基于伸缩窗口和等级调整的 SNM 改进方法*

陈 爽[1,2,3], 刁兴春[1,2], 宋金玉[1], 曹建军[2], 丁晨路[2]
(1. 解放军理工大学 指挥信息系统学院, 南京 210007; 2. 总参第六十三研究所, 南京 210007; 3. 吉林陆军预备役步兵第 47 师, 吉林 吉林 132000)

摘 要: 对基本邻近排序算法(basic sorted-neighborhood method,SNM)进行分析,指出其不足,提出了 SNM 算法的一种改进方法。采用变步长伸缩窗口,动态改变检测窗口大小,避免漏配问题,并减少不必要的比较。采用动态调整等级法,根据记录相似度调整字段等级,并通过等级法将字段等级转换为权重,解决了人为赋予固定权重主观性强、不准确的问题。通过对实际系统中的数据进行测试,验证了方法的有效性和优越性。同时,这两种方法适用于大多数基于排序—合并的相似重复记录检测方法,提高了相应方法的效率和准确度。

关键词: 数据清洗; 相似重复记录; 变步长伸缩窗口; 动态调整等级; 基本邻近排序算法
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2013)09-2736-04
doi:10.3969/j.issn.1001-3695.2013.09.044

Amelioration method of SNM based on flexible window and ranking adjusting

CHEN Shuang[1,2,3], DIAO Xing-chun[1,2], SONG Jin-yu[1], CAO Jian-jun[2], DING Chen-lu[2]
(1. *Institute of Command Information System*, *PLA University of Science & Technology*, *Nanjing* 210007, *China*; 2. *The* 63*rd Research Institute of PLA General Staff Headquarters*, *Nanjing* 210007, *China*; 3. *Jilin Army Reservist Infantry* 47*th Division*, *Jilin Jilin* 132000, *China*)

Abstract: This paper introduced the basic SNM and analyzed its deficiency, and put forward a amelioration method of SNM. To avoid missing comparison and reduce excrescent comparison, adopting changing flexible window method, which changed the size of window dynamically. Solved the problem of subjectivity and inaccurate with fixed field weight offer by man-made by using dynamic adjusting ranking method, and according the record similarity to adjust the rank of field. And it transferred the ranking of field to weight by rank-based weights method. The experiments on the data from the information system show the effectiveness and advantage of the proposed method. At the same time, these two methods are the same with other approximately duplicate records examine methods which based on sorted-merge, advancing the efficiency and accuracy.

Key words: data cleaning; approximately duplicate records; changing step flexible window; dynamic adjusting ranking; SNM algorithm

0 引言

数据质量问题能够较大程度地影响决策支持的可靠性,但因数据规模庞大,并且数据来自多数据源,很多现实中的数据难以保证其质量。因此,数据挖掘技术必须考虑重复数据、缺失数据、异常数据、逻辑错误数据和不一致数据等数据质量问题[1]。其中,重复数据比较普遍,也是容易发生的问题。当对几个数据源的数据进行集成时,多条记录代表同一实体的现象经常存在,这些记录即为重复记录。在信息管理系统或数据仓库中,重复记录导致数据冗余,并且浪费网络带宽和磁盘存储空间。检测并清除相似重复记录,能保证信息管理系统和数据仓库的可用性,提高工作效率,为决策支持打好基础。

相似重复记录问题很早就得到科研工作者的关注,很多技术和清洗算法被提出,并能有效地对相似重复记录进行清洗。文献[2,3]采用编辑距离、Cosine 等文本相似函数对记录进行重复检测,这些方法基于代表同一实体记录的大部分字段内容相似或相同的思想。其主要缺点是难以识别同一对象的多种表达形式,降低了检测精度,如缩写、单词顺序等。文献[4,5]提出了分割法,将某一字符串分割成几个组成部分来处理,在一定程度上能够解决同一对象多种表示形式的问题。同时,其他一些方法也被用于重复记录检测。文献[6]利用依赖图的概念,计算数据表中的关键属性,根据关键属性值将记录集划分为小记录集,在每个小记录集中进行相似重复记录检测。重复记录合并时,保留一条主记录,其他重复记录的信息合并到主记录后被删除。该方法适用于大数据量的重复记录检测,提高了重复记录检测的效率和准确度。文献[7]介绍了复杂数据实体识别的概念和应用,分别就 XML 数据、图数据和复杂网络上实体识别技术进行了讨论,介绍了相应的技术原理。文献[8]提出了基于内码序值聚类的相似重复记录检测方法,采用字符内码序值对记录进行排序,聚类相似重复记录,把大数据集聚集成多个小数据集,在小数据集中分别检测相似重复记录。

SNM 算法[9]是较早被提出的排序—合并方法。因为其思

收稿日期: 2012-12-26; **修回日期**: 2013-03-04 **基金项目**: 中国博士后科学基金特别资助项目(201003797);江苏省博士后科研资助计划项目(0901014B);解放军理工大学预研基金项目(20110604)

作者简介:陈爽(1985-),男,硕士研究生,主要研究方向为数据工程(shuangchen1985@126.com);刁兴春(1964-),男,研究员,博导,主要研究方向为数据质量管理、数据库应用;宋金玉(1964-),副教授,硕导,主要研究方向为数据库技术;曹建军(1975-),博士(后),主要研究方向为数据质量控制与数据清洗、进化算法;丁晨路(1986-),女,工程师,主要研究方向为数据库技术、计算机应用.

想简单、效果明显、易于实现,得到了广泛应用。排序—合并方法的基本思想是:选定排序关键字,根据关键字对记录进行排序,再比较邻近记录,判断其是否为相似重复记录。

SNM 算法采用滑动窗口,可以减少比较次数,大大提高了检测效率,但其存在两个主要缺点:

a)比较窗口大小难以确定。窗口过大将导致没有必要的记录比较,增加时间消耗;窗口过小就会出现漏配现象,降低检测精度。

b)字段权重固定不变,难以保证准确性。字段的权重多为人为给定,并且字段权重在检测过程中固定不变,存在一定的主观性。

针对 SNM 方法的不足,本文提出了基于 SNM 的改进方法,采用变步长伸缩窗口,确定相似重复记录检测中合适的窗口大小,在避免漏配的同时,减少不必要的记录比较;采用动态调整等级方法,调整记录中各字段的等级,并用等级法[10]将字段等级转换成字段权重,提高检测精度。改进方法在相似重复记录检测的准确度和效率上都进行了优化。实验结果表明,该方法是有效、可行的。

1　SNM 算法基本思想

a)选取排序关键字。选取记录中的关键字段或属性值字符串,作为记录排序的关键字。

b)记录排序。根据选定的排序关键字对整个数据集进行排序,使相似重复记录在位置上更接近,为下一步的重复检测作准备。

c)相似重复记录检测。在排序后的数据集上滑动一个大小为 W 的固定窗口,窗口内的第一条记录与其余的 $W-1$ 条记录进行比较。比较过程中,一般通过相似度比较算法计算字段相似度,再利用字段权重加权平均得到记录相似度。将记录相似度与记录相似度阈值比较,判定两条记录是否为相似重复记录。比较结束后,第一条记录从窗口取出,第 $W+1$ 条记录进入窗口。此时窗口中的第一条记录与其他的 $W-1$ 条记录进行比较。如此反复进行,直到数据集中最后一条记录比较完毕。最后,得到数据集中所有的相似重复记录。图 1 为 SNM 算法示意图。

2　改进方法

2.1　变步长伸缩窗口

传统的 SNM 算法在比较记录是否相似时,只对窗口内的 W 条记录进行两两比较。该方式节省了比较时间,特别是数据规模较大时将大大提高比较效率。但窗口 W 的大小难以确定,只能通过实验人为估计。本文采用变步长伸缩窗口,根据记录相似度值确定窗口是扩大还是缩小,以及扩大或缩小的比例。

为讨论方便,首先给出文中使用的相关定义。

定义 1　窗口扩大阈值。当记录相似度大于该值时扩大比较窗口大小。

定义 2　窗口缩小阈值。当记录相似度小于该值时缩小比较窗口大小。

定义 3　窗口扩大/缩小步长。窗口扩大或缩小的大小称为窗口扩大/缩小步长。

记录在比较前一般先按关键字排序,排序后相邻的记录很有可能是相似重复记录。设窗口的初始值为 W,当窗口内的第一条记录 R_1 与最后一条记录 R_w 的相似度 $\mathrm{sim}R_{1+w}$ 大于窗口扩大阈值 sim_l 时,此时 R_1 与窗外下一条记录 R_{w+1} 相似的可能性较大。所以应该扩大窗口,使记录 R_1 与更多的记录进行比较。同时,根据记录相似度比例确定合适的扩大步长。窗口扩大步长为

$$W_{\mathrm{step}} = W \times \frac{\mathrm{sim}R_{1+w} - \mathrm{sim}_l}{1 - \mathrm{sim}_l} \tag{1}$$

扩大后的窗口为 $W_l = W + W_{\mathrm{step}}$。同时,为减少不必要的比较,窗口最大值不超过排序字段中具有相同属性值的记录数的最大值。

同理,当 $\mathrm{sim}R_{1+w}$ 小于窗口缩小阈值 sim_s 时,记录对的相似度较低,应缩小窗口,减少不必要的比较。此时的缩小步长为

$$W'_{\mathrm{step}} = W \times \frac{\mathrm{sim}_s - \mathrm{sim}R_{1+w}}{\mathrm{sim}_s} \tag{2}$$

缩小后的窗口为 $W_s = W - W'_{\mathrm{step}}$。为避免漏配,窗口最小值不低于排序字段中具有相同属性值的记录数的最小值。

2.2　动态调整等级法

进行相似重复记录检测时,根据字段的重要程度,为字段赋予不同的权重。计算两条记录的相似度时,首先计算对应字段的相似度,并加权平均得到记录相似度。合理地为各字段分配权重是关键而重要的步骤,多数方法通过领域专家打分方式为字段赋予固定权重。文献[11]提出了模糊综合评判的属性重要性权值确定方法,解决了专家打分方法确定属性权值主观性强的问题。本文采用动态调整等级法,根据记录相似度动态合理地调整字段等级,并通过等级法将字段等级转换为权重。

下面先给出有关定义。

定义 4　相似重复记录相似度向量集合 M。当前窗口中,相似度接近 1 的记录相似度向量集合,该记录对表示同一对象。

定义 5　非相似重复记录相似度向量集合 N。当前窗口中,相似度接近 0 的记录相似度向量集合,该记录对表示不同对象。

动态调整等级法基于以下思想:a)若两条记录的相似度较高,接近 1,说明两条记录为相似重复记录,因为记录相似度由字段相似度加权平均得来,所以较高的等级就该赋予字段相似度大的字段,较低的等级赋予字段相似度小的字段;b)若两条记录的相似度较低,接近 0,说明两条记录为非重复记录,此时,较高的等级应该赋予字段相似度小的字段,较低的等级赋予字段相似度大的字段;c)若两条记录相似度不接近 0 或 1,各字段的等级不变。

根据以上分析,动态调整等级算法如算法 1 所示。算法首先分别将当前窗口内符合条件的记录相似度向量分到集合 M 和 N 内(1 ~ 4 行);根据集合 M 中的向量计算相似度向量 $S_M = \langle s_{m_1}, s_{m_2}, \cdots, s_{m_n} \rangle$,$n$ 为字段个数(5 ~ 11 行);根据集合 N 中的向量计算相似度向量 $S_N = \langle s_{n_1}, s_{n_2}, \cdots, s_{n_n} \rangle$(12 ~ 18 行),调整后的相似度向量为 $FS = \frac{|M|}{|M|+|N|} \times S_M + \frac{|N|}{|M|+|N|} \times S_N$。根据相似度向量 FS 调整字段等级,并用等级法将字段等级转

换为字段权重。

算法 1　动态调整等级算法 AdjustRanking

输入:当前窗口 W 内记录。

输出:字段权重($fw_1, fw_2, \cdots, fw_n$)。

1. for $\forall i, j = 1 \sim W, i \neq j$
2. if $|\text{recordsim}_{ij} - 1| < \delta$ //δ 为一定小的值
$\langle \text{sim}_{ij1}, \text{sim}_{ij2}, \cdots, \text{sim}_{ijn} \rangle \in M$
3. if $|\text{recordsim}_{ij} - 0| < \delta'$ //δ' 为一定小的值
$\langle \text{sim}_{ij1}, \text{sim}_{ij2}, \cdots, \text{sim}_{ijn} \rangle \in N$
4. end for
5. for $i = 1 \sim n$
6. $p_i = \sum_{j=1}^{|M|} v_{ij}$ //v_{ij} 为 M 中第 j 个向量第 i 个字段的相似度
7. end for
8. $S_1 = \sum_{i=1}^{n} p_i$
9. for $i = 1 \sim n$
10. $s_{mi} = p_i / S_1$
11. end for
12. for $i = 1 \sim n$
13. $q_i = \sum_{j=1}^{|N|} (1 - v_{ij})$ //v_{ij} 为 N 中第 j 个向量第 i 个字段的相似度
14. end for
15. $S_2 = \sum_{i=1}^{n} q_i$
16. for $i = 1 \sim n$
17. $s_{ni} = q_i / S_2$
18. end for
19. $fs_i = \frac{|M|}{|M| + |N|} \times s_{mi} + \frac{|N|}{|M| + |N|} \times s_{ni}$
20. Sort($fs_1, fs_2, \cdots, fs_n$) //对字段按相似度由大到小排序
21. Adjust($r_1, r_2, \cdots, r_n$)
//根据排序后的字段相似度调整各字段等级
22. 由 RC 等级法生成字段权重($fw_1, fw_2, \cdots, fw_n$)

2.3　相似重复记录检测改进方法

本文主要采用变步长伸缩窗口和动态调整等级法实现相似重复记录的检测,提出了 SNM 的改进方法。其具体内容如下:

a)记录排序。选定关键字段或属性组合,作为记录排序关键字。根据确定的关键字,对记录进行排序,使相似重复记录在位置上靠近。

b)噪声字段删除。记录的相似度由字段相似度加权平均得到,而某些字段对记录相似度的计算是无用的,甚至影响处理的效率和准确度,此类字段称为噪声字段。例如人员信息表中的“性别”字段,此字段只有男、女两个类别,因此该字段的相似度只有较高或较低两种情况。这将导致记录相似度增高或降低,影响记录对相似与否的判断。并且其对动态调整等级方法的影响更大,个别字段的相似度过大或过小,将使得该字段等级过高或过低,这与实际是不符的。因此,在计算记录相似度时有必要去掉这类噪声字段。

c)变步长伸缩窗口。在计算记录相似度时,根据记录相似度大小,按比例调整比较窗口的大小。当窗口内记录与最后一条记录的相似度大于某一阈值时,扩大比较窗口,增加比较的记录数;反之,缩小比较窗口。

d)动态调整等级。在记录比较过程中,如果记录相似度接近 1 或 0,按动态调整等级法对字段等级进行调整,并运用等级法将等级转换为权重。

e)记录相似度计算。本文采用编辑距离实现字段相似度的计算,并对字段相似度加权平均得到记录相似度。同时,采用传递闭包算法[12],提高比较效率,减少比较次数。

3　实验与分析

为验证本文提出方法的有效性,将改进方法与 SNM 方法进行对比实验。实验环境:Core™2Duo CPU,2.4 GHz,1 GB 内存,Windows XP 操作系统,数据库软件为 Oracle9g,程序用 Java 语言编写。

3.1　评价指标

为检验方法的有效性和优越性,主要采用常用的查全率、查准率及调和均值 F 值三个指标对方法效果进行测试。

3.2　实验数据集

实验数据采用某信息管理系统中的人员信息表,共包含了 26 254 条记录,记录了人员的身份证号码、姓名、所属单位编号等信息。实验中分别对其中的 500 条、1 000 条、5 000 条、1 万条、1.5 万条、2 万条、2.5 万条记录进行测试,并借鉴其他文献中的方法,为每个数据集生成 2% 的相似重复记录。本文分两组进行实验,分别为实验 1 和 2。实验 1 中记录相似度阈值设为 0.75,初始窗口为 20;实验 2 中记录相似度阈值设为 0.75,初始窗口为 50。两种方法对数据集各字段的初始等级设置相同。

3.3　实验结果与分析

SNM 和 SNM 改进方法在相同环境下进行实验,并对同一数据集进行测试。实验 1 的结果如图 2 ~ 4 所示。

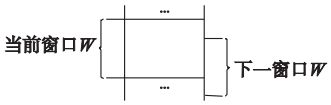

图1　SNM算法示意图

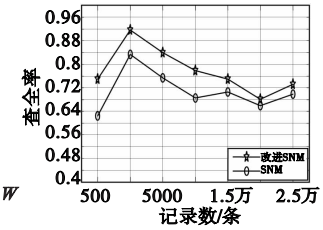

图2　SNM与改进SNM方法查全率对比

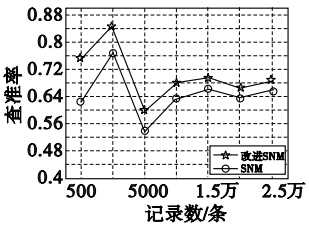

图3　SNM与改进SNM方法查准率对比

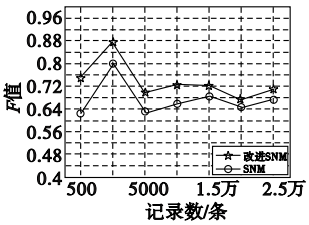

图4　SNM与改进SNM方法F值对比

从图 2 中可以看出,SNM 改进方法的查全率高于 SNM 方法,提高了相似重复记录检测能力,解决了因窗口过小出现漏配和字段权重分配不合理相似度计算精度不够,导致部分重复记录不能识别出来的问题。

从图 3 中可以看出,SNM 改进方法提高了相似重复记录检测的查准率,增强了检测相似重复记录的能力,减少了误被当成相似重复记录而被检测出的记录数。

从图 4 中可以看出,SNM 改进方法的综合评价指标 F 值也优于 SNM 方法,说明改进方法的整体性能也得到了提高。

实验 2 的结果如图 5 ~ 7 所示。

从图 5 中可以看出,尽管此时比较窗口较大,基本可以囊括所有相似重复记录,所有数据集上,改进的 SNM 方法查全率

仍不低于 SNM 方法,主要在于字段权重的动态调整,提高了相似度的计算精度。数据集为 1.5 和 2 万条记录时,两种方法的效果相当,主要由于窗口大小为 50 时已足够用于相似重复记录的检测,基本不会出现漏配的现象,但在记录为 2.5 万条时改进方法仍优于 SNM 方法,因为改进方法的记录相似度计算精度要优于 SNM 方法。

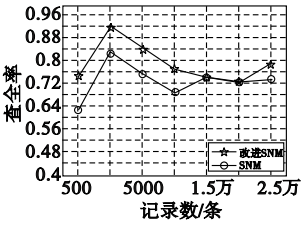

图5　SNM与改进SNM方法查全率对比

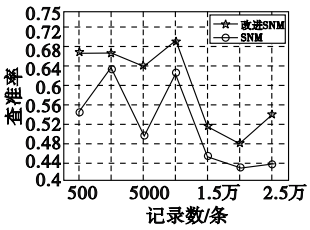

图6　SNM与改进SNM方法查准率对比

从图 6 中可以看出,窗口设为 50 后,改进 SNM 方法的查准率仍然高于 SNM 方法。由实际数据可知,窗口为 50 时,很多记录比较是没有必要的,此时改进方法能够缩小比较窗口,减少将非相似重复记录误认为是相似重复记录的情况,同时采用动态调整等级法,提高了检测的准确性。另外,从图中可以看出,针对不同数据集 SNM 方法查准率的波动较大,检测结果不够稳定。改进 SNM 方法除了 1 ~ 1.5 万条记录时查准率与 SNM 方法一样存在波动,其查准率整体比 SNM 方法稳定。

从图 7 中来看,比价窗口为 50 时,改进 SNM 算法的整体性能也是优于 SNM 算法的。

实验 2 中的程序运行时间对比如图 8 所示。从图 8 中可以看出,由于改进 SNM 算法采用了变步长伸缩窗口,减少了不必要的记录比较,节省了相似重复记录的检测时间,提高了比较效率。

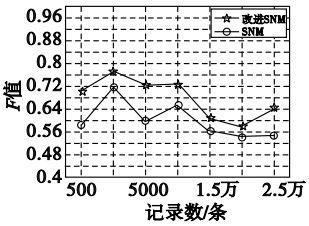

图7　SNM与改进SNM方法 *F*值对比

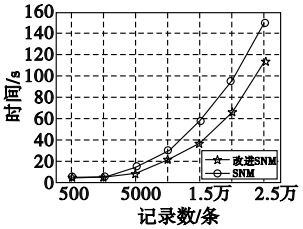

图8　SNM与改进SNM方法运行时间对比

4　结束语

本文详细阐述了 SNM 算法,并指出其存在的不足,提出了相似重复记录检测的改进方法。采用变步长伸缩窗口,在避免漏配的同时,减少了不必要的比较,根据记录相似度确定合理的窗口改变步长的大小。采用动态调整等级方法,解决了人为设定固定权重的不足,提高了检测精度。并且还实现了噪声字段删除、等级、编辑距离、传递闭包等方法。通过对实际系统中数据的测试,验证了改进方法的有效性和优越性。同时可以看到,本文实现的 SNM 和改进 SNM 方法的各检测指标偏低,这主要是由于排序方法不科学、字符匹配函数不准确造成的,也是下一步有待改进和完善的地方。

参考文献:

[1] 韩京宇,徐立臻,董逸生. 数据质量研究综述[J]. 计算机科学,2008,35(2):1-5.

[2] BILENKO M, MOONEY R J. Adaptive duplicate detection using learnable string similarity measures[C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC:ACM Press,2003:39-48.

[3] CHANDEL A, HASSANZADEH O, KOUDAS N, *et al*. Benchmarking declarative approximate selection predicates[C]//Proc of ACM SIGMOD International Conference on Management of Data. [S. l.]: ACM Press,2007:353-364.

[4] SARAWAGI S, COHEN W W. Semi-Markov conditional random fields for information extraction[C]//Advances in Neural Information Processing Systems. 2004.

[5] VIOLA P, NARASIMHAN M. Learning to extract information from semi-structured text using a discriminative context free grammar [C]//Proc of the 28th Annual International ACM SIGIR Conference on Research and Development in information Retrieval. [S. l.]:ACM Press,2005:330-337.

[6] 庞雄文,姚占林,李拥军. 大数据量的高效重复记录检测方法[J]. 华中科技大学学报,2010,38(2):8-11.

[7] 王宏志,樊文飞. 复杂数据上的实体识别技术研究[J]. 计算机学报,2011,34(10):1843-1852.

[8] 鲁均匀,李星毅,施化吉,等. 基于内码序值聚类的相似重复记录检测方法[J]. 计算机应用研究,2010,27(3):874-878.

[9] HERNANDEZ M A, STOLFO S J. Real-world data is dirty: data cleansing and the merge/purge problem [J]. Data Mining and Knowledge Discovery,1998,2(1):9-37.

[10] 陈伟,王昊,朱文明. 一种提高相似重复记录检测精度的方法[J]. 计算机应用与软件,2006,23(10):29-30.

[11] 肖满生,周浩慧,王宏. 基于模糊综合评判的相似重复记录识别方法[J]. 计算机工程, 2010,36(13):51-53.

[12] HERNANDEZ M A, STOLFO J S. The merge/purge problem for large databases[C]//Proc of ACM SIGMOD International Conference on Management of Data. [S. l.]:ACM Press,1995:127-138.

(上接第 2735 页)

[14] 周津慧,王宗,杨宗奎,等. 基于模糊评价方法的软件质量评价研究[J]. 系统工程与电子技术,2004,26(7):988-991.

[15] SOMMERVILLE I. Software Engineering[M]. 9th ed. 程成,等译. 北京:机械工业出版社,2011.

[16] NIELSEN J. 可用性工程[M]. 刘正捷,译. 北京:机械工业出版社,2004.

[17] ISO/IEC 9126,Software engineering:product quality is an international standard for the evaluation of software quality[S].

[18] LI Tong. An approach to modeling software evolution process [M]. 北京:清华大学出版社,2008.

[19] 肖刚,古辉,程振波,等. 实用软件文档写作[M]. 北京: 清华大学出版社,2005.

[20] 李安贵,张志宏,孟艳,等. 模糊数学及其应用 [M]. 北京:冶金工业出版社,2005.

[21] 石柱,何新贵. 软件质量的模糊属性及其表示[J]. 计算机工程设计,2001,22(4):1-4,20.

[22] SWANSON M, BARTOL N,SABATO J,*et al*. Security metrics guide for information technology systems [R]. [S. l.]:National Institute of Standards and Technology Special Publication,2003.