文献：

相关检索结果1：

申请（专利）号: 2012101444570，名称：一种图片聚类的方法及装置

该专利文献通过对图片的视觉特征，包括全局特征以及局部特征进行了两次聚类，第二次聚类在第一次聚类的基础上进行切割。

技术要点比较：

1．该专利根据图片的内容，即视觉特征进行图片聚类，而本发明中利用图片上下文的特征进行聚类。

2．该专利的二次聚类把大的类切割成小的类，而本发明从小的类聚合成大的类，利用每次扩展概念向量进行特征的筛选，过滤噪声数据。

3．本发明采用的概念向量表示方式能为每一类标注描述概念，而基于图片内容的聚类方式无法提供概念描述。

相关检索结果2：

申请（专利）号: 2013106111554，名称：一种基于聚类紧凑特征的海量图像检索系统

该专利文献通过图像的局部特征对图像库中的图像进行聚类。搜索的时候通过查询关键词先检索到图片聚类然后返回相应的图像。

技术要点比较：

1．该专利根据图片的局部特征生成聚类紧凑特征，进行图片聚类，而本发明中利用图片上下文的特征进行聚类。

2．该专利通过图像聚类来提高检索的速度，而本发明通过把搜索结果进行聚类并概念化以提供区分各个类别的搜索结果。

相关检索结果3：

申请（专利）号: 201210545637X，名称：一种基于分层聚类的均衡图像聚类方法

该专利文献利用图片聚类的方式减少搜索时所需要遍历的图片数量。图片聚类基于图像高维特征数据。

技术要点比较：

1．该专利根据图片的高维特征，进行图片聚类，而本发明中利用图片上下文的特征进行聚类。

2．该专利通过图像聚类减少检索时需要遍历的图片，采用的图像聚类方式是层次聚类，而本发明基于三种不同的上下文特征，通过三层聚类的方式提升聚类的精度。

相关检索结果4：

申请（专利）号: 201210163641X，名称：图像聚类方法

该专利通过拍摄设备获取图像的时间数据和位置数据，并利用时间和位置以及速度数据作为特征进行聚类。

技术要点比较：

1．该专利主要针对拍摄图像进行聚类，而本发明针对网页图片进行聚类。拍摄的图像没有上下文信息，而网页图片不一定是拍摄图像，大部分没有拍摄时间和位置。两者的特征有所不同。

2．该专利基于事件序列进行聚类，而本发明基于概念向量。概念向量可以用于描述概念的生成。

相关检索结果5：

申请（专利）号: 2009801523973，名称：使用基于内容的过滤和基于主题的聚类将图像布置到页面中

该专利基于设备捕获到的图片的内容，即视觉特征，按照不同的主题聚类，并且把聚类的结果映射到相应的相簿中。

技术要点比较：

1．该专利利用图片的视觉特征聚类，而本发明利用网页图片的上下文进行聚类。

2．该专利将图片通过图片布局到不同的页面上，而本发明为用户提供分类的搜索结果以及相应的描述概念。

相关检索结果6：

申请（专利）号: 2010105171639，名称：图像聚类方法和系统

该专利采用参数估计的方式建立图像的有向图，并且以分割有向图的方式进行图像聚类。有向图的分割形成多个子图，而每个子图的图像归为一个类。

技术要点比较：

1．该专利利用图的方式进行聚类，图像库表示成一个有向图。本发明通过从小到大的方式聚合图片形成图片类，每一层聚类考虑不同的图像上下文特征。

相关检索结果7：

申请（专利）号: 2005800393866，名称：图像聚类方法和系统

该专利利用时间地点特征对图像按照事件进行聚类，采用的聚类算法根据不同的时间范围进行不同层的聚类。

技术要点比较：

1．该专利的多层聚类中的层是不同时间范围，而本发明的层是按照不同特征所定义的层。

2．该专利按照事件序列进行聚类，而本发明按照不同的实体区分不同的图片类。

**发明内容**

本发明针对现有技术中的不足，提供了一个对搜索引擎返回的网页图片进行实体聚类的系统，本发明解决的技术问题包括：

1. 抽取图像上下文信息，并把上下文信息表示为概念空间中的向量，为图像相似度的计算提供特征。
2. 由于某些图像存在上下文信息量不足的情况，本发明提供一种扩展上下文信息的机制，把上下文的概念向量通过维基百科或者其他知识库进行扩展。
3. 由于不同的特征跟图片的相关度不同，相关度越高的特征的置信度越高，本发明为了有效利用不同相关度的特征来提高聚类的精度，依次对图片的概念向量进行扩展，并且聚类。

发明的对搜索引擎返回的网页图片进行实体聚类的系统包括离线系统和在线系统两部分，其中：

离线系统，用于对所有图片所在的源网页进行预处理，包括抽取网页元数据，把原网页文本和元数据概念化成一组带权概念的集合（概念向量）。概念化后的元数据和网页内容供在线系统查询使用。

在线系统，用于接收用户查询，提交到搜索引擎并接收返回的多页图片结果。对于每一个页的返回结果，找到源网页的概念化元数据和文本，并在概念化的文本中抽取用户查询关键词的上下文（查询上下文）以及图片上下文。在线系统分别利用元数据，上下文，以及通过维基百科对上下文进行概念扩展后的扩展上下文进行三层聚类，并为每一个类别自动标注相关的描述性概念，以供用户了解每一个类别的实体。

所述离线系统包括元数据抽取和概念化两个模块，其中：

元数据抽取模块，包括对URL中有效词条的抽取，图片ALT属性。对URL有效词条的抽取，利用二类分类器对有效和无效词条进行分类，并返回有效词条。图片ALT属性可以直接从HTML源代码获得。

概念化模块，包括对元数据和图片原网页文本的概念化。概念化通过把元数据和文本中的词映射到维基百科的概念上，使元数据和文本转化成带权概念的集合，以计算相似度，供聚类算法使用。每个概念的权值为该概念对图片的重要性，其定义如下：

其中， 为概念对图片的重要性，包括两部分的乘积：概念在图片上下文出现的频率，以及反向上下文频率。其中反向上下文频率反比于概念出现过的上下文的数量。

所述在线系统包括文本上下文抽取和三层聚类算法两个模块，其中：

文本上下文抽取模块，在已经概念化的原网页文本里抽取上下文信息，包括图片上下文的抽取和查询上下文的抽取。图片上下文和查询上下文皆通过一个固定大小的窗口截取，比如图片或者查询关键词前后50个概念。抽取的文本上下文形成一个概念向量，以用于计算图片相似度。

三层聚类算法模块，包括元数据聚类，文本上下文聚类，以及上下文概念扩展聚类三个模块， 其中：

第一层聚类，通过元数据概念化后的概念向量进行聚合层次聚类，获得类内精度高的聚类结果，并且合并每个类里所有图片的概念向量作为类的概念向量。其中，聚合层次聚类算法利用类的概念化进行类的相似度计算。类的概念化通过把类中的图片的概念向量进行相加，并且去除向量中值比较低的概念，得到高精度的类概念。类的概念化用如下公式定义：

其中，为概念，为类，为类中图片，为概念对图片的重要性。

第二层聚类，向每个图片的概念向量中加入概念化上下文的概念向量，更新所有第一层聚类后得到的类的概念向量，并进一步对这些得到的类进行聚合层次聚类。

第三层聚类，把每个图片的向量替换成扩展的概念向量，更新所有第二层聚类后得到的类的概念向量，并进一步对这些概念向量进行聚合层次聚类。其中，向量的扩展利用维基百科的概念描述页面，把相关的概念加入到图片的概念向量中，并且更新每个类的概念向量。其更新定义为如下公式：

其中， 为概念对概念的维基百科描述页面的重要性，为当前类概念向量中的概念。此上下文扩展过程通过选取值最大的前个概念对噪声数据进行过滤。

用三层聚类后得出的类概念向量给每个图片类标注相关的描述概念：选取每个类的概念向量中值最高的前几个概念用于描述该类所代表的实体。

与现有技术相比，本发明创造性地利用三种不同的特征，和对应的三层聚类算法，对图片进行聚类，并且为每一个类提供概念标注，使得图片搜索结果更好地按照不同实体组织起来，并且每个实体类具有高精度，不同实体之间具有明显的区分度。本发明把整个框架分成了在线和离线两个部分，大大减小了在线聚类的时间开销。

**附图说明**

通过阅读参照以下附图对非限制性实施例所作的详细描述，本发明的其它特征、目的和优点将会变得更明显：

图1示出本发明的系统框架图；

图2示出本发明的三层聚类算法示例图。

**具体实施方式**

下面结合附图对本发明的实施例作详细说明，本实施例在以发明技术方案为前提下进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

本实施例的任务是对用户输入的查询关键词“bean”，获取搜索引擎图片搜索结果，对结果中的不同“bean”的实例进行聚类，以辨别不同的实体，并为每个不同的“bean”提供前不同的概念标注。

如图1所示，本发明的离线系统的元数据抽取模块对本实施例“bean”相关的所有原始网页进行元数据上下文抽取。如某网页的URL为：

“http://domain.com/53C316-C2oJ5/mr\_bean.jpg”

元数据抽取模块通过分割符将词分开，并利用二类分类器将有效字符检测出来。如：“mr bean”。离线系统的概念化模块对“bean”的元数据以及相关网页进行了概念化，得到元数据概念向量和文本概念向量。

当接收到用户的查询关键词“bean”后，在线系统的文本上下文抽取模块从概念化的文本中找到图片和查询关键词“bean”的位置，并且抽取前后50个概念作为文本上下文概念向量。利用元数据概念向量和文本上下文概念向量，在线系统进行三层聚类。

如图2所示，在线系统的三层聚类模块首先按照元数据概念向量计算图片相似度并进行聚合层次聚类（图片1和图片2的概念向量皆包含概念“Mr. Bean”，而图片3和图片4皆没找到有效的元数据概念）。在聚合层次聚类中，类间的相似度用类的概念向量来计算。系统从第一层聚类的结果计算出类的概念向量，如图片1和图片2形成了一个类，此类的概念向量包含概念“Mr. Bean”。

第二层聚类在第一层聚类的基础上通过扩展图片的概念向量进行进一步聚类。如图2中图片1和图片2形成的类的概念向量加入了概念“Rowan Atkinson”，图片3的概念向量加入了“Rowan Atkinson”以及“Comedy”，图片4加入了“Blackadder”。 由于扩展后的向量拥有更多共同的概念，在线系统经过第二次层次聚类合并一些相似的类，得到更为大的类。如图2中图片1,2,3形成了新的类，并且把类的概念向量扩展为“Mr. Bean”，“Rowan Atkinson”，“Comedy”。

第三层聚类首先对各个类或者图片的向量用维基百科进行扩展，如图2中图片1,2,3组成的类的概念向量中加入了“Blackadder”，图片4加入了“Rowan Atkinson”。通过基于维基百科的扩展，类向量之间拥有更高的相似度。在线系统通过第三次层次聚类去进一步聚合一些原来由于信息量不足而没有合并的类。如图2中的图片4通过扩展向量可以合并到包含图片1,2,3的类中。

在三层聚类算法结束后，在线系统分开不同的类别，把所有实体及其图片呈现给用户。每个实体用对应概念向量中最有代表性的概念（值最大）的前几个概念来描述。比如图2中的类可以用“Mr. Bean”，“Rowan Atkinson”，“Comedy”，“Blackadder”等概念来描述关于名为憨豆先生的美国喜剧演员的图片。

以上对本发明的具体实施例进行了描述。需要理解的是，本发明并不局限于上述特定实施方式，本领域技术人员可以在权利要求的范围内做出各种变形或修改，这并不影响本发明的实质内容。

**说 明 书 附 图**

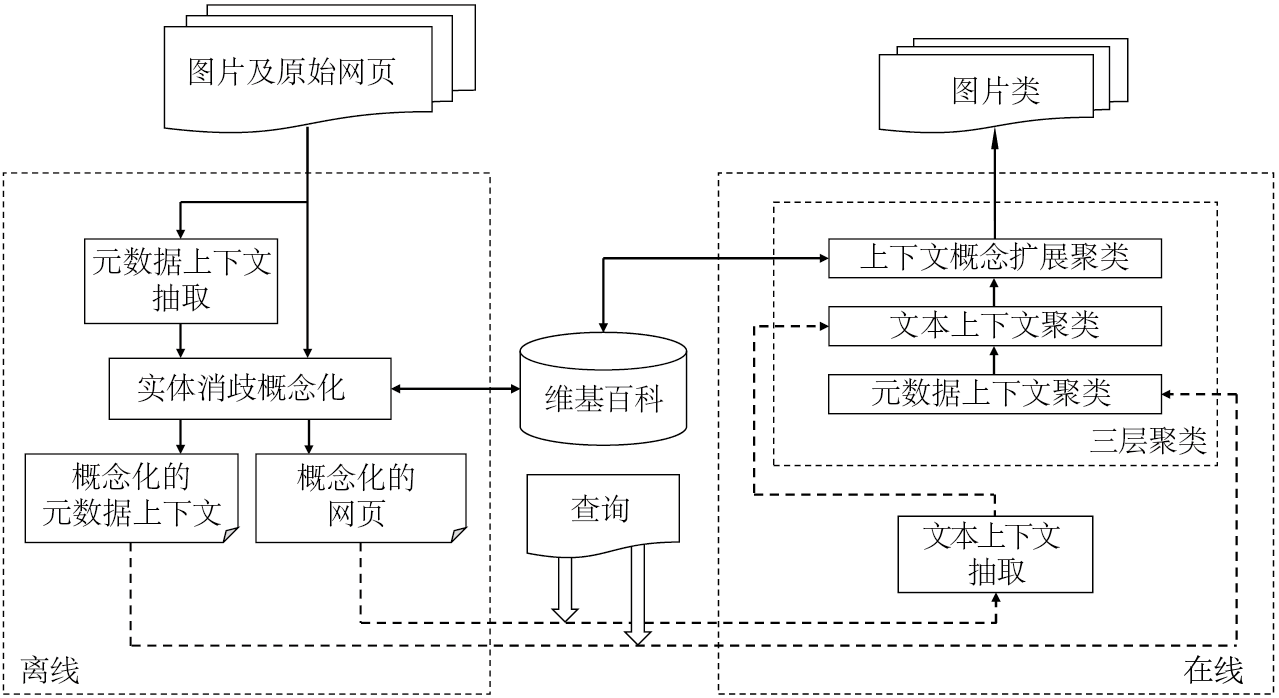


图 1

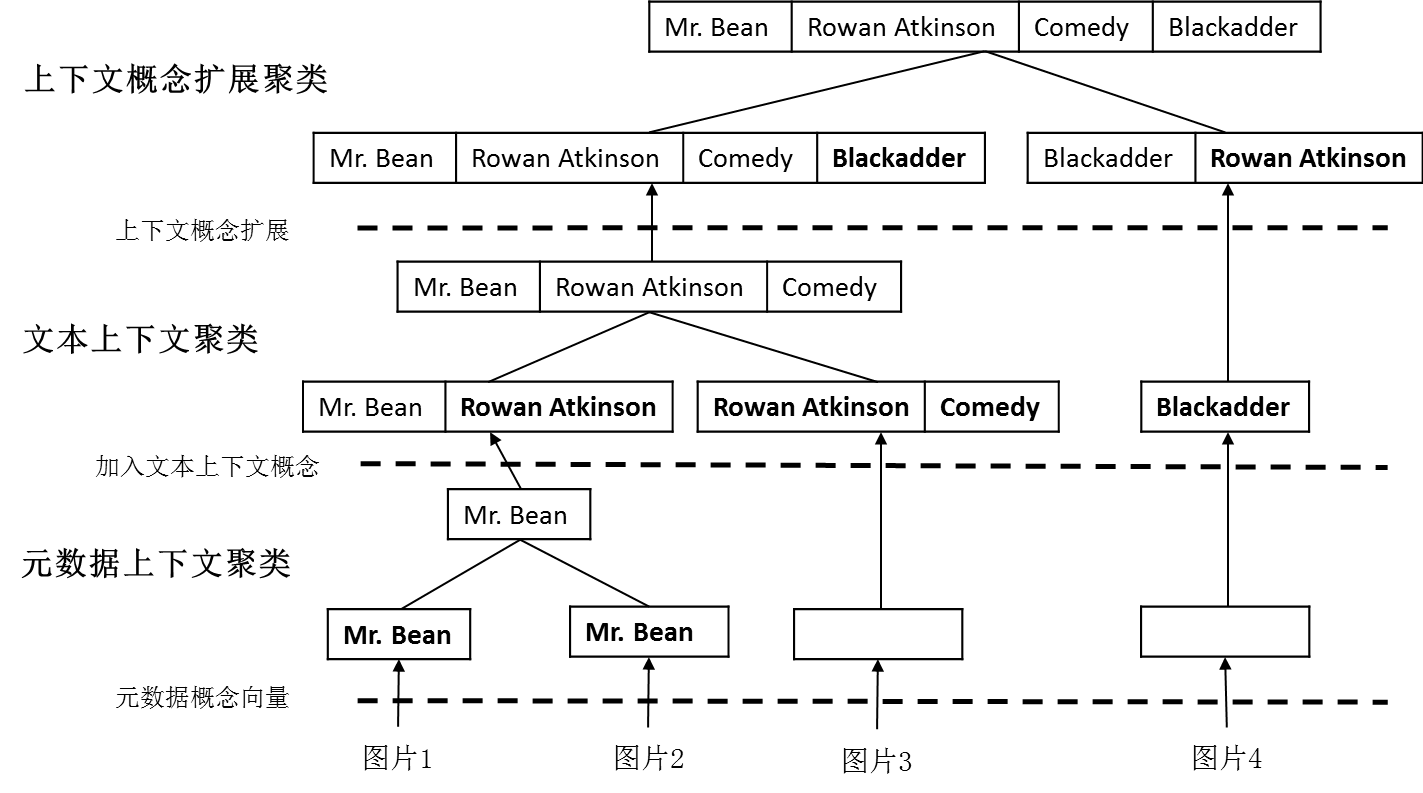


图 2