第5章 Web页面逻辑链接块研究

链接块是 Web 页面中广泛存在的一种区块结构。在现有的涉及链接块的相关研究和应用中，存在两类主要问题：一是仅针对物理结构上的链接块，甚至仅针对 Block 级别的 HTML 元素链接块；二是链接块的发现与识别等处理都是建立在 Web 页面标签树的基础之上，这一方面影响了 Web 页面的处理速度，另外也无法应付当前互联网上多样化而又不规范的 Web 页面。针对目前相关方法的不足，本章提出逻辑链接块的概念和逻辑链接块的发现方法和判别方法。该方法无须解析网页标签树或者 DOM[62]树，相关处理程序健壮性大幅度提高，且不再受限于某些特定的 Block 层次的链接块，既可以处理 Block 层次的链接块，也可以处理 Inline 层次的链接块，而且还可以处理跨越块级元素的链接块。实验结果表明：本章提出的方法可以有效地发现页面中的逻辑链接块，判别规则简单，为链接块的识别及文本提取提供了一种新的方式，在对链接块粒度要求不高的场合具有广泛的应用前景，在其他 Web 信息处理和挖掘领域中也具有一定的实际应用价值。

5.1 引言

万维网是一个通过页面之间的链接构建起来的超大型复杂网络。链接在 Web 信息组织和展示、页面导航等方面发挥着巨大的作用。网络爬虫依靠 Web 页面之间的链接实现互联网的遍历爬行，互联网用户则正是依靠页面之间的链接实现同主题内容的“聚合”阅读。Web 页面中的链接往往以不同的粒度块来组织，粒度块越精细则所含链接的主题相关性越高；随着粒度块的增大，链接块的主题“内聚”性逐渐减弱。如图 5-1 所示，当粒度要求极

其精细时，则其将被划分为三个链接块；而当精细度要求不高时，则可以将其视为一个链接块，整个链接块的用途即“导航”。在针对链接块的相关研究中，根据研究目的不同，对链接块的粒度精细要求也将不同。在专门针对链接块进行分析的研究中，对链接块粒度的要求往往较为精细，而在其他非链接块研究中，如 Web 页面文本提取中，则对链接块的粒度要求不高。

图 5-1　链接块的粒度

在技术实现上，视觉上的分块往往也对应着块（block）级标签元素（block-level elements）[63]，目前的涉及链接块的应用和研究基本仅针对该实现方式。然而，由于网页设计技术和实现的多样性，视觉上的分块，在实现方式上并不总是采用 Block 类型标签实现，也有可能采用内联类型标签（inline elements）[63]实现，这也就意味着无法准确地预知设计者使用何种方式实现链接块，或者需要建立在对 HTML 标签属性的精细解析基础之上，这给基于海量 Web 数据的一些自动化应用带来了诸多麻烦。

5.2　相关研究及存在的问题

Web 页面链接块的研究历史悠久，对 Web 页面进行分块或者信息提取的方法众多。文献[64]将 Web 页面的抽取方法总结为基于 Wrapper、模板、机器学习、视觉布局特征、HTML 特征等五类。这五类方法同样可以适用于 Web 页面链接块的分块。其中，Wrapper 和模板法的通用性较差，且一般需要人工参与，并需要更新维护，极为耗时费力，鉴于此有研究人员提出了无须模板支持或人工监督的 Wrapper 算法[65-67]，并取得了较好的效果；机器学习的方法需要借助合适的训练集和适量的特征[68]，且难以完全脱离人工监督；利用视觉布局特征的方法的典型代表即 VIPS[69]，该方法虽然准确率高，但是对网页的解析要求过于精细，计算消耗大，面对大量非规范化的网页时健壮性难以保证，且在当前普遍采用 CSS[70]来控制各页面标签属性视觉呈现的情况下，还需要另行解析相关 CSS，最终导致解析任务量大，程序健壮性欠缺；基于 HTML 特征的相关方法多偏向一些启发式规则[71-74]或一些统计规律，通用性有待提高。此外，也有研究者提出了其他的一些方法，例

如，利用模糊神经网络实现页面分块的方法[75]、MSS 页面分块方法[76]等。虽然相关研究方法多种多样，各有特点，然而经过分析总结可以发现：目前关于 Web 页面链接块的发现和识别相关算法基本都是基于标签树[73, 77-81]，而 DOM[82]是一种构建标签树最为常见的方式；其他方法则都以 HTML 标签树或 DOM 为基础[83-84]。

另外，在对 Web 页面进行分块的相关研究中，有相当一部分基本都仅仅针对块级层次的 HTML 标签元素，如 div、table、tr、td 等，其中由于 table 功能的多样性和强劲性[85]，早期网页布局修饰和内容组织几乎对 table 不可或缺，相应的，部分文献也仅考虑了针对 table 布局的网页[81]，且未能很好地区分用于布局的 table 和用于内容组织的 table。Son 等[81]专门研究了基于 table 设计的网页，对 table 的两种作用做了区分并分别识别，实验证明了所提出方法的先进性。但仅针对 table 的处理方式局限性太大，目前的网页设计基本都是 table 和 div 共存，Uzun 等[77]同时考虑了这两种情况，先根据 div 和 td 获得分块信息，其次结合决策树生成抽取规则，取得了良好的效果，特别是在抽取速度上获得了和手工规则相当的性能；Wang 等[71]则提出了 BSU 概念，并基于此采用聚类和启发式规则两种方法实现页面信息抽取，比采用基于 div 和 table 的方法结果更好。

现有的对链接块进行分块的算法，尤其是基于标签树的各种方法需要 Web 页面遵从较好的规范，这种规范既包括 HTML、XHTML 等标签语法规范（如标签的配对关系），也包括语义设计方面的规范（如通过浏览器渲染后在视觉上呈现块状的内容在实际的代码中也会通过块级元素 div、table 等来呈现，视觉上的标题通过 h1、h2 等标签来呈现等）。但实际上，海量的 Web 页面中，有相当数量的 Web 页面并不遵从 HTML 等标签语法规范和语义设计规范。虽然 HTML 标签语法上的不规范性可以通过一些现有的或自行设计的 Web 页面规范化程序进行矫正，但并不能保证 100％的正确率；语义设计规范问题的矫正难度则更大。这就决定了基于标签树的各种方法仅能在设计规范或易于矫正的 Web 页面中获得良好的效果，在非规范化 Web 页面中则显得捉襟见肘。

由于在已有的 Web 页面处理相关研究中，一般只将块级标签对应的代码块称为块，这种处理方式虽然极大地提高了诸多 Web 页面处理的效果。然而在面对纷繁复杂的 Web 页面时，在某些情况下，这种处理方式可能带来两种后果：误判或无法检出。例如在很多 Web 页面中，存在着并非块级的广告，在页面正文抽取等研究领域，按传统的块级处理方式，无法检出这些

广告链接，如图 5-2 所示。

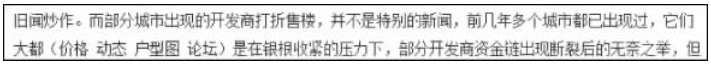

图 5-2　非块级的嵌入式广告链接

本章基于现有链接块识别方法的不足，提出标签距离和逻辑链接块的概念，并基于标签距离提出了逻辑链接块的发现和判别方法，并实现逻辑链接块的识别和判别。逻辑链接块的识别在对链接块精细力度要求不高的场合具有良好的应用场景。

5.3　方法及原理

为了后文表述方便，首先定义如下概念。

在 Web 页面的 HTML 代码中，两个标签之间存在着两类距离：代码距离和文本距离。分别定义如下。

代码距离：任意两个标签之间的代码距离即介于前一个标签的标签结束符“>”和后一个标签的标签开始符“<”之间所有内容的长度。在本章的计算中，将先去除各标签的属性然后才执行代码距离的计算，例如，“< div id＝"main"> ABC ”经过去除标签属性得到“< div > ABC ”。

文本距离：任意两个标签之间的文本距离即介于前一个标签的标签结束符“>”和后一个标签的标签开始符“<”之间所有文本的长度。

但是在计算文本距离时，遵从如下规则：①英文等字符以单词为统计单位，即一个单词长度计为 1；②中文等字符以单个字为统计单位，即一个汉字长度计为 1；③数字以一个完整数字为统计单位，即一个完整数字长度计为 1，例如“北京 2008”的长度计为 3；④日期字符串以日期整体为统计单位，即一个完整日期长度计为 1，例如“今天是 2014 年 3 月 8 日”的长度计为 4；⑤标点符号与汉字统计规则一样，但是若相邻的若干标点符号相同，则长度只计 1。

链接距离：即 Web 页面中相邻两个链接之间的距离。链接距离可以采用如下两种度量方式之中的一种：

（1）代码距离：即前一个链接的“”与后一个链接的“< a >”之间的代码距离。

（2）文本距离：即前一个链接的“”与后一个链接的“<a>”之间的文本距离。

逻辑块：即由不少于一个且相邻或相近的标签所构成的连续代码区域。逻辑块可能是一个标签块，也有可能是几个相邻或相近的标签块合并构成，且被包含在逻辑块中的各个标签并不要求都是完整的，被包含在逻辑块中的各个标签也不必是块级标签。如图 5-3 所示，A 与 B 为相邻的兄弟标签，构成逻辑块；A1 与 A2 均是 A 的相邻子标签，构成逻辑块；A2 与 B1 虽然隶属于不同的父标签，但 A2 与 B1 相近，通过 A 的后半部分代码和 B 的前半部分代码，可以最终使得 A2 和 B1 成为一个连续的代码区域，故也是逻辑块。

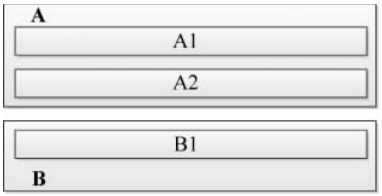

图 5-3　逻辑块

逻辑链接块：设某逻辑块中的链接数为 C_{link}，逻辑块内各相邻链接之间的距离为$(d_1, d_2, \cdots, d_{C_{\text{link}}-1})$，若该逻辑块满足如下条件，则称该逻辑块为逻辑链接块。

$$\begin{cases} C_{\text{link}} \geqslant C_t \\ \max(d_i) < d_t \end{cases} \tag{5-1}$$

其中，C_t 为链接块中最小链接数；d_t 为相邻链接之间所允许的最大值。C_t 决定了所识别出的逻辑链接块的大小，而 d_t 则表示链接块将链接纳入链接块的能力。

逻辑链接块的发现可以通过对 Web 页面的 HTML 代码从前往后扫描，对发现的链接，逐个计算与其相邻的链接的距离，当距离低于阈值 d_t 时，则记录链接数，并继续往后扫描，直至遇到相邻链接之间距离超过 d_t 时，判断当前积累的链接数是否超过 C_t，若超过，则表明一个链接块发现完毕，重新开始下一个链接块的发现过程。该逻辑链接块发现方法的好处在于，无须标签树支持，也就意味着无须在标签树解析或 DOM 解析上耗费大量的计算资源，从而也就避免了解析纷繁复杂且缺乏规范的 HTML 时的各种问题。

对逻辑链接块识别结果的评价，本章采用**链接覆盖率**（Link Coverage Rate，LCR）和**代码覆盖率**（Code Coverage Rate，CCR）两个指标，其表达如下：

$$\mathrm{LCR} = \frac{C_{\mathrm{BlockLinks}}}{C_{\mathrm{PageLinks}}} \tag{5-2}$$

$$\mathrm{CCR} = \frac{L_{\mathrm{Block}}}{L_{\mathrm{Page}}} \tag{5-3}$$

其中，$C_{\mathrm{BlockLinks}}$表示包含在逻辑链接块中的链接总数；$C_{\mathrm{PageLinks}}$则指 Web 页面中的链接总数；L_{Block}表示所识别的逻辑链接块代码长度总和；L_{Page}表示 Web 页面代码长度。

5.4 实验设计及结果分析

5.4.1 实验目的

下述实验的目的是验证上述所提出的逻辑链接块发现和判别方法的有效性，且探讨该方法在处理索引型和内容型 Web 页面时的效果与特性。

5.4.2 实验方案

实验所用原始 Web 页面数据是通过程序从互联网中随机爬取，然后对随机爬取的 Web 页面采用两种取样方式：

（1）**人工筛选**。人工筛选的 Web 页面数据来自于 5 家国内门户网站，即网易、新浪网、中国新闻网、中华网、凤凰网，每个网站均选取 16 个索引页（即门户首页或者各子频道首页）和 40 个内容页，共计 280 篇。

（2）**随机抽取**。随机抽取的索引页为 46 个，内容页为 256 个。鉴于 Web 页面文本抽取是逻辑链接块的可能最主要的潜在应用，故在筛选内容页时，尽量选择了多种不同类型的 Web 页面，如既有长篇幅的也有短篇幅，既有纯文字页面也有视频图片页面。

实验分为两组进行，每组实验又分别使用代码距离和文本距离作为链接之间距离的度量，试验在不同参数配置下索引页和内容页的链接块识别情况。为下文表述方便，将采用文本距离时的链接距离阈值记为 d_t^{t}，而将采用代码距离时的链接距离阈值记为 d_t^{c}。

第一组：设定 $C_t=3$，使用文本距离的情况下，$d_t^{\mathrm{t}}=\{5,10,\cdots,60\}$；使用

代码距离的情况下，$d_t^c=\{10,20,\cdots,120\}$。

第二组：设定 $C_t=\{2,3,\cdots,12\}$，使用文本距离的情况下，$d_t^t=40$；使用代码距离的情况下，$d_t^c=80$。

5.4.3 实验结果与分析

1. d_t^t 对 Web 页面链接块的影响

对任意 Web 页面，不难想象，随着相邻链接之间距离阈值 d_t^t 的增大，相邻超链接被归属到同一逻辑链接块中的可能性越大，各个逻辑链接块也将越大，在链接总数确定的情况下，逻辑链接块数也将越少。相应的，各链接块累计覆盖的链接数和代码区域也将越多，即链接覆盖率和代码覆盖率也将越高。图 5-4 中的实验数据证实了这一点。

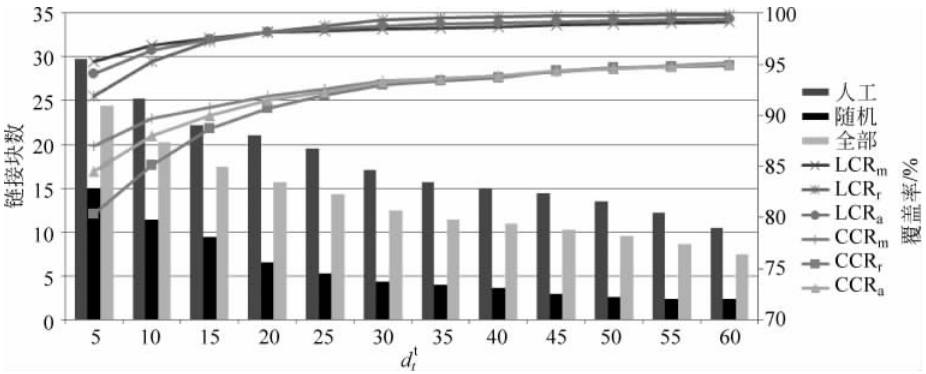

图 5-4 d_t^t 对逻辑链接块的影响——索引页

LCR_m、LCR_r、LCR_a 分别表示对实验攫取的 Web 页面数据分别采取人工筛选数据、随机抽取数据以及全部数据三种方案的链接覆盖率；CCR_m、CCR_r、CCR_a 分别表示对实验攫取的 Web 页面数据分别采取人工筛选数据、随机抽取数据以及全部数据三种方案的代码覆盖率。

由图 5-4 可见：①虽然索引页中包含数量可观的链接数，但由于索引页中纯文本数极少，在无纯文本或仅被极短文本间隔的区域，所有的链接都将被归属到同一个逻辑链接块中，故采用文本距离作为链接距离时，逻辑链接块数极少，尤其是当 d_t^t 增大时这种现象更为明显。②人工取样数据由于都来自于门户网站，Web 页面大而结构复杂，所需呈现的内容多，栏目多，也就

导致链接数多；而随机组中的网页绝大多数属常规大小，所需呈现的内容少，栏目简单，从而链接数也少。另外，由于随机组中 Web 页面相对较小，出现长文本的情形也更为少见，故其逻辑链接块数明显少于人工组。③索引页中，当 $d_t^t=5$ 时，链接覆盖率即超过 90%，这表明索引页中长度大于 5 的纯文本数量很少，这也正是我们平时所熟知的情况。④当 $d_t^t<20$ 时，人工组和随机组的代码覆盖率存在差异，与链接覆盖率曲线稍有不同，这主要是因为：当链接覆盖率提高到某个较高水平时，“孤立”链接或链接块数将越来越少，此时提高 d_t^t 的值，其主要作用不再是将“孤立”链接或链接块纳入逻辑链接块而增加链接覆盖率，而是将由于 d_t^t 偏小而被某些较长文本分割开来的那些小逻辑链接块合并为更大的链接块，表现为一种对链接之外的其他代码的“吞噬作用”；在合并的过程中，一方面使得逻辑链接块数更少，另外一方面由于多个链接块的合并，将原本属于逻辑链接块间的中间地带整体纳入进新的逻辑链接块，该过程中虽然基本不会或很少导致新的链接被归属到逻辑链接块而提高链接覆盖率，但逻辑块间中间地带代码的纳入，却能显著提高代码覆盖率。⑤通过对比图 5-4 中的链接覆盖率曲线和代码覆盖率曲线可知，$d_t^t\geqslant 20$ 时，链接覆盖率基本维持不变；而 $d_t^t\geqslant 45$ 时，代码覆盖率也将维持不变。这也就意味着，在索引页中，当 $20\leqslant d_t^t<45$ 时，d_t^t 增加而带来的主要贡献表现在对非链接代码的吞噬；而在 $d_t^t\leqslant 25$ 时，d_t^t 的增加则同时吞噬了链接及链接之间的代码，从而呈现出链接覆盖率和代码覆盖率的同步上升。⑥相对而言，随机组的逻辑链接块数更易受到 d_t^t 的影响，其主要原因在于：首先随机组中 Web 页面链接数总体偏少，一般在几十至数百个，人工组中的门户 Web 页面则一般都包含上千个链接；其次，在随机组中的 Web 页面中较长的纯文本极少，d_t^t 的增加将使得原本较小的逻辑链接块迅速聚合为较大的逻辑链接块，逻辑链接块数大幅度降低，因而就导致了随机组中的逻辑链接块数的波动更为明显。

与索引页相比，针对内容页的实验结果存在如下显著不同。①逻辑链接块数显著减少。这主要是因为内容页所承担的作用不同所致。索引页承担着导航的作用，包含着尽可能多的链接，而内容页则着重呈现某一个主题内容，该主题可能为文本、图片、视频等，这些主题元素占据了大量的篇幅，链接数量大幅减少，从而导致最终的逻辑链接块数大幅减少。当 d_t^t 足够大时，Web 页面中的逻辑链接块数基本维持在 2～3 个，其中相当多的页面链接数为 2，即正文内容前后的链接分别被划分为一个逻辑链接块。②人工组和随机组的实验结果差异不显著。在索引页的实验结果中，随机组的逻辑

链接块数远小于人工组，但在内容页上，却并无太大差异。可见，从内容页的角度看，人工组和随机组中的 Web 页面，均具有相似的结构特征和文本特征。③代码覆盖率显著降低。这主要是因为在内容页中，非链接块占据了相当大的篇幅，且内容页的规模远小于索引页所致。④在逻辑块发现的过程中，Web 页面正文文本能够被很好地保留下来，少数文本极其短的页面例外，这表明基于逻辑块识别的方法是可以应用于 Web 页面文本提取的。⑤对于内容页文本中零星出现的孤立链接，由于其间距离过远而不会被纳入链接块，即文本块的完整性不受影响；而对于嵌入在文本中小区块广告链接（如图 5-2 所示的情况），由于链接之间距离短而会被纳入逻辑链接块中。这在基于块级元素的链接块识别中是无法达到的。不过若某些孤立链接恰好离嵌入在文本中的广告区块较近，则也有可能发生误判的情况，在 $d_t^{\downarrow}$ 较小时该情况出现概率较低，随着 $d_t^{\downarrow}$ 的增大，该情况出现的概率将增加，这种现象有待进一步研究。针对内容页的实验结果如图 5-5 所示。

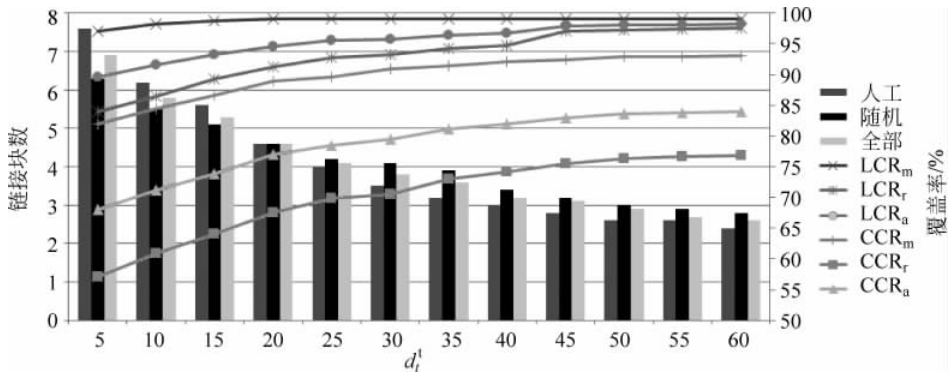

图 5-5　$d_t^{\downarrow}$ 对逻辑链接块的影响——内容页

2. $d_t^{\leftrightarrow}$ 对 Web 页面链接块的影响

在采用文本距离作为链接距离时，仅计算了相邻链接之间的文本，这就导致在文本偏少或者较短的 Web 页面中，即使相邻链接间存在大量代码，但若无文本，则由于缺少文本的分割作用，它们仍将被归属到同一个逻辑链接块中。而在采用代码距离作为链接距离时，代码和文本同时对逻辑链接块的分割起作用，这也就意味着，采用代码距离作为链接距离时，Web 页面将被划分为更多的逻辑链接块；与此同时，链接块间的中间地带也将增多，这将导致代码覆盖率的降低。实验证实了上述分析成立，结果如图 5-6 所示。

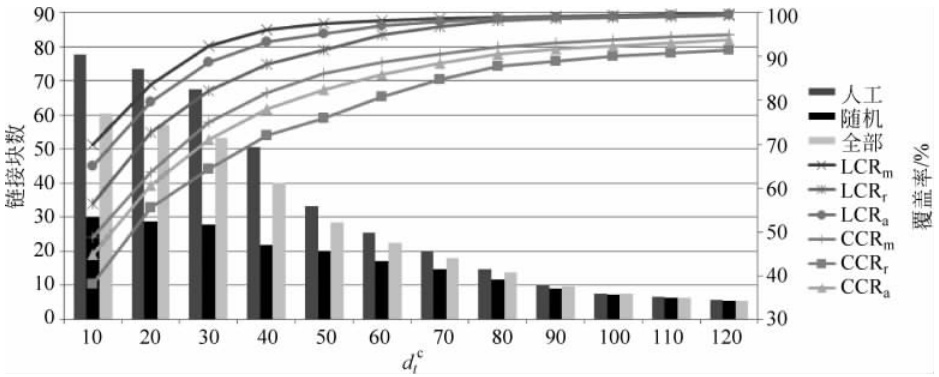

图 5-6　d_t^c 对逻辑链接块的影响——索引页

从图 5-6 可见：当 d_t^c 较小时，人工取样数据所包含的链接块数远多于随机组，而随着 d_t^c 的增大，其间的差距逐渐缩小，当 $d_t^c>90$ 时，这种差别几乎不再存在。这也就意味着，在索引页中，无论是人工组中的门户网站索引页面，还是随机抽取的常规索引页面，相邻链接之间的代码距离基本都在 90 以内。由于 d_t^c 越小，其对 Web 页面的分割作用越“精细”；反之 d_t^c 越大，其分割作用越“粗糙”，更易凸显 Web 页面的宏观结构特性。可见，无论 Web 页面规模的大小如何，都存在一定的宏观结构相似性。这种特性在针对内容页的实验中同样存在。

与采用文本距离方式的实验结果类似，针对内容页的代码覆盖率显著低于索引页，其他方面则无显著差异。

3. 从文本距离谈 C_t 对 Web 页面链接块的影响

逻辑链接块中链接数阈值 C_t 决定了一个逻辑块要成为一个逻辑链接块所必需的最小链接数。在 d_t（d_t^c 和 d_t^t）确定的情况下，C_t 越小，则在逻辑链接块的扫描发现过程中，各个逻辑块更容易达标而成为链接块，也将包含更多的链接，链接的纳入，必将相应地吸纳更多的链接间代码；反映在曲线上，即链接覆盖率和代码覆盖率的高位。反之 C_t 越大，各个逻辑块则更难以被认定为链接块，诸多的逻辑块虽然包含了链接，但由于数目上小于 C_t 从而被舍弃，其结果就是更多的链接将被排除在逻辑链接块之外，相应的也将有更多的代码未能纳入逻辑链接块；反映在曲线上，即链接覆盖率和代码覆盖率的低位。同时由于诸多“准链接块”被舍弃，也将导致逻辑链接块总数的衰

减。实验结果证实了上述结论，如图 5-7 所示。

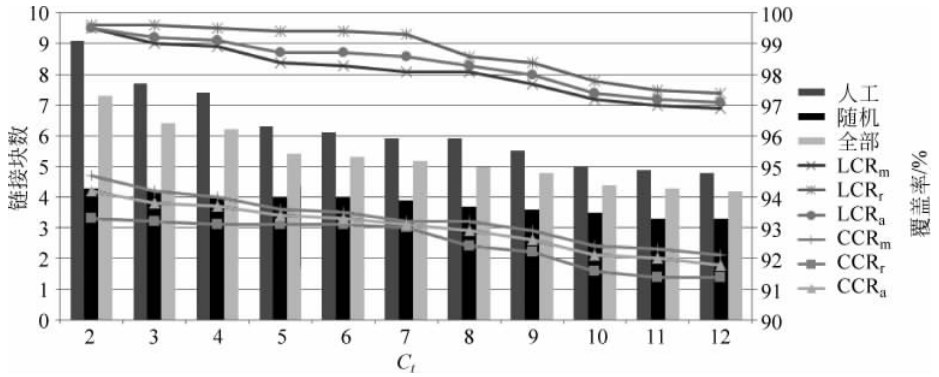

图 5-7　C_t 对逻辑链接块的影响——索引页

从图 5-7 可见，人工组的逻辑链接块数受 C_t 的影响较为显著，而随机组的逻辑链接块数基本无大幅度变化。这主要是因为在本实验中 $d_t^L=40$，而绝大多数的随机组 Web 页面中较少存在长度超过 40 的纯文本，这也就导致了无论 C_t 取值如何，整个 Web 页面被划分为逻辑链接块时的分界点较为固定，即那些数量很少且长度超过 40 的纯文本充当了分界点角色。不难推断，若 d_t^L 较小时，充当这种分界点角色的纯文本就逐渐增多，此时逻辑链接块数也将呈现为较大波动。实验数据证实了这一推断。

针对内容页的实验结果(如图 5-8 所示)，与索引页相比其最大不同表现在：①逻辑链接块数量少，基本都在 4 个以下。这主要是因为在内容页中纯

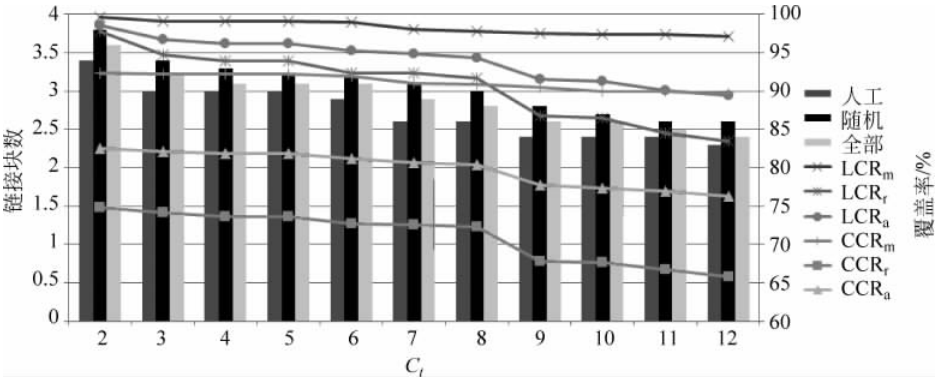

图 5-8　C_t 对逻辑链接块的影响——内容页

文本几乎集中呈现，即使文本中偶尔会出现孤立的超链接，但往往由于这些超链接与其他超链接距离过远而无法纳入到链接块中。这恰好能够维持文本块的完整性。倘若将 d_t^L 继续增大，则可能导致文本块被部分甚至全部归属到链接块中。②逻辑链接块数随着 C_t 的增大一直呈现平缓的下滑，不再出现类似上述在 C_t 较小时链接块数波动较大的情况。这主要是因为在内容页中的长文本数量和位置基本固定所致。在 d_t^L 较大且确定的情况下，无论 C_t 如何变化，对逻辑链接块划分起决定性作用的都是那些长文本。在内容页中的长文本集中呈现，这也就决定了当 d_t^L 增大到某个值时，绝大多数的内容页将被划分为两个链接块：正文之前作为一个链接块，正文之后作为一个链接块。该结论在实验中得到证实。③当 C_t 较小时，链接覆盖率与索引页基本持平，而随着 C_t 的增大，链接覆盖率与索引页的差距逐渐增大。这主要是因为：在索引页中，链接分布比较密集而均匀，而在内容页中，例如，在某些正文中零星分布少量链接的页面中，特别是在某些附带评论的博客页面和论坛页面中，链接呈现为一种相对“离散”分布。这样当 C_t 较小时，零星散落的链接只要距离不致太远，或者能以小团簇形式（典型的如博客或论坛中每条回复周围关于发帖人个人信息的一些链接）出现，它们仍能被认定为逻辑链接块；随着 C_t 的增大，越来越多的以团簇形式存在的小链接区域由于无法满足最小链接数阈值 C_t 的要求，且相邻链接团簇又因为被某些较长的文本切断而被排除在链接块之外。这种情况在索引页中是极其少见的，因而导致了这一现象。④代码覆盖率远低于索引页。其本质原因在于内容页中存在着大篇幅的主题文本块，这些文本块基本不会被纳入逻辑链接块中，从而也就导致了内容页的代码覆盖率显著低于索引页。⑤人工组的链接覆盖率显著高于随机组。其主要原因正如③中所述，博客页面或论坛页面中的往往由于部分篇幅较长帖文对页面的分割作用所致。这些长帖文的存在，将导致部分包含的链接数低于 C_t 的逻辑块未被认定为链接块，从而丢弃了大量的链接，这一现象在人工组的门户新闻页面中几乎不存在。最终造成链接覆盖率的降低，相应的也使得代码覆盖率降低。⑥人工组的代码覆盖率显著高于随机组。其主要原因在于：第一，人工组的页面往往较随机组中的页面更长，然而从其所包含的正文长度而言，则两者并无显著差异，依据代码覆盖率的计算表达式不难看出这将导致整体篇幅短的内容页其代码覆盖率也更低。第二，正如⑤中所述，部分篇幅较长帖文对页面的分割作用导致了代码覆盖率的降低。

4. 从代码距离谈 C_t 对 Web 页面链接块的影响

对索引页实验结果而言，采用代码距离和采用文本距离的方式相比较，主要差异体现在四个方面：①采用代码距离时的链接块数更多。原因与前文一致，不再赘述。②代码覆盖率和链接覆盖率更低。原因同前文，不再赘述。③随机组和人工组的逻辑链接块数差异不明显。

对内容页实验结果而言，采用代码距离和采用文本距离的方式相比较，主要差异与索引页基本相同。

5.5 结论

本章提出的逻辑链接块，扩展了常规链接块的范畴；本章的逻辑链接块发现方法，避开了传统链接块识别所不可或缺的标签树解析或者 DOM 解析过程，从而也就无须在标签树解析或 DOM 解析上耗费大量的计算资源，同时避免了解析纷繁复杂且缺乏规范的 HTML 时的各种问题；另外，链接块的判别规则简单，无须复杂计算，在对 Web 页面进行一次扫描即可同时完成逻辑链接块的发现与判别。本章的方法分析速度快，抗干扰性强，能更好地适应设计不规范的 Web 页面，且不要求链接块内的链接主题内聚性高，这也就决定了该方法在 Web 页面文本抽取方面有着潜在的应用价值，在其他对链接块精细粒度要求不高的 Web 信息处理和挖掘领域中也具有广泛的应用前景。