文本知识发现:基于信息抽取的文本挖掘

周雪忠　吴朝晖
(浙江大学计算机系　杭州310027)

Knowledge Discovery in Text: A Survey

ZHOU Xue-Zhong　WU Zhao-Hui
(Dept. of Computer, Zhejiang University, Hangzhou 310027)
Email: {zxz; wzh}@cs.zju.edu.cn

Abstract In the general context of Knowledge Discovery, Knowledge Discovery in Text (KDT), which uses Text Mining techniques to extract and induce hidden knowledge from unstructured text data, surges in the data and natural language processing research. KDT is a multi-discipline of Artificial Intelligence, Machine learning, Natural Language Processing and Information Extraction & Information Retrieval. This paper presents an overview of Text Mining with a stressing on its IE (Information Extraction)-based induction and specific sublanguage fields oriented practices.

Keywords Text mining, Text data mining, Information extraction, Knowledge discovery in text, Sublanguage

1. 引言

大家熟知,所谓"数据丰富但知识缺乏"的现状导致了数据挖掘(Data Mining)技术研究的兴起,数据挖掘又称数据库知识发现(Knowledge Discovery in Databases)是从海量的结构化信息中抽取或挖掘隐含信息和知识的重要方法和途径[1]。数据挖掘技术已相当成熟。因为除了结构化的数据之外,在数字化信息中更多地存在大量自由、非结构化或半结构化的文本信息如新闻文章、电子书本、电子图书馆藏、Web页面内容、Email、文档数据库[1]等,显然手工处理需要花费大量的人力物力,并且具有不确定性。所以出现了从文本中发现知识(Knowledge Discovery in Texts)[22]的巨大需求。文本挖掘(Text Mining or Text Data Mining)就是KDT的方法,它从文本集中挖掘和发现隐含的归纳知识如关联知识[21]、时间序列信息,甚至科学文献的创新推断和假设[13,14]等。先于文本挖掘技术发展起来的是计算语言学、信息抽取和信息检索等技术及相关的机器学习,人工智能方法。本文将就文本挖掘的技术发展、文本挖掘的相关基础方法及应用做比较全面的分析和探讨,以明确文本挖掘关键技术和应用趋势。

2. 文本挖掘及其过程

文本知识发现(Knowledge Discovery in Texts)就是从文本集(Texts)中发现和挖掘归纳性的知识如有用的模式、模型、趋势、规则等知识(和KDD中的知识概念一样)[22],这种文本知识发现技术也即文本挖掘技术,是人工智能、机器学习、自然语言处理、数据挖掘及相关自动文本处理[4]如信息抽取、信息检索、文本分类等理论和技术相结合的产物,它得到了越来越多的研究人员的关注[24]。由于大量文本数据库(Text Databases)的存在,文本挖掘成为KDD领域的一个热点研究方向。文本挖掘技术与知识获取(Knowledge Acquisition)概念难以辨别,不过知识获取(KA)注重能应用于一定智能任务的知识的抽取和表示;而文本挖掘(Text Data Mining)包括通过分析已有的文本数据集实现一些创新的知识的(自动)推断,它的价值在于能自动处理比人多得多的文本数据和信息[23]。

文本挖掘过程一般包括文本的预处理,文本模型表示,信息或文本特征属性抽取,文本分类(聚类)或结果集的数据挖掘等步骤。

文本预处理　任何对文本的处理的技术研究都会采用一定的规则对文本进行必要的预处理,如英文文本的去词根处理、例外词处理以及文本的分段、分句等。文本的预处理具有语言相关性(比如中文文本和英文文本的预处理方法就显然不会相同),好的预处理方法是通过分析具体的语言特点发现的。

文本的表示模型　现有的文本处理系统或研究基本上都是基于离散的词表示为基础的文档表示模型,经典的文本表示模型就是基于词布袋法(bag-of-words)的向量空间表示模型(vector space model),或是词的二值表示[16],或是词频加权法[17],或记住词的位置,或采用n元(n-gram)词表示[18]等。TFIDF (Terms Frequency and Inverse Document Frequency)[19]是基于词向量的文本处理的一种传统学习算法。文档集D中的第i个词在文档d中的权重w(i,d)计算方法如下:

$w(i,d) = tf(i,d) \cdot idf(i)$

其中 $tf(i,d)$ = 第i个词在文档d中出现的次数;

$idf(i) = \log(|D|/df(i))$;

$df(i)$ = 词I至少出现一次的文档数。

TFIDF是为了文本检索而采用的文本分类算法。它的基本假设就是词在文本中出现频率是其权重评价的唯一标准。一般与文本预处理中的例外词处理相结合能达到相当好的效果,但其缺点是很明显的。因此有些研究人员对基于词的文本表示进行了一定程度的改进,提出了基于概念的文本表示方式,即用概念向量代替词向量。一个概念可以由多个词术语组成,这种概念文本表示考虑了一定的词语义和词之间的一定的依赖关系,提高了文本处理的精确度,但同时文本处理的复

周雪忠　博士,从事数据挖掘、文本挖掘、分布式数据库等研究;吴朝晖　教授,博士生导师,从事数据挖掘、工作流、分布式数据库、生物认证技术等研究。

杂度将显著增加。潜在语义索引(Latent Semantic Indexing)是基于潜在语义分析(Latent Semantic Analysis)的术语-文本表示方法[20],它应用于信息检索,以实现一定程度的基于词的语义检索。

特征选取和信息抽取 采用一定的文本表示模型对文本进行建模后,还有根据不同的目标采用特征选取方法来降低维度(对于基于向量空间的文本表示模型)或去除一些噪音数据,具体的特征选取方法一般都以词汇分析为基础,子特征选取方法在机器学习中有过滤和包裹器两种常用的方法,在文[10]中采用基于特征积分为基础的简单特征抽取方法。

信息抽取在非结构化的自然语言文本中定位相应的结构化数据单元,从而使自由文本数据成为相应的结构化数据。目前流行的信息抽取系统都是基于机器学习的自动或半自动系统。文本挖掘从非结构化的自由自然语言文本集中发现隐含知识,要在掌握一定的文本"内容脉络"的基础上才能真正实现,所以把信息抽取作为文本挖掘的前期步骤,基于信息抽取的文本挖掘系统是研究趋势所在[25]。把以演绎推理为基础的信息抽取技术和以归纳推理为主导的文本挖掘技术相结合能更有效地实现文本知识发现,信息抽取作为一种重要的文本处理技术将在第三节中作详细的讨论。

信息检索和文本分类 文本聚类(Clustering)和分类(Text Classification or Categorization)是一般的浅层文本处理的技术,有些专家[15]认为文本分类不应该归结为文本挖掘的目标之一,由于大量文本数据组织、开发和检索的需要,文本分类在文本信息处理的过程中起着至关重要的作用,高效的信息检索(Information Retrieval)多以文本聚类和分类为基础。文本分类是信息检索的基础。

信息检索(Information Retrieval)以帮助用户更有效地搜索信息为目的。它不能从现有的文本信息中发现新的知识[15],信息检索一般研究文本的关键字向量空间表示如布袋法(bag-of-words)即采用一定的检索算法来提高检索效率。文本挖掘技术可以为更精确的内容语义检索提供服务。

数据挖掘 文本挖掘的前几个部件的处理结果是结构化的数据。分析和处理结构化的数据,发现知识是数据挖掘的目标,因此文本挖掘的最后步骤就是基于结构化数据的知识挖掘。可以根据不同的文本挖掘目标如关联知识发现、趋势预测、序列知识发现等采用不用的数据挖掘算法[1]。

3. 文本挖掘关键技术分析

由以上的文本挖掘的处理流程分析可见,文本挖掘的关键技术是信息抽取、文本分类和机器学习等。下面对文本挖掘中机器学习、信息抽取等技术的研究情况进行分析和总结,并阐述了次语言(sublanguage)效应在文本挖掘应用中的重要性。

1. 信息抽取

信息抽取从自然语言文本中定位目标数据单元,把非结构化的自由文本转化成符合应用要求的结构化数据,即抽取自由文本的数据填充预先定义的结构化模板[26]。信息抽取是一种浅层文本理解技术,它涉及计算语言学和自然语言处理,早期的系统面向具体的应用领域以减少运算复杂度[26],随着信息抽取技术研究的深入,机器学习被采用来实现自适应性和领域无关性,Dayne Freitag 和 Andrew McCallum 采用隐马尔科夫模型(HMM)于文本信息的建模和自动抽取[7],同时为了减少学习样本文本的手工标识的劳动量,Ellen Riloff (1996)等人学习从未经标识的文本中自动抽取和发现信息模式。由于信息抽取的特点使得基于规则学习的系统成为主流[28],对自然语言的理解程度和领域知识结合紧密程度决定了信息抽取的能力。由于数据库仓库技术和超大型数据库技术的出现及实际计算机存储中大量文档数据的存在,信息抽取作为非结构化数据和数据库之间的"桥梁"处理技术,对信息知识化的实现十分关键和重要。

信息抽取(IE)不同于信息检索(IR),两者的目标和实现方法都不一样,信息抽取的目标是从单个文档中抽取相关的数据、信息,而信息检索从文档集合中检索相关的文档;从实现方法来分析,信息抽取是从自然语言处理和计算语言学的基于规则的系统发展起来的,而信息检索更多的是和信息理论、概率统计理论有关。信息检索对文本数据的建模是非常简单的,词袋表示法(bag-of-words)或其改进表示法是信息检索文本表示的最基本方法,而信息抽取需要对文本进行一定的语义分析。总之,信息抽取和信息检索具有一定的互补性,Robert Gaizauskas 及其合作者结合这两种方法从 Web 上搜索信息并进行信息结构化[27],它们的不同特性反映了演绎和归纳方法的差异。

MUC(Message Understanding Conference)是推动信息抽取技术研究的关键因素,MUC 通过给出共同的信息集检测不同系统对新的、未知信息的处理情况,来理解和比较不同的信息理解(Message Understanding)系统的性能差异,现在由 DARPA 资助,它的会议目标已经转移到了制定评价信息抽取和信息理解的标准方案上来,到目前为止已经开了七次会议(MUC-7)。

2. 机器学习及统计学习理论的运用

文本挖掘是多学科交叉的新兴研究领域,机器学习一直是信息智能化处理的关键理论和技术,是文本挖掘的基础。

一个自动化的文本挖掘系统要求具有自适应的功能,所以在现有的文本挖掘研究和系统中普遍采用了机器学习的方法,各种机器学习方法渗透文本挖掘的特征向量选取、分类、聚类,直至最终的知识发现。传统的 Naive Bayesian 分类器及 Lam,W. 等人的 Bayesian Network Induction 方法[17]、k-最近邻域(k-nearest neighbor)[32]、概念学习(Inductive Concept learning)、增强学习(Reinforcement learning)[31]等各种机器学习方法等广泛地运用在自动文本处理中。总之,除了早期的一些实验系统没有采用学习方法之外,机器学习在文本处理和文本挖掘中得到了广泛的运用,是文本处理取得一定成果的关键所在。

文本挖掘的对象自然语言文本是离散的词串流的集合,词的分布具有统计规律性和高维特征性,所以统计学习方法特别是支持向量机[8]将是文本挖掘的学习方法新的研究方向。统计学习的基本观点是把机器学习问题看作利用有限数量的观测来寻找待求的依赖关系的问题[9]。它表示的机器学习的函数估计模型如图2所示。

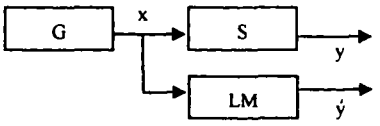

图2 统计机器学习函数估计模型

在模型表示的机器学习过程中,学习机器 LM 观察数据对(x,y)(训练集)。在训练之后,学习机器必须对任意输入 x 给出输出 $\hat{y}$,学习的目标是能够给出输出 $\hat{y}$,使之接近训练器的响应 y。其中

(1)产生器(G),产生随机向量 $x \in R^n$,它们是从固定但未

知的概率分布函数 F(x)中独立抽取的。

(2)训练器(S),对每个输入向量 x 返回一个输出值 y,产生输出的根据是同样固定但未知的条件分布函数 F(y/x)。

(3)学习机器(LM),它能够实现一定的函数集 f(x, α),α∈Λ,其中 Λ 是参数集合。

学习的问题就是从给定的函数集 f(x, α), α∈Λ 中选择出能够最好地逼近训练器响应的函数。在统计学习模型中,贝叶斯方法得到了相当广泛的应用,如贝叶斯分类器运用于文本分类,但它对强先验信息的依赖性,使得其不具备良好的可推广性。支持向量机(SVM)是统计学习的最新研究热点。SVM 的基本实现思想是:通过某种事先选择的非线性影射把输入向量 x 映射到一个高维特征空间 Z,在这个空间中构造最优分类超平面。也就是 SVM 采用输入向量的非线性变换;在特征空间中,在线性决策规则集合上按照正规超平面权值的模构造一个结构;然后选择结构中最好的元素和这个元素中最好的函数,以达到最小化错误率的界目标等步骤,实现了结构风险最小化归纳原则。SVM 能有效地应用于模式识别[5]、特征选取[6]、自动文本处理如文本归档[30]等任务。

3. 文本挖掘中的次语言(sublanguage)

自然语言处理(Natural Language Processing)一直是人工智能研究领域的一个难点,在开放的语言和文字环境中,由于自然语言的丰富、多义的语法、语义、语用层次信息,使得基于自然语言彻底理解基础上的自然语言处理或其相关技术研究举步维艰。

纽约大学(NYU)的 Linguistic String Project 从1965年就开始自然语言处理的研究,到1975年提出了次语言(sublanguage)方法论,在1986年 Ralph Grishman(负责纽约大学 Proteus Project 研究)发表文章阐述自然语言处理中的次语言(sublanguage)现象[34],次语言指的是特定的主题或领域中表现的特殊形式的自然语言,如天气预报、飞机维修指南、有关病理的科学文章、医院的 X 光线报告和房地产广告等[33]。这样从通用的语言环境到具体的应用领域环境语法、语义范畴的缩小,能大大地提高自动语言理解或处理的效率和正确率。Ralph Grishman(2001)还就利用次语言现象提出了高效的信息抽取方法——自适应信息抽取[33]。

另外,实际的工程项目与应用领域相结合,利用一定的领域背景知识,是自然语言处理及其相关领域的研究趋势所在。文本挖掘技术处理的对象是自然语言数据,所以充分地利用次语言效应能更好地实现系统的目标。

4. 文本挖掘技术应用及趋势

随着文献资料的积累和商业文档的大量增加,不管是在科学研究领域还是商业领域都有文本挖掘技术的用武之地。文本挖掘技术具有广泛的应用领域和巨大的社会需求,IBM 针对商业机构的大量在线有价值文档开发了文本挖掘应用平台 TextMiner。他们认为许多公司都拥有大量且在不断增加的在线文档如:包含产品或服务信息的客户的 email、公司内部的有价值的技术报告和工作文档、有关商业环境和商业竞争对手情况的新闻资料等,TextMiner 具有从文本中抽取关键信息、按主题组织文档、从文当集中获取突出的主题、对相关文档的强有力、灵活的检索等功能[35]。在1996年 Feldman 等人开发一个实验文本知识发现系统,从路透社(Reuter)的新闻资料库中发现关联知识[36];赫尔辛基大学的 Helena Ahonen(1997)等人采用数据挖掘的方法从文本中发现有意义的习语及同现术语(co-occurring terms)[3];在医学领域的应用的成功例子是 Swanson and Smalheiser (1997)的工作[14],通过半自动地分析生物医学文献的题名,获得了一个完全创新、有价值的医学猜测:偏头痛与人体镁元素的缺乏有关。简言之,他们的方法是通过确定一组报告变量 A 和 B 之间关系的文献和一组 B 和 C 之间关系的文献,并且确定在现有的文献中不存在描述 A 和 C 之间的关系,这样 A 和 C 之间的关系就是潜在的新知识。由于医学领域的次语言(sublanguage)效应及医学领域文献资源的重要作用,现今多项重要的文本挖掘项目都面向医学,其中两个著名的项目是:

1. 加州伯克利分校信息管理和系统系的 Hearst M. A. 主持的 LIDNI Project

LIDNI 是一个辅助专家系统,它的目标是提供交互式界面辅助研究工作者从大量的文献资料中发现重要的信息或知识。该系统主要有两种类型的工具:a)对文本集上一系列查询及其相关操作的支持工具;b)对结果文档集中概念的相关性的统计和可视化的支持。系统的整个操作是一个专家用户与计算机互动的过程,系统适时的给专家建议和意见来执行下一步,当然专家可以采用或忽略这些信息。LINDI 的一个研究重点之一是运用于分子生物学中现代新兴基因工程中的基因序列串功能发现的自动化。LINDI 对生物医学文献进行文本挖掘,形成一定的假设或推理以指导基因与疾病的对应关系的研究。

2. 纽约大学的 Linguistic String Project 70年代后的主要研究方向 MLP(Medical language Process),MLP 把叙述性的临床医学文档转换成(结构化)的语义表示,在此基础上可进行各种有意义的医学归纳知识发现和研究。

同时,Internet 上丰富的非结构化文档资源也是文本挖掘的重要的目标,随着电子商务及其各种商业服务、各种数字化信息的 Web 化,Internet 上累积了大量的文本数据,如新闻、招聘信息、产品广告及大量的在线科学文献等。随之出现了各种高效的搜索引擎如 Google、Yahoo、AltaVista、Excite 等,它们的关键技术之一就是文本的表示和文本归档。而且由于 Internet 上大量的信息具有巨大的潜在商业价值,近来的研究热点 Web Mining 就是由此应运而生的。

5. 中文文本挖掘技术特点

中文信息处理一直是我国计算机专家或语言学者研究的一个热点。目前在字处理如中文输入法、汉字编码、汉字识别等方面取得了相当大的成绩。同时在机器翻译,语料库建设,中文分词和词性标注,中文的语法、语义分析等方面都有一定的建树,不过在以上各个领域中都存在几乎难以突破的障碍:语言的歧义和多义现象。彻底的语法,语义分析的目标是自然语言理解,而文本挖掘可以说是避开以演绎推理为基础的自然语言理解这个一直存在于计算机领域的难题,另辟蹊径的有效方法,文本挖掘的思想是通过机器学习的方法,在对自然语言载体数据部分分析的基础上以归纳推理发现自然语言文本中特定的隐含目标知识,而不强求对自然语言文本的完全透彻理解。

对中文信息的文本挖掘要求原先的连续"字符流"数据被分割成连续的"词流"数据,因为词流数据是文本挖掘的最基本条件(英文文本数据或其他拉丁语言数据以空格为词界的特性使其自动满足了这个要求)。所以对中文文本的挖掘要首先实现其自动分词的功能,而必要的词性标注可以使文本挖

掘具备更高的精确度。由于一般文本挖掘的目标是面向应用的，因此领域启发式知识或知识库的使用能提高分词和标注的精度，已达到文本挖掘所需的要求。

结论和展望 文本挖掘技术是一种综合机器学习、数据挖掘、自然语言处理和其他各种文本处理方法如信息抽取和文本分类的文本自动处理技术。文本挖掘的目标是发现隐含的归纳知识如关联知识、序列知识及完全创新的科学推断和假设等。基于信息抽取和数据挖掘算法的文本挖掘技术是文本知识发现的技术研究趋势，与具体的次语言(sublanguage)领域相结合，从而发现和提炼更具价值的应用领域知识模式是文本挖掘的应用趋势。

参 考 文 献

1 Han Jiawei. Micheline Kamber (2000). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers. All rights reserved

2 Tukey J W. Exploratory Data Analysis, Addison-Wesley, 1977

3 Ahonen H, Heinonen O, et al. Mining in the Phrasal Frontier. PKDD-97, Trondheim, Norway, June 1997

4 Salton G. Automatic text processing. Reading, MA: Addison-Welsley, 1989

5 Burges C. A tutorial on SVM for pattern recognition. Data Mining and Knowledge Discorvery

6 Chapelle O, Vapnik V. Model selection for SVM. Advances in Neural Information Processing Systems, MIT Press, 2000

7 Freitag D, McCallum A. Information Extraction with HMMs and Shrinkage. In: Proc. Workshop on ML and IE, AAAI-99, 1999

8 Cortes C, Vapnik V. Support-vector networks. Machine Learning. Nov. 1995

9 Vapnik V N. The Nature of Statistical Learning Theory. Springer. New York, 1995

10 Mladenic D. Feature subset selection in text -learning. In: Proc. of the 10th European Conf. on Machine Learning ECML98, 1998

11 Sager N. Sublanguage: Linguistic Phenomenon, Computational Tool. In [Grishman and Kittredge 1986].

12 Swanson D R. Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Retrieval, 1987, 38 (4): 228～233

13 Swanson D R, Smalheiser N R. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neuroscience Research Communications, 1994, 15: 1～9

14 Swanson D R, Smalheiser N R. An interactive system for finding complementary Literatures: a stimulus to scientific discovery. Artificial Intelligence, 1997, 91: 183～203

15 Hearst M A. Untangling text data mining. In: Proc. of ACL '99: the 37th Annual Meeting of the Association for Computational Linguistics, New Brunswick, NJ: The Association for Computational Linguistics

16 Armstrong R, Freitag D, Joachims T, Mitchell T. WebWatcher: A Learning Apprentice for the World Wide Web. In: Proc. of the AAAI 1995 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford

17 Lam W, Low K F, Ho C Y. Using Bayesian Network Induction Approach for Text Categorization. In: Proc. of the 15th Intl. Joint Conf. on Artificial Intelligence IJCAI97 (pp. 745～750)

18 Sorensen H, McElligott M. PSUN A Profiling System for Usenet News. In: Proc. of CIKM'95 Intelligent Information Agents Workshop, Baltimore

19 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988, 24: 513～523

20 Papadimitriou H, Raghavan P, Tamaki H, Vempala S. Latent Semantic Indexing: A probablistic analysis. In: Proc. of Symposium on Principles of Database Systems (PODS). ACM Press, 1998

21 Feldman R, Hirsh H. Finding Associations in Collections of Text. In: Machine Learning and Data Mining, R. S. Michalski, I. Bratko, M. Kubat, eds. John Wiley & Sons, NY 1997

22 Kodratoff Y. Knowledge Discovery in Texts: A Definition, and Applications. Proc. ISMIS'99, Warsaw, June 1999a

23 Stephen, Potter. A survey of knowledge acquisition from natural language. Artificial Intelligence Applications Institute, Division of Informatics, University of Edinburgh

24 Feldman R. In: Proc. of the Sixteenth Intl. Joint Conf. on Artificial Intelligence (IJCAI-99) Workshop on Text Mining: Foundations, Techniques and Applications. 1999

25 Nahm U Y, Mooney R J. Text Mining with Information Extraction To appear in the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002

26 Cowie J, Lehnert W. Information Extraction. Communications of the ACM, 1996, 39(1): 80～91

27 Gaizauskas R, Robertson A M. Coupling IR and IE: A New Text Technology for Gathering Information from the Web. In: Proc. of RIAO 97: Computer-Assisted Information Searching on the Internet, Montreal, Canada, 1997. 356～370

28 Califf M E, Mooney R J. Relational learning of pattern-match rules for information extraction. In: Working Papers of ACL-97 Workshop on Natural Language Learning, 1997

29 Holt J D, Chung S M. Efficient Mining of Association Rules in Text Databases. 1999 ACM 1-581

30 Taira H, Haruno M. Feature Selection in SVM Text Categorization. In AAAI-99

31 Rennie J, McCallum A K. Using Reinforcement Learning to Spider the Web Efficiently. ICML-99 Workshop: Machine learning in Text Data Analysis

32 Yang Y, Pedersen J. A Comparative Study on Feature Selection in Text Categorization. ICML-97 (pp. 412-420)

33 Grishman R. (2001). Adaptive Information Extraction and Sublanguage Analysis. IJCAI-2001 Workshop on Adaptive Text Extraction and Mining

34 Grishman R, Kittredge R. Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Lawrence Erlbaum Assoc., Hillsdale, NJ, 1986

35 Text Mining Technology Turning Information Into Knowledge. A White Paper from IBM, 1998

36 Feldman R, Dagan I, Kloegsen W. Efficient Algorithm for Mining and Manipulating Associations in Texts. 13th European Meeting on Cybernetics and Research, 1996

(上接第48页)

环境：733MHz 主频、1G 内存的 Itanium 工作站，操作系统为 Red Hat Linux 6.2。使用 edge profiling 的编译器 ORC 和没有采用 edge profiling 的 ORC 测试结果表示，edge profiling 对所有程序的性能都有一定的提高，性能提高百分比为：2%(vpr)到20%(gap)，平均达到5%。

参 考 文 献

1 Albert G. A Transparent Method for Correlating Profiles with Source Programs. 2nd Workshop on Feedback Directed Optimization, Haifa, Israel, Nov. 1999

2 Anderson J, Berc L, et al. M. Continuous Profiling: Where Have All the Cycles Gone? ACM Trans. on Computer Systems, 1997, 15(4): 357～390

3 Ball T, Larus J. Efficient Path Profiling. In: Proc. 29thAnnual IEEE/ACM Intl. Symp. on Microarchitecture, Dec. 1996. 46～57

4 Conte T M, Menezes K N, Hirsch M A. Accurate and practical profile-driven compilation using the profile buffer. In: Proc. 29th Annual International Symposium on Microarchitecture, Dec. 1996. 36～45

5 Young C, Smith M. Better Global Scheduling Using Path Profile. In: Proc. 31st Annual International Symposium on Microarchitecture, Dec. 1998. 115～123