基于标引技术的特定领域 XML 文本自动生成*

刘桐菊　于浩　赵铁军

哈尔滨工业大学计算机科学与技术学院　哈尔滨 150001

E-mail: ltj,yu,tjzhao@mtlab.hit.edu.cn

摘要: XML 语言的一个突出的优点就是可以成功的解决资源共享问题，给人们的科学研究带来了广阔的发展前景。针对目前手工完成 XML 转换这一现状，本文将自动标引技术引入，先对文献进行标引，提取出关键词、主题词、相关人物、机构等重要信息，然后自动生成 XML 文本。进行自动标引时，采用了改进的 TFIDF 算法，针对金融领域进行了试验，给出了结果并对后期工作进行了展望。

关键词: XML，自动标引，分词，加权

XML Documents Automatic Generation of Given Field Based On Marking Technology

Liu Tongju,Yu Hao,Zhao Tiejun

School of Computer Science and Technology ,Harbin Institute of Technology, Harbin 15001

E-mail: ltj,yu,tjzhao@mtlab.hit.edu.cn

Abstract: XML is one of the most promising solutions to the problem of resource share. However, the conversion of current texts into XML is usually completed manually. This paper introduces marking technology to facilitate its automation, focusing on identification of key word, topic-word, related-people and related-organization from the literatures. TFIDF is adopted in this process and the results are saved as XML. Experiment on financial texts indicates the effectiveness of the proposed method.

Keywords: XML, Automatic Marking,Word Segmentation

1　引言

网络现已成为信息的流通渠道，如何将网络所包含的大量、无序的信息即时、准确的提取、整理、组织成便于查询、存储的形式已成为研究、开发的焦点。目前，人们很积极的将自己的数据转换成 XML 的格式，因为这种格式的数据能够满足人们的便于检索查询的需求，

*本文研究受到国家 863 计划资助(项目号 2002AA117010-09)。

优点如下[6]：

(1) 它本身就是一种数据库，便于查询检索，但它不要求安装特定的数据库环境，易于移植；

(2) 标记含有语义信息，这些信息可以用来进一步限定检索词，从而提高检索的准确率，标记可以由用户自行定义，可以满足特殊用户的特殊要求，这是HTML语言所做不到的；

(3) 这种格式的数据生命期长，一是因为编辑器很多，notepad，vi等都可以，不会随着时间的流逝，某种编辑器的消亡而无法阅读，二是如果数据的某个片断被破坏，其它部分不受影响，依然能被人们读出、读懂；

(4) 有很好的数据通用性，它不受平台的限制，一经建立，处处适用，而且操作简便，不要求用户安装特定的应用程序就能够被使用；

(5) XML是一种WEB语言，此种格式的数据具备HTML的便于网上传输和浏览的优点。

目前XML转换都是人工完成的，任务量很大，无法适应海量信息的要求，所以自然想到应该用计算机来解决这个问题。可是，计算机的智能程度远没达到人脑的程度，它无法正确识别出每个词或句子属于哪个标记定义的范围，所以，根本就没有办法用有语义信息的标记进行标识。考虑到人们检索的习惯，都是用有代表性的关键词进行检索，所以本文采用的解决方法是从文章中抽取出能够代表文章特征的词汇，如人名、机构名、关键词、主题词等词汇(由于是在经济范畴做的，所以抽取上述信息作为关键信息)，然后用XML格式将其分别标识出来。从大量的文献中提炼出有用的信息作为检索的依据，这一过程称为文献的标引。以往，文献的标引多是人工进行的，不但耗费大量人力，而且效率极低，无法适应庞大信息量的要求，成为信息加工过程中的瓶颈。随着计算机技术的发展，利用计算机的高速运算能力来对文献进行加工处理，就成为解决这一问题的重要途径。目前的自动标引技术比较成熟，得到的结果比较理想，所以，这种做法有很大的可行性。

2 算法设计

从分词角度看，信息标引技术有基于分词和非分词两种方法，基于分词的方法准确率高，但速度较慢，非分词的方法则恰恰相反，二者各有千秋。

目前的信息标引系统，基本上都采用基于分词的方法，如参考文献[2][3]，但这些系统也存在一些可以改进的地方：比如文献[3]，它的分词基本上全是基于词典的，这样不但影响人名、机构名等未登陆词的识别精度，而且词典的构造难度也增大了；权值函数只考虑了词频信息，根本没有涉及到位置、词长等一些非常重要的因素。文献[2]在权值函数设计上比较全面，但没有未登陆词的识别，所以结果不是很理想。本文采用改进的TFIDF算法来设计权值函数。对于未登陆词的识别和分词算法选取了文献[1][4]的算法。

本系统对准确率要求较高，因此，在抽取的过程中采取了基于分词的方法。另外，XML格式的数据有利于备份、网上发布等优点，所以采用了此种存储方式，为了方便用户的浏览，给他们配上了几个可供选择的样式单。

2.1 加权公式的设计（TFIDF 方法）

本文采用 TFIDF 算法来计算标引词的权值，权值=tf*IDF,其中，tf 是相对词频，IDF 是反文献频率。相对词频是沿用绝对频率加权法的思想，用词语在特定文献中的出现频次作为评价词语表达文献主题重要性的一个方面。因为无法为绝对频次定一个统一的标准来筛选标引词，所以采用词语在特定文献中的相对频率值（某词语在特定文献中的出现次数与文献中出现次数最多的词的出现次数之比值），作为 tf 值。词频是评价词语表达文献主题的一个方面，但是词语在某一文献中的出现次数并不能直接代表它在该文献中的重要程度，比如：的，了等高频词，根据信息熵的理论，稀有词含有的信息量更大，所以，用反文献频率作为权值，采用加权相对词频作为标引词的权值。

权值公式如下：

权值=tf*IDF= (f(W)/MAX f(W))*log2(N/n)

其中——f(w)为绝对词频；

——N 表示文献集合中的文献量

——n 表示文献集合中包含词 k 的文献数。

2.2 对公式的改进

由于 TFIDF 算法只考虑了词频信息，导致准确率方面不尽人意，文献[3]中的系统有此缺点。所以，要对公式进行改进，在计算权值时考虑词的位置和词长两个重要信息。

2.2.1 位置因子

考虑人们撰写文章的习惯以及文章各部分反映内容主题的重要程度不同，对以下几个位置赋予不同的权值[5]：

标题	包括主标题、副标题和小标题。这是因为标题更能反映相关部分的主题。
摘要	摘要相对于标题而言，它反映文献的主题信息量较大，一般情况下能够完全反映文献讨论的主题，但是，如果仅仅利用摘要作为标引源，则很难在里面抽出 5-6 个重要的词作为标引词，另外，有些文献没有摘要，因此，将它作为一个标引源，给它赋予了一个较高的位置权值。
段首句	国外有学者对科技文献的 200 个段落进行了主题句的分析，分析结果为：85%的段落主题句是段落的第一句，7%为段尾句。
参考文献	参考文献是一些相关的文献的标题，也可以在侧面反映文章的主题。
正文	除上述几个部分外的其余部分，统统定义为正文，正文含有的信息多，但篇幅也大，所以，对于正文中的词给定的权值并不是很高。

把正文的位置因子设为 1，其他位置因子和正文位置因子的比值设为它的位置因子。经过从五万多篇正确标注的语料统计得出的各个位置因子如下：标题为 1.8；副标词为 1.6；小标题为 1.5；摘要为 1.4； 段首句为 1.3； 参考文献为 1.3。引入位置因子改进信息量因子的公式，在计算 f(W)时，对每个词乘上它的位置因子，就得到了每个词的位置因子和，用 F(W)表示。权值公式改为：权值=TF*IDF=(F(w)/MAX F(w))*IDF

2.2.2 词长因子

汉语是以词为表意单位，词之间没有分隔标记，因此汉语存在分词的问题，同时词长也可以作为一个加权因子，据统计，在汉语中，特征词一般是词长较长的词，词长较长的词在表意能力以及区分能力上都要强于词长较短的词。如："社会主义市场经济体制"要比"市场经济体制"的表意更明确，区分能力也更强。

词长因子定义如下：

词长因子=W/MAXW

其中，W 是词 k 的长度；

MAXW 表示文本中词长最大的词的词长；

在词长因子的定义式中，用的是相对长度而非绝对长度，这同样是为了制定一个统一的筛选标引词的标准。

综合以上几个方面，最终得到的公式为：权值=TF*IDF*词长因子。

2．3　XML 框架

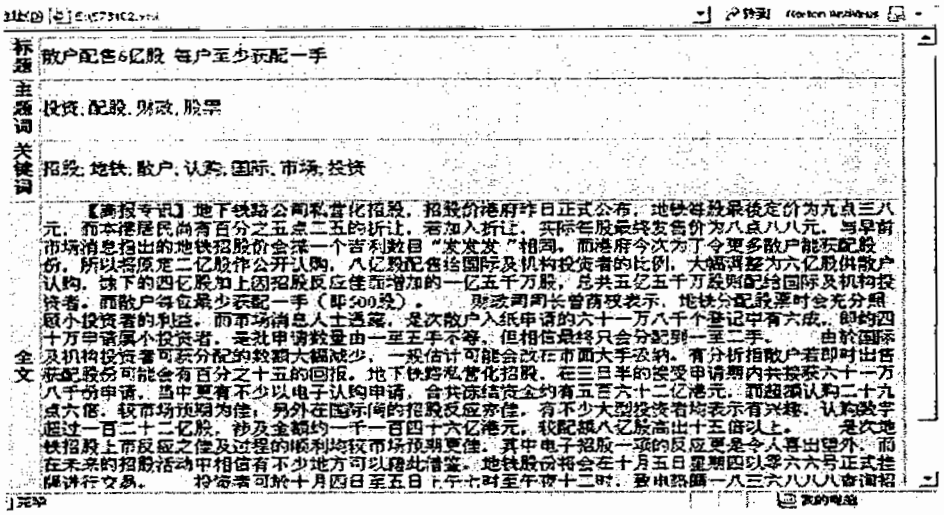
XML 的标记含有语义信息，且可以根据需要进行定义，由于要生成经济领域的 XML 文本，所以在不违反命名规则的情况下定义了标题、主题词、关键词、相关人物、相关上市公司、相关机构、全文这些标记，形式如下：

```
<?xml version='1.0' encoding='GB2312'?>
<标引文章>
<标题>…… </标题>
<主题词>…… </主题词>
<关键词>…… </关键词>
<相关上市公司>……</相关上市公司>
<相关人物>……</相关人物>
<相关机构>……</相关机构>
<全文>……</全文>
</标引文章>
```

XML 对语法的要求是很严格的，在每个标记对中间的是用户需要存储的内容，XML 语言要求其间不可以包含<、>、'、"、&等字符，需要用相应的字符串<、>、'、"、&进行代替。

采用 XML 的好处是在检索时，我们可以仅针对除<全文></全文>标识外的其余部分进行匹配，在返回给用户时，仅把全文返回，这样，既节省了检索时间，也提高了检索准确率，而且可以把用户不关心的部分对用户进行隐藏起来。

为了便于用户查看，在这里给出了相应的样式单，可以根据用户的不同要求进行改变，这是 XML 优于 HTML 的很重要的一个方面，在这里采用了表格的形式，仅显示标题，主题词，关键词和全文这四部分，如下图所示：



图一：IE 显示的标引结果

2．4 后处理

由于词典无法将主题词、关键词全部收录在内，对于没有加到主题词词典中的关键词，可能会被切分为若干部分，那么，它的各个部分都可能会被抽取出来，试验结果也证明如此，而且此类现象比较严重。比如图一中，标引词有“地铁”“招股”两个词，可在原文里，“地铁招股”一词出现多次。后处理的方法是对标引结果中的任意两个词进行排列组合，然后到原文搜索，将出现次数高的“新词”也作为关键词，追加到结果中。

3. 算法描述

系统的流程图如图二所示。下面介绍一下各个部分的功能：

（一）系统数据和词典初始化：这部分主要是把分词词典和关键词词典及一些统计数据读入内存，为后边的分词及词频统计做准备。

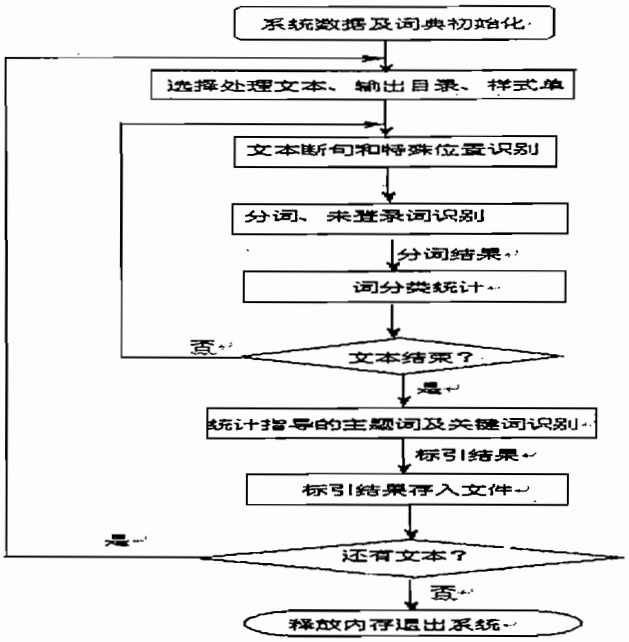
（二）选择处理文本、输出目录和样式单：这部分由用户选择需要进行标引的文章，可以选择单个文件进行处理，然后就可以对结果进行调整，也可以选择一个目录下的所有文件进行批处理，此时就不提供修改功能。结果文件输出的位置和选用或不选用样式单，都有用户根据需要自行选择。

（三）文本断句及特殊位置识别模块：主要功能是对文本进行分类。识别的主要成分包括标题，副标题，小标题，摘要，段首句，参考文献，正文等成分。

（四）分词及未登陆词识别模块主要是采用了文献[1][1][1]的方法，不再加以论述。

（五）词分类统计模块：计算权值 TF（即每个词的位置权值之和），将词排在不同的队列里，按照 TF 值进行排序。

（六）统计指导主题词关键词的识别：按照公式计算最后的权值，排序，输出高于阈值的标引词。



图二：系统流程图

4．实验结果分析后续工作

从人工标注的文本中随机抽取 400 篇进行开放测试，测试的标准为：如果人工标注的关键词有一个包含在系统标注的结果中，我们就认为系统标注的结果是正确的。这样测得正确率为 79%，比算法改进前的 70%左右的准确率有所提高。

本系统在测试阶段取得了令人满意的结果，正准备用于经济领域文献的整理系统中去。但从实验结果来看，从标引方面看，受目前语料规模、分词研究、未登陆词识别研究这一系列条件的限制，本系统还有很大的改进空间。

作为今后要做的工作，主要是解决大规模语料的标引速度以及存储管理问题，另外，对于 HTML 文本到 XML 文本的转换也可以按照此法来进行，考虑特殊字体、链接等信息。这两部分工作正在开展。

参考文献

[1] 吕雅娟、赵铁军："基于分解与动态规划策略的汉语未登录词识别"， 中文信息学报，2001。

[2] 薛翠芳，郭炳炎等："汉语文本特征词的抽取方法"，情报学报，2000。

[3] 肖明：基于《中国分类主题词表》的 www 科技信息资源自动标引设计方案。

[4] 赵铁军，吕雅娟等："提高汉语自动分词精度的多步处理策略"，中文信息学报， 2001。

[5] 苏新宁，邹晓明："文献信息自动标引研究"，现代图书情报技术，2000。

[6] Elliotte Rusty Harold：《 XML 实用大全》，中国水利水电出版社，2000。