本栏目由西安交通大学出版社协办 7/2004

新闻专题的高效组织和生成新方法

谭 浩[1] 贾自艳[2] 史忠植[3]
(中国科学技术大学研究生院,硕士研究生[1] 北京 100039;新华社国内部,编辑[1] 北京 100803;中国科学院计算技术研究所,博士生[2] 北京 100080;中国科学院计算技术研究所,博士生导师、研究员[3] 北京 100080)

〔摘要〕 为了解决对新闻文档人工分类繁琐的问题,本文采用文本挖掘方法对新闻专题进行组织和生成,对新闻事件进行探测与跟踪,然后进行相关新闻专题的归纳。本文还特别对生成的新闻事件如何进行组织和管理,以及得到专题事件的来龙去脉,做了较为详尽的描述,同时对新闻事件的检索方法做出了一些探索性的工作。
〔关键词〕 新闻专题 数据挖掘 〔中图分类号〕 C93 C94 〔文章编号〕 1000-7857(2004)07-0048-04

HOW TO ORGANIZE AND GENERATE NEWS TOPICS WITH GREAT EFFICIENCY

TAN Hao[1] JIA Zi-yan[2] SHI Zhong-zhi[3]
(1.University of Science and Technology of China, Beijing 100039, China;
2.Institute of Computing Technology ,Chinese Academy of Sciences, Beijing 100080, China
3.Institute of Computing Technology ,Chinese Academy of Sciences, Beijing 100080, China)

Abstract: In this thesis, the author uses the text-mining method of organizing all news formats datums to organize and generate special news topics. To solve the difficulties of manually categorizing news documents, the topic generation is related to the news event probing and tracking. Deeper researches on how to organize and manage generated news events, the detailed procedure of obtaining special events as well as the retrieval of news events have been done in this thesis.With this method , we can improve system efficiency and accuracy of organizing news topics.
Key Words: news topics, data mining

新闻专题是指围绕某一个突发的新闻事件或某一个广泛受关注的问题提供详细、深入的资料。这样的专题信息目的明确、信息丰富,让人一目了然地清楚整个新闻事件的前因后果和来龙去脉,能够较好地满足读者的需要。但通常情况下,这些新闻专题都是经过专业人员加工处理的,即人工归纳到一起。本文研究的目的,是借鉴文本挖掘技术、文本分类和聚类技术,实现对新闻资料的自动组织、生成专题,以满足网络用户检索新闻信息的需要。专题的生成涉及到新闻事件的探测以及对新闻事件的跟踪。本文特别对生成的新闻事件如何进行组织和管理和得到新闻专题的来龙去脉,以及新闻事件的检索方法,做了一些探索性的工作。

1 系统框架和概念层次

新闻专题生成、组织和检索系统能够将报道同一主题的新闻文档自动地聚类在一起,这样可以实现高效的管理和检索。设计该系统最初的目的是为了改善耗时并且代价昂贵的人工组织和管理新闻事件的过程, 同时将体现事件来龙去脉的查询结果,使用友好界面呈现给用户,而不像传统的网络搜索引擎技术只是将检索到的文档排列起来。通过借鉴人工分类网上新闻文档的方法, 这个系统能够从大量语料中分离出描述同一主题的文档,同时能够根据新闻事件的模板自动追踪新来的文档,并将其分配到相关事件的文档中。

该系统讨论的主要数据是新闻语料,因此,如果在设计算法时考虑到新闻语料的特点,那么系统的性能将会得到提高。例如,新闻的标题很重要,一般来讲通过它读者就可以了解该篇新闻报道了什么内容。一篇好的新闻报道应该是即时、准确和全面的。本文将简单介绍时间获取算法(When)、来龙去脉获取算法(How)。至于事件发生的原因以及相关意义和影响, 因为比较难获得,将会是未来工作的重点。

2.知识学习和关键信息的获取

因素的影响,同一时期不同国家又有着不同的模式和实践。因此,应结合我国经济发展的特点和当前我国工程专业结构调整的现状,以及我国独有的文化、教育发展状况建立我国的工程认证制度,进一步明确我国工程专业认证的发展方向、服务面向及实现途径。这样建立起来的工程认证制度才有可能更为公平和公正,才更可能符合我国实际。

参 考 文 献

[1]李延成.美国高等教育认证制度:一种高等教育管理与质量保障模式[J]. 高等教育研究,1998(6)
[2]王庆华.英美工程教育专业认证制度比较研究[A]. 北京:北京航空航天大学高教所,2004

(责任编辑 肖庆山)

2.1 文档知识表示和学习

经过对文本进行分词等预处理后,新闻文档被表示成为一系列的词(向量空间模型——VSM),同时每个词对该文档都有不同的支持度。因此,如何计算词的权重就很重要。同时为了降低其它相关学习算法的复杂度,需要采取某些策略进行特征选择,以减少文档空间的大小。

向量空间模型把文档简化为项的权重为分量的向量表示,把分类过程简化为空间向量的运算, 使得问题的复杂性大大减低。此外,向量空间模型对项的权重评价、相似度的计算都没有作出统一的规定,只是提供一个理论框架,可以使用不同的权重评价函数和相似度计算方法,使得此模型有广泛的适用性。经过权值评价后的特征项,由于有些特征项对应的权值太小,不能很好地表示特征,因此把所有特征项按从大到小排序,只保留前面的若干项。同时,为了保证有足够的特征表示,一般要保留占总权值85%~90%的特征项。

2.2 事件模板知识学习

新闻事件生成时,最重要的参考知识就是事件模板(也叫事件空间 $Know_m$,就像文档空间一样,也是由一组词和权重对组成的。$Know_m = \{(term_1, weight_1)... (term_n, weight_n)\}$)。可以通过一组报道同一主题的文档训练得到事件模板。

事件模板实现新闻专题生成的基本思想,是以人工分类为基础而形成的,即针对一个给定的专题文档类,每来一篇新的文档,就和此文档的权重较高的词相比较,如果符合,则归入专题,否则弃之,并调整专题文档阈值。

2.2.1 事件特征权重计算

事件模板的学习过程依赖于文档向量空间模型的学习过程。根据前面对事件模板知识的介绍,模板知识最关键的问题是计算词对事件的支持度,即事件特征权重的计算,记为 $weight_{t_i,NE}$,计算公式如下所示:

$$weigh_{t_i},t_{NE}=\sum_{D_j} weigh_{t_i},t_{D_j}$$

$weigh_{ti,Dj}$ 是词 t_i 对文档 D_j 的支持度,D_j 是属于特定事件的文档

$$weight_{t_i,NE}=\frac{weight_{t_i,NE}}{MAX\{weight_{t_i,NE}\}}$$ 公式(1)

对 $weight_{t_i,NE}$进行归一化。

2.2.2 事件特征选择

事件特征的选择对新闻事件的专题生成是一个重要环节。由于海量的新闻数据分布极为广泛,所以每个事件模板中含有大量的词,并且它们的权重通常较低。通过分析,笔者发现权重较大的词将更能代表该专题的特征。因此,对于新闻事件来讲,特征挑选过程是必须的,这样可以提高知识学习的效率。本文使用下面两种方法对事件模板进行特征挑选。

(1)前 n 条原则(Top n) 这个方法可以通过选取前 n 个特征进行特征选取,这样可以达到限制模板知识的大小的目的。首先需要以降序对事件特征权重($weigh_{t_i,NE}$)进行排序,然后挑选前 n 条词作为事件特征。

(2)阈值限制(Threshold =Ω) 根据这个原则,挑选 $weight_{t_i,NE} \geq \Omega$ 的词作为事件特征。为了保证事件模板描述事件的准确性,本文主要使用阈值(Threshold = Ω)对事件进行特征挑选,辅助参考前 n 条(Top n)原则。这里设定阈值 threshold = 0.1。

2.2.3 事件模型学习算法

通过以上对文档进行向量空间模型学习,得到了与文档相关的一些知识,并在对事件的特征选择基础上,笔者提出了事件模型学习算法,具体步骤如下:(1)首先对报道同一主题的训练文档进行预处理;(2) 使用文档向量空间模型学习算法得到与文档相关的一些知识;(3)根据公式 1 计算词对事件的支持度;(4)根据前面提到的原则对事件进行特征挑选;(5) 对模板知识建索引或者重建索引以得到索引知识,便于后面的事件查询。

2.3 专题来龙去脉学习

针对大量新闻文章的表达相对固定,可采用类机械式的方法生成多篇文本的综合性的文摘。通过阅读它可以了解整个事件的来龙去脉。目前,对于文摘的处理,应用搜索引擎依然存在令人不满意的地方。因为大部分停留在一篇文章的处理上,虽然产生了文摘,利于用户的查询,但是由于一系列未经处理排序的摘要,还是带给用户一些麻烦。因此,在生成文摘的同时,不但要关注每一篇文章,而且还要注重文摘之间的相互关系,也就是说应对多篇的文摘进行纵向比较,以形成一个线索。

Andrew McCallum 提出建立领域搜索引擎的概念,据此笔者提出一种建立主题文摘的方法, 来避免用户查看大量文摘的繁琐,并达到过滤同内容不同主题的重复信息的文摘处理。对于用户感兴趣的主题,集中地形成一篇经过处理(比如,相关内容的、关联事件的前后排序;过滤内容相同出处不同的网页,等等)的多篇摘要的集合,展示给用户,也就是说将同一概念下的多篇文本,集中地生成一篇摘要集。这样不但可提供给用户一个简单的结果,而且用户不必频繁地翻看链接。如果要进一步了解详尽的内容,可以进入文章查看实情。如此即可获得一举两得的效果。下面对摘要的生成算法进行一定的语言描述。

2.3.1 定义句权, 构造文摘关键句 由于本文采用的是类机械式文摘选取的原则,因此必然存在统计方面的信息。生成的文摘采用文章中的原句构造。首先是在每篇文章中选择比较重要的主题句。由于每篇文章都经过了编码,因此,编码后的文章向量含有文章中重要关键词的词频,根据这些词就可排列出代表文章内容的概念。可按照如下的原则对文章中的句子进行句权统计。

(1) 主题句应含有较多的关键词 不含关键词的句子肯定不是主题句;而含有关键词少的句子,表达主题的能力也较弱。

$W=k\times\frac{N}{S}$, 这里 k 是一个系数,N 是特征词数,S 是句子中总的词数。W 为句子的权值。

(2)有关键词的句子的权重与关键词的权重相统一 由于关键词本身权重不同,因此包含不同关键词的句子,在相同的关键词个数下,句子的权重由关键词的重要程度决定。

$W=\sum_{k=1}^{j} w_k$,j 表示一句话中出现的关键词数,W_j 表示关键词的词权,W 为句子的权值。

(3)题句的重要程度与分句的个数成反比

$W=k'\times\frac{N}{S\times S_1}$,N 是特征词数,S 是句子中总的词数,$S_1$ 是分句的个数,k′是加权系数。

(4)修改分句的结构 一句话中含有多个分句,分句作为整个句子的有益补充更能说明句子的意图,但是,在表达上不可避免地存在某些分句过于冗长或者对主题意义的表达不是很重要的语句。在这种情况下,生成摘要的时候可以对分句进行适当的删改。在后面的实验结果比较中可以看出修改的有效性。

2.3.2 通过句子整理构造文摘 生成多篇文档的文摘需要将加权后的文档句子按以下步骤进行整理。(1)按权值将各个主题

句排序。由于新闻类的文章在表达上先后顺序与前面的权重的高低基本上保持一致,因此各句子排列好了后基本上不会出现表达的问题。(2)过滤重复文章。大量内容雷同的网页造成网民对同一事件相同内容的重复转载,因此在生成多篇摘要的同时,应通过内容的比较过滤重复内容。(3)按照事件发生的顺序排列各篇文摘。在生成多篇文摘时,或者按照时间的顺序,或者按照事件发展的顺序进行排列。由于时间作为主导性的因素,所以本文中采用以时间为线索排列多篇文摘。

2.3.3 动态调整文摘 在组合多篇文章的文摘过程中,可以根据用户的需要,按照文章的百分比输出摘要。在摘要的输出中,要根据用户的需要输出任意比例的摘要。其中的比例可以在对单篇的处理过程中完成。主题句提取时应按照先后顺序将超出比例的句子直接过滤,这样便可为组合文摘解决比例的问题。按比例地调整文摘的长度,即可动态地调整输出的效果,从内容上则做到了从概括到具体的展示效果。

3 新闻专题的生成

如何自动地生成新闻专题是该系统最重要的任务。根据事件模板可以自动对新闻报道进行跟踪,这样可以将报告同一主题的新闻组织在一起,生成新闻事件。事件模板可以通过对指定文档进行训练而得到,或者由专家直接给定模板知识。新闻专题生成的过程如下:(1)得到事件的模板知识;(2)得到文档时间和文档的向量空间模型知识;(3)计算文档和事件之间的相关性($score_{d,NE}$),通过它可以决定该文档属于哪个事件;(4)根据文档之间的相似度度量,对事件进行去冗余处理,该过程主要是基于新闻的 5 个要素(When,Who,Why,Where,How)来实现的;(5)重复执行步骤 1 到步骤 5,直至得到好的结果。

上面的步骤 5 可实现对事件模板的不断修正。由于新闻事件是动态变化的,所以事件模板也应是动态的。例如“美国 9·11 恐怖事件”,开始时所有关于该事件的报道都是基于恐怖发生过程的;接下来是救助报道以及美国政府和世界人民对此的反映;然后就是对事件起因的推测和调查,这时“拉登”一词即频繁出现在新闻报道中。

这里采用传统的余弦夹角公式来计算文档向量和事件模板向量之间的相似度,公式如下:

$$sim_{d,NE}=\frac{\sum_{t_i\in know_d\cap know_m}(weight_{t_i,d}\times weight_{t_i,NE})}{\sqrt{\sum_{i=1}^{\|know_d\|}weight_{t_i,d}^2}\times\sqrt{\sum_{i=1}^{\|know_m\|}weight_{t_i,NE}^2}} \quad \text{公式(2)}$$

其中,$t_i\in know_d\cap know_m$ 表明词 t_i 是文档向量和事件模板向量的共同元素,$weight_{t_i,d}$ 是词 t_i 对文档 d 的支持度,而 $weight_{t_i,NE}$ 是词 t_i 对事件 NE 的支持度,$\|know_d\|$ 是文档向量的大小,而 $\|know_m\|$ 是事件模板向量的大小。

因为事件模板的不精确性,有些文档不能很好地定位到它所属的事件。例如,在处理“美国恐怖事件”和“中美撞机事件”时,就很难将文档分派给正确的事件。这是因为两个事件有很多共同的重要特征词,例如“美国”、“飞机”等。因此,对于新闻文档的分类来说,仅仅考虑向量之间相似度计算是不够的。前面已经提到,时间是新闻文档中最重要的要素,所以在事件生成过程中如果考虑到时间因素将会提高分类精度。一般来讲,对于某个事件的报道将会持续一段时间,所以文档与某个事件之间的时间距离越小,理论上它属于该事件的可能性就越大。例如,前面提到的“美国恐怖事件”和“中美撞机事件”这两个例子,它们发生在不同的时间,如果考虑到时间距离,那么报道这两类事件的文档就很容易区分开。本文定义文档和事件之间的时间距离如下:

$$Dis_{time}(d,NE)=\min\{|time_d-time_{NE_b}|,|time_d-time_{NE_e}|\} \quad \text{公式(3)}$$

其中,$time_{NE_b}$ 是事件开始时间,$time_{NE_e}$ 是事件最近时间,$time_d$ 是文档时间。

得到时间距离后,就可以修正前面的相似度计算公式。

$$score_{d,NE}=\frac{sim_{d,NE}+\theta\times(1/dis_{time}(d,NE))}{2} \quad \text{公式(4)}$$

其中,θ 是用来调整时间因素的影响。

相似度得到后,就需要选择相应的规则来判断某个文档应该属于哪个事件。与事件特征挑选相似,也可使用下面的两个原则来解决文档所属问题。(1)阈值判断(threshold =θ)。如果文档和事件之间的相似度度量小于θ($score_{d,st}\le\theta$),那么该文档不属于此事件。(2)第一名原则(Top 1)。文档属于相似度度量最大的一个。

4 专题检索

专题检索模块是该系统最重要的组成部分。大多数提供新闻专题的网站,都是将人工组织的那些新闻事件排列在一起,如果用户想了解某个事件的来龙去脉,它需要依次打开属于某个事件的文档。这对用户来说是耗时且极为不便的。该系统为用户提供了友好界面,方便其对事件的检索。与一般的搜索引擎相似,它提供了两种类型的检索方式,即事件目录检索方式和关键词检索方式。

专题检索的流程如下。(1)用户输入查询字符串。该字符串能够最大程度地表明用户的意图。(2)解析查询字符串。将查询字符串解析成一组关键词,每个词的权重给 1,这样就得到查询知识,表示为 $know_q=\{(term_1,1)\cdots(term_n,1)\}$。(3)事件检索过程。根据相似度计算公式 2,计算查询知识($know_q$)和事件模板知识($know_m$)之间的相似度度量。(4)将检索到的事件根据相似度以降序排列。(5)根据某些原则,挑选出用户想要的事件。(6)根据输入参数决定结果的显示方式,用户输入的参数将会决定查询结果的显示方式。(7)显示结果。友好的人机交互界面可以方便用户对查询结果的阅读。

5 实验结果

本文的实验数据是 3 000 多篇新闻文档,包括的事件有“美国恐怖事件”(NE1)、“中美撞机事件”(NE2)、“厦门远华特大走私案”(NE3)、“江泽民访美”(NE4)、“布什访华”(NE5)等事件,将这些新闻文档放在一起组成原始的语料。根据本文所介绍的各种知识学习算法以及工作流程,可以对原始的语料进行分类并得到相应的新闻专题,同时可以对组织好的事件进行查询。

事件生成和检索的评价方法很相似,一般用回收率和正确率来评价。假设检索到的相关信息用 a 表示,检索到的不相关信息表示为 b,而未检索到的相关信息用 c 表示,未检索到的不相关信息表示为 d。其中,检索到的文档是指被系统正确分类的文档,相关文档是指人工判断与事件相关的文档。本文中使用的效率评价公式从下表总结可得:回收率=$\frac{a}{a+c}$,正确率=$\frac{a}{a+b}$。

表 1 未考虑时间距离的实验结果

	回收率	正确率
NE1	0.67	0.72
NE2	0.75	0.59
NE3	0.87	0.91
NE4	0.34	0.54
NE5	0.56	0.48

高级内部信用风险度量模型方法的比较

杨蕴石[1] 徐枞巍[2]

(北京航空航天大学经济管理学院,博士研究生[1] 北京 100083; 合肥工业大学,校长、教授、博士生导师[2] 合肥 230009)

〔摘要〕 本文对目前国际上比较著名的信用风险管理模型 CreditMetricsTM、KMV 系统、CreditRisk+、CreditPortfolio View 及其方法进行了一系列比较,研究了模型建立的理论基础以及模型之间的相互关系,阐述了现代信用风险度量技术和方法的特点和发展趋势。这 4 个新型信用风险度量模型的框架和思路,将对我国金融机构信用风险管理提供有益借鉴。

〔关键词〕 商业银行 信用风险度量模型 风险价值 〔中图分类号〕 C93 F8

〔文章编号〕 1000-7857(2004)07-0051-003

AN ANALYSIS ON THE MODELS FOR THE MEASUREMENT OF CREDIT RISKS

YANG Yun-shi[1] *Xu Cong-wei*[2]

(1.Beijing University of Aeronautics and Astronautics, Beijing 100083, China;
2. Hefei University of Technology, Hefei 230009, China)

Abstract: Various credit risk assessment approaches and models have been put forward in the recent years. The paper has introduced the most famous models including CreditMetricsTM, CreditRisk$^+$, KMV and CreditPortfolio View, and analyzed the differences, advantages and disadvantages among them. These models, with their unique framework and reasoning clues, will have reference values to the country's credit risk management of financial institutions.

Key Words: commercial bank, credit risk measurement model, value at risk

20 世纪 80 年代以来,受债务危机的影响,各国银行普遍重视对信用风险的管理和防范。1997 年爆发的亚洲金融危机,使各大银行和监管机构更加重视信用风险、市场风险和操作风险的综合管理和量化分析,国际金融机构开发了一系列成功的信用风险量化管理模型。目前,国际上对信用风险管理的研究,集中体现在以下几个比较流行的风险管理系统:JP Morgan(1997)银行开发的信用度量制(CreditMetricsTM)系统,KMV 公司开发的 KMV 系统(KMV,1993), Credit Suisse First Boston(CSFB,1997)银行开发

表 2 考虑时间距离的实验结果

	回收率	正确率
NE1	0.94	0.98
NE2	0.99	0.97
NE3	0.89	0.93
NE4	0.98	0.92
NE5	0.92	0.94

上述实验结果表明, 在未考虑文档和事件之间的时间距离时,很难将 "美国恐怖事件"和"中美撞机事件"的相关报道区分开,这是因为它们的模板知识的很多特征都是相似的;同样"江泽民访美"和"布什访华"也是很难区分的。如果在计算文档和事件之间的相似度时,考虑了它们之间的时间距离,将会得到令人欣喜的结果(见表 2)。

关于新闻专题来龙去脉生成算法的性能评价需要人工来完成,实验结果表明,它能够很好地反映出事件的发生、发展的全过程。例如,用户输入"美国"、"恐怖"、"事件"这几个字符串去检索相关新闻专题的来龙去脉, 同时输入相关参数来决定显示方式,该例子包括了 1 142 篇相关文档。

参 考 文 献

[1]史忠植. 知识发现[M]. 北京:清华大学出版社,2002

[2]王继成,孙颖,张福炎.文本挖掘——数据挖掘研究的新课题[J]. 兰州大学学报(自然科学版),1999,35(专辑):314~318

[3]傅伟鹏. 文本聚类和文本摘要的研究[A].合肥:中国科学技术大学,2002

[4]陈建华,包煊. WEB 挖掘系统的设计与实现[J]. 计算机工程,2002,28(8):141~143

[5]严威,赵玫. 开发中文搜索引擎汉语处理的关键技术[J]. 计算机工程,1999,25(4):38~38

[6]林鸿飞,庄恩贵,姚天顺.中文文本挖掘中数字特征的抽取和表示[A]. 1999 青岛—香港国际会议文集

[7]谢春丽,崔志明. web 文本挖掘中特征提取的设计与实现[J]. 微机发展, 2003,13(2):77~79

[8] 李晓黎. WEB 信息检索与分类中的数据采掘研究[A]. 中国科学院计算机技术研究所,2001

[9]朱华宇,孙正兴,张福炎.一个向量空间模型的中文文本自动分类系统[J]. 计算机工程, 2001,27(2): 15~17

[10]李俊杰. 非受限域中文自动文摘系统研究与实现[D].哈尔滨:哈尔滨工业大学,1995

[11]孙春葵,李蕾,杨晓兰,钟义信. 基于知识的文本摘要系统研究与实现[J]. 计算机研究与发展,2000, 37(7)

(责任编辑 肖庆山)