多粒度时间文本数据的周期模式挖掘算法

孟志青 楼婷渊 胡 强
(浙江工业大学经贸管理学院 杭州 310023)

摘 要 大规模文本数据挖掘是大数据分析的重要分支,也是近年来的一个研究热点。研究了多粒度时间文本数据周期模式挖掘算法,首先提出了时间粒度转换、多粒度时间间隔等概念,然后建立了文本数据的周期模型,给出了一个多粒度时间文本下的周期模式挖掘算法,最后对大量病毒文本文献数据进行了实验,表明了提出的算法可以挖掘一些有效的周期模式,讨论了周期宽松度对支持度和置信度的影响。该研究为大文本数据分析提供了一种新的方法。

关键词 多粒度时间,文本数据,数据挖掘,周期模式
中图法分类号 TP311 **献标识码** A

Periodicity Algorithm of Textual Data Mining with Multi-granularity Time

MENG Zhi-qing LOU Ting-yuan HU Qiang
(College of Business and Administration, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract The large-scale text data mining is an important branch of the big data analysis and is also a hot research topic in recent years. This paper studied algorithm of the textual periodicity data mining with multi-granularity time. First, the concepts of granularity conversion and multi-granularity time interval were presented. Then, a periodic pattern of textual data and an algorithm of the periodic pattern to textual data with multi-granularity time were proposed. Finally, by testing virus textual data, the proposed algorithm shows that some efficient periodic patterns are obtained. The influence of the periodic range on the degree of support and confidence were discussed. This paper provided a new method for the big text data analysis.

Keywords Multiple granularity, Textual data, Data mining, Periodic pattern

1 引言

在数据挖掘中把含有时间维度的数据称为时态数据,时态数据挖掘不仅可以挖掘有关状态方面的信息,而且可以挖掘行为方面的信息,提示出时间上的相关关系,其中的部分时态关系可以进一步转化为因果关系。文本时态数据挖掘是针对文本时态数据的模式发现过程,研究文本时态数据中所隐含的变化模式,包括时态关联挖掘、序列模式挖掘、序列的趋势分析和周期模式挖掘等,例如科学文献中隐含时间的发展规律。

近年来,大数据挖掘是一个非常热的课题,大数据挖掘是指大量的非结构数据挖掘,如网络上大量文本数据(QQ聊天记录、微博、文献数据库)的挖掘,其中时间是文本数据的重要属性,如新闻的时间、聊天记录的时间、文献发表的时间等等,文本数据被加上了时间维就产生了时态文本数据,事实上大量的科学文献就是以时态文本形式出现的,例如对美国科学文献数据库公司(ISI)的许多研究主要集中在信息计量学方面,主要研究集中于期刊文献数据统计规律,即通过期刊各种指标研究文献影响,如SCI影响因子说明期刊的影响力,文献的引用率说明文献的重要和影响程度,但这些都是统计规律,而对文献的内含的内容之间的特征联系与规律的知识发现却研究很少,如某种研究领域随时间的兴衰过程是什么,科学家集中的研究兴趣时段的变化特征是什么?是否所有的研究领域都有周期规律?显然发现期刊文献的内在发展特征和变迁,对于揭示科学研究发展规律具有重要意义。另一个重要的例子,2012年奥巴马竞选总统团队时针对Facebook文本进行了数据挖掘,直接找出支持总统竞选的潜在选民特征和时间规律,解决了快速掌握选民意向、降低竞选成本等关键性问题,对成功竞选起到了重要作用。

在时态文本挖掘中,文本周期模式挖掘,即为在文本时态数据中找出文本中重复出现的特征模式。现实生活中周期现象十分普遍,例如四季的更替、潮水的涨退、动物的繁殖、农作物的生长、股票价格波动和网络流量的变化等。但是大量文本中的周期模式却是很难发现的,人们一旦发现并理解了周期性规律,便能避害趋利,因为生活中80%的数据以非结构化的文本形式存储,所以时态文本数据的周期模式挖掘是一个很有意义的研究方向。

近年来,有关时态数据周期挖掘的研究国内外已经取得了一些成果。Bettini C[1,2]首次提出了时态型的概念,研究了多时间粒度的知识发现问题。孟志青[3-5]系统地阐述了时态型、时间粒度等相关数学定义,并研究了时态数据关联规则挖掘的若干性质问题,进一步探讨了时态数据近似周期的挖掘问题,时态型模型的思想是将时间按一定长度划分后,将数据按时态型进行变形,实施上许多时态数据在原来的时间属性

孟志青(1962—),男,教授,主要研究方向为数据挖掘、神经网络;楼婷渊(1988—),女,硕士生,主要研究方向为数据挖掘。

上很难找到规律，但经过变形后就能够发现其规律，例如经济中许多非平稳数据，经过差分变形后，数据变为平稳数据，可以得到精确的预测模型。程昱[6]给出了多粒度时间的定义，研究了多粒度时间格式的数学表示和性质，通过数据变形找到股票中的部分周期规律。上述学者主要针对结构化数据展开研究，而基于多粒度时间与非结构化文本数据相结合的研究却很少。本文将以一个医学文本时态数据库为研究对象，提出一个文本时态数据的时态周期模式挖掘算法，数值实验表明了所提出的算法可以发现文本之间的周期模式。

本文第 2 节介绍了多粒度时间相关概念，给出了多粒度时间文本数据模型；第 3 节介绍了多粒度时间文本数据的周期模型；第 4 节对 2009 年的生物医学研究文献数据进行了实验研究，并对实验结果展开了讨论；最后给出了总结和展望。

2 时态文本数据模型

2.1 文本表示

在对文本数据进行特征表示前需要进行预处理，我们首先对文本数据进行无用词过滤处理和 Stemming 处理（词干化处理）。在进行无用词过滤处理时，首先查阅停用词表（stopwords），过滤掉一些不会影响文本内容的词条，例如“it”、“a”、“the”介词、冠词等。然后，统计每个词条的频数，设置词条的文本频率阈值，滤掉高频词和生僻词。

文本特征表示是将非结构化或半结构化的文本数据转化为挖掘工具可以处理的中间形式，这里我们用向量空间模型（VSM）来进行特征表示，一篇文本 D 可以用一系列的特征词条来表示 $D=\langle D_1,D_2,\cdots,D_n\rangle$，其中 D_i 是表示特征词条的项，$1\leqslant i\leqslant n$，项 D_i 被赋予一定的权重 W_i，表示它们在文本 D 中的重要程度，简记为 $D=\langle W_1,W_2,\cdots,W_n\rangle$，这时我们说项 D_i 的权重为 W_i，$1\leqslant k\leqslant n$。可以把 $D_1,D_2,\cdots,D_n$ 看成一个 n 维的坐标系，而 $W_1,W_2,\cdots,W_n$ 为相应的坐标值，因而 $D=\langle W_1,W_2,\cdots,W_n\rangle$ 被看成 n 维空间中的一个向量。被广泛使用的权重计算公式为 $W=TF\times IDF$，TF 表示度量词与给定文档之间的关联度，IDF 指包含特征项文档数在全部文档数中的倒数。完整的权重计算公式为：

$$W=(1+\log(1+\log(freq)))\times\log\frac{1+N}{n}$$

式中，N 表示全部文本文档集合的数量，n 表示全部文本文档中包含特征项的文档集合个数，$freq$ 表示特征项出现的频率。

但是这样得到的向量维数很高，对于后面的处理非常不方便，因此我们需要进行特征抽取，降低数据维度。特征抽取一般是构造一个评价函数，对每个特征向量进行评估，选取评估分值高的、预定数目的最佳特征向量作为特征子集。

2.2 时态文本模型

时态文本数据的特征就是文本数据含有时间属性，为了能更深入地理解时态文本数据，我们首先介绍一与时间相关概念的定义。文献[3]已经对时态型、时间粒度做了精确的数学定义。根据时态型的定义，我们可以把现实世界中的时间看成是两端无限的实数轴 T，轴上的每一点代表现实世界中的某一时刻，时态型 μ 是对时间数轴 T 的一个划分，我们可以用秒、分、小时、日、周、月和年等来划分时间数轴 T，每个时态因子 $\mu(t)$ 是一个绝对时刻的集合。为了讨论简单，本文把文献[3]中所讲的粗时间粒度也作为时间粒度，我们把日常生活中的小时、天、周、季度、年作为时间粒度，如某篇文献的发表时间为 2008 年 12 月 5 日。也可以自己定义时间粒度，如将两个小时看成一个时间粒度。

在建立时态文本数据模型之前，首先根据文献[7]中提到的时间模板概念提出如下多粒度时间表示公式：

$$T=(\mu_n:f(\mu_n),\cdots,\mu_2:f(\mu_2),\mu_1:f(\mu_1))$$

式中，μ_i 表示时间粒度，如年、月、日、小时等，$f(\mu_i)$ 是一个正整数集合，表示对应时间粒度的取值范围，若 μ_i 为小时，则 $0\leqslant f(\mu_i)\leqslant 24$。

定义 1（粒度转换） 若 μ,υ 是两个不同的时间粒度，且对于 μ 的任何一个时态因子 $\mu(t)$ 存在 n 个 $t_1,t_2,\cdots,t_n$，使得 $\mu(t)=\sum_{i=1}^{n}\upsilon(t_i)$ 成立，则时间粒度 μ 可转换成时间粒度 υ，表示为 $Shift(\mu(t))^{\upsilon}=n$。

如将时间粒度天转换为小时可表示为 $Shift(天)^{小时}=24$。若 $\mu(t)$ 为月，υ 为日，$\mu(t)=1,3,5,7,8,10,12$，则 $n=30$，若 $\mu(t)=2$，则 $n=28$，否则 $n=31$。

定义 2（时间间隔） 若 $(t_1',t_2',\cdots,t_n')$，$(t_1'',t_2'',\cdots,t_n'')$ 符合多粒度时间格式，且 $(t''_1,t_2'',\cdots,t_n'')<(t_1',t_2',\cdots,t_n')$，$Shift(\mu_{i-1}(t))^{\mu_i}=m_i$，则时间间隔 $\Delta T=(\Delta t_1,\Delta t_2,\cdots,\Delta t_n)=(t_1',t_2',\cdots,t_n')-(t_1'',t_2'',\cdots,t_n'')$，$\Delta t_i=t_i'-t_i''$，若 $\Delta t_i<0$，则 $\Delta t_{i-1}=\Delta t_{i-1}-1$，$\Delta t_i=\Delta t_i+m_i$。

如多粒度时间格式（年，月，日）的两个时间（2012,5,20）和（2012,6,12）的多粒度时间间隔为（0,0,22）。

性质 1 $T=(t_1,t_2,\cdots,t_n)$ 和 $T'=(t_1',t_2',\cdots,t_n')$ 是两个多粒度时间，且 $T>T'$，$\Delta T=(\Delta t_1,\Delta t_2,\cdots,\Delta t_n)$ 是从 $(t_1,t_2,\cdots,t_n)$ 到 $(t_1',t_2',\cdots,t_n')$ 的多粒度时间间隔，则 $(t_1',t_2',\cdots,t_n')+(\Delta t_1,\Delta t_2,\cdots,\Delta t_n)=(t_1,t_2,\cdots,t_n)$，即 $T'+\Delta T=T$；$(t_1,t_2,\cdots,t_n)-(\Delta t_1,\Delta t_2,\cdots,\Delta t_n)=(t_1',t_2',\cdots,t_n')$，即 $T-\Delta T=T'$。

我们假设 D 表示文本文档的集合，记为 $D=\{D_1,D_2,\cdots,D_n\}$，A 表示文本数据中有限个特征项的集合，由于每篇文档的特征项不一定都相同，因此 A 有多种特征项组合方式，记为 $A=f\{A_1,A_2,\cdots,A_m\}$，表示集合 $\{A_1,A_2,\cdots,A_m\}$ 的某个不确定子集，如 $A_{D_i}=f_{D_i}\{A_1,A_2,\cdots,A_m\}$ 表示文本文档 D_i 的特征项为集合 $\{A_1,A_2,\cdots,A_m\}$ 中的某个子集，$T=(t_1,t_2,\cdots,t_n)$ 表示一个多粒度时间。

定义 3 多粒度时间下的文本数据模型是一个二元组 (A,T)，表示文本数据的特征项在多粒度时间 T 处发生，如表 1 所列。

表 1 多粒度时间下的文本数据模型

	A_1	A_2	…	A_m	T
D_1	W(1,1)	W(1,2)	…	W(1,m)	T_1
D_2	W(2,1)	W(2,2)	…	W(2,m)	T_1
⋮	⋮	⋮	⋮	⋮	⋮
D_n	W(n,1)	W(n,2)	…	W(n,m)	T_1

其中每一行表示一个文本，每一列表示文本数据的一个特征词条，$W(i,j)$ 表示文本 D_i 中特征词条 A_j 的权重，T 表示文本的时间属性。

定义 4 文本文档 D_i 的特征项 A_{D_i} 在多粒度时间 T_i 处发生，记为 $E(A_{D_i},T_i)=1$，否则记为 $E(A_{D_i},T_i)=0$。我们称 (A_{D_i},T_i) 为事件，$E(A_{D_i},T_i)$ 为事件的状态函数。

3 时态文本数据周期模型构造

在数学领域将周期定义为一个函数，对于一个函数 $f(x)$，如果存在非零常数 T，使得当 x 取定义域内的任意值时 $f(x+T)=f(x)$ 成立，则函数 $f(x)$ 就称为周期函数，非零常数 T 则称为函数 $f(x)$ 的周期。借鉴数学中周期函数的定义给出多粒度时间下的文本数据周期定义：

定义 5 $(A_{D_i}, T_i, \Delta T)$ 表示文本文档 D_i 的特征项 A_{D_i} 在多粒度时间 T_i 处出现后，经过多粒度时间间隔 ΔT，该特征项在数据库中的其他文本文档中重复出现，若满足下列条件且时间间隔 ΔT 是固定不变的，则称 $(A_{D_i}, T_i, \Delta T)$ 为严格周期模式，ΔT 为周期模式 $(A_{D_i}, T_i, \Delta T)$ 的一个周期。

1) (A_{D_i}, T_i) 存在，即在时态文本数据库的时间范围内，确实存在特征项为 A_{D_i} 的文本文档 D_i；

2) 若 (A_{D_i}, T_i) 存在，假设与文本文档 D_i 的特征项相同的文档为 D_j，则 $(A_{D_i}, T_i+\Delta T)$ 也必定存在，除非 $T_i+\Delta T$ 超出了时态数据库的时间范围。

定义 6 假设时态数据库的时间范围为 $[T, T']$，周期事件 $(A_{D_i}, T_i, \Delta T)$ 在 $[T, T']$ 内重复发生的次数称为周期发生频数，记为 $Sum(A_{D_i}, T_i, \Delta T)$。

$$Sum(E(A_{D_i}, T_i, \Delta T))=\sum_{i=1}^{n} E(A_{D_i}, T_i, \Delta T)$$

式中，$i=1,2,\cdots,n$。

上述定义的严格周期模式在理想情况下是非常有效的，但是在实际生活环境中，周期往往不是固定不变的，而是在某一范围之内波动，由此我们引出宽松周期模式的概念。

定义 7 $(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])$ 表示文本文档 D_i 的特征项 A_{D_i} 在多粒度时间 T_i 处出现后，经过多粒度时间间隔，时间间隔范围为 $[\Delta T-\varepsilon, \Delta T+\varepsilon]$，该特征项在数据库中的其他文本文档中重复出现，则称 $(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])$ 为宽松周期模式，只要在 $[\Delta T-\varepsilon, \Delta T+\varepsilon]$ 范围内的多粒度时间都可以称为周期模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])$ 的一个周期。

宽松周期模式与严格周期模式的区别在于给时间间隔 ΔT 增加了一个宽松度 ε，合理的宽松度可以使很多因为细小的偏差或者噪音而被排除的周期模式能够保存下来，增加了周期模式挖掘的效果。ε 的值越大，可以挖掘出的周期模式越多，周期模式准确率越低，ε 的值越小，可以挖掘出的周期模式越少，周期模式准确率越高。当 $\varepsilon=0$ 时，宽松周期模式就变成严格周期模式了。

定义 8 周期模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])$ 在 $[T, T']$ 内重复发生的次数称为宽松周期事件发生频数，表示 $[T, T']$ 内所有多粒度时间满足宽松周期模式的个数，记为 $Sum(E(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon]))$。

$$Sum(E(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon]))=\sum_{i=1}^{n} E(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])(i=1,2,\cdots,n)$$

定义 9 假设存在 $\varepsilon_1>\varepsilon_2$，则称周期模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon_1, \Delta T+\varepsilon_1])$ 覆盖模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon_2, \Delta T+\varepsilon_2])$，又称模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon_1, \Delta T+\varepsilon_1])$ 为模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon_2, \Delta T+\varepsilon_2])$ 的父模式。

在医药文献数据库中，感冒病毒研究出现的情况如表 2 所列。

表 2 感冒病毒研究情况

序号	时间	感冒病毒研究	感冒病毒研究	严格周期模式	$\varepsilon_1=1$	$\varepsilon_2=2$
1	9月1日	E	E			
2	9月2日					
3	9月3日	E	E	√	√	√
4	9月4日					
5	9月5日	E	E	√	√	√
6	9月6日					
7	9月7日					
8	9月8日					
9	9月9日	E	E			√

从表 2 可以得出感冒病毒研究的严格周期模式为(感冒,(9,1),(0,2))，表示感冒病毒在 9 月 1 日出现第一篇研究文献后，以时间间隔 2 天重复出现。若取宽松度 $\varepsilon_1=1$，则感冒病毒研究的宽松周期模式记为(感冒,(9,1),[(0,1),(0,3)])，表示在 9 月 1 日出现第一篇感冒病毒研究后，隔 1～3 天重复出现第二篇研究文献。若取宽松度 $\varepsilon_2=2$，则感冒病毒研究的宽松周期模式记为(感冒,(9,1),[(0,0),(0,4)])，表示感冒病毒在 9 月 1 日出现第一篇研究文献后，以时间间隔 0～4 天重复出现第二篇感冒病毒研究。严格周期模式和两种宽松周期模式结果见表 2，从表中不难看出 $\varepsilon_2=2$ 时产生的周期模式覆盖了全部 $\varepsilon_1=1$ 时产生的周期模式和所有严格周期模式，并且 $\varepsilon_2=2$ 时产生的周期模式发生频数大于 $\varepsilon_1=1$ 时产生的周期模式发生频数和所有严格周期模式发生频数。

定义 10(支持度和置信度) 严格周期模式 $(A_{D_i}, T_i, \Delta T)$ 的支持度为：

$$\text{support}(A_{D_i}, T_i, \Delta T)=\frac{\sum_{i=1}^{n} E(A_{D_i}, T_i, \Delta T)}{n}$$

同理，宽松周期模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])$ 的支持度为：

$$\text{support}(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])=\frac{\sum_{i=1}^{n} E(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])}{n}$$

严格周期模式 $(A_{D_i}, T_i, \Delta T)$ 的置信度为：$\text{confidence}(A_{D_i}, T_i, \Delta T)=\frac{\sum_{i=1}^{n} E(A_{D_i}, T_i, \Delta T)}{\sum_{i=1}^{n} E(A_{D_i})}$；

同理，宽松周期模式 $(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])$ 的置信度为：

$$\text{confidence}(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])=\frac{\sum_{i=1}^{n} E(A_{D_i}, T_i, [\Delta T-\varepsilon, \Delta T+\varepsilon])}{\sum_{i=1}^{n} E(A_{D_i})}$$

式中，n 表示在时态数据库中所有文本文档集合的数量，$\sum_{i=1}^{n} E(A_{D_i})$ 表示在时态数据库中，出现与文本文档 D_i 的特征项相同的所有文本数据集合的数量。

性质 2 父模式的支持度大于子模式的支持度，父模式的置信度大于子模式的置信度。

例如，分析表 2 不难得到：$\varepsilon_1=1$ 时的周期模式(感冒,(9,1),[(0,1),(0,3)])的支持度为 support(感冒,(9,1),[(0,1),(0,3)])=2/9，置信度为 confidence(感冒,(9,1),[(0,1),(0,3)])=1/2。$\varepsilon_2=2$ 时的周期模式(感冒,(9,1),[(0,0),

(0,4)])的支持度为 support(感冒,(9,1),[(0,0),(0,4)])=1/3,置信度为 confidence(感冒,(9,1),[(0,0),(0,4)])=3/4。$\varepsilon_2=2$ 的周期模式(感冒,(9,1),[(0,0),(0,4)])为 $\varepsilon_1=1$ 时周期模式(感冒,(9,1),[(0,1),(0,3)])的父模式,可以看出父模式的支持度和置信度均大于子模式的支持度和置信度。

4 周期模式挖掘算法

本文借鉴蚁群算法的思想来解决周期模式发现问题,蚁群算法的基本原理是蚂蚁在寻找路径的途中会释放一种特殊的物质——信息素进行信息传递,而且蚂蚁在运动途中能感知到这种物质,并以此指导自己的运动方向,途中信息素浓度越高,蚂蚁选择这条路径的概率就越大。本文参考蚂蚁在途中释放信息素这一特点来解决周期模式挖掘问题。

下面给出相应的算法:

我们假设若事件以时间间隔 ΔT 重复发生一次,则释放信息素,为了计算方便,设 ΔT 处释放的信息素为常量 C,因为我们研究的是宽松周期模式$(i_i,X,[\Delta T-\varepsilon,\Delta T+\varepsilon])$,则 $\Delta T-\varepsilon$ 和 $\Delta T+\varepsilon$ 处也释放信息素 $C'(C'<C)$。事件以时间间隔 ΔT 重复发生的次数越多,留下的信息素浓度越会逐渐增加,但是信息素又有一定的挥发性,我们用 ρ 来表示信息素的挥发系数,ρ 越大,信息素挥发越快。则信息素浓度的计算公式为

$$\tau(i+1)=C(C')+(1-\rho)\times\tau(i) \quad (1)$$

式(1)表示某一时刻的信息素为该时刻释放的信息素加上上一时刻未挥发掉的信息素。具体步骤如下所示:

(1)创建初始值为空的表 L,L 中包括事件发生的可能时间间隔 ΔT、对应的信息素浓度 τ 和重复发生的次数 N;

(2)事件以时间间隔 ΔT 发生,若 ΔT、$\Delta T-\varepsilon$ 和 $\Delta T+\varepsilon$ 对应的信息素浓度 τ 为空,则 ΔT 对应的信息素浓度增加常量 C,$\Delta T-\varepsilon$ 和 $\Delta T+\varepsilon$ 对应的信息素浓度增加常量 C',若 τ 不为空,则按式(1)进行计算信息素浓度,事件发生次数 N 加 1;

(3)按步骤(3)依次处理所有 ΔT;

(4)选取信息素浓度最高的 ΔT 作为事件发生的周期模式。

下面根据表 2 感冒病毒研究情况为例进行计算,设 $\rho=0.4$,$C=1$,$C'=0.5$,$\varepsilon=1$,计算过程如表 3 所列。

表 3 计算过程

	ΔT	τ(信息素)	N
9月3日发生	(0,2)	1	1
	(0,3)	0.5	1
	(0,4)		
9月5日发生	(0,2)	1+(1−0.4)*1=1.6	2
	(0,3)	0.5+(1−0.4)*0.5=0.8	2
	(0,4)		
9月9日发生	(0,2)	1.6	2
	(0,3)	0.5+(1−0.4)*0.8=0.98	3
	(0,4)	1	1

信息素浓度最高的时间间隔为(0,2),感冒病毒研究以时间间隔 2 天重复发生两次。

5 实验及结果分析

我们对 2009 年 MEDLINE 上发表的部分医学文献数据进行分析,结合文献的摘要和标题提取得到相关病毒关键字,选取其中 8 种病毒进行研究,其数量分布规律如图 1 所示。

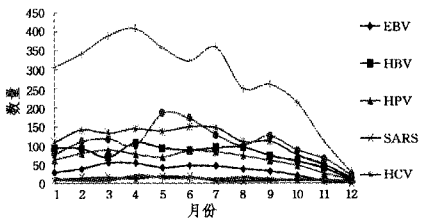

图 1 病毒研究文献分布图

对上述病毒出现的周期现象进行了实验,我们使用(年,月,日)作为实验的多粒度时间格式,设置支持度阈值为0.01,置信度阈值为 0.5,所得的实验结果如表 4 所列。

表 4 实验结果

关键字	周期模式	支持度	置信度	ε
Human Immunodeficiency Virus(HIV)	(HIV,(2009,1,1),(0,0,1))	0.29497706	0.72881862	0
Hepatitis C Virus (HCV)	(HCV,(2009,1,2),[(0,0,1),(0,0,3)])	0.15274088	0.95399698	1
Human Papillomavirus (HPV)	(HPV,(2009,1,1),[(0,0,0),(0,0,2)])	0.06073412	0.66358839	1
Hepatitis B Virus (HBV)	(HBV,(2009,1,2),[(0,0,1),(0,0,3)])	0.10722048	0.97155361	1
Ebstein Barr Virus(EBV)	(EBV,(2009,1,1),[(0,0,1),(0,0,3)])	0.04141512	0.82850242	1
West Nile Virus (WNV)	(WNV,(2009,1,5),[(0,0,2),(0,0,6)])	0.01642115	0.57352941	2
SARS	(SARS,(2009,3,9),[(0,0,3),(0,0,7)])	0.01086694	0.52222222	2
Influenza Virus	(Influenza,(2009,),[(0,0,0),(0,0,2)])	0.11217907	0.71904025	1

根据表 4 可知,自 2009 年 1 月 1 日出现第 1 篇 HIV 病毒研究文献后,每隔 1 天就会出现 HIV 病毒研究文献,HCV 病毒自 2009 年 1 月 2 日出现第 1 篇研究文献后,每隔 1～3 天就会出现第 2 篇 HCV 病毒研究文献,依此类推,WNV 病毒自 2009 年 1 月 5 日出现第 1 篇研究文献后,每隔 2～6 天就会出现第 2 篇 WNV 病毒研究文献。

另外,本文还研究了宽松度 ε 对周期模式支持度和置信度的影响,我们将 HPV 和 INFLUENZA 的宽松度 ε 做了调整,所得结果如图 2、图 3 所示。

图 2、图 3 显示宽松周期模式满足性质 1,即父模式的支持度大于子模式的支持度,父模式的置信度大于子模式的置信度。宽松度为 3 时 HPV 病毒研究文献周期模式(HPV,(2009,1,1),[(0,0,0),(0,0,4)]) 的支持度和置信度均大于宽松度为 1 时周期模式(HPV,(2009,1,1),[(0,0,0),(0,0,2)])的支持度和置信度,即 HPV 病毒研究文献以时间间隔 0～4 天重复出现的概率大于以时间间隔 0～2 天重复出现的概率。

通过对医学文献周期模式的挖掘研究可以较好地预测医学领域的未来研究趋势,总结当前研究热点,若挖掘连续多年的医学文献周期模式还可以归纳出病毒研究的发展规律。从表 4 实验结果可以看出 2009 年的研究热点是 HIV 病毒,其中 HPV 病毒和感冒病毒仍然是医学研究者们关注的焦点,

(下转第 262 页)

计算方法如式(4)和式(5),结果如表7所列。

$$正确率=\frac{实验得到的正确数据数量}{实验得到的所有数据数量}*100\% \quad (4)$$

$$召回率=\frac{实验得到的正确数据数量}{人工标注的数据数量}*100\% \quad (5)$$

表7 正确率与召回率

识别的数据	正确率	召回率
基本块	82.3%	78%
依存关系	89%	90.5%

分析识别错误的基本块,主要有以下几个原因:

(1)分词时,把某些词分成了两个词。

(2)这两个词的词性与原来词的词性不一致。

(3)在把中科院的词性标注集转化为清华树库中的词性标注集时,两个词性标注集合的大小不一样,某些词的词性转化不太标准。

(4)由CRF得到的训练模板还有待改善。

分析识别错误的基本块间的依存关系,原因如下:

若某个基本块内有多个词与其他基本块有依存关系,则无法判断哪个词是该基本块的核心词,也就不能准确地识别该基本块与其他基本块的关系。默认的是选择第一与其他基本块有依存关系的词,该词所对应的依存关系为整个基本块所对应的依存关系。

结束语 本文使用BIO标记集来标注基本块的边界,以识别基本块,由词之间的依存关系来识别基本块间的依存关系,从实验结果看,取得了一定的效果。本文以后的研究工作是改善CRF的特征模板;识别基本块的核心词,把核心词与其他基本块的依存关系作为其所在基本块与其他基本块的依存关系。

参考文献

[1] 宇航,周强.汉语基本块标注系统的内部关系分析[J].清华大学学报,2009,49(10):1708-1711

[2] 袁里驰.基于依存关系的句法分析统计模型[J].中南大学学报,2009,40(6):1630-1635

[3] 陈亿,周强,宇航.分层次的汉语功能块描述库构建分析[J].中文信息学报,2008,22(3):24-31

[4] Steven A. Parsing by Chunks[M]. Robert Berwick, Steven Abney and Carol Tenny, eds. Principle-Based Parsing, Kluwer Academic Publishers, 1991:257-278

[5] 周强.汉语基本块描述体系[J]:中文信息学报,2007,21(3):21-27

[6] 李素建,刘群.汉语组块的定义和获取[C]//语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集.2003

[7] 唐怡.用于常识推理的中文句子语义知识抽取[D].厦门:厦门大学,2010

[8] 宗成庆.统计自然语言[M].北京:清华大学出版社,2007

[9] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]// Proceedings of the 18th International Conf. on Machine Learning. 2001:282-289

[10] 王昕,王金勇,刘春阳,等.基于CRF的汉语语块分析和事件描述小句识别[R].北京:中文信息学会,2009

[11] 程勇,孙承杰,刘远超,等.基于CRFs的级联中文组块识别[R].北京:中文信息学会,2009

[12] 周强.汉语句法树库标记体系[J].中文信息学报,2004,18(4):1-8

[13] http://ir.hit.edu.cn/demo/ltp

(上接第254页)

SARS病毒却已经逐渐淡出人们的视线。以每天都有HIV病毒研究文献发表的频率预测HIV病毒仍然是未来几年医学领域的研究热点。

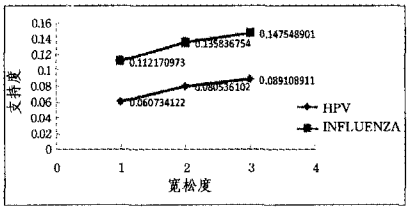

图2 宽松度与支持度的关系

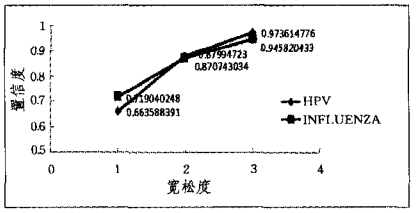

图3 宽松度与置信度的关系

结束语 本文针对文本数据的时间属性,论述了多粒度时间下文本数据的周期模式概念,并在此基础上建立了时态文本数据模型与其周期模型。分别提出了粒度转换和时间间隔定义,给出了宽松周期模式的概念,提出了一个多粒度时间下的文本周期模式挖掘算法,通过实例分析得到了宽松周期的支持度和置信度。本文研究对多粒度时间下的大文本数据的周期模式挖掘具有重要的意义。

参考文献

[1] Bettini C. Testing complex temporal relationships involving multiple granularities and its application to data mining[J]. ACM, 1996,12(4):86-88

[2] Bettini C, Wang S X, Sushil J, et al. Discovering frequent event patternswithmultiple granularities in time sequences [J]. IEEETransactions on Knowledge and Data Engineering, 1998, 10(2):222-237

[3] 孟志青.时态数据采掘中的时态型与时间粒度研究[J].湘潭大学自然科学学报,2000,22(3):1-4

[4] 孟志青.时态关联规则采掘的若干性质[J].计算机工程与应用,2001,37(10):42-44

[5] 姜华,孟志青,肖建华,等.一种时态近似周期的数据挖掘研究[J].软件技术与数据库,2006,32(22):61-63

[6] 程昱.时态数据周期挖掘理论与算法的研究[D].湘潭大学,2005

[7] Li Ying-jiu, Wang X, et al. Discovering Temporal Patterns in Multiple Granularitiesp [C]// TSDM' 00 Proceeding of the First International Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining-Revised Papers. 2007:5-19