

# WWW 页面的文档分类技术

吴赣 程学旗 余智华

北京 2704 信箱智能中心软件部, 北京 100080

wugan@ncic.ac.cn

**摘要:** 人们从 INTERNET 网上查找信息的时候, 习惯按照一定的类别来获取自己感兴趣的信息。目前, 大多数网站在提供信息的时候, 采用的都是手工分类方式, 工作量比较大, 效率不高。本文对 WWW 页面的特点进行了分析, 并利用向量空间模型对网上的 HTML 页面进行了自动分类, 取得了一定的效果。

**关键词:** 向量空间模型 文本分类 页面分类

## Document categorization for the World Wide Web

Wu gan Cheng xueqi Yu zhihua

Software Department of NCIC, 2704 Box, Beijing 100080

wugan@ncic.ac.cn

**ABSTRACT:** When browsing in INTERNET, people are used to get information according to its category. Now, most websites provide information by manual classification, which costs more workload, but has low efficiency. This paper analyses the character of HTML categorization, and introduces method of categorizing HTML homepages by VSM.

**KEYWORDS:** Vector Space Model(VSM), Document Categorization, HTML Categorization

### 1 引言

随着 INTERNET 的普及和流行, 人们所能获得的信息量迅速增长, 为了更好的让用户选择自己所需要的信息, 许多信息服务站点都提供了信息分类浏览的功能。但是目前大多数站点的分类工作是由人工完成的, 投入大、信息更新慢, 很难适应 INTERNET 数据量大、变化快的特点, 使得用户在查找自己需要的信息的时候很不便。

向量空间模型 (VSM) 是 Salton 等人于 60 年代末提出来的, 是一种简便高效的文本表

示模型，已经被成功的运用于文本自动分类等研究领域。利用文本分类和页面分类的相似性，可以在页面分类中利用文本分类的已有成熟的方法和技術，来实现页面分类的目的。

用户面对 Internet 时，不是担心信息量少，而是信息量太多以后不知道如何找到自己需要的信息，页面分类工作对信息过滤、信息检索、站点自动分类等工作都有非常重要的帮助作用，为用户快速准确的找到自己所需要的信息提供了有力的工具。同时，自动分类技术能有效的提高站点运转效率，充分发挥网络信息速度快，含量大的特点。因此有必要考虑用计算机来提高页面分类的效率。

曙光公司以提供各种 Internet 信息服务技术为目标，进行了许多相关工作，包括利用向量空间模型帮助 Internet 信息站点进行 WWW 页面和站点的自动分类工作，在这里主要介绍其中的页面分类的工作。

## 2 向量空间模型与传统文本分类简介

向量空间模型简而言之就是把文本表征成由特征项构成的向量空间中的一个点，通过计算向量之间的距离，来判定文本之间的相似程度。采用该模型的文本分类方法一般步骤是：先通过对训练语料的学习对每个类建立特征向量作为该类的表征，然后对每一个新的文档，也求出其特征向量，然后依次计算该向量和各个类的特性向量的距离，选取距离大小符合域值的类别作为该文本所属的最终类别。这里涉及到的关键问题是特征项的选取和权重的计算。

目前，一般采用词做为文本的特征项，由于高频词的统计意义不大，所以在统计过程中一般都滤去。而权重计算的目的是要正确突出每个词在文章中的重要程度，一般来讲某个词在某文本中经常出现，而在其他文本中不常出现，就说明该词对该文本或该类文本更具有代表性，应该具有更高的权重。

向量空间模型计算向量距离的时候一般采用向量的夹角余弦来表示，计算公式如下 ( $V_1$   $V_2$  表示两个文档)，它表明两个文档之间相同的词越多并且这些词的权重越高，则其距离越近。

$$\text{Sim}(V_1, V_2) = \frac{\sum_{i=1}^n w_i \cdot v_{1i} \cdot v_{2i}}{\sqrt{(\sum_{i=1}^n w_i^2 \cdot v_{1i}^2)} \sqrt{(\sum_{i=1}^n w_i^2 \cdot v_{2i}^2)}} \quad (1)$$

传统的采用向量空间模型的文本分类方法已经取得的了不少成果。目前传统文本分类的封闭测试查全率能达到 85%左右，准确率能达到 90%左右，甚至更高，开放测试的查全率和准确率也都有 80%以上，甚至 90%。在具体实现上，主要差别是在特征向量的选取上和特征项的权值计算上，这两点也是向量空间模型最关键的地方。另外，传统的文本分类的应用对象主要面向一些大型数据发行单位，具有很大的针对性，文本比较规范，统一。

### 3 文本分类和页面分类的异同

文本分类和页面分类归根到底都是对文本信息的分类，都存在着文本信息的表示、分类信息的获取等。正是基于这样的共性使得我们可以借鉴文本分类中成熟的技术来处理页面分类问题。但是文本分类和页面分类又有不同，主要集中在如下几个方面：

1. 一般来说文本分类的类别比较少，一般在几十类左右；而页面分类的类别往往比较多，达到上百类是很正常的。
2. 文本分类中的文本风格比较一致，信息来源相对封闭，准确性容易提高；而网上信息的来源五花八门，没有什么固定的风格，信息来源相对是开放的，准确性不容易提高。
3. 文本分类的类别体系相对比较容易确定，主要可以依据应用领域或公众认可的分类标准。而页面分类由于面对的是用户的兴趣，很难有一个统一的标准，只能依据适用原则。
4. 文本分类中的类别一般没有层次关系；而页面分类为了便于用户浏览和选择，一般都要求类有层次关系。
5. 文本分类中的类体系一旦确定就基本不动，而网上的信息是不断的变化，有人在后台负责维护，并且往往会根据实际情况做一定的调整。另外，根据不同的站点要求，相应的类别体系也会相差很大。

考虑到以上不同，页面分类需要解决以下几个问题：

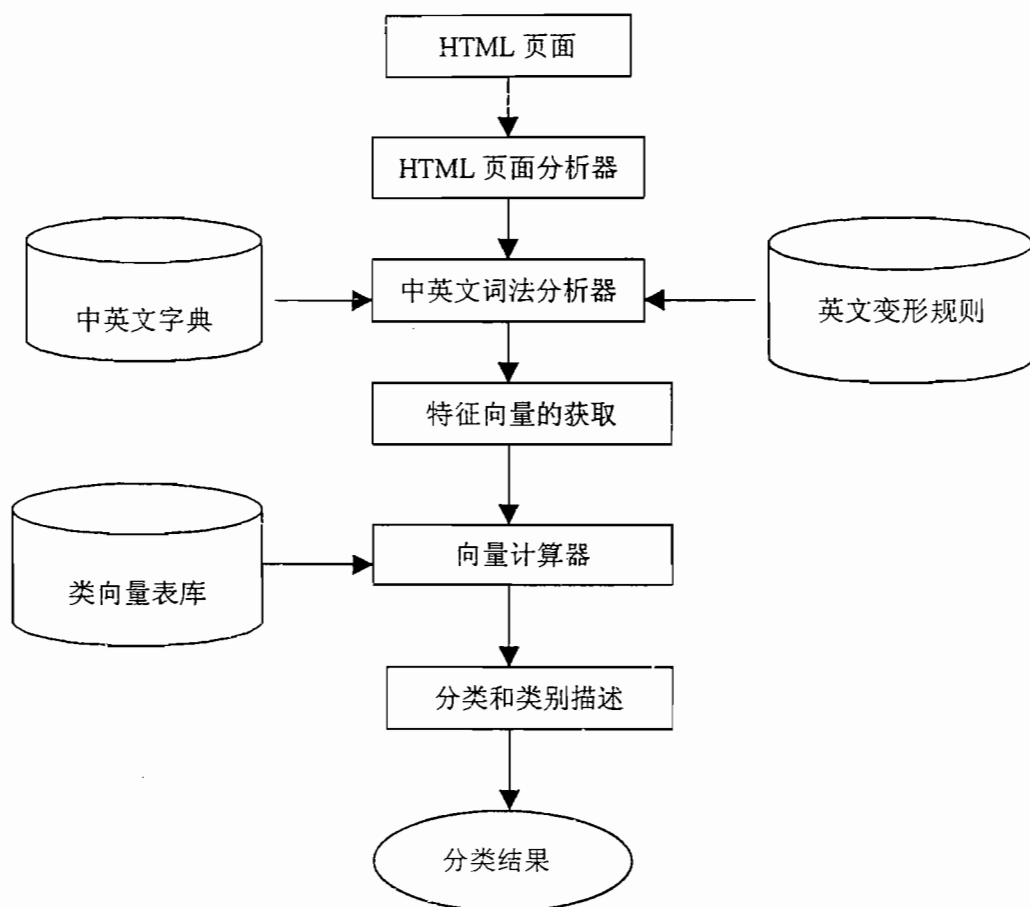
1. 抽取 Web 页面中的文本信息。
2. 算法的高效性。类别的增多对算法的效率提出了更高的要求，在计算过程中可以进行一定的优化处理。
3. 一个适用的分类体系。所谓适用是指类信息区别明显，计算机对这些类可操作，而不管这个类是否符合某个分类标准。比如：“美容美发”类是用户可能关心的类别，但不属于任何那个分类标准。
4. 支持带层次的分类体系。网上的分类信息一般是带结构的，比如：“体育类”-->“棋牌”-->“桥牌”。
5. 支持英文文本的分类。网上的不但有中文信息，还有英文信息，因此有必要提供英文的分类功能。
6. 提供方便的接口，以便网站工作人员的维护。
7. 充分利用 HTML 文本中的 TAG 以及 META 信息。
8. 充分利用页面的内部布局结构来获取隐藏在内容之外的分类信息。

应当指出的是，自动页面分类应该是一种计算机辅助的方法，网上信息太过浩繁，希望计算机能一劳永逸的帮助人解决分类问题是不现实的，计算机主要是帮助人们做那些类别特征区别比较明显的工作。

## 4 页面分类的有关处理技术

1. HTML 文本的分析。由于信息来源于网上，所以大量的信息是 HTML 文本，在进行词法处理前，必须删除其中的 HTML 标记。目前，网上有不少关于 HTML 语言分析的程序可以利用，我们做的工作主要是根据自己的需要添加相应的处理模块和根据 HTML 版本的更新增加处理的 HTML 标记。
2. 词法处理。包括中文词法和英文词法。前者的主要工作是分词，由于分词是个频繁的操作，所以效率问题是最重要的，相对而言，由于该计算模型不涉及深层的语义理解，所以精度的要求相对低些。另一个问题是领域词的收集，由于页面分类涉及的面广，因此有许多领域词不会出现在通用的词库中，而这些领域词对该领域页面的辨识往往是比较重要的，所以必须进行领域词的收集，考虑到每个领域专门收集工作量比较大，因此采用在选取训练语料时，用工具对语料中的领域词进行收集，并在此基础上扩展的做法。英文词法的处理主要是词形变换，因为英文的词形变换很活跃，而我们希望同一个词的不同词形变化的词频统计是在一个词上，而不是每个变形统计自己的频率。比如我们希望一个词单数和复数的频率统计在一起，而不是分开。因此有必要对英文进行词形变化后在进行统计。词法处理还包括停用词的标注。
3. 计算效率。为了提高计算效率，在计算机内部为每个词分配了一个整数索引，使得内部对字符串的操作转化为对整数的操作，而整数操作的速度快于字符串操作，占用的内存也少且长度相同。同时设计一个计算缓冲区，使得每个词都直接进行映射计算，大大加快了计算速度。
4. 分类体系。不同分类要求的分类体系也不同，由于没有依据的标准，只能是从现有网站上参照已有的一些类，再根据自己的需要进行删改。我们目前的类别设定主要是面向家庭的需要，同时考虑到类类之间的可区别性。目前进行的工作中共有 13 个大类，60 多个二级子类以及更多的三、四级子类。类的划分和删改在工作过程中还需不断调整。注意在这里页面分类已经和站点分类放在一起说了，事实上，一个站点类往往对应页面分类中的一个较高层次，比如“体育”类，因此在站点分类中，页面分类接口必须保证有关“体育”的小类信息都能正确的叠加到“体育”大类中。另外，一些类并不需要，也不适合计算机处理，比如：“报纸类”，做为用户浏览，这样的一个类是需要的，但让计算机去分“报纸类”和其它页面分类中的类的差别，从而在此基础上自动分出某个站点是“报纸类”则是不容易的，至少用这样的分类办法效果不会好，因为这个站点内一定包含了各种类的页面，会让计算机无所适从的。同时，各大著名报刊的网站是相对固定的，完全可以人工收集。总之，给用户浏览的类是根据用户需要组织的，但那些类适合计算机做，那些应该人做是应该有所区别的。
5. 模板接口。系统允许用户自己定义显示模板，用户只需要在模板中要显示分类或检索结果的地方留下标记，系统将自动用输出结果替换。这样方便了用户的页面设计需求，提高了灵活性。

## 5 页面分类实现流程图



其中 html 页面是由页面采集器从网上采集信息后，系统直接从本地记录库中读取采集信息，并根据该信息从缓存库中获得相应的页面文档。

整个处理过程如下：HTML 页面分析器从页面中抽取文本信息，并把该文本送入中英文词法分析器，如果遇到中文串则进行分词，如果遇到英文串则先查字典，如果不在字典中，则进行词形变化，如果变化后词不在英文字典中则抛弃，然后对词法处理结果进行统计，并根据域值进行筛选，根据筛选的结果和设定的域值来决定是采用中文向量计算还是英文向量计算或者是两者计算结果按权重叠加。最后根据计算的结果进行排序，按减序列出所有认为有效的类别。该分类结果可以提供给上层应用使用，由上层应用决定数据的取舍。

## 7 结束语

实现系统的核心速度非常快，在 PII/266 64M Windows NT 4.0 机器上，在类的总向量长度为 4 万左右时，1 秒钟可以处理大小不超过 5K 的纯文本 20 篇左右。系统中文字典含词 8 万条左右，英文字典含词 4 万 6 千条。运行表明系统分类速度快于系统的采集速度，说明在一个包括采集，分析，处理，浏览功能的完整信息服务平台上，该分类方法不会成为系统运行的瓶颈。

采用向量空间模型对页面进行分类，虽然取得了一定效果，但是页面分类还存在着如下一些问题：

1. 页面信息提取不够。HTML 标记语言包含了丰富的信息，<title>、<H1>等标记都标明了其信息的与众不同，目前对这些标记的处理考虑的还太少，基本上都是看成一般的文本。

2. 页面过于短小。许多 HTML 页面都太短，以至于被滤去或者特征值的抽取非常少导致计算的结果也非常小，为上层的应用带来一些不便。

3. 页面类型各有不同。真正有分类价值的页面是那些包含实际内容的页面，而不是索引之类的页面。而目前是所有的页面都参与分类，页面类型的确定要根据其所处的站点链接结构中的地位而定，这也是个复杂的问题。

进一步的问题还等待我们去想办法解决。

### 参 考 文 献

- [1] 黄萱筭 大规模中文文本的检索、分类与摘要研究 复旦大学博士论文 1998 5
- [2] 邹涛 王继成 黄源 张福炎. 中文文档自动分类系统的设计与实现. 中文信息学报. 1999 3
- [3] 余智华. WWW 站点分析与分类. 计算所硕士论文 1999 5