

基于命名实体及关系的网页文本关联分析方法

席 峰

(西北工业大学 航天学院, 陕西 西安 710072)

摘 要: 针对传统关联分析技术应用于网页文本分析上存在的问题, 提出一种基于命名实体及实体关系的网页文本关联分析方法. 该方法以命名实体和实体关系作为特征来代替传统高频词, 首先采用基于向量相似度比较的修正策略来提取网页文本中的命名实体, 然后分析 Maxfpminer 算法并对其进行改进, 利用改进的 Maxfpminer 算法对网页文本进行关联分析. 实验结果表明, 该方法分析得到的知识模式的有效性和可读性均优于传统方法.

关键词: 命名实体; 实体关系; 关联分析; 相似度; 特征向量

中图分类号: TP331

文献标识码: A

文章编号: 1000-7180(2011)08-0051-05

WebPages Association Analysis Based on Named Entities and Relations

XI Feng

(Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Aiming at problems of traditional association analysis methods on webpages, this paper presents a new method of webpages association analysis based on named entities and relations. The new method uses named entities and relations, instead of traditional words, to express text feature. First, the new method adopts modify strategy based on vector similarity comparison to extract named entities on webpages. Second, it analyzes Maxfpminer algorithm and improves it. At last, the new method uses the improved algorithm to complete webpages association analysis. Results of experiment show that validity and readability of the new method's result excel to traditional methods.

Key words: named entity; entity relation; association analysis; similarity; feature vector

1 引言

随着 Internet 的不断发展, 网络数据的规模呈指数级增长. 如何快速、准确地获得有价值的网络信息, 并从这些海量数据中发现感兴趣的知识, 是数据挖掘研究中的一大热点^[1]. 在海量的网络数据环境下, 如何快速准确地从中获取到感兴趣的信息, 以及各种信息之间的联系, 成为当今研究的热点之一. 将数据挖掘方法应用到网络数据中, 利用关联分析对网页文本进行关联分析, 已经在深度挖掘网页文本内容及获取感兴趣站点有着迫切的需要.

网页文本关联分析就是利用关联分析算法对搜集到的网页文本集进行关联规则的挖掘. 主要包括

以下过程: (1) 搜集网页文件集; (2) 文件的结构化; (3) 文本的特征化; (4) 文本特征关联分析; (5) 输出关联规则.

本文针对以上问题, 利用网页文本中的命名实体及实体关系作为关联分析的特征, 同时在提取网页中的命名实体及实体关系上采用新的策略, 并对关联分析方法进行了改进. 实验结果表明, 本文基于命名实体及实体关系的网页文本关联分析方法是一种有效的关联分析方法.

2 关联分析方法

关联分析是网页文本关联分析中的一个关键技术, 它是利用关联规则来深度挖掘蕴含在数据之中

的知识的一种技术. 一个事务数据库中的关联规则可以描述如下:

设 $R = \{I_1, I_2, \dots, I_m\}$ 是一组由不同的项目组成的物品集, W 是一组事务集. W 中每个事务 T 是一组物品, 即 $T \subset R$. 假设有一个物品集 A , 一个事务 T , 如果 $A \subset T$, 则称事务 T 支持物品集 A .

关联规则可以表示为如下一种形式: $A \rightarrow B$, 其中 A, B 是两组物品, $A \subset I, B \subset I$, 且 $A \cap B = \emptyset$.

关联规则属性的两个主要参数的定义如下:

(1) 置信度(confidence): 设 W 中支持物品集 A 的事务中, 有 $c\%$ 的事务同时也支持物品集 B , $c\%$ 称为关联规则 $A \rightarrow B$ 的置信度.

(2) 支持度(support): 设 W 中有 $s\%$ 的事务同时支持物品集 A 和 B , $s\%$ 称为关联规则 $A \rightarrow B$ 的支持度.

给定一个事务数据库, 关联分析问题就是通过用户指定最小支持度(minsup)和最小可信度(minconf)来寻找合适关联规则的过程.

传统关联分析方法主要包括 Apriori 算法、FP-growth 算法及 maxfpmminer 算法. Apriori 算法^[2]主要思想是重复扫描数据库并在执行过程可能产生大量的候选集. FP-growth 算法^[3]则不产生候选频繁项集, 它采用分而治之的策略, FP-growth 只需对数据库进行两次扫描, 而且避免了产生大量候选集的问题. 但它在挖掘频繁模式时, 需要递归地生成条件 FP-tree, 在支持度阈值较小时, 即使对于很小的数据库, 也将动态的生成和释放数以十万计的条件 FP-tree, 将耗费大量时间和空间. Maxfpmminer 算法是建立在 FP-growth 算法基础上的一种最大频繁项集挖掘算法, 它只生成最大频繁项集, 其遍历空间大大缩小, 运算时间也大大减少, 适用于高维数据^[4].

3 网页文本关联分析的主要问题

关联分析对于事务数据库而言, 是在大量事务数据之中发现项集之间有趣的关联和相关联系. 对于网页文本而言, 如果将每一文本视为事务数据库中的一个事务或关系数据库中的一条记录, 将每一文本中的特征集视为一个事务中的项集或一个记录中的属性集, 网页文本集中的特征关联分析就可以转化为类似事务数据库中事务项的关联分析. 但实际上在对网页文本集实施特征关联分析需要考虑以下一些问题:

(1) 事务数据库中的事务或关系数据库的记录

都具有结构化的数据, 这是关联规则挖掘展开的基础. 但对于网页文本而言, 这里我们主要考虑 HTML 文件, 它属于半结构化, 直接对 HTML 文件进行关联分析是较困难的, 因此需要将 HTML 由半结构化数据转变到结构化数据.

(2) 对于事务数据库或是关系数据库而言, 被挖掘的最小单位是项或属性. 项与项之间, 属性与属性之间的区别不仅存在于语义中, 还体现在存储结构中, 处理时很容易对它们进行区分. 而对于网页文本则不同, 特征虽然以词或者词组的形式出现在文件中, 但词或词组并不等同于文件的特征, 需要对文本的词/词组进行预处理, 提取合适的特征以形成文件的特征集.

(3) 网页文本的数量对于事务数据库或关系数据库而言是巨大的, 网页文本集的特征集数量与事务数据库的项集或关系数据库的属性集比较而言, 也是很大的, 这对于关联分析算法的效率要求较高. 如果用传统的关联分析算法产生的关联规则的规模将非常庞大. 因此需要在现有的数据挖掘算法中作一定改进, 可以适应特征关联分析的需要.

为了解决问题(1)和(2), 需要对网页文本集实施关联分析前, 进行两个预处理网页文件的结构化和特征化. 为了解决问题(3), 需要一个高效的关联分析算法.

传统的文本特征化采用的就是“词袋法”, 即以文本中出现的词和词组作为文本特征. 基于 tf-idf 加权模式的权重计算方法虽然能够缩小文本特征集的规模, 但对于巨量的网页文本集来说, 特征集仍然太大^[5]. 且由于选择特征词的时候只考虑词频, 很容易丢失一些信息量丰富但出现频次并不高的词语. 这样的特征化方法生成的关联规则冗余度也相当高. 因此, 本文提出利用命名实体及实体关系作为关联分析的特征, 利用高效的改进 maxfpmminer 对海量网络数据进行分析.

4 基于命名实体及关系的网页文本关联分析方法

新方法主要在文本特征化和选择高效关联分析算法两方面对传统方法进行改进. 一是在文本特征化方面, 利用命名实体和实体关系构建文本特征集. 相比传统方法, 该方法提取的特征能够更加充分地描述一篇文本的主要信息和结构. 二是在选择高效的关联分析算法方面利用采用最大频繁项集挖掘算法——Maxfpmminer 算法, 并提出一种基于向量相似

度的修正策略,利用其对分析结果进行冗余处理。

4.1 文本特征向量构建

要在网页文本集中进行关联分析,首要任务就是将网页文本特征化。命名实体是网页文本中的主要信息元素。例如,“新华社电俄罗斯紧急情况部5月3日证实,一架亚美尼亚客机当天凌晨在俄罗斯南部城市索契黑海沿岸坠毁,机上有112人,目前还没有人员伤亡的确切消息。”这段文字中有“新华社”、“俄罗斯紧急情况部”、“5月3日”、“亚美尼亚”、“当天凌晨”、“俄罗斯”、“索契”、“黑海”和“112”等命名实体。这些命名实体指示了这段文字描述的主要信息内容——时间、地点、人物和机构等。因此,本节提出利用这些命名实体作为文本特征是合理可行的。而且,存在于文本中的命名实体本来就不是孤立的,大部分实体之间都是具有一定的关系。例如,“索契”和“俄罗斯”具有地理包含关系。因此,进行网页文本关联分析时,同时将命名实体和实体关系作为特征更加符合网页文本的结构特点,从而能够挖掘到更深层次的知识模式。

定义 $T = \{t_1, t_2, \dots, t_n\}$ 为网页文本集合, t_i 为第 i 个文本向量; $E = \{e_1, e_2, \dots, e_m\}$ 为从网页文本集中提取的命名实体集; $R = \{\text{relation}_1, \text{relation}_2, \dots, \text{relation}_n\}$ 为网页文本集中提取的实体关系集合,则利用提取的命名实体和实体关系构造网页文本 $t_i (t_i \in T)$ 的向量为

$$t_i = (e_{i1}, e_{i2}, \dots, e_{ik}, \text{relation}_{i1}, \text{relation}_{i2}, \dots, \text{relation}_{il}),$$

式中, $e_{ik} \in E, 0 < k < k+1, \text{relation}_{ih} \in R, 0 < h < l+1$ 。

建立一个所有命名实体和实体关系形成的索引,则所有命名实体和实体关系都得到一个唯一的编号。然后将文档向量以这些索引号来表示。则

$$t_i = (s_{i1}, s_{i2}, \dots, s_{ik}), s_{ij} \in (0, m+n+1), 0 < j < k+1.$$

这样,每一个文本都形成一个由其中的命名实体和实体关系构建的特征向量,我们把每个特征向量称为一个事务。所有这些事务形成一个基于命名实体和实体关系的事务集,然后可以采用 Maxfpminer 算法在此事务集中进行挖掘。

4.2 Maxfpminer 关联分析算法

利用最大频繁项集挖掘算法 Maxfpminer 算法对已经构建好的文本特征集进行挖掘,下面给出算法的伪代码。

Algorithm: Maxfpminer(T)

//根据 FP-tree 挖掘最大频繁模式树 MAXT

//输入: T 为当前命名实体和实体关系特征集形成的

FP-tree

//全局变量: Head 为当前链接项列表

MAXT 为最大频繁模式树

//输出: 最大频繁模式树 MAXT

(1) If T 只包含单一路径 then

(2) 将 $T \cup \text{Head}$ 直接加入到 MAXT 中;

(3) else 对 Head-Table 中每一项 i 作如下处理:

(4) 将 i 添加到 Head 中;

(5) 构造 Head 的条件模式库 $H\text{-base}$;

(6) if $H\text{-base}$ 不为空 then

(7) item-list = { $H\text{-base}$ 中的频繁项 };

(8) if item-list $\cup$ Head 没有包含在当前 MAXT 的路径中 then

(9) 由模式库 $H\text{-base}$ 构建条件 FP-tree T' ;

(10) 递归调用 Maxfpminer(T', MAXT);

(11) 从 Head 中删除 i ;

经过这一步骤, Maxfpminer 算法挖掘出了由命名实体和实体关系构成的最大频繁项集。然而此时的频繁项集的冗余还比较多,下面提出一种基于向量相似度的修正策略,对分析结果进行处理。

4.3 基于向量相似度的修正策略

冗余问题一直是关联规则挖掘在实际应用中的一大障碍。大量的冗余规则使得挖掘出来的知识模式的可读性非常有限。因此,必须对挖掘的关联规则进行处理,以便更好地理解所挖掘的规则。

利用 Maxfpminer 算法挖掘特征集合中的频繁项集,已经大大缩小了其频繁项集的规模,为了更好地去除冗余,提高规则的有效性和可阅读性,本节提出一种基于向量相似度比较的关联规则修正策略来对 Maxfpminer 的分析结果进行处理。该策略与已有方法相比,运算复杂度低,更具有实用性。

为了方便表述,将作为网页文本集特征的命名实体集合和实体关系集合统称为特征集合。

频繁项集集合: $L\text{Set} = \{L_1, L_2, \dots, L_M\}$, M 为生成的频繁项集总数; 特征集合: $E = \{e_1, e_2, \dots, e_N\}$, N 为频繁项集集合中的特征总数; 频繁项集向量: $L_i = (e_{i1}, e_{i2}, \dots, e_{ik}), e_{il} \in E, 0 < l < k+1$;

因为频繁项集的项数是不一致的,为了便于比较它们的相似度,将其表示为同维的向量: $L_i = (e'_{i1}, e'_{i2}, \dots, e'_{iN})$ 。

$$\text{其中, } e'_{is} = \begin{cases} w_s & \text{若 } L_i \text{ 包含 } e_{is}, 0 < s \leq N \\ 0 & \text{else} \end{cases},$$

N 为频繁项集集合中的特征总数; 词频权值 $w_s =$

$(0.5 + 0.5 * \text{freq}_i / \text{maxfreq})$, 其中 freq_i 为此特征出现的频次, maxfreq 为所有特征在所有频繁项集中出现次数的最大数。

定义相似度为

$$\text{SIM}_{ij} = \frac{\sum_{k=1}^N e'_{ik} * e'_{jk}}{\sqrt{\sum_{k=1}^N e'^2_{ik}} * \sqrt{\sum_{k=1}^N e'^2_{jk}}} \quad (1)$$

设定一个阈值 threshold , 如果 $\text{SIM}_{ij} > \text{threshold}$, 则称频繁项集 L_i 和 L_j 是相似的。

将频繁项集集合映射到同维的向量空间, 形成频繁项集向量, 并对向量的每一维根据词频赋予权重。然后根据余弦夹角公式计算频繁项集之间的相似度。设定阈值 threshold , 删除相似度大于阈值 threshold 的频繁项集。

基于命名实体及关系的关联分析算法具体步骤如下:

- ① 提取网页文本集中的所有命名实体, 存入命名实体库;
- ② 在此标注好命名实体的网页文本的基础上提取实体关系, 并存入实体关系库;
- ③ 利用网页文本集的所有命名实体和实体关系作为特征形成每个网页文本的特征向量;
- ④ 利用 Maxfpminer 算法对网页文本集形成的向量空间进行关联分析;
- ⑤ 利用基于向量相似度的修正策略对关联分析结果进行冗余处理。

5 实验结果及性能分析

实验数据为网上随机下载的 100 篇包含多个主题的网页, 提取其文本内容后进行实验。首先进行命名实体提取, 将每篇文本由其中提取的不重复的命名实体构成一个事务向量。以下是所提取的命名实体数及所形成的平均事务长度(表 1)。

表 1 网页关联分析实验数据描述表

网页文档数	命名实体数	平均事务长度
100	751	35

5.1 不同文本特征对比实验

支持度阈值设置为 5%~10%, 分别用两种方法构造文本特征集。方法 1: 基于传统的“词袋法”; 方法 2: 基于命名实体和实体关系。然后利用 Maxfpminer 算法(有策略)进行关联分析, 频繁项集数目对比的实验结果如图 1 所示。

两种特征集构造方法在 5% 的支持度下得到的

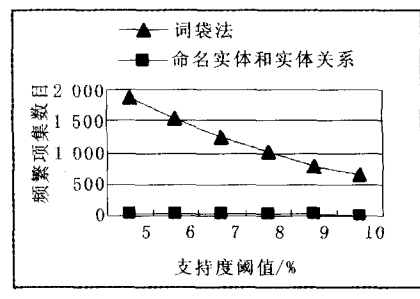


图 1 两种特征集构造方法生成的频繁项集数目对比图
频繁项集的结果分别如图 2 和图 3 所示。

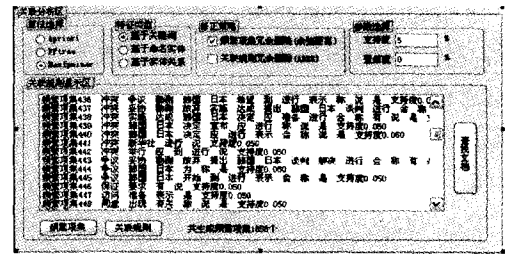


图 2 基于“词袋法”的算法测试结果图

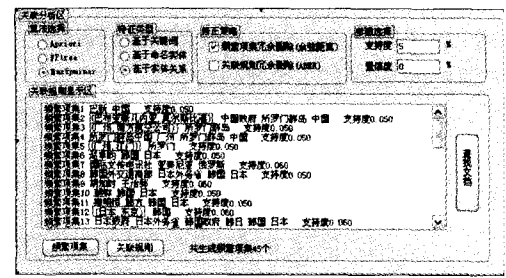


图 3 本节算法测试结果图

从图 2 中给出的曲线图可以看出, 基于命名实体和实体关系的网页文本分关联分析算法产生的频繁项集规模要明显小于“词袋法”。从图 2 和图 3 的测试结果图中可以更直观地看出: 图 3 中, 使用“词袋法”构造文本特征集, 产生了 1 858 条频繁项集, 规模非常大。从内容上来看, 分析得出的频繁项集冗余度都很大, 可读性比较差。图 3 中, 使用命名实体和实体关系构造文本特征集, 产生的频繁项集仅为 45 条, 规模大大缩小, 并且基本完整地保留了有价值的信息内容, 结构化程度和可读性相比“词袋法”都大大提高。

5.2 不同关联分析算法对比实验

利用结合实体关系和命名实体一起构造网页文本特征集, 然后分别利用 FP-growth 算法, Maxfpminer 算法(无策略)和 Maxfpminer 算法(有策略)在支持度 5%~10% 下对新的文本特征集进行关联分析, 进行比较实验。生成的频繁项集数目的实验结果对比如图 4 所示。

FP-growth 算法, Maxfpminer 算法(无策略)在

5%的支持度下分析得到的频繁项集的结果分别如图5和图6所示。

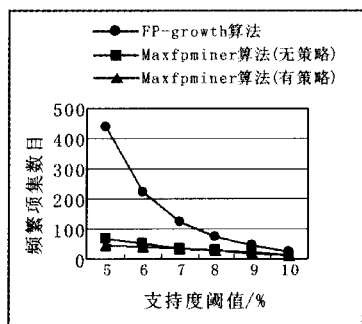


图4 三种算法生成的频繁项集数目对比图

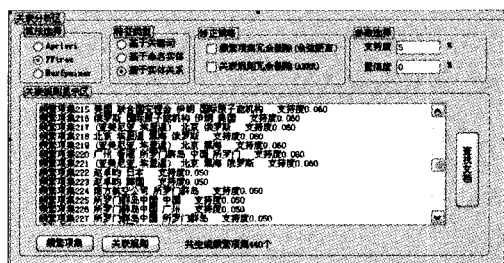


图5 FP-growth 算法测试结果图



图6 Maxfpminer 算法(无策略)测试结果图

从图5可以看出,在相同的文本特征集下,有策略的Maxfpminer算法生成的频繁项集规模要小于FP-growth算法和无策略的Maxfpminer算法。通过观察对比图3、图5和图6,即对比三种算法在5%支持度下的分析结果,FP-growth生成的频繁项集有440条,冗余度大,可读性较差。无策略的Maxfpminer算法生成了65条频繁项集,大大缩小了频繁项集的规模,降低了冗余度。有策略的Maxfpminer算法仅生成了45条频繁项集,进一步缩小了频

繁项集规模,冗余度也降到了最低。

6 结束语

本文提出一种基于命名实体和实体关系的网页文本关联分析方法。该方法以命名实体和实体关系作为文本特征来构建网页文本特征向量;利用一种基于向量相似度的频繁项冗余处理策略对Maxfpminer算法进行改进,并将这种改进的Maxfpminer算法应用于网页文本集的关联分析。经过实验比较,与传统的利用关键词作为文本特征,然后利用传统关联分析算法进行关联分析的方法相比,本章方法关联分析的结果减少了大量冗余项,提高了网页文本关联分析结果的有效性和可阅读性,生成的知识模式的结构化程度大大提高,为在海量网页文本中挖掘有用知识模式提供了一种行之有效的新思路。

参考文献:

- [1] 黄晓斌. 网络信息挖掘[M]. 北京: 电子工业出版社, 2005.
- [2] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]// Proceeding of the 20th international Conference on very large database. Santiago, Chile: Morgan kaufman Pub Inc, 1994:487-499.
- [3] Gosta Grahne, Jianfei Zhu. Fast algorithms for frequent itemset mining using FP-trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2005(10): 1347-1362.
- [4] 陈慧萍, 王建东, 叶飞跃. MAXFP-Miner: 利用FP-tree快速挖掘最大频繁项集[J]. 控制与决策, 2005, 20(8):887-891.
- [5] 李琳. 在Web文本集中进行关联规则挖掘及相关算法的研究[D]. 西安: 西安交通大学, 2002.

作者简介:

席峰 男, (1980-), 硕士研究生. 研究方向为网络数据挖掘、知识发现。