基于竞争分类的 Web 信息抽取

李向阳[1],陆建江[2],张亚非[1]
(1. 解放军理工大学通信工程学院,江苏南京 210007;2. 东南大学计算机科学与工程系,江苏南京 210096)

摘 要: 提出一种基于竞争分类的 Web 信息抽取方法,以信息片段与样本之间的相似度作为竞争力,通过信息片段对信息模板槽的竞争实现对信息片段的分类和噪声信息的过滤,直接从分类的角度抽取 Web 信息. 较之基于规则的信息抽取方法,竞争分类法只需用户提供少量的标记样本. 测试表明,该方法在没有用户提供特征提示信息的情况下,抽取信息仍能保持较高的准确率,适应性强;对数据项缺失和顺序变化较大的数据源,竞争分类法也具有较好的健壮性.

关键词: 信息抽取; 竞争分类; 特征提取; 包装器生成
中图分类号: TP311 **文献标识码**: A **文章编号**: 0372-2112 (2004) 11-1915-03

Web Information Extraction by Competing Classification

LI Xiang-yang[1], LU Jian-jiang[2], ZHANG Ya-fei[1]
(1. *Institute of Communications Engineering, PLA University of Science & Technology, Nanjing, Jiangsu* 210007, *China*;
2. *Department of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu* 210096, *China*)

Abstract: A competing classification method is presented to extract Web information. The method uses similarity between information fragments and samples as competing ability. It classifies fragments and filters out noise information through competition of fragments for template slots. It needs far less tagged samples than those using rules to extract information. Experiments show that the method keeps high precision of information extraction without any feature clues provided by users. Therefore it is adaptive. The competing classification method is also robust in dealing with data sources having missing items and items of various orders.

Key words: information extraction; competing classification; feature extraction; wrapper induction

1 引言

随着计算机网络的普及,Web 已经成为人们获取信息的主要渠道之一. 但用于表达 Web 页面信息的 HTML 标记语言不包含任何语义,不适合作为一种数据交换方式由计算机来处理. 如何从非结构化的、无语义的 Web 文档中抽取信息是 Web 数据源集成所面临的重大挑战之一.

目前,从 Web 页面抽取信息的方法主要是基于规则,这些规则一般集成到包装器中. 许多系统就是用包装器抽取 Web 信息的,例如,WIEN[1]用归纳法从样本中学习规则,然后生成包装器;SoftMealy[2]、STALKER[3]、RAPIER[4]和 WHISK[5]等系统也都利用集成的规则来抽取信息. 另一方面,分类法在信息抽取中也有所应用,例如 SRV[6,7]就是从分类的角度来提取特征规则.

一般来讲,基于规则的方法要手工编制规则,或需通过大量标注样本的学习来生成规则. 编制规则和标注样本需用户的大量参与,系统的适应性较差. 本文提出一种竞争分类法,直接从分类角度抽取 Web 页面信息,在用户标记样本较少的情况下,仍能保持较高的抽取准确率. 测试表明,对数据项缺失和顺序变化较大的数据源,该方法也具有较好的健壮性.

收稿日期:2004-01-06;修回日期:2004-07-10
基金项目:国家自然科学基金(No. 60303024)

2 分类法抽取 Web 信息的总体框架

Web 页面信息可分为显示信息和解释信息. 将标记分隔开的显示信息称为信息片段,从分类的角度看,Web 信息抽取就是判断信息片段类别的分类过程:Web 页面在含有抽取内容的同时,也含有无须抽取的内容,信息抽取首先需要判断哪些是要抽取的信息,哪些是无须抽取的信息,即将信息片段分为目标信息和噪声信息两类. 其次,信息抽取还要将目标信息放入信息模板槽(slot)中,在分类中,这是一个加上具体标签的过程.

一般来讲,分类是训练和测试的两阶段过程. 用分类法实现 Web 信息抽取也有这样的两阶段,如图 1 所示.

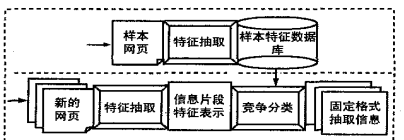

图 1 Web 信息抽取的两阶段

分类中,以信息模板中的每个槽作为一个类别. 在训练阶段,先由用户在样本网页中指出目标信息片段所对应的信息槽,即标记其类别. 然后,用特征抽取算法构建这些信息片段的特征,将信息片段表

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

示成特征量,与类别信息一起存入数据库,形成样本特征库.在测试阶段,当一个新页面需要抽取信息时,先自动分隔出网页中所有的信息片段,将其作为候选的目标信息.抽取每个信息片段的特征,并用特征量来表示这些候选的目标信息.竞争分类法通过比较新信息片段与样本之间的相似度,决定信息片段的类别,从而实现噪声信息的过滤和目标信息的分类,达到信息抽取的目的.

3 信息片段的特征提取

分类中首先要解决的是分类对象的表示问题,也就是特征选择.SRV[6,7]将 Web 信息特征分为简单特征和关系特征.除此之外,基于 DOM 的信息抽取[8]将目标信息在 DOM 层次结构中的路径作为信息抽取的"坐标",本文将这种路径信息称为结构特征.

结构特征 同一类页面往往具有基本相同的 HTML 结构,目标信息所处的位置具有一定的稳定性,表现出一定的结构特征.这种结构特征可用 DOM 树中的一种路径表达式[8]来表示.

关系特征 目标信息与其前后的信息往往有一定的关联,其之前的信息片段(称为前导词)通常具有很强的提示作用.在同类网页中,同类别的目标信息具有相同的前导词,可用前导词作为关系特征.

简单特征 不同类别的目标信息之间有一定的区别,比如英文人名中一般含有大写字母.这种特征有时要通过用户的提示信息来构建.

特征提取过程为:先分隔出信息片段,得到隐含特征的信息单元;然后由这些信息单元构造具体的 3 个特征.简单特征的特征项随着数据源的变化而有所不同,检测即可获得.关系特征取为前导词,无须进一步处理.结构特征的获取较为复杂,可用算法 1 实现.在算法 1 中,HTML 标记分为成对出现的开端标记(opening tag)和闭端标记(closing tag),路径表达式用路径标记、路径序号来表示,其中,路径标记表示从树的根部到当前位置所经历的 HTML 标记,路径序号表示各 HTML 标记的索引.

算法 1 提取路径表达式

```
GetCurPath(curTag,&preTag ,&curRoad,&curMark)
{ //curRoad,curMark 为当前的路径标记、路径序号
  IF curTag ISA OpenTag //当前标记是开端标记
    curRoad.Add(curTag) //增加路径的标记
    IF preTag ISA OpenTag //前一标记是开端标记
      newMark = 0,curMark.Add(newMark)
    else //前一标记是闭端标记,则修改路径序号值
      MarkValue = curMark.GetLast() + 1
    curMak.SetLast(MarkValue)
  IF curTag ISA CloseTag // 当前标记是闭端标记
    curRoad = curRoad.RemoveLast() //移除末端标记
    IF preTag ISA CloseTag //前一标记是闭端标记
      curMark = curMark.RemoveLast() //移除末端序号
  preTag = curTag //重置前一标记 }
```

4 用竞争分类算法抽取 Web 信息

为将信息片段划分成信息模板中每个槽对应的类别,要通过结构特征、关系特征、简单特征来定义样本与信息片段之间的相似度.

路径表达式分为路径标记、路径序号两个特征量.设样本与信息片段的路径标记分别为 $RdA=(a_1,a_2,a_3\cdots a_n)$, $RdB=(b_1,b_2,b_3\cdots b_m)$,两者的相似度为:

$$RdS(RdA,RdB)=\sum_{i=0}^{(m>n)\,?\,m:n}(a_i=b_i)\,?\,1:0/(m>n)\,?\,m:n$$

其中,$(m>n)\,?\,m:n$ 表示当 $m>n$ 条件成立时,取值为 m,否则取值为 n.同理,路径序号的相似度 $MkS(MkA,MkB)$ 可用类似的公式计算.

为两个特征分量分配相等的权重,故两个结构特征 PA 和 PB 的相似度为:

$$PhS(PA,PB)=\frac{1}{2}(RdS(RdA,RdB)+MkS(MkA,MkB))$$

设样本与信息片段的前导词分别为 RA、RB,关系特征相似度为:

$$RtS(RA,RB)=(RA=RB)\,?\,1:0$$

表示若 RA 与 RB 相同,则相似度为 1,否则相似度为 0.

设样本与信息片段的简单特征为 $SA=(a_1,a_2,a_3,\cdots,a_m)$, $SB=(b_1,b_2,b_3\cdots b_m)$,简单特征相似度为:

$$SpS(SA,SB)=\frac{1}{m}\sum_{i=1}^{m}(a_i=b_i)\,?\,1:0$$

为三个特征分配相同的权重,样本 A 与信息片段 B 之间的相似度为:

$$Sim(A,B)=\frac{1}{3}(PhS(PA,PB)+RtS(RA,RB)+SpS(SA,SB))$$

Web 页面中通常含有大量的无固定特征的噪声信息,一般的分类算法存在非此即彼的现象,容易将噪声信息作为一个类别信息来抽取,信息抽取的准确率差.竞争分类法可避免这个问题,其基本思路是将信息模板的每个槽作为一个竞争目标,每个信息片段作为一个竞争者,竞争的胜者获得填充信息模板槽的机会,而败者则被淘汰.通过一个竞争者对不同竞争目标的竞争实现对该竞争者的分类,而多个竞争者对同一竞争目标的竞争并淘汰败者的过程则实现了噪声信息的过滤.竞争者对一样本的竞争力可定义为竞争者与该样本之间的相似度.由于待分类对象一般与最相似样本具有相同类别的可能性最大,竞争者对竞争目标的竞争力可定义为竞争者与竞争目标的所有样本之间的最大相似度.

Web 页面中有时会出现抽取项缺失的现象,这时竞争分类法会抽取最相似的信息片段来填充.若不予阻止,竞争分类算法会降低抽取的准确率.考虑到其他信息对一个类别的竞争力将远低于正确的目标信息对这个类别的竞争力,通过设置最低相似度的方法,可避免错误地填充缺失项.竞争分类法的过程如下:

算法 2 基于竞争分类的 *web* 信息抽取算法

```
for every Slot[t] in Template
  max-Similarity[t] = min-Similarity[t]
  for every Fragment[i] in a Web page
    for every Case[j] belonging to Slot[t]
```

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

```
if Similarity[i, j] > max-Similarity [t]
    max-Similarity[t] = Similarity[i, j]
    Slot[t] = Fragment[i]
```

5 测试结果

测试中所使用的两个数据源与文献[8]相同:IMDB (http://www.imdb.com/top_250_films)和 IAFDB (http://www.isi.edu/info-agents/RISE/repository). IMDB 中数据项的组织较为规范,很少有数据项缺失和顺序变化较大的情况,而 IAFDB 数据源则刚好相反.与文献[8]类似,由于抽取内容与所要抽取内容数量相同,准确率(Precision)与召回率(Recall)同值,以准确率作为唯一的测试指标.

首先将 IMDB 数据分为不相交的两部分,按文件名排序,取前 50 篇为训练集,后 200 篇为测试集.从训练集中随机选择 1 个网页,标注后构成训练样本.以此样本作为训练阶段的输入,然后抽取测试集内的信息,统计正误情况.为避免偶然性,选择不同网页重复 5 次,统计平均值.随后,分别以 2,3,4,5 个样本作为训练数据.分别用 G、L、S 来表示结构特征、关系特征和简单特征.取不同特征组合进行分类,分别以 G、L、LS、GL、GS 和 GLS 作为分类依据抽取数据项,结果统计如表 1,在 IAFDB 上的测试结果如表 2.

表 1　IMDB 上测试结果

样本页数	不同特征组合抽取信息的平均准确率(%)						DOM(%)
	G	L	LS	GS	GL	GLS	
1	51.79	93.85	95.62	76.16	99.68	98.85	83
2	58.62	96.09	97.58	75.51	99.59	99.45	89
3	59.67	95.20	96.96	77.31	97.83	98.11	90
4	62.90	95.71	97.58	78.14	97.97	98.23	92
5	64.35	97.91	98.94	79.95	99.67	99.45	95

表 2　IAFDB 上测试结果

样本页数	不同特征组合抽取信息的平均准确率(%)						DOM(%)
	G	L	LS	GS	GL	GLS	
1	51.63	100	92.75	79.75	94.25	91.75	61
2	64.13	100	92.75	79.25	93.50	92.75	64
3	56.13	100	93.25	80.25	93.50	92.85	65
4	57.25	100	93.38	79.00	92.88	92.63	64
5	57.00	100	93.63	78.35	92.50	92.50	66

从表 1、表 2 的结果看,较之基于 DOM 的方法,竞争分类法通常具有更高的准确率.3 个特征中,除简单特征须由用户给出提示信息外,其余均可从 Web 数据源中自动提取.较之全部特征,结构特征与关系特征组合的抽取效果并不差,说明在无用户干预时,该方法能保持较好的抽取效果,适应性强.从特征组合看,所有以关系特征为分类依据的信息抽取都保持了较高的准确率,优于相应的基于 DOM 方法的抽取结果,表明关系特征是 Web 信息抽取中的关键性特征.从数据源的差异看,数据项的缺失和顺序变化对竞争分类法的影响要小于基于 DOM 的方法,因此,竞争分类法具有更好的健壮性.

6 结论

提出一种基于竞争分类的 Web 信息抽取方法,直接从分类的角度抽取 Web 信息.较之基于规则的信息抽取方法,竞争分类法只需用户提供少量的标记样本.测试表明:该方法在没有用户提供特征提示信息的情况下,用自动提取的特征作为分类依据,抽取信息仍能保持较高的准确率,适应性强.较之基于 DOM 的方法,对数据项缺失和顺序变化较大的数据源,竞争分类法也具有更好的健壮性.

参考文献:

[1] N Kushmerick. Wrapper induction for information extraction[D]. Washington: University of Washington, 1997.

[2] C-H. Hsu, M-T Dung. Generating finite-state transducers for semi-structured data extraction from the web[J]. Information systems, 1998, 23(8): 521 - 538.

[3] I Muslea, S Minton, C Knoblock. STALKER: Learning extraction rules for semi-structured, web-based information sources[A]. AAAI-98 on AI and information integration [C]. Madison, Wisconsin: AAAI/MIT Press, 1998. 74 - 81.

[4] M Califf, R Mooney. Relational learning of pattern-match rules for information extraction[A]. Proceedings of the 16th National Conference on Artificial Intelligence (AAAI99) [C]. Orlando, Florida: AAAI Press, July 1999. 328 - 334.

[5] S Soderland. Learning information extraction rules for semi-structured and free text[J]. Machine Learning, 1999, 34(1-3): 233 - 272.

[6] D Freitag. Multi-strategy learning for information extraction[A]. J Shavlik. Proceedings of the 15th International Conference on Machine Learning (ICML-98)[C]. Madison, Wisconsin: Morgan Kaufmann, July 1998. 161 - 169.

[7] D Freitag. Information extraction from HTML: Application of a general machine learning approach[A]. Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)[C]. Madison, Wisconsin: AAAI/MIT Press, July 1998. 517 - 523.

[8] 李效东,顾毓清. 基于 DOM 的 Web 信息抽取[J]. 计算机学报, 2002, 25(5): 1 - 8.

作者简介:

李向阳　男,1974 年生于江苏沭阳,博士生,讲师,主要研究方向:信息抽取.

陆建江　男,1968 生于江苏无锡,博士后,副教授,主要研究方向:数据挖掘,Web 内容分析,自然语言处理.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net