

文章编号:1001-9081(2005)02-0301-04

基于虚拟表示模型的 Web 页面模块化设计方法

熊 茜,朱征宇,朱庆生
(重庆大学 计算机学院,重庆 400044)
(remember230@sina.com)

摘 要:在虚拟网页技术基础上,借鉴模块化程序设计思想,提出了 Web 页面模块化设计方法。将虚拟网页技术与模块化相结合,可显著改变信息的组织与存储方式,具有支持模块级的网页设计复用、快速重组、扩展与更新等显著特点。

关键词:网页;虚拟网页;模块化设计;网页模块

中图分类号: TP311.11 **文献标识码:** A

Modularization design method for Web pages based on virtual description model

XIONG Qian, ZHU Zheng-Yu, ZHU Qing-sheng

(College of Computer Science and Engineering, Chongqing University, Chongqing 400044, China)

Abstract: Refer to the idea of Modularization program design method, this paper presented a new idea of Modularization design method for the Web pages base on its virtual description model. The combination of the new model and the modularization method greatly changed the way to organize and storage Web information. The new technique bring us many advantages on the design of Web page such as the easy reuse of a module, the quick recombination of a Web page, the easy update for the content of a Web page, the easy expanding of a Web page, etc.

Key words: Web page; virtual Web page; modularization method; module

0 引言

目前,HTML 网页等数据资源都是被看作孤立文件单个地进行设计和建立,没有很好的数据和设计重用方法。当增添新网页时,可能需要从头设计。这种设计方式耗费了大量的重复劳动,严重降低了开发效率。不仅使原来已有网页中的数据和设计信息难以重用,而且可能带来新的数据和设计冗余,增加网站信息维护和扩展的困难。

传统的模块化程序设计技术已经非常成熟,它使程序结构更加清晰,代码得到重用,其中参数的传递使其功能更加灵活。我们在文献[1]的基础上提出了一种切实可行的基于虚拟表示模型的 Web 页面模块化设计方法。文献[3]也曾从另一个角度简略探索了实际 HTML 网页的模块化设计问题,但本文方法在设计灵活性、模块化程度、模块复用性、以及网页内容动态更新等方面都更具特色和更为成熟。

1 虚拟网页技术简介

文献[1]从关系数据库模型的基表与视图技术得到启发,并借鉴其他半结构化数据模型^[2,4,5]的研究成果,通过建立素材库和定义扩展标记图 ETG(Extended Tag Graph)给出了网页的虚拟表示模型,它在网页的结构化检索中起着重要作用^[6]。

虚拟网页的设计过程颠覆了传统的网页设计方式。把设

计 Web 页面(HTML 文档)所需的所有基本信息(包括文本和属性)放入一个素材库中,再基于此素材库来定义一个虚拟网页,间接地表示 HTML 文档,而不是采用 HTML 语言去直接编写一个实际的 HTML 文档。这样的一个虚拟网页,只包含了网页的结构信息。当服务器响应客户浏览器对虚拟网页发出的调用请求时,在素材库中按指针搜寻内容,并动态生成 HTML 文档返回给浏览器。这一过程对用户来说是“透明”的,即所谓“虚拟”。

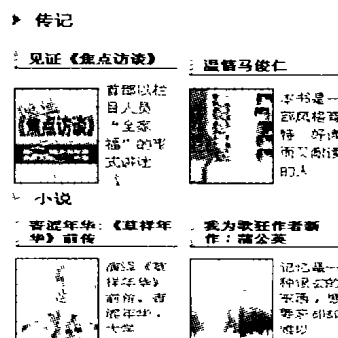


图 1 “网上书市”网页

如图 1 设有实际网页“网上书市”,其 HTML 文档如下(<table> 的上标数字仅用于区分不同的 HTML 段落,以方便后面模块设计的说明):

```
<html>
<head> <title> 网上书市 </title> </head>
```

收稿日期:2004-07-21;修订日期:2004-10-14

作者简介:熊茜(1981-),女,硕士研究生,主要研究方向:Web 服务;朱征宇(1959-),男,副教授,博士,主要研究方向:电子商务、Web 智能检索、Web 个性化服务;朱庆生(1956-),男,教授,博士生导师,主要研究方向:图像处理、多媒体技术。

```

<body>
...
<tr>
...
<table1>
<tr><td>传记</td></tr>
<tr>
<td><table10>
<tr><td>见证*焦点访谈*</td></tr>
<tr><td><img src = .../jdft.jpg></td>
<td>首部以栏目人员...</td></tr>
</table10></td>
<td><table11>
<tr><td>温情马俊仁</td></tr>
<tr><td><img src = .../mjr.jpg></td>
<td>本书是一部...</td></tr>
</table11></td>
...
</tr>
</table1>
<table2>
<tr><td>小说</td></tr>
<tr>
<td><table20>
<tr><td>青涩年华:《草样年华》前传
</td></tr>
<tr><td><img src = .../qcnh.jpg></td>
<td>演绎《草样年华》...</td></tr>
</table20></td>
<td><table21>
<tr><td>我为歌狂作者新作:蒲公英
</td></tr>
<tr><td><img src = .../pgy.jpg>
</td><td>记忆是一种很玄的东西...
</td></tr>
</table21></td>
...
</tr>
</table2>
...
</body>
</html>

```

则该页面可采用如下虚拟网页表来描述(@代表指针,指代如表1所示的素材库^[1]中具体内容或属性值,后跟数字代表在素材库中的编号):

```

<html>
<head><title>@0</title></head>
<body>
...
<tr>
...
<table1>
<tr><td>@10</td></tr>
<tr>
<td><table10>
<tr><td>@11</td></tr>
<tr><td>@12</td><td>@13</td>
</tr>
</table10></td>
<td><table11>
<tr><td>@14</td></tr>
<tr><td>@15</td><td>@16</td>
</tr>
</table11></td>
...
</tr>
</table1>
<table2>
<tr><td>@30</td></tr>

```

```

<tr>
<td><table20>
<tr><td>@31</td></tr>
<tr><td>@32</td><td>@33</td>
</tr>
</table20></td>
<td><table21>
<tr><td>@34</td></tr>
<tr><td>@35</td><td>@36</td>
</tr>
</table21></td>
...
</tr>
</table2>
...
</body>
</html>

```

表1 素材库片段:内容素材

POINT ID	CONTENT
0	网上书市
...	...
10	传记
11	见证*焦点访谈*
12	
13	首部以栏目人员...
...	...
30	青涩年华:《草样年华》前传
31	
...	...

2 Web页面的模块化设计方法

受模块化程序设计思想的启发,若把编写一张网页的HTML文档描述文件视为编写一种程序,那么其中大量重复的版块就是一种子程序,如标头、广告、表格和版权信息等描述。若将它们提取出来看作独立模块在模块库内存储,则对其进行的添加、更新、删除等操作将同时体现在引用它们的所有网页里。因此,提出了一种全新的基于虚拟表示模型的Web页面模块化设计方法。下面通过对“网上书市”虚拟网页进行分析设计的过程,详细说明该技术的基本思想。

2.1 网页模块的引入

在“网上书市”虚拟网页中,因“传记”、“小说”两部分内容描述相对独立,可以抽取出来分别做成网页模块。为了描述对网页模块的引用,我们将HTML标识符集进行扩充,引入<module>作为模块标识符,并使用标示属性ModuleID来指明所引用网页模块的唯一标识。那么该虚拟网页可精简如下:

```

<html>
<head><title>@0</title></head>
<body>
...
<tr>
...
<module id = 1>
<module id = 2>
...
</body>
</html>

```

而该虚拟网页所引用的两网页模块文件(其显示片段分别如图2(a)、(b)所示)均存放在的模块库中,“传记”和“小

说”的模块文件如下:

```
<table1>
  <tr><td>@10</td></tr>
  <tr>
    <td><table10>
      <tr><td>@11</td></tr>
      <tr><td>@12</td><td>@13</td></tr>
    </table10></td>
    <td><table11>
      <tr><td>@14</td></tr>
      <tr><td>@15</td><td>@16</td></tr>
    </table11></td>
  </tr>
</table1>
<table2>
  <tr><td>@30</td></tr>
  <tr>
    <td><table20>
      <tr><td>@31</td></tr>
      <tr><td>@32</td><td>@33</td></tr>
    </table20></td>
    <td><table21>
      <tr><td>@34</td></tr>
      <tr><td>@35</td><td>@36</td></tr>
    </table21></td>
  </tr>
</table2>
```

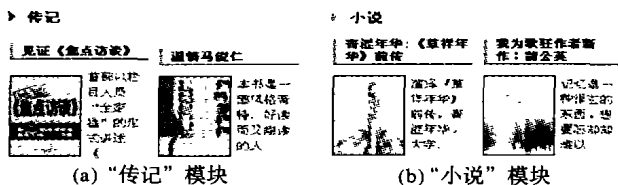


图2 网页模块文件

2.2 网页模块的嵌套

由模块化程序设计中允许子程序嵌套子程序、扩展标记图中存在节点子树嵌套节点子树,我们联想到网页模块嵌套的问题。随着网页复杂度的增加,模块嵌套深度便随之增加,如像主页类的复杂网页可能会有多层嵌套。

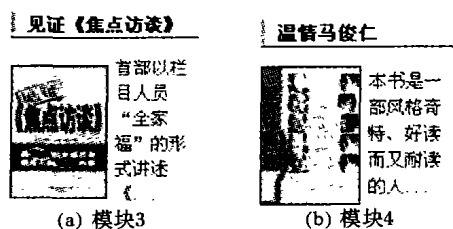


图3 网页模块

“传记”模块中,含有相对独立的“见证《焦点访谈》”和“温情马俊仁”两书的有关内容描述。故在设计该模块时,可以进一步地将它们抽取为更小的网页模块3和4(其显示片段如图3(a)、(b)所示),并保存在模块库中。于是,通过引用子模块3和4,“传记”模块的设计可进一步地简化如下:

```
<table1>
  <tr><td>@10</td></tr>
  <tr>
    <td><module id=3></td>
    <td><module id=4></td>
  </tr>
</table1>
```

描述模块3和4的模块文件如下:

```
<table10>
  <tr><td>@11</td></tr>
  <tr><td>@12</td><td>@13</td></tr>
</table10>
<table11>
  <tr><td>@14</td></tr>
  <tr><td>@15</td><td>@16</td></tr>
</table11>
```

2.3 支持参数传递

无参数的模块仅能支持虚拟网页集合中大量完全相同的子标记树的重用,而带参数传递的模块则可以支持大量标记结构相同而内容不同的子标记树的重用,从而使模块复用方式更加灵活。我们为 module 标记增设了参数列表属性 ParaList,支持从模块外部传入参数。

上例中,模块3和4的结构相同,只是其中的书名、封面和简介三部分内容不同。为避免对模块3和4的重复设计,可通过使用参数传递来精简设计。若编写带参数的模块7存于模块库中,并允许以形式 <module id=num ParaList={p1, p2, p3, ...}> 在其他模块中进行引用(“#”号后的数字代表在参数列表中的位置,如“#1”将被属性参数值 p1 代替,“#2”将被 p2 代替,依次类推),则可将“传记”模块的设计简化如下:

```
<table1>
  <tr><td>@10</td></tr>
  <tr>
    <td><module id=7 ParaList={@11 @12 @13}></td>
    <td><module id=7 ParaList={@14 @15 @16}></td>
  </tr>
</table1>
```

带参数的网页模块7的模块库片段如下:

```
<table>
  <tr><td>#1</td></tr>
  <tr><td>#2</td><td>#3</td></tr>
</table>
```

进一步地,为避免对“传记”和“小说”模块的重复设计,可编写带参数的模块8存于模块库中,从而可将“网上书市”虚拟网页的设计进一步地简化为:

```
<html>
  <head><title>@0</title></head>
  <body>
    ...
    <tr>
      ...
      <module id=8 ParaList={@10 @11 @12 @13 @14 @15 @16}>
      <module id=8 ParaList={@30 @31 @32 @33 @34 @35 @36}>
    </tr>
  </body>
</html>
```

带参数的网页模块8的模块库片段:

```
<table>
  <tr><td>#1</td></tr>
  <tr>
    <td><module id=7 ParaList={#2 #3 #4}></td>
    <td><module id=7 ParaList={#5 #6 #7}></td>
  </tr>
</table>
```

特别地,参数传递也分为“形参”和“实参”两种。传递

指针则为“形参”，而传递文本之类的实际内容则为“实参”。如可将上例中的指针@10和@30分别改写为文本串“传记”和“小说”，则为“实参”传递。

顺便指出，由于允许传递“形参”，故对素材库中内容的修改直接体现在传入的参数上。不仅真正实现了内容与结构的分离，而且直接支持对网页内容的动态更新而不必修改网页，这对于数据的快速更新、降低冗余、保持一致性特别有效。

关于“网上书市”页面的虚拟网页模块化设计过程(从图4的演化)看出，采用带参数的网页模块引用，使网页模块设计更为精简，组成结构更为清晰，设计工作更为简化，网页的扩展和维护也更为方便和高效。仅需设计上述的虚拟网页文件与网页模块7和8，即可完成“网上书市”页面的设计。

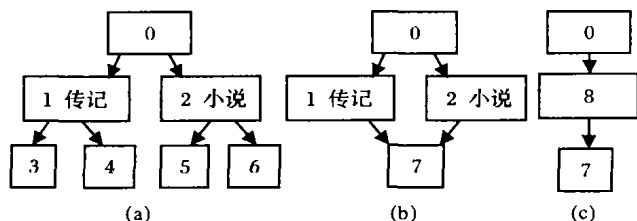


图4 “网上书市”模块调用图及演化过程

2.4 支持“变脸”引用

众多商业网站为满足特定的用户视觉需求，可能需要经常变换版式的设计信息。但只求一种风格的改变，并不希望变动网页里的内容信息，称之为“变脸”。

通过允许在网页模块调用中加入版式属性参数 Pattern，就可以达到这种效果。为允许在同一网页模块中包含多种版式设计信息，我们允许在模块描述文件中使用类似程序设计中的“Case”控制结构。在模块解析器对虚拟网页中包含模块进行解析时，将根据模块引用过程中传递参数值的不同而动态地转向到不同版式。通过这种模块级的设计控制，可使用户在得到不同的视觉享受时仍然可以浏览到相同的内容信息。这一功能对设计追求风格多变的网站应用来说特别有用。

例如，在设计“传记”模块时，我们可以设计多种版式来显示两本书的信息，以满足不同用户的浏览习惯。如可设计以横排方式显示(图5(a))的模块9(用于 Pattern = 1 情形)和以竖排方式显示(图5(b))的模块10(用于 Pattern = 2 情形)，并将它们存于模块库中。进而在“传记”模块的设计中，通过加入“Case”控制结构实现对横、竖排版式的选择，将其修改为(其模块调用图如图6所示)：

```
<table>
  <tr><td>@10</td></tr>
  <tr>
    <case 1><module id=9></case>
    <case 2><module id=10></case>
  </tr>
</table>
```

支持多版式的模块文件的模块库片段如下：

```
<td><module id=7 ParaList={@11 @12 @13}></td>
<td><module id=7 ParaList={@14 @15 @16}></td>
<tr><module id=7 ParaList={@11 @12 @13}></tr>
<tr><module id=7 ParaList={@14 @15 @16}></tr>
```

当更上层的“网上书市”虚拟网页引用该“传记”模块时，可通过引用形式<module id=1, Pattern=num>来实现“传记”模块的“变脸”(num为1或2)。模块解析器在解析虚拟网页中对该

模块的引用时，需消除该模块中的“Case”控制结构，将根据当前 Pattern 值是1或2选择采用模块9或10进行替换。

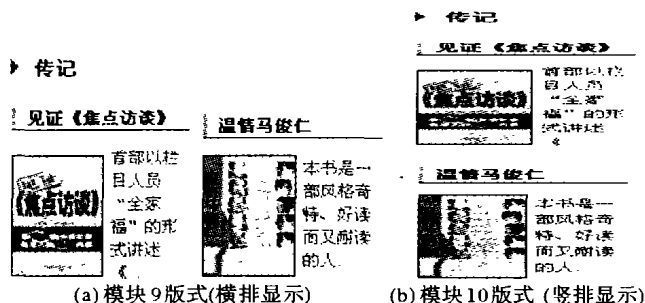


图5 多种显示方式

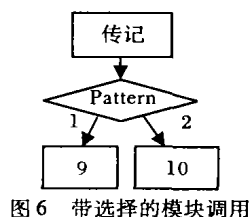


图6 带选择的模块调用

当然，显示风格的变化并不仅仅局限于内容的格式排列，其他一些网页风格设计技巧也可得到体现。通过修改模块调用中的版式属性 Pattern 的值或参数列表属性 ParaList 的值(如传递不同的颜色、字体、位置、表格属性等参数值)，甚至不需修改网页，而仅仅调整素材库中内容，即可使网页能够在瞬间达到绚丽多变的效果，可有效缩短用传统方式重新规划设计新版式的时间。

2.5 支持 HTML 语言的通用格式

模块和其他页面元素一样，也可以有诸如 name、bgcolor、width、height 等属性，因此，应当为 Module 标记再增加 AttributeList 显示参数列表属性，使网页模块的设计符合 HTML 语言的通用格式。例如可在“网上书市”虚拟网页的设计中可采用如下方式引用“传记”模块：

```
<module id=1 ParaList={@10} bgcolor="#00FF" width="440">
```

3 模块引用的形式化文法

通过以上实例的分析，基于虚拟表示模型的网页模块化设计方法所蕴涵的强大功能和作用。下面总结给出网页模块引用的形式文法：

```
Module_TAG := <module id=ModuleID [ Pattern =
  @PatternPointer] [ ParaList=Parameters] [ AttributeList] >
Parameters := Parameter{ Parameter,...}
Parameter := @Pointer | ContentValue
ContentValue := Text | Date | Number | ...
AttributeList := name[ AttributeList] | bgcolor[ AttributeList]
  | width[ AttributeList] | height[ AttributeList] | ...
```

其中，module 为模块标识符，“<”和“>”分别为该标记的首、尾标识。ModuleID 为模块的标识，@PatternPointer 为可选版式指针，Parameters 为可选参数列表，@Pointer 为素材指针，AttributeList 为可选属性列表。

4 系统框架

在文献[1]给出的虚拟网站系统框架基础上，通过增加模块库和模块解析器等，可改进得到新的支持虚拟网页模块化设计的系统框架，如图7。注意，在设计新的模块化虚拟网

(下转第308页)

由于词分布的改善有限,词分布只能作为重排序的一个辅助量。因为归一化的词频统计(也就是方法2)的效果较好,因此,将词分布应用在了方案二返回的有序文集的前30个文档上。选择的原因是:在不同的选择下,将词分布方法应用在前30个文档上,可能产生新的质量更好的文集。这是一个经验数值。按方法3的方式来应用词分布这个特征量进行排序。

在应用了词分布这个特征量后,采用下面的公式来计算重排序列的优劣:

设在一个有 s 个文档的序列中,相关文档的序号为: N_1, N_2, \dots, N_m , 显然有 $m < s$ 。 $P = (N_1 + N_2 + \dots + N_m) / m$ 。

如果相关文档全部排在序列的前面, P 值会最小。也就是说, P 值越小, 序列越优。因此, 这个值可以正确的反映出序列的优劣。具体的计算结果见图5(横坐标为样本号)。

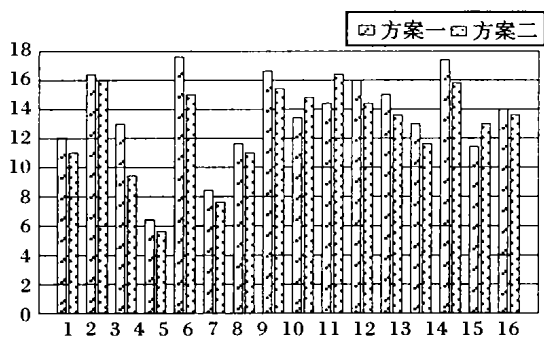


图5 各排序方案的30页面的P值

图5中方案一为归一化的词频统计的质量评价结果,

方案二为应用了词分布的质量评价结果。在多数情况下,方案二优于方案一。在少数情况下,应用词分布并没有改善有序文集的命中率。

实验使用了一台配置为 AMD AthlonXP 1700 +, 256MB 内存的计算机。每个样本数据的计算时间小于 10s。这个时间的花费在实际的应用中应该是可以接受的。

4 结语

应用词频统计和词分布统计方法,特别是词分布后的归一化词频统计可以有效地提高搜索引擎的返回的页面集的准确率,也即提高了有序页面集的相关文本的命中率。这样处理后的结果使用户可以在更短的时间内找到有用的资料,提高了人们的工作效率。将该算法应用在智能知识搜索引擎中可以较好地实现 Internet 上知识信息的采集工作。

参考文献:

- [1] BRIN S, Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine[J/OL]. <http://www-db.stanford.edu/~backrub/google.html>, 1997.
- [2] iProspect, iProspect's Search Engine User Attitudes Survey Results [DB/OL]. <http://www.iprospect.com/>, 2004.
- [3] Ed Greengrass. Information Retrieval: A Survey[J/OL]. <http://www.csee.umbc.edu/cadip/>, 2000.
- [4] 陆玉昌, 鲁明羽, 李凡, 等. 向量空间中单词权重函数的分析和构造[J]. 计算机研究与发展, 2002, 39: 1205.
- [5] WEBB N. Statistics for Natural Language Processing[DB/OL]. <http://citeseer.ist.psu.edu/>, 2002.

(上接第304页)

站时,不再是孤立、单个地设计每个网页,而是基于模块库和素材库上,按照虚拟化和模块化设计方法、通过模块设计与复用进行虚拟网页设计。

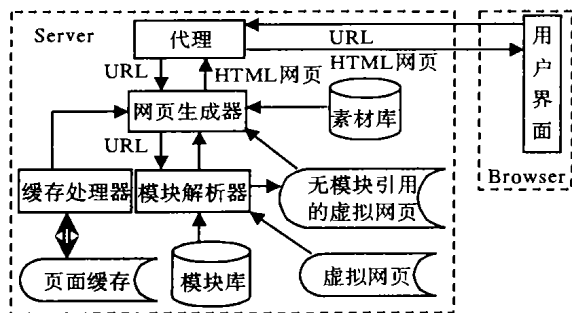


图7 支持虚拟网页模块化设计的系统框架

系统处理过程说明:

当代理模块截获到用户的 URL 浏览请求时,通过网页生成器将 URL 传递给模块解析器,模块解析器基于模块库消出(该 URL 对应的)虚拟网页中所有模块引用得到无模块引用的虚拟网页,网页生成器在基于素材库动态生成 HTML 网页,最后代理模块将 HTML 网页返回给用户浏览器。对于那些被频繁浏览的热点网页,我们增加一个页面缓存装置,临时存放动态形成的 HTML 网页,可明显减小网页动态生成的额外开销,提高对浏览器的响应效率。

虽然虚拟网页模块化设计方法确实在许多方面提高了 Web 性能,但对用户来说访问是“透明”的,即与访问其他普通网站并无什么明显不同。

5 结语

虚拟网页模型与模块化思想的结合是 Web 页面设计的又一创新。虚拟网页模块化设计技术具有如下显著特点:1) 模块级设计共享与细粒度数据共享。不仅可以通过设计独立模块实现设计模块复用,而且可以通过传递和修改 Module 标记的属性参数实现对网页段落风格变换和动态信息的传递。2) 新的网页模块共享机制对于解决网页资源表示中的数据冗余、数据和设计模块重用、网页内容动态更新等问题非常有效。3) 通过对网页模块的修改、更新、扩展和灵活引用,使得模块级的网页快速维护与更新、灵活重组与扩展更加方便易行。新的网页表示与设计方法融合了 HTML 语言中 CSS、Frame 和基于数据库的动态网页等技术的许多优点,而且在许多情形的网页设计更为方便、灵活和快捷。

参考文献:

- [1] 朱征宇, 朱庆生, 王茜. 基于扩展标记图的虚拟网页技术[J]. 计算机科学, 2001, 28(11): 80-83.
- [2] 陈滢, 徐宏炳, 王能斌. 基于标记图的 Web 数据模型[J]. 计算机学报, 1999, 22(3): 306-312.
- [3] 张宏森, 朱征宇. 基于模块的网页设计技术[J]. 计算机应用研究, 2002, 19(2): 49-50.
- [4] ABITEBOUL S, QUASS D, MCHUGH J, et al. The Lorel query language for semistructured data[J]. Int J Digit Libr, 1997, 1: 68-88.
- [5] 王静, 孟小峰. 半结构化数据的模式研究综述[J]. 计算机科学, 2001, 28(2): 6-10.
- [6] 朱征宇, 王亮, 赵银春, 等. 基于扩展标签图的网页信息重组技术[J]. 计算机科学, 2004, 31(5): 56-60.