整合文本挖掘的商务智能系统结构研究

葛育祥,熊 励
(上海大学 国际工商与管理学院,上海 200444)

摘 要:针对当前商务智能系统对文本类资源的处理功能缺失,引入文本挖掘的概念,分析文本挖掘的过程、特点和处理方法。在此基础上设计了一个具有文本挖掘功能的商务智能系统架构,并对其中的一些关键技术,如数据预处理技术、文本聚类算法和文本分类算法等文中挖掘算法进行详细分析,以期对当前商务智能系统的功能的扩展有所帮助。

关键词:商务智能;文本挖掘;数据挖掘;数据预处理
中图分类号:G202;TP18 **文献标识码**:A **文章编号**:1673-629X(2009)04-0001-04

System Structure Study of Business Intelligence Integrated Text Mining

GE Yu-xiang, XIONG Li
(College of International Business and Management, Shanghai University, Shanghai 200444, China)

Abstract: Points out business intelligence's disadvantage in text mining. Then introduces the text mining, analyzes its process, traits and technique, and reveales the framkwork of business intelligence integrated text mining and analyzes its key technologies, such as data preprocessing, text classification, text clustering, in order to improve business intelligence's function.

Key words: business intelligence; text mining; data mining; data preprocessing

0 引 言

随着企业信息化的深入发展和数据、信息、知识的不断增加,商务智能应运而生,并伴随数据仓库、联机分析处理(OLAP)和数据挖掘等技术的日趋成熟而快速发展。从起初人们对商务智能的引入到现在作为信息化的一个热点领域,吸引广大不同专业背景的学者对其进行研究;从开始商务智能项目实施的数据驱动模式到现在的业务驱动模式,这些都印证了我国商务智能快速发展的态势。但在发展过程中,人们也逐渐认识到商务智能并非万能,如果能更好地突破现有瓶颈,商务智能的发展前景将更为广阔。其中对文本数据源处理的缺失,就是值得关注的一个方面。

1 商务智能

当前数据处理系统的目标已变成商务领域信息处理的快速化、成本降低和自动化。决策者认识到企业生存依赖于有效的信息,如使用ERP,CRM等软件工具辅助企业管理与决策,但结果是这些系统产生了大量的数据,而这些数据都需要进一步的分析才能充分地发挥自己的价值;另一方面,随着全球化的继续推进和组织的日益分散化,使得认清市场趋势和收集竞争者的信息变得尤为重要,这就要企业对市场变化做出快速反应,但这些信息分布在许多系统中,甚至在不同的国家,这就使得有效利用数据变得异常困难。有效的信息处理是维持竞争优势的一个决定性因素,而商务智能正是对这些问题做出的有力回答。

1.1 商务智能定义

商务智能BI(Business Intelligence)于1989年由Gartner Group的Howard Dresner首次提出,它描述了一系列的概念和方法,并通过基于事实的支持系统来辅助商业决策的制定。之后随着商务智能的深入发展和人们对其认识的加深,不断有不同的商务智能的定义出现。文中采用2007年Gartner商务智能峰会对BI的重新定义,即BI为一个伞状的概念,它包括了分析应用、基础架构和平台和良好的实践。数据仓库、数据标准等平台已经涵盖在BI范畴里,BI已不再仅仅是前端展现工具。商务智能已经开始成为一种用于描述企业范围内使用数据、分析信息、制订决策和管理绩效的原则的术语。而且组织应该用全面的绩效提升来衡量商务智能的成功。衡量BI的成功与否,不再是数据组织的有序、数据的ETL过程的更完美,不再是异

收稿日期:2008-07-05
基金项目:上海市教育科研基金项目(07ZS22)
作者简介:葛育祥(1984-),男,硕士研究生,研究方向为商务智能和数据挖掘;熊 励,博士后,教授,研究方向为商务智能与决策支持等。

构数据的集成能力，也不再是数据变换和数据归约的强大功能，而是 BI 是否有助于促进企业业绩的提升。此外，BI 分析型系统应该强调和形成效果，也就是说，BI 必须要促进和业务或某一方面业务的顺利展开，提升业绩。BI 的核心在于应用，这也是 BI 实践者在工作中的真实体会。

1.2 商务智能的缺陷

当前商务智能的发展取得了很多可喜的成就，如实施模式从数据驱动转变到业务驱动，利用 ETL 提高数据质量，技术实现开始围绕业务需求设计等。当前市场上 BI 产品很多，有 Microsoft 的 SQL Server 2005，Business Objects 的 Business Objects XI 3.0，SAS 的 SAS®等。特别是 2007 年，信息软件产品领域的巨人 IBM 和 SAP 分别通过巨资收购 Cognos 和 Business objects 来大举进军商务智能市场，突显 BI 市场的广阔前景。但分析这些产品和总结当前理论研究来看，商务智能还存在很多不足，其中不能充分对文本类资源进行分析与处理是急需解决的。

因为当前企业中充斥着这种文本资料，如各种文书、技术报告、E－MAIL、市场报告等。过去的 40 几年里，每年在联机医学文献分析和检索系统中出版的摘要以五倍的速度增长。而且超过 1200 万甚至更多的在线资源都是全文本的文章。除了这些，还有专利、内部报告和其他潜在可获取的公开资源。尽管有一小部分信息是以结构化的形式存在于数据库中，但 80% 的信息是以自然语言组织的非结构和手写资源[1]。如何将这些资源也充分地整合到商务智能的数据源中，进行和结构化数据一样的分析和处理，是要深入研究的，而文本挖掘作为数据挖掘的一个分支为解决问题提供了一个很好的方法。

2 文本挖掘

文本挖掘(TM，Text Mining)是以计算语言学、统计数理分析为理论基础，结合机器学习和信息检索技术，从文本数据中发现和提取独立于用户信息需求的文档集中的隐含知识。它是一个从文本信息描述到选取提取模式，最终形成用户可理解的信息知识的过程[2]。

2.1 文本挖掘的发展

文本挖掘之前，用信息抽取技术(Information Extraction，IE)进行非结构化的信息挖掘。但随着 IE 系统的发展，人们认为它更适合利用精确的查询相匹配概念和文字找出关系。IE 系统的主要优势在于以下几点：查询的精确，输出结果的透明和直接进入数据库或真实地显示出来。“文本挖掘”这个词通过类似于传统的数据挖掘系统一样被应用到这些系统中。针对 IE 对自然语言处理的不足，文本挖掘可以通过统计共现方法处理自然语言。文本挖掘的过程(见图 1)与数据挖掘的一般过程有所不同。

文本具有有限的结构，有的甚至没有结构，此外计算机不能直接处理人类的自然语言，所以对文本数据源要进行数据预处理。数据预处理主要包括分词技术(英文文本则需要 Stemming 技术)和特征表示和特征提取。因为中文词与词之间没有固有的间隔符(空格)，需要进行分词处理[3]。

2.2 文本挖掘的特点

文本挖掘的对象是书籍、研究论文、Web 网页等非结构化的数据源。这些数据不同于数据仓库里结构化的数据，一般只有几百个特征数目，文本数据转换为特征矢量后特征数目将至少也达到几万。所以尽管作为数据挖掘的一个分支，文本挖掘有自己的一些特点和处理方法。文本数据有其特殊性，不能直接表示出来，之前要经过特征提取，就是找出最能代表文本的文本特征存储起来，处理数据就是处理这些文本特征的集合，而不是原来的文本，典型的文本表示模型有向量空间模型(VSM，Vector Space Model)。用向量空间模型提取得到的特征向量的维数一般都有几十万维，这些特征向量并不全是对结果有用的，而且高维会增加机器学习的时间，所以特征选择就势在必行。特征选择一般会利用一些评价函数，对特征进行评价，然后进行取舍[4]。

文本经过分词、特征表示和特征提取后就可进行挖掘了。对于非结构化问题，一条途径是发展全新的数据挖掘算法直接对非结构化数据进行挖掘，由于数

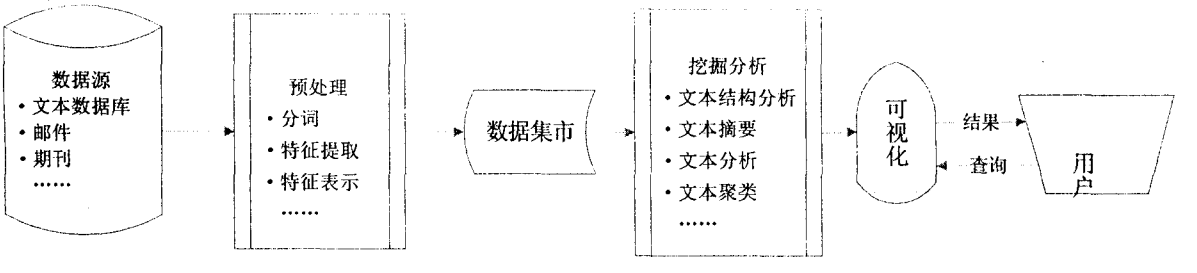

图 1 文本挖掘过程

据非常复杂,导致这种算法的复杂性很高;另一条途径就是将非结构化问题结构化,利用现有的数据挖掘技术进行挖掘,目前的文本挖掘一般采用该途径进行处理。对于语义关系,则需要集成计算语言学和自然语言处理等成果进行分析[5]。常用的文本挖掘技术有:文本结构分析、文本摘要、文本分类、文本聚类、文本关联分析、分布分析和趋势预测。

3　整合文本挖掘功能的商务智能系统结构

结合以上的介绍,将文本挖掘整合到商务智能中是可能的,具体的体系结构见图 2。

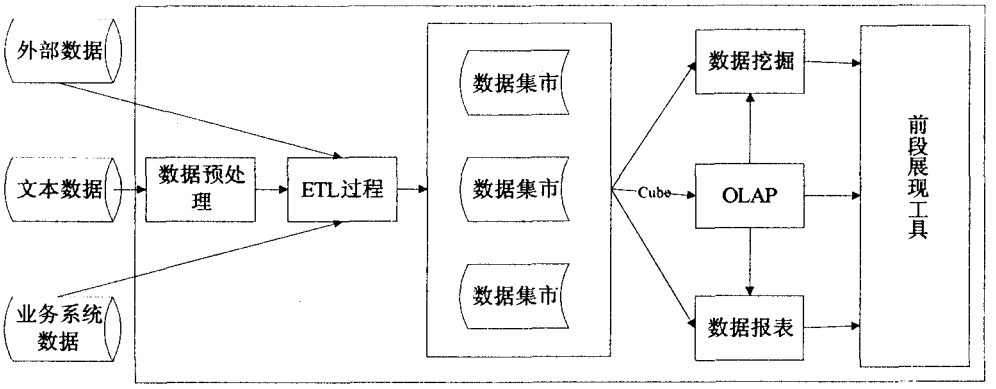

图 2　整合文本挖掘的商务智能体系结构

整合以后的商务智能主要有三个数据来源:外部数据,文本数据和业务系统数据。外部数据,如竞争者动态,市场需求等被收集起来分析企业的竞争环境;业务系统数据则是对企业内部各信息系统存在的数据的总称,如 ERP,CRM 中的数据,这些数据被用来跟踪企业的内部业务过程,分析自身竞争能力;而文本数据则是对这些数据的补充。对这些非结构化以及半结构化的数据进行分析挖掘,可以发现产品存在的问题,客户的需求以及忠诚度,竞争对手的动向以及行业的发展趋势等。

这三种数据源,文本数据需要进行数据预处理,经过分词、特征表述和特征选择和必要的语义处理,对结果数据可以用 XML 进行结构化表示,或进一步倒入关系型数据库中。不过考虑了去其他两个数据源的数据标准化问题,需要 ETL 工具对其和来自另外两个数据源的数据进行抽取、转换、清洗、过滤和转载,以统一的格式或标准存入数据仓库中。

数据仓库的应用可以包括三个方面:OLAP、数据挖掘和数据报表。从数据仓库中可以直接生成报表。通过对数据仓库中的数据进行切片、钻取和旋转等方式,可以完成决策支持需要的查询及报表。而通过数据挖掘可以发现隐藏在数据中的潜在规则[6]。最后将这些分析的结果通过前段展现工具展示出来。

4　关键技术分析

4.1　数据预处理技术

现在大多数文本分类器都使用向量空间模型对文本进行表示,它们不考虑词条在文本中的位置、次序以及文本机构,运用自然语言处理对训练文本进行分词。抽取出训练样本中代表其特征的元数据,即文本特征 t_i,这样整个文本就用它的特征项 $t_1, t_2, \cdots, t_n$ 表示。然后对每一特征定义一个权值 $w_1(d), w_2(d), \cdots, w_n(d)$ 为其特征对应的权值,从而使训练文本映射为一个特征向量:$V(d) = (t_1, w_1(d), t_2, w_2(d), \cdots, t_n, w_n(d))$,$w_i(d)$ 一般定义为 t_i 在文本 d 中出现频率数 V_{TFi} 的函数[7]:

$$w_i(d) = f(V_{TFi}(d)) \tag{1}$$

常用的权值函数公式有:

① 布尔函数 $\Psi = \begin{cases} 1 & V_{TFi}(d) \geq 1 \\ 0 & V_{TFi}(d) = 0 \end{cases}$

② 平方根函数 $\Psi = \sqrt{V_{TFi}(d)}$

③ 对数函数 $\Psi = \log(V_{TFi}(d) + 1)$

④TFIDF 函数 $\Psi = V_{TFi}(d)\log(\frac{N}{N_i})$

其中 N 为所有文档数目;N_i 为所含词条 t_i 的文档数目。

下面对于特征项的选取方法——特征评估函数进行介绍[8]。

对文档特征所采用的特征项选取算法通常是构造一个评价函数,对特征集中的每一个特征进行独立的评估。这样每个特征都获得一个评估分。然后对所有的特征按照其评估分的大小进行评估。具体函数为:

(1)信息增益(Information Gain)。

$$\text{InfGain}(F) = P(W)\sum_i p(C_i/W)\log\frac{P(C_i/W)}{P(C_i)} + P\overline{W}\sum_i p(C_i/\overline{W})\log\frac{P(C_i/\overline{W})}{P(C_i)} \tag{2}$$

(2) 期望交叉熵(Expected Cross Entropy)。

$$\text{CrossEntry}(F) = P(W)\sum_i P(C_i/W)\log\frac{P(C_i/W)}{P(C_i)} \tag{3}$$

(3) 互信息(Mutual Information)。

$$\text{MutualInfo}(F) = \sum_i P(C_i)\log\frac{P(C_i/W)}{P(C_i)} \tag{4}$$

(4) 文本证据权(Weight of Evidence for Text)。

$$\text{WeightofEvidText}(F) = P(W)\sum_i P(C_i) \mid \cdot \log\frac{P(C_i/W)(1-P(C_i))}{P(C_i)(1-P(C_i/W))} \mid \quad (5)$$

(5) 词频(Word Frequency)。

$$\text{Freg}(f) = TF(W) \quad (6)$$

F 为对应于单字 W 的特征，$P(W)$ 为单字 W 出现的概率，$\overline{W}$ 意味着单字 W 并不出现，$P(C_i)$ 为第 i 类值的出现概率，$P(C_i/W)$ 为单字 W 出现时属于第 i 类的条件概率，$TF(W)$ 为单词在文档集中出现的次数。

4.2 文本挖掘技术

文本挖掘的关键技术有文本分类和聚类，下面对这两种技术进行分析。

4.2.1 文本分类

文本分类[9]是一种有指导机器学习问题，它需要事先定义一些主题类别。然后根据文本的内容自动将每篇文档归入其中的一个类别，这样用户即可以根据自己的所需来选择信息。从数学角度来看，文本分类其实就是一个映射的过程，它将未标明类别的文本映射到已有的一个或多个类别中。

分类系统一般分为训练和分类两个阶段，具体过程如下：

(1) 训练阶段：首先需要确定类别的集合 C，这些类别可以是层次式的，也可以是并列式的。再选择适量具有代表性的文档组成训练文档集合 S，确定训练文档集 S 中的每个训练文档 s_j 所属的类别 C_i，然后抽取训练文档 s_j 的特征，得到特征向量 $\boldsymbol{V}(s_j)$，最后，统计训练文档集 S 中所有的文档特征向量 $\boldsymbol{V}(s_j)$，以此确定代表 C 中每个类别的特征矢量 $\boldsymbol{V}(C_i)$。

(2) 分类阶段：对于测试文档集合 T 中的每个待分类文档 d_k，计算特征矢量 $\boldsymbol{V}(d_k)$ 与每个 $\boldsymbol{V}(c_j)$ 之间的相似度 $\text{sim}(d_k, c_j)$ 然后选取相似度最大的类别作为 d_k 的类别，如 d_k 与这些类别之间的相似度超过某个预定的阈值。也可以为 d_k 指定多个类别，如果 d_k 与所有类别的相似度均低于阈值。可将该文档放在一边，由用户来做最终决定。对于类别与预定义类别不匹配的文档而言，这是合理的，也是必需的，但是如果这种情况经常发生，就需要修改预定义类别，并重新进行训练与分类过程。分类算法很多。主要有朴素贝叶斯分类(Native Bayes)、向量空间模型、决策树、支持向量机、后向传播分类、遗传算法、基于案例的推理、K— 最临近、基于中心点的分类方法、粗糙集、模糊集以及线性最小二乘等。

4.2.2 文本聚类

文本聚类与分类的不同之处在于，聚类没有预先定义好的主题类别，它的目标是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能地大，而不同簇间的相似度尽可能地小。利用文本聚类技术将搜索引擎的检索结果划分为若干个簇，用户只需要考虑那些相关的簇，大大缩小了所需要浏览的结果数量。目前，有多种文本聚类算法，大致可以分为两种类型：以 G－HAC 等算法为代表的层次凝聚法，以 k－means 等算法为代表的平面划分法。下面就层次凝聚法给出详细说明。

对于给定的文档集合 $D = \{d_1, d_2, \cdots, d_n\}$，层次凝聚法的具体过程如下[10]：

① 将 D 中的每个文档看作是一个具有各成员的类 $C_i = \{d_i\}$，这些类构成了 D 的一个聚类 $C = \{c_1, \cdots, c_i, \cdots, c_n\}$；

② 计算 C 中每对类 (c_i, c_j) 之间的相似度 $\text{sim}(c_i, c_j)$；

③ 选取具有最大相似度的类对，$\arg\max_{c_i \in C} \text{sim}(c_i, c_j)$，并将 c_i 和 c_j 合并为一个新的类 $c_k = c_i \cup c_j$，从而构成了 D 的一个新的聚类 $C = \{c_1, \cdots, c_{n-1}\}$；

④ 重复上述步骤，直至 C 中剩下一个类为止。

该算法构造出一棵生成树，包含了类的层次信息以及所有类内和类间的相似度，但是在每次合并时，需要全局地比较所有类间的相似度，并选择出最佳的两个类，因此运行速度较慢，不适合于大量文档的集合。

5 结束语

文中研究了整合文本挖掘的商务智能体系结构，分析了其中的一些关键技术，试图扩大当前商务智能的功能，使其能对文本进行处理。基于文本挖掘的商务智能系统是使商务智能能进行文本挖掘的一种新方法。而且随着商务智能的快速发展和商务智能市场的日趋成熟，整合文本挖掘功能的商务智能产品将是今后的一个热点。随着中文抽取系统和文本挖掘技术的发展，整合文本挖掘的商务智能系统将得到长足的发展。

参考文献：

[1] Hale R. Text mining: getting more value from literature resources[J]. Editorial, 2005, 10(6): 377－379.

[2] 薛为民，陆玉昌. 文本挖掘技术研究[J]. 北京联合大学学报，2005(12): 50－63.

[3] 袁军鹏，朱东华，李 毅，等. 文本挖掘技术研究进展[J]. 计算机应用研究，2006(2): 1－4.

[4] 张 燕，寒 枫，楚红涛. 文本挖掘技术简述[C]//2006 年研究综述与技术论坛专刊. [s.l.]: [s.n.], 2006: 138－140.

(下转第 8 页)

3 性能分析

在服务器端的数据调度策略中，假设数据项访问概率服从 Zipf 分布，描述为：

$p_i = \frac{(1/i)^\theta}{\sum_{i=1}^{N}(1/i)^\theta}$，其中 θ 为斜率(Skewness)。经过大量仿真试验表明 TOSA 是近似最优调度策略。在相同长度的数据项长度以及信道带宽情况下，TOSA 算法常用的 GREEDY 算法的比较，如图 3 所示，仿真参数表如表 4 所示。

表 4 仿真参数

参数	含义
广播数据项的个数	10 000
信道数	2～5
数据项的长度	512Bytes

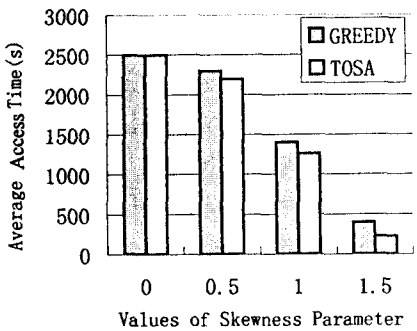

图 3 平均访问时间比较

从图 3 中可知，在同等条件下，随着数据项访问倾斜率的不断增加，TOSA 越能显示其优越性。同时 TOSA 还考虑到了数据项长度的变化，更进一步地适应了将来不断变化的需要。

在移动客户端的索引策略中，假设索引树中，叶子节点数为 n，则

● 在传统索引树中：

平均谐调时间：$TTT = [\log_2 n] + 1$，其中 $[\log_2 n]$ 为索引树的高度。

额外开销：$TSP = n + \sum_{k=1}^{L}[\frac{n}{2^k}]$，$L = [\log_2 n]$。

平均访问时间：$TAT = n + \sum_{k=1}^{L}[\frac{n}{2^k}] + Data + 1$，其中 Data 为等待时间。

● 在 Huffman 索引树中：

平均谐调时间：$HTT = (WPL + n)/n$，WPL 为带权路径长度。

额外开销：$HSP = 2n - 1$

平均访问时间：$HAT = 2n - 1 + Data + 1$，其中 Data 为等待时间。

虽然 Huffman 树的 WPL 不能使用公式计算，由 Huffman 树的 WPL 的最优性可以得出：$HTT \leqslant TTT$，又因为 $HSP \leqslant TSP$，$HAT \leqslant TAT$，故 Huffman 索引树优于传统索引树。从而大大减少了移动客户机的平均电能消耗以及平均访问时间。

4 结束语

文中综合考虑了在移动环境中从服务器端和移动客户端同时提高数据广播效率，尽量减少移动客户端的等待时间和节约电能。在服务器端，文中使用了 TOSA 数据调度策略；在移动客户端，文中使用了 Huffman 树索引技术。并证明了其优越性。

在移动环境中，移动客户端如何同时请求多个数据项，在索引策略中如何减少索引树的高度，如何提高广播数据的安全性等等问题，有待进一步研究。

参考文献：

[1] Shivakumar N, Venkatasubramanian S. Efficient indexing for broadcast based wireless systems[J]. Mobile Network and Application, 1996(12): 433－446.

[2] Lo S－C, Chen L P. Optimal index and data allocation in multiple broadcast channels[C]//In Proceedings of the 16th International Conference on Data Engineering. San Diego, CA, USA: [s. n.], 2000.

[3] Lee G, Yeh M S, Lo S C, et al. A Strategy for efficient access of multiple data items in mobile environments[C]//In Proceedings of 3rd International Conference on Mobile Data Management. Singapore: [s. n.], 2002: 71－78.

[4] Imielinski T, Viswanathan S, Badrinath B R. Energy efficient indexing on air[C]//the 4th International Conference on Extending Database Technology. [s. l.]: [s. n], 1994: 254－258.

[5] Zheng Baihua, Wu Xia, Jin Xing, et al. TOSA: a near－optimal scheduling algorithm for multichannel data broadcast[M]. New York, NY, USA: [s. n.], 2005: 29－37.

(上接第 4 页)

[5] 程红莉，周 宁，肖 爽. 文本驱动的商务智能研究[J]. 情报科学，2007(10)：1525－1529.

[6] 朱德利. SQL Server 2005 数据挖掘与商务智能完全解决方案[M]. 北京：电子工业出版社，2007：10－11.

[7] 陈思睿，张 永，杨志勇. 基于粗糙集的特征选择方法的研究[J]. 计算机工程与应用，2006(21)：159－161.

[8] 梁开健. 基于 DCSSM 的文本特征提取及文本挖掘研究[J]. 自动化技术与应用，2005，24(5)：54－56.

[9] 王珍珍. 关于文本挖掘中文本分类与文本聚类的研究[J]. 计算机与信息技术，2007(6)：55－56.

[10] 黄迎春，李晓晔，邓文新. 文本挖掘技术的研究[J]. 齐齐哈尔大学学报，2006，22(3)：53－55.