使用衰减概念和遗传规划的聚焦网络爬虫

摘 要 网络信息的持续快速增长是许多论文研究的主题。本文中，我们介绍了一种新的针对网页抓取的优化方法。使用遗传规划提高相似度测量的准确性。这种算法适用于包括标题和正文在内的网页的不同部分。因此，爬虫使用这种优化的相似度测量去遍历页面。为了提高爬虫的准确性，我们使用衰减概念去限制爬虫根据搜索条件所采集到的有效网页。衰减计算按搜索条件给每个页面打分。这将有效的减少在对网页编历时层次的深度。这个分值可以根据页面在搜索条件下的相似性进行修改。在这种情况下，我们使用三种计算设置该值。实验的结果表明，使用遗传规划与动态衰减阈值结合能得到最佳的准确性。

关键词 聚焦网络爬虫；遗传规划；衰减概念；相似度空间模型

1. 简介

互联网包含了超过1150万网页并且这一数字还在迅速增长。根据最近的统计，60%的用户会搜索特定主题，并总使用大众化的商业搜索引擎以获得他们的结果[2,3]。这样的用户没有使用通用搜索引擎的能力。

事实上，许多搜索引擎并不包含所有网页[1]。因此，我们需要更有效的方法,来更准确的收集数据。最常见的方法之一是将爬虫限制在几个指定的主题上。通过这种方式，爬虫能够检索到绝大多数寻常的页面。这种方法称为聚焦爬虫[4]。下图说明了普通爬虫和聚焦爬虫的区别。

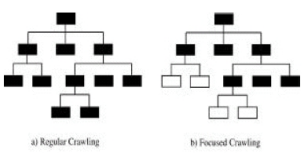


图1.聚焦爬虫和普通爬虫的区别

网络就像一个社交网络。在这样的网络中，页面之间的链接不是没有意义的。页面中的每个链接都表示了页面之间的语义关系。爬虫使用这些链接有效地获取相关的网页。我们可以通过以下两种方式：

一种方法是使用基于内容的相似度打分。在这种方式中，爬虫会估计当前页和主题的相似度并为其打分。若分值高于预定义的阈值，则可获取整个页面的链接。

另一种方法是给页面的链接排名。我们按所需的排位分数来获取链接。由于我们在爬虫中使用预定义的阈值，因此会导致不可预测的噪音。在本文中，我们使用衰减的概念来减少噪音引起的不准确的影响[5]。

2.文献综述

最早的聚焦爬虫之一在文献[6]中被讨论。Chakrabarti在1999年描述了聚焦爬虫实现的经验[7]。聚焦爬虫包含了两种算法来维持抓取范围在所需范围内：

（1）分析算法被用于判断由目标URL指出的网页相关性与质量。

（2）网页搜索算法确定访问目标URL的最佳顺序[9]。

任何爬虫的基础都是根据查询来确定页面的相似度。以这种方法，页面被映射到向量空间计算相似度。这种方法仍被认为是一种有效的解决办法[6]。在另一篇文献中，用到隐式的本体论概念去建立更准确的相似度测量。他们使用本体论来丰富具有相似内容主题的文本。这样的做法已经被证明比单纯使用相似度概念更加准确[7,8]。

在结构方法中，网页的链接分析被广泛用于构建关系图。例如，页面排名算法常被使用[9,10]。网页排名算法计算尚未提取的URL的分数并最先检索具有最佳分数的URL。页面排名算法根据重复率以及在其他页面中的重要性对它们进行排序。根据这种方法，爬虫将在抓取一个不相关的页面时停止[11]。这将导致，可能存在与搜索主题不完全相关的页面。

一些聚焦爬虫依赖于使用以前的爬虫数据库。这样的数据库包含了连接和USL的相关信息。已经对于这样基于内容和链接的算法组合的使用进行了一些研究[12]。 一些研究人员已经使用概念图来构建基于链接的图[11]。在这种情况下，他们使用元搜索建立连接到特定页面的页面。

在一些情况下，遗传规划已被用于提高基于内容的分类的准确性。作为遗传算法的扩展，遗传规划也已经被应用于数据分类。Castillo [14]开发了一种用于文档分类的多策略分类器系统。将不同类型的分类器（例如：朴素贝叶斯，决策树）应用于文档的不同部分（例如：标题，引用）。 遗传算法应用于特征选择以及用于组合不同分类器的输出。 在最近的研究中，Zhang [15-18]提出了一个基于遗传算法的分类框架，智能地融合来自多个来源的证据，以改进文本文档的分类到预定义的类别[19]。

2.1 相似度测量

在他的研究中，Chen [19]引入了三种相似度测量方法：词袋，余弦和Okapi。 这样的基于内容的相似度测量已经应用于网页的内容。这三种相似度测量已经广泛应用于科学研究活动，特别是在文本分类领域[15-18]。为了计算这些相似度测量，文档需要被表示为向量，像在向量空间模型中[20]。 假设我们有一个有 个不同索引项的集合，一个文档可以表示如下：，其中表示文档中分配给项的权重。

词袋测量，两个文档和之间的相似度可以这样计算：

词袋函数

其中对应文档中出现的词语集合。

对于余弦测量，两个文档之间的相似度可以这样计算[21]：

余弦函数

对Okapi测量，两个文档之间的相似度可以这样计算：

Okapi

其中是文档中的词频，是整个集合中的文档频率，是整个集合中的文档数，是文档的长度，是集合中所有文档的平均长度。[19]。

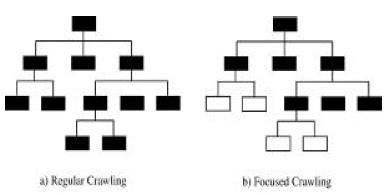
通过这些方程，我们可以推导出余弦相似度测量是对称的，而词袋和Okapi的相似度测量则不是[19]。

2.2 证据组合

Chen在网页的不同部分，包括正文和标题，使用了上述相似度测量的不同组合[19]。

表一 不同的证据

|  |  |
| --- | --- |
| 基于内容  的证据 | 在标题使用词袋 |
| 在标题使用余弦 |
| 在标题使用Okapi |
| 在正文使用词袋 |
| 在正文使用余弦 |
| 在正文使用Okapi |



2.3基于遗传算法的分类

上一节中显示的每种类型的证据都表示为文档×文档矩阵，并且用作基于遗传算法的分类框架的输入矩阵[19]。 矩阵定义为：



在这个矩阵中,是基于前面部分提到的一种相似度测量得到的和两个文档间的相似度值。遗传规划将尝试找到一个最佳非线性函数与矩阵相结合, 其中是证据类型的数量。通过这种非线性函数的组合得到的计算结果是定义为的输出矩阵：



是文档对之间的相似度矩阵。为了充分利用中表示的信息去预测在分类过程中文档的类标签，我们引入了一种基于最近邻分类器KNN的方法[22]。与相比，更密集，更精确，并能产生更好的分类结果[19]。

3.遗传规划

基于生物遗传和进化的原理，遗传规划（GP）是遗传算法（GA）的扩展，它是一组人工智能搜索算法，有很强的能力去有效遍历一个非常大的搜索空间，并找到近似的全局最优解，而不是局部最优解。遗传规划已被广泛使用，并证明在解决优化问题方面是有效的，如财务预测，工程设计，数据挖掘和运营管理[23]。GP能够解决传统方法不能轻易找到答案的复杂问题[19]。

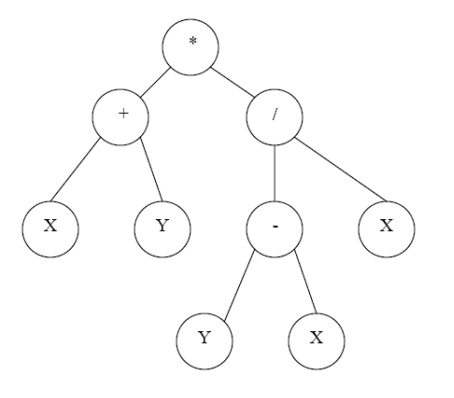


图1. GP树结构示例[19]

GP通过代复一代，不断变化的人口搜索“最佳的”解决方案。它的工作原理是反复对个体人口应用遗传变换（例如繁殖，交叉和突变），在后代中产生更多样化和更有表现力的个体。直接拷贝复制操作符，或使用一个更合适的词，将一些个体克隆到下一代中。选中复制的概率应与其适合性成正比。因此，越好的解决问题的方案，进入下一代的概率越高。虽然繁殖会保持人口中最好的个体，但交叉和突变会引入变换，并提供变种进入新的一代。交叉算子随机选择两组个体，在每两组中选择每个组中的最佳（根据适合度）个体作为亲本，交换每个亲本中随机选择的基因片段，并产生两个“孩子”。因此，“孩子”可以获得其优秀亲本的最佳片段，所以可能超过它们，从而提供对问题的更好的解决方案。由于父母是从“竞争”中选取出来的，良好的个体更有可能用于产生后代。突变操作随机地改变个体的基因代码。使用这些遗传操作，后代使个体在最后一代保持最佳适应性，并使用“更新鲜的空气”，为目标问题提供创造性的解决方案。更好的解决方案是通过继承和重组旧的，或是通过模拟达尔文进化更好地变异得到的[19]。

为了将GP应用于网页分类问题，需要定义GP系统的几个必需的关键组件。表2列出了这些基本组件及其说明。

|  |  |
| --- | --- |
| 表2. GP在分类中的应用 | |
| 终端 | 证据类型 |
| 功能 | +，-，\*，sqrt |
| 适应度函数 | Macro F1 |
| 遗传操作 | 复制，交叉，突变 |

3.1适应度函数

适应度函数的选择对最终分类结果有很大的影响[24]。一个优秀的相似度函数被定义为具有高值的相似度函数。当它应用于C类的文档y时，它对来自C类的文档进行排名，使与y具有更大相似度的那些文档排在第一位。适应性越高，适应度函数越好。我们选择使用的适应度函数是Macro F1。该算法使用kNN来预测文档的类别标签。其详细程序如下所示[19]。

使 R = 0, P = 0, T = 0

for （测试集合中的每个文档y） do

查找与y最相似的文档k

根据kNN算法使用文档k作为k最邻近点来预测文档y的类别

if 这是一个正确的预测 then

R = R + 1, P = P + 1

end if

T = T +1

end for

使 p = P/T, r = R/|C|

F=2p\*r/(p+r) (F 代表 Macro F1)

4.建议方法  
 如我们所提到的，估计期望的主题和页面之间内容的相似度的应用是每个聚焦爬虫的重要部分。这种方法的缺点之一就是如何确定最佳阀值。考虑低阀值可能导致输入大量不相关的页面。这些无效的页面会导致较低的效率。另一方面，考虑到为了符合用户的标准设定这样高的阀值可能导致丢失许多有效页面。

通过考虑最佳的可能阀值，我们可能要面临另一个问题。在这种情况下，由于与期望阀值的微小差异，可能存在大量的无效页面[5]。

我们提出了一个衰减概念来克服阀值的缺点。对于每个页面，我们在0和1之间设置一个变量。这些变量会显示衰减概念。每个变量值接近1的页面表示有更优的相似度。每个子页面以一个百分比减少继承亲本的衰减。我们将这个值设置为其亲本衰减的一半[5]。

当衰减值减小到小于阀值时，衰减概念的使用会被应用。在这种情况下，爬虫停止遍历新页面。实际上，我们使用两个阀值。这种方式下，子页面的衰减值总是减少。这样的假设是不正确的，因为我们可能面对一个与期望的搜索查询完全相关的页面。在这种情况下，如果相似度值大于阀值，则将衰减值重置为1。换句话说，对于每一个页面，我们都有两个值。一个是页面和查询之间的相似度值，另一个是放在1,1/2,1/4,1/8,...中的衰减值。此外，这种方法中具有三个阀值。主阀值或表示查询和页面之间的相似度。衰减阀值或中小于阀值的衰减值会导致爬行停止。对于具有高于[5]的相似性的页面，复位阀值或Tr使衰减值重置为1。

4.1 聚焦爬虫的爬行框架

在本节中，我们将介绍我们聚焦爬虫的高级步骤。

步骤1：发现最佳相似度函数。 从DMOZ[25]中获得的计算和非计算文档的数据集合将被用作训练和验证集合。分析这些网页的内容，并根据不同测量的相似度进行计算，例如词袋，余弦[21]和Okapi[26]。 GP将被用来去发现这些相似度测量组合的最佳相似度函数。新发现的相似度函数可以更精确地表示这些网页之间的相似度关系。发现的最佳相似度函数将用于分类步骤[19]。

步骤2：初始化。 在该步骤中，起始URL指向的网页由爬虫提取，形成基本集。这通常是元搜索引擎的第一个结果。

步骤3：分类。对于每个读取的网页，GP发现的最佳相似度函数将由KNN分类器决定这是否是计算相关的网页。如果是的话，这个网页将留下并被放入收藏。否则，此网页将被丢弃[19]。

步骤4：广度优先搜索。 使用广度优先搜索算法来获取新的网页。留下的相关网页的导出链接将被收集并放入爬行队列中。选择广度优先搜索的原因是它不是本地搜索算法（如最佳优先搜索），并且它不具有局部搜索算法的自然限制。 虽然广度优先搜索可能会增加爬行时间，但仍然被认为是解决由本地搜索引起的问题的好方法，因为爬行时间不是构建特定领域集合[19]时的关键因素。我们应用我们的衰减概念到每个页面。每个不符合预定义的阀值的页面都将导致爬行停止。

步骤5：元搜索。 将一些搜索引擎的前10个结果相组合并放入爬行队列中。元搜索步骤将尝试从整个搜索空间获取各种相关的URL，并且它不会受相关网络社区之间的界限，因为它不必根据超链接找到相关的页面[19]。

步骤6：终止.重复步骤3-5，直到本地收集库中的网页数量达到要求。

5.结果和实验

为了展示我们基于GP的分类器的能力，我们使用DMOZ [25]资源中的30％和10％。 在30％的数据集中，我们考虑了25％作为验证，其余作为测试。 我们将基于GP的分类器与简单基于SVM和基于组合的SVM进行比较。 对于10％，我们设置7％作为验证。

基于内容的SVM和基于组合的SVM之间的区别在于内核组合。 Joachims等人 [27]指明在他们的研究工作中如何通过复合内核组合不同的相似度测量。 他们的方法包含了一系列“内核保存”操作和简单易于理解的内核的组合，来构建日益匹配的特征空间。为了将这个方法应用于我们的应用领域，我们开始发现，我们每一种类型的证据都可以表示为正文档×文档矩阵，如第5节所述。因此，我们可以将每种类型的证据表示为内核矩阵。用于我们的最终特征空间的内核矩阵通过我们的初始内核矩阵的线性组合获得。

对于算法的评估，包括精度，召回和F的因素都基于以下参数计算[28]。

|  |  |  |
| --- | --- | --- |
| 表3.评价参数 | | |
| 页面 | 分配给 | 不分配给 |
| 属于 |  |  |
| 属于外的其他集群 |  |  |

精度显示算法的准确性，而召回代表搜索算法的完整性。

精度

召回

也可以如下计算：

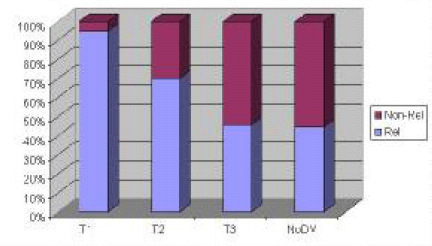


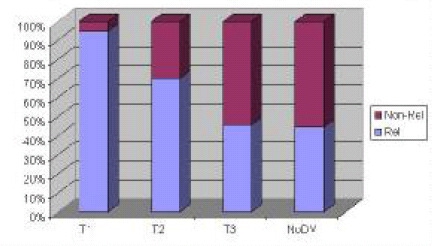
我们已经显示了10％和30％数据集的。我们在[19]中设置GP属性。对于三个阀值，我们设= 0.3和= 0.15和= 0.3。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 表4. 比较评价 | | | | | |
| 等级 | GP+元搜索+衰减 | GP+元搜索 | GP | 基于内容的SVM | 基于组合的SVM |
| 10% | 65% | 64% | 62% | 54% | 55% |
| 30% | 72% | 69% | 66% | 58% | 56% |

作为已知，可行的方法是使用GP与元搜索并用衰减概念去限制爬行页面。

在另一个实验中，为了确定我们的衰减限制的不同阀值，我们从[5]中得到启发。考虑到爬虫不同的刚度水平。他们设置Tr有三个连续宽松的，平均的，刚性的值0.1,0.2和0.3。



 图2. 爬行不同Tr 值时的准确性[5]

如图所示，在刚性方法中，存在最不相关的页面。 它显示，在宽松方法中，一半的抓取页是不相关的。 在另一个实验中，他们展示了在每种方法中消耗多少堆栈内存。

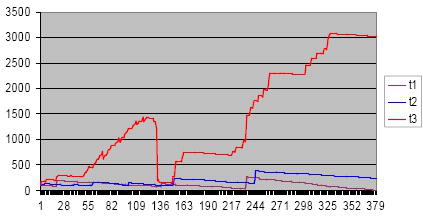


图4. 不同方法的堆栈大小[5]

显然，图中的快速倾斜表现出了相关页面的发现度。而下降表示缓存集中的一些不相关的页面。 在这份数据中，显然刚性方法节省了最多的内存[5]。

6.结论

在比较了普通网络爬虫和聚焦网络爬虫之后，我们为聚焦爬虫引入了一种新的方法。为了实现减少效率低下的目的，提出了一种新方法，利用遗传规划最有效地发现用于估计页面之间的相似度测量的最佳组合。这种遗传规划方法在网页之间获得了最好的相似度测量。 该方法根据不同的相似度测量，应用于网页的标题和正文。结果显示GP方法超越基于SVM的方法。甚至基于组合的SVM具有最小的准确度。此外，我们使用衰减概念方法来限制爬虫获取不相关的页面。 我们的衰减方法在遍历时动态地对页面计分，并且可以在遍历时消除或修改它们。以这种方式，爬虫可挑选出最相关的网页。

研究结果表明，与其他算法相比，本文所提出的方法可产生更好的准确性和更高的效率。