通用抽取引擎框架:一种新的Web信息抽取方法的研究

宫继兵[1,2] 唐 杰[2] 杨文军[3]
(燕山大学计算机科学与工程系 秦皇岛 066004)[1] (清华大学计算机科学与技术系 北京 100084)[2]
(中石油规划研究院信息中心 北京 100083)[3]

摘 要 大规模的网络视频信息既为用户信息分享带来了方便,同时也为国家监管部门带来了新的挑战。考虑到效率问题,在线视频监管则主要考虑视频描述信息。主要研究了网络视频描述信息的抽取问题,提出了一种新的Web信息抽取方法:通用抽取引擎框架,其主要包括对视频描述信息抽取问题的形式化描述和用户感知的视频网站逻辑模型。该方法在国家某部委的视频监管项目中已得到应用,并取得了很好的效果。实验结果表明,该方法的扩展性、通用性和抽取准确率大大优于其他方法。

关键词 通用抽取引擎框架,网络视频监管,视频网站逻辑模型,Web信息抽取,抽取模式产生算法
中图法分类号 TP391 **文献标识码** A

General Extraction Engine Framework: Research of a New Approach for Web Information Extraction

GONG Ji-bing[1,2] TANG Jie[2] YANG Wen-jun[3]
(Department of Computer Science and Engineer, Yanshan University, Qinhuangdao 066004, China)[1]
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)[2]
(Information Center, Planning and Engineering Institute Petrochina Corp. Ltd, Beijing 100083, China)[3]

Abstract The large size of video collection not only provides an easy way for users to share information, but also brings a big challenge for managing them, in particular online monitoring. A critical requirement to monitor the video information is to accurately and adaptively identify the key information describing the video, which is also the first step for video analysis and video search. In this paper, we focused on the extraction problem of the video information from different websites. Specifically, we proposed an engine framework for information extraction. We formally defined the description model in the framework and implemented a customizable engine for information. The proposed framework has been applied to a real-world application of a national department and obtains promising results. Experimental results show that the proposed approach can effectively extract the video information and it significantly outperforms the baseline methods.

Keywords General extraction engine framework, Internet video monitoring, Logical model of video website, Web information extraction, Algorithms for generating extraction patterns

近几年来,网络视频的数量以爆炸方式增长,网络视频的分析和监管应用就变得更加重要,特别是在防止网络色情信息对青少年的毒害,在收集恐怖、分裂国家以及反党等阴谋破坏活动的信息方面具有特别重要的实际意义。而对视频描述信息的抽取则是以上这些应用的首要前提,这就涉及到Web信息抽取问题。视频描述信息是指用户所关心的网络视频所有属性信息构成的集合,如视频名称、视频URL和视频评论等。需要特别说明的是,视频描述信息一定不包括视频所播放的声音和图像信息。图1示意了某个视频网站中一个视频播放页面中所包含的视频描述信息。

Web信息抽取研究已成为热点问题[1-4],其一般方法通常采用一种归纳学习方法[5],从给定的训练样本网页中学习到抽取规则。文献[5]提出一种新的可适应性Web信息抽取方法,该方法首先通过聚类方法来获取商品在网页中频繁出现的关键词组,然后利用网页的DOM树结构来确定包含这些关键词的信息块,从而实现Web信息的自动抽取。文献[6]给出一种基于重复模式的Web内容抽取方法,通过使用一种叫做后缀树的数据结构,分析页面结构中所包含的重复模式,进而从模式的实例中抽取出对应的数据记录。还有人专门针对新闻网站采用WICCAP方法进行信息抽取[4]、基于机器学习的学术信息抽取(如:Arnetminer系统[16])以及其他很多Web信息抽取系统[4,7-11],但这些系统要么没有真正反映视频网站的逻辑结构[12,13],要么针对视频描述信息的抽取精度不够[14,15]。

到稿日期:2010-02-05 返修日期:2010-05-12 本文受国家863高技术研究发展计划(No. 2009AA01Z138)和新教师基金(No. 20070003093)资助。

宫继兵(1975—),男,博士生,讲师,CCF会员,主要研究方向为文本挖掘和语义Web,E-mail:gongjibing@gmail.com;唐 杰(1977—),男,博士,副教授,主要研究方向为社会网络挖掘、文本挖掘和语义Web;杨文军(1977—),男,博士,工程师,主要研究方向为Web服务和数据挖掘。

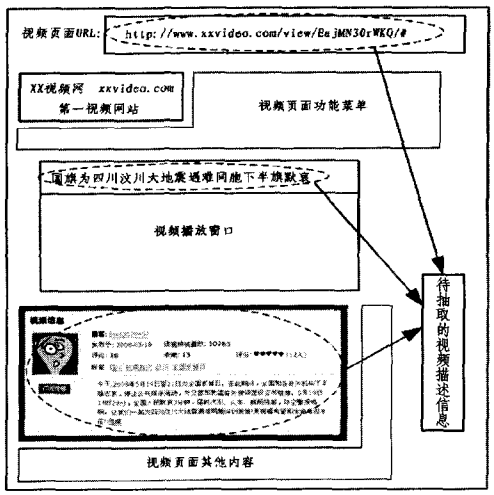

图1 视频描述信息示意图

目前,已有的Web信息抽取方法在抽取效率、准确率以及覆盖率上还存在着巨大挑战。具体表现在:(1)一些Web信息抽取方法大多还处在手工编写抽取模式阶段[3],虽然抽取准确率和覆盖率较高,但视频网站页面格式的不断改变会这些方法带来致命的灾难,其扩展性和维护性很差;(2)虽然也存在一些较成熟的Web信息(半)自动化Web信息抽取方法和技术[1,3],但由于视频网站具有高娱乐性的特点,其Web页面格式不仅千差万别,而且各个网站也在不断改版升级,所包含的视频页面又不是机器可读的和可理解的,这就导致这些(半)自动化信息抽取方法不适用于网络视频信息的抽取任务,在效率和准确率上偏低。

针对以上需求和挑战,本文提出了一种新的网络视频描述信息抽取方法:通用抽取引擎框架(General Extraction Engine Framework,GEEF)。在这种方法中,首先对抽取问题进行了形式化定义,然后从用户感知的角度建立了视频网站的一般逻辑模型,并设计和实现了带限定的非连续抽取模式产生算法。算法通过学习并辅助产生抽取模式,既克服了手工方法的扩展性和维护性差的缺点,还解决了自动化方法抽取效率和准确率偏低的问题。所提出的通用抽取引擎框架方法采用了XML技术,因此具有很好的扩展性和通用性。扩展性体现在方法可以抽取视频网站中任意指定的视频描述信息,通用性表现在它不仅适合任何视频网站的视频描述信息的抽取任务,稍作修改还可以适用于其他任何网站的特定信息的抽取。本方法已在国家某部委的视频监管项目中得到应用,并取得了很好的效果,目前正在推广当中。实际项目运行结果表明,本文方法在扩展性、通用性和抽取准确率上大大优于其他方法。

本文的主要贡献在于:(1)提出了一种新的Web信息抽取方法:通用抽取引擎框架。该方法不仅能准确、高效地抽取网络视频描述信息,还具有很好的可扩展性和可维护性。(2)从用户感知的角度刻画了视频网站,并提出了视频网站的一般逻辑模型。该模型不仅是设计通用抽取引擎框架方法的基础,还使得本方法更合理和更容易被理解。(3)在引入并扩展了关联规则挖掘算法的基础上,提出了带限定的非连续候选抽取模式产生算法(LMEP-CC),其能够为本文所提Web信息抽取方法产生有效的候选抽取模式。此算法不仅提高了本文方法的自动化程度,还保证了抽取任务的高效完成。(4)本文所提出的方法不仅克服了已有手工方法的扩展性和维护性差的缺点,还解决了自动化方法抽取效率和准确率低的问题,已在国家部级项目中取得了很好的实际应用效果。

本文第1节给出了视频描述信息抽取问题的形式化定义;第2节从用户感知的角度刻画和抽象出网络视频网站一般逻辑模型;第3节是本文方法的核心内容,给出了带限定的非连续候选抽取模式产生算法(LMEP-CC)的描述及说明;第4节在已提出的网络视频信息逻辑模型和抽取算法的基础上,建立了通用抽取引擎框架(GEEF),并通过一个实例来解释和说明本文的Web信息抽取方法;第5节给出了评估抽取方法的标准,对本文所提出的方法进行了实验和评估,还与已有方法进行了对比分析;最后给出了结论。

1 视频信息抽取问题的形式化定义

视频播放页面(Page of Playing Video,简称视频页面)是由HTML标签、JavaScript或VBScript脚本组成的非结构化的数据,抽取任务的目标则是从视频页面获得XML格式的结构化的视频描述信息。由此,抽取的问题转化为从非结构化数据到结构化数据转化的问题。

设所有待抽取的视频网站构成的集合为 $W=\{w_1,w_2,\cdots,w_v\}$,每个视频网站可以看作是视频页面构成的集合 P,$P=\{p_1,p_2,\cdots,p_n\}$,视频页面中待抽取的每个独立的原子性的描述信息被称作视频元信息(Metadata),所有预定义的元信息构成集合 $M=\{m_1,m_2,\cdots,m_k\}(1\leqslant k\leqslant p)$,其中 p 为要抽取的元信息的最大个数。一般情况下,不同的视频网站被定义抽取的元信息的数目和种类不尽相同,元信息的个数可多可少,最少的情况只有视频URL,最多的情况下个数为 p,因此一个视频网站中待抽取的元信息组成的集合必是 M 的一个子集。用 $A=\{a_1,a_2,\cdots,a_n\}$ 表示所有视频网站待抽取的元信息组的集合,则 $A\subseteq\rho(M)$。集合 A 中的任意元素 a 也是一个集合,它是一组定义好的待抽取的元信息名称,对应该视频网站中一类视频页面。a 与集合 M 的关系为 $a\subseteq M$。最后,设 V 为某一个视频网站中所有被提供的视频构成的集合,$V=\{v_1,v_2,\cdots,v_n\}$。

任何一个视频都有一个元信息集合与之对应。因此,从集合 V 到集合 A 存在一一映射关系 $f(v)=a$,其中 $v\in V,a\in A$,即任何一个视频都有唯一的一组视频描述信息来说明。从视频集合 V 到视频页面集合 P 也存在一一映射关系 $f(v)=p,v\in V,p\in P$。这样,忽略视频网站的其他辅助功能(如查询、用户管理和收费等功能),可将抽取问题形式化定义为三元组,即 $W=(V,P,A)$,其中元组 A 和元组 V 满足映射函数 $f(v)=a$,元组 P 和元组 V 满足映射函数 $f(v)=p$。此外,元素 a 来源于元素 p,其中 $a\in A,p\in P$,即元信息(视频描述信息)a 是从对应视频页面 p 中抽取得到的,并由此定义了视频描述信息抽取操作。三元组之间的关系如图2所示。

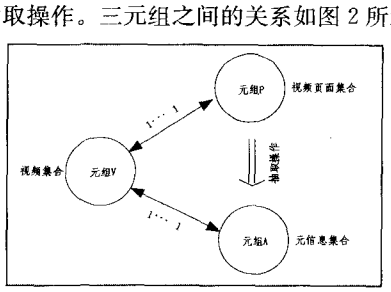

图2 问题形式化描述中三元组之间的关系

进一步定义视频元信息抽取操作如下：视频元信息的抽取问题归属于文本挖掘问题，文本都是由各类字符的任意组合而成，设 V 是包含所有可见和非可见字符的字母表，由形式语言可知，集合 $V^*=\bigcup_{i=1}^{\infty}V^i=\{\omega\mid|\omega|\geqslant 0\}$ 是字母表 V 上所有串的集合。每个视频页面内容是一个字符串，前面已经定义视频页面的集合为 $P=\{p_1,p_2,\cdots,p_n\}$。由于任何一个视频页面 p 都是一个非空和非结构化的字符串，因此 $P\subset V^*$。而引擎文件(模板)也是由字符串组成，设已分析的、存放在通用引擎框架文件中的所有模板的集合为 E，$E=\{e_1,e_2,\cdots,e_m\}$。同理，$E\subset V^*$。抽取结果是一组元信息(Metadata)构成的集合，它虽然有多个项，但也是字符串组成的。设所有视频描述信息的抽取结果对应集合为 D，$D=\{d_1,d_2,\cdots,d_k\}$，$D\subset V^*$。由此，定义代数系统 $F=\{V^+,\Rightarrow\}$，其中运算$\Rightarrow$定义为 $d=e\Rightarrow p$，其中 $p\in P$，$e\in E$，$d\in D$。由于 $P\subset V^*$，$E\subset V^*$，并且 $D\subset V^*$，因此运算$\Rightarrow$是封闭的。

因为集合 P 中任何一个页面都要被抽取并生成唯一一个对应的抽取结果(即一组元信息 d)，而集合 E 中所有模板都是针对集合 P 中视频页面的，对于任意的 e 都必须对应 P 中一类视频页面，因此在集合 P 中“具有相同的抽取模板 e”是一个等价关系。该等价关系对集合 P 的等价划分恰为 m 个。这里用 $[p]_R$ 表示关系 R 对 P 的一个等价划分，则 $P=[e_1]_R\cup[e_2]_R\cup\cdots\cup[e_m]_R$。这个结论可用图 3 表示。

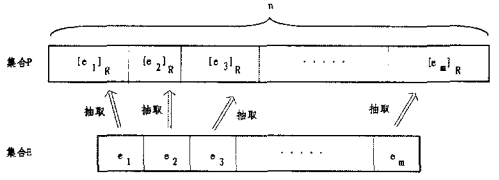

图 3　集合 P 和集合 E 之间的关系

综上所述，所定义的运算 $d=e\Rightarrow p$ 能够清晰刻画视频元信息的抽取操作，其中 $p\in P$ 为视频播放页面，$e\in E$ 为通用抽取引擎，$d\in D$ 为抽取后的结构化元信息。该运算表示通过抽取引擎 e 从视频页面 p 中抽取出视频描述信息 d。

2　视频网站逻辑模型(GEEF-LM)

每个视频网站中视频描述信息的物理格式和表现形式都是完全不同的，但从抽取的任务和目标来看，它们却具有完全相同的逻辑结构。这里提出了用于构建通用抽取引框架的视频网站一般逻辑模型(Logical Model of General Extraction Engine Framework，GEEF-LM)，它是人们对网站视频信息的感知形式，而不是网页的物理格式和结构。

视频网站的逻辑表示的本质是一个树，树中每个节点表示一个用户可感知的逻辑概念[4]。树中单个树节点的组织形式形成了目标视频网站的逻辑框架。在 GEEF-LM 模型中，用于信息抽取的用户将会在此逻辑结构上进行操作。为了适合于抽取任务，每个树节点都包含一个被称为“映射”的元素，该元素将节点和实际网站中对应的数据联系起来。

3　带限定的非连续抽取模式产生算法(LMEP-CC)

在文献[2]中提出了一套用于 Email 头部特征的学习和检测算法。受此启发，本文尝试从数据中尽量发现元信息抽取模式，并利用这些被发现的模式来构建通用抽取引擎框架。本文应用一个非连续的模式产生方法来从视频页面字符串中获得初步的元信息抽取模式。该方法是关联规则挖掘算法的一个扩展。学习得到的元信息抽取模式例子如下所示：

“发布于：*(.+?)<”

带限定的非连续元信息抽取模式，其产生方法包含两步：抽取模式产生和模式分级。第一步，产生所有覆盖正实例的所有可能的模式(对于格式良好的视频网页，本文将大部分包含正则表达式原始数据的元信息抽取模式视为正实例，而其他的则为负实例)。第二步中，根据它们的信任分数对这些抽取模式进行分级，通过简单排序将最高信任分数的抽取模式列为候选，然后人工验证并进行修改调整。非连续模式可以包含通配符，本文使用“(.+?)”将抽取模式中的正则表达式补充完整。该方法令 P_i 表示在第 $i(1\leqslant i\leqslant k)$ 个迭代中所产生模式的集合。初始化过程中，令 P_1 为单词的集合，算法通过结合 P_{i-1} 和在 P_1 中单调递归地建立 P_i 中的模式。

由此，整个算法包含两个子算法：带限定的非连续元信息抽取模式学习算法(LMEP-CC)和非连续元信息抽取模式发现算法(FNCP)。算法 LMEP-CC 中的“锚”是视频页面中成对出现的抽取模式所在的边界标志，如“〈li〉…〈/li〉”，“〈b\s*(.+?)〉…〈/b〉”或“〈div\s*〉…〈/div〉”等。

算法 LMEP-CC　带限定的非连续元信息抽取模式学习算法

输入：P 是所有输入视频页面的集合，E_1 是集合 P 中视频页面字符串的集合

输出：所有元信息抽取模式的集合 $E=\bigcup_{j=2}^{i}E_j$

```
1: 初始化 P 和 E1
2: WHILE(i<=K)
3:   Ei=FNCP(Ei-1,E1);
4:   FOR(对于 Ei 中每个元素 e)
5:     IF(e 不满足限定条件 C)
6:       Ei= Ei-{e};
7:     ENDIF
8:     IF(e 的出现频率小于阈值 T)
9:       Ei= Ei-{e};
10:    ENDIF
11:    IF(e 不包含“锚”)
12:      Ei= Ei-{e};
13:    ENDIF
14:  ENDFOR
15:  IF(Pi 为空)
16:    跳转到 19 行;
17:  ENDIF
18: ENDWHILE
19: 输出∪j=2 Ej
```

算法 FNCP　非连续元信息抽取模式发现算法

输入：初始的元信息抽取模式集合 E_1，待查找的抽取模式所在集合 E_{i-1}

输出：已找到的非连续元信息抽取模式 E_i

```
1: WHILE(i<=|Ei-1|)
2:     IF(对于 E1 中元素 e1 存在)
3:       ei=ei-1e1;
4:     ENDIF
5:     IF(P 中存在元素 ei)
6:     Ei=Ei∪{ei};
7:       ENDIF
8:     ei'=ei-1{(.+?)}e1;
9:     IF(P 中存在元素 ei')
```

10: $E_i = E_i \cup e_i'$;
11: ENDIF
12: ENDWHILE

4 通用抽取引擎框架(GEEF)

前面已经对网络视频描述信息抽取问题进行了形式化定义,并从用户感知的角度建立了视频网站的一般逻辑模型,还提出了非连续的带限定条件的候选抽取模式产生算法,下面应用这些来构建通用抽取引擎框架(GEEF),具体描述这种新的视频描述信息抽取方法。

构建解决视频描述信息抽取问题的通用抽取引擎框架(GEEF)大致包含4个阶段:(1)确定所要抽取的视频描述信息列表;(2)在已建立的视频网站逻辑模型中选择并给出对应的"映射"元素结构;(3)应用带限定的非连续候选抽取模式产生算法产生与每个元素对应候选的候选抽取模式;(4)调整候选抽取模型并调用启发式抽取规则以形成完整的通用抽取引擎框架。这里的启发式规则指的是对初步抽取结果的合并或选择操作。

实际应用中所要监控的(视频)元信息包括"节目URL"、"节目名称"、"点击数"、"评论数"、"上传时间"、"上传者"、"节目分类"、"转载源"、"内容介绍"和"页面标题"。根据本文第1节中的定义,所有预定义的视频元信息构成集合 $M=\{m_1, m_2, \cdots, m_k\}(1\leqslant k\leqslant p)$,其中 p 为要抽取的元信息(Metadata)最大个数。这里,$M=$ {"url","title","uploadedTime","uploadedBy","category","hits","numOfComments","description","OriginatedAt","pageTitle"},$k=10$。一个视频网站中待抽取的元信息组成的集合必是 M 的一个子集。用 $A=\{a_1, a_2, \cdots, a_n\}$ 表示所有视频网站待抽取的元信息组的集合,这里 $A=$ {{"url","title","uploadedTime","uploadedBy","category","hits"},{"url","title","OriginatedAt","pageTitle"},{"url","category","hits","numOfComments","description","OriginatedAt","pageTitle"},…},不难看出 $A\subseteq\rho(M)$。集合 A 任意元素 a 也是一个集合,并且 $a\subseteq M$,同样符合本文第1节中的抽取问题形式化定义。

图4是以土豆视频网为例给出了通用引擎框架一个实例。该实例是树状结构,并与本文第2节中视频网站逻辑模型保持严格的一致性。该树状结构的每个"叶节点"是GEEF-LM模型中的"映射"元素的具体实现,内容包括元信息名称、由正则表达式构成的抽取模式以及自定义的启发式抽取规则。

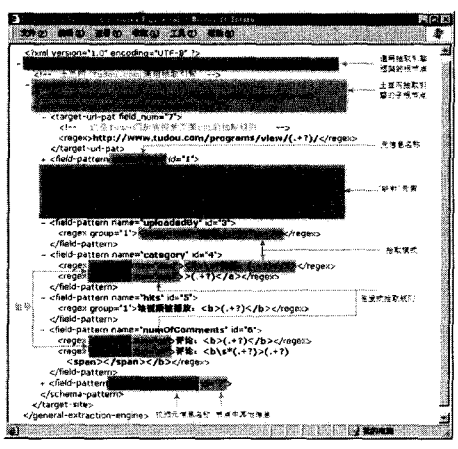

图4 通用抽取引擎实例(Tudou.com)

图4所代表的实例符合本文第1节中抽取问题的形式化定义 $W=(V,P,A)$,其中元组 P 是由正则表达式(如"http://www.tudou.com/programs/view/(.+?)/")所代表的一类视频页面的集合,元组 V 则是 P 中每一个元素中所包含的视频所构成的集合,元组 A 是所有视频网站待抽取的元信息组的集合,它由图4中所有"叶节点"所代表的抽取元信息名称组成,$A=$ {{"title","uploadedTime","uploadedBy","category","hits","numOfComments","description"}},由于 $M=$ {"title","uploaded-Time","uploadedBy","category","hits","numOfComments","description"},有 $A\subset\rho(M)$,进而 $A\subseteq\rho(M)$,与本文第1节的形式化定义相符。

5 实验与对比分析

5.1 实验结果及评估标准

本文的实验就是长达6个月的该部级项目实际运行测试。应用已实现的通用抽取引擎框架方法对国内184个大的视频网站进行抽取。由于篇幅所限,表1中只给出了访问量和规模较大的16个视频网站的抽取准确率和覆盖率。

表1 视频网站列表及抽取准确率和覆盖率

网站名称	视频页面个数	VWEE-Rate	VWEC-Rate
土豆网	9700	99.3%	99.5%
优酷	8923	94.3%	95.8%
菠萝网	6980	98.5%	96.4%
爆米花	8290	97.3%	81.5%
华聚网	6090	92.4%	82.8%
六间房	6875	95.6%	74.4%
酷溜网	6759	97.2%	83.5%
SeeHaHa	8019	98.9%	84.6%
偶偶视频	6100	96.5%	70.1%
我乐网	6582	94.9%	75.8%
比酷网	9295	91.3%	83.8%
大旗视频	7926	96.4%	82.3%
tom宽频	6500	96.7%	86.4%
大视网	5890	93.3%	72.5%
中国播客	7869	98.2%	77.6%
第九频道	8769	91.5%	65.2%

本文采用以下两个常用的指标对本文方法进行评估。

(1)视频网站抽取准确率(VWEE-Rate):在一次抽取任务中,视频网站所有视频页面抽取准确率的平均值。即 $R=\frac{1}{n}\sum_{i=1}^{n} r_i$,$n$ 为该网站已被抽取的视频页面个数,r_i 为每个页面的抽取准确率。一般情况下,通过抽样的方法来确定网站抽取准确率。

(2)视频网站抽取覆盖率(VWEC-Rate):在一次抽取任务中,视频网站所有视频页面覆盖率的数学期望值。即视频网站抽取覆盖率 $C=\frac{1}{n}\sum_{i=1}^{n} q_i/p_i$,其中,$n$ 为被抽取的视频页面个数,q_i 为每个视频页面被抽取出的域的个数,p_i 为视频页面中要抽取的元信息个数。本试验中 $n=9700$(不同的网站 n 值不同),$q_i=9$,p_i 则由具体的抽取情况而定。C 是等概率情况下每个视频页面抽取覆盖率的数学期望。

从表1所列实验结果可以看出,网站的抽取准确率是抽取好坏的标准。根据前面提出的抽取规则,示例中的这些网站都能表现出很好的抽取效果。而网站抽取覆盖率则是衡量网站规范程度的标准,由于网页中某些页面不规范(如没有提供全面的元信息或格式不统一)而导致比值有所下降,这也反

映了网站规范程度具有差异性这一实际情况。

由于 FNCP 是 LMEP-CC 的子算法，这里只需评估 LMEP-CC 算法的时间复杂度和空间复杂度。由于 LMEP-CC 算法第一层外循环中顺序嵌套两个内循环，E_i = FNCP(E_{i-1},E_1)和 FOR(对于 E_i 中每个元素 e)，因此其时间复杂度为 $O(K.(|E_{i-1}|)+|E_i|))O(k\cdot(|E_{i-1}|+|E_i|))$，而 E_i 是通过 E_{i-1} 中发现并去除非连续元信息抽取模式中逐渐建立的。不失一般性，这里取集合 E_{i-1} 或 E_i 元素个数的平均值或数学期望 $\overline{|E_{i-1}|}$ 或 $\overline{|E_i|}$，记为 $\bar{E}$。最后其时间复杂度为 $2O(k\cdot\bar{E})$，其空间复杂度为 $|E|=|\bigcup_{j=2}^{i}E_j|=\sum_{j=2}^{i}|E_j|$。

5.2 对比分析

本文仍以土豆视频网为例(共抽取 9700 个视频页面)将通用抽取引擎框架方法、一般方法和 WICCAP 方法做了对比实验，实验结果如图 5 所示。

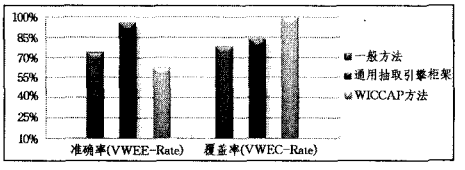

图 5　3 种抽取方法的实验对比

由图 5 可以看到，从抽取准确率上来说，本文的方法有显著提高，说明本文方法抽取效果很好，因此也达到了监管项目的严格要求。本文抽取方法准确率高于一般方法的根本原因是通用引擎框架的建立和对抽取模式及规则的大幅改进。而 WICCAP 自动抽取方法[4]虽然适用于所有网站，但正是由于这种普遍性，也导致了处理某一类型网站的精确度不高，这在本次实验中也得到体现。换句话说，要做到高精度 Web 信息抽取，不能忽略领域专有的知识和规则这一事实。但是，从图 5 中还可以看到，WICCAP 方法的覆盖率却是最高的，原因在于自动抽取无法精确定位专业领域信息，无论抽取内容的好坏，在域信息存在的情况下，都能够给出每一个要抽取域的内容。而由于一般方法和本文方法在原理上基本相同，因此在覆盖率上基本相差无几。

结束语　针对网络视频应用的需求和已有抽取方法所面临的各种挑战，本文提出了一种新的 Web 信息抽取方法：通用抽取引擎框架。它不仅能够高效、准确地抽取网络视频描述信息，还具有很好的扩展性、应用性和通用性。本文从用户感知的角度提出了视频网站的一般逻辑模型，模型不仅有效地刻画了视频网站的本质，还为搭建通用抽取引擎框架提供了有效的设计依据。本文对网络视频描述信息的形式化描述为抽取任务找到了理论依据的同时，也为产生抽取模式算法提供了设计基础。本文提出了用于产生带限定的候选抽取模式的算法，算法在时间复杂度和空间复杂度上都满足抽取任务的性能要求，如实时性和准确性等。本文还通过实例来说明搭建通用抽取引擎框架的过程和该框架所包含的具体内容。通过长达 6 个月的实际项目的运行测试，实验结果表明本文方法在扩展性、通用性和抽取准确率上大大优于其他方法。本文下一步的研究目标是通过概率模型或分类学习方法建立自动更新引擎机制，进而完成网络视频描述信息的全自动抽取。

参 考 文 献

[1] Tang Jie, et al. Information Extraction: Methodologies and Applications[M]// Prado H A, Ferneda E. Ed. Emerging Technologies of Text Mining: Techniques and Applications. Hershey, USA: Idea Group Inc., 2007: 1-40

[2] Tang Jie, et al. Email Data Cleaning[C]// Proc. of International Conference on Knowledge Discovery and Data Mining. 2005: 489-498

[3] Muslea I. Extraction patterns for information extraction tasks: a survey[C]// Proc. of Workshop on Machine Learning for Information Extraction. 1999: 1-6

[4] Liu Zehua, et al. Towards building logical views of websites[J]. Data & Knowledge Engineering, 2004, 49: 197-222

[5] 李朝，彭宏，等. 基于 DOM 树的可适应性 Web 信息抽取[J]. 计算机科学，2009，36(7)：202-204

[6] 高强，张敬之，等. 基于重复模式的 Web 信息抽取[J]. 计算机科学，2007，34(4)：210-213

[7] Adelberg B. NoDoSE-a tool for semi-automatically extracting semistructured data from text documents[C]// Proc. of ACM SIGMOD International Conference on Management of Data. 1998: 283-294

[8] Arasu A, Garcia-Molina H. Extracting structured data from Web pages[C]// Proc. of ACM SIGMOD International Conference on Management of Data. 2003: 1-12

[9] Ashish N, Knoblock C A. Semi-automatic wrapper generation for Internet information sources[J]. ACM SIGMOD Record, 1997, 26(4): 8-15

[10] Baumgartner R, Flesca S, Gottlob G. Visual Web information extraction with Lixto[C]// Proc. of 27th International Conference on Very Large Data Bases. 2001: 1-10

[11] Embley D W, et al. Conceptual model-based data extraction from multiple-record Web pages[J]. Data & Knowledge Engineering, 1999, 31(3): 227-251

[12] Gottlob G, Koch C. Monadic datalog and the expressive power of languages for Web information extraction[C]// Proc. of 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2002: 1-12

[13] Zhou Mingjian, Gao Ji, Li Fei. Ontology-based information extraction from Web sources[J]. Journal of Computer-aided Design & Computer Graphi, 2004, 16(4): 535-541

[14] Laender H F, et al. A brief survey of Web data extraction tools [J]. SIGMOD Record, 2002, 31(2): 84-93

[15] Crescenzi V, et al. RoadRunner: towards automatic data extraction from large Web sites[C]// Proc. of 27th International Conference on Very Large Data Bases. 2001: 1-10

[16] Tang Jie, Zhang Jing, Yao Limin, et al. ArnetMiner: Extraction and Mining of Academic Social Networks[C]// Proc. of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 990-998