计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
45(3):542~547, 2008

一种层次化的检索结果聚类方法

张 刚 刘 悦 郭嘉丰 程学旗
(中国科学院计算技术研究所 北京 100190)
(gangzhang@ict.ac.cn)

A Hierarchical Search Result Clustering Method

Zhang Gang, Liu Yue, Guo Jiafeng, and Cheng Xueqi
(*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Search result clustering can help users quickly browse through the documents returned by search engine. Traditional clustering techniques are inadequate since they can not generate clusters with highly readable names. In order to improve the performance of the search result clustering and help user to quickly locate the relevant document, a label-based clustering method is used to make the search result clustering. A multi-feature integrated model is developed to extract base-cluster labels, which combines the DF, query log and query context features together. Using the extracted labels, some basic clusters are built. In order to setup a hierarchical clustering structure, a basic cluster relation graph is built based on these basic clusters. A hierarchical cluster structure is generated from the basic cluster relation graph using the graph based cluster algorithm (GBCA). To evaluate the search result clustering method, a test-bed is set up. P@N and F-Measure are introduced to evaluate the extracted labels and the document distribution in clusters. The experiment shows that the integrated label-extraction model is very effective. The more feature is used, the higher P@N can be gained. Compared with the STC and Snaket clustering method, GBCA outperforms the STC and Snaket in cluster label extraction and F-Measure.

Key words information retrieval; search result clustering; hierarchical clustering; text clustering; clustering

摘 要 检索结果聚类能够帮助用户快速地浏览搜索引擎返回的结果.传统的聚类方法由于不能生成有意义的类别标签因此是不适合的,为了改善检索结果层次化聚类的效果,采用了基于标签的聚类算法,提出了将DF、查询日志、查询词上下文特征融合的类别标签抽取算法,并以抽取的标签构造基础类别图,通过GBCA算法构建层次化聚类结果.实验证明了多特征融合模型的有效性;GBCA算法在类别标签抽取和F-Measure两个评价指标上都比STC和Snaket算法有很大的提高.

关键词 信息检索;检索结果聚类;层次化聚类;文本聚类;聚类

中图法分类号 TP391

以Google、百度为代表的搜索引擎在用户输入一个查询后,返回一个"相关"结果的列表,然而这个检索结果列表往往并不能让用户感觉满意.一方面由于查询歧义的原因,搜索引擎返回的结果并不都是用户需要的信息,用户需要顺序浏览列表来找到真正相关的结果;另一方面,对于在搜索引擎返回的

收稿日期:2006-07-27;修回日期:2007-07-06
基金项目:国家"九七三"重点基础研究发展规划基金项目(2004CB318109,2007CB311100)

大量结果,用户通常只选择浏览 Top10 的检索结果,由于检索结果没有进行合理的总结与组织,而仅仅是简单的罗列,Top10 的检索结果可能是不全面的,因此用户获取到的信息可能是不全面的. 检索结果聚类可以很好地解决这两方面的问题,一方面对于有歧义的查询,通过检索结果聚类,可以按照不同语义将检索结果聚成不同类别;另一方面,检索结果聚类能够对检索结果进行全面的分析处理,可以给出一个全面的关于被查询对象的介绍.

在以往的研究中,检索结果的聚类方法可以分为两类:基于文档的方法和基于标签的方法[1]. 基于文档的方法通过传统文本聚类方法,把搜索引擎返回的文档聚成多个类别,然后从各类别中抽取出合适的标签来标注各个类别,相关的工作包括文献[2-4]等. 基于标签的方法首先从文档集中抽取有代表性的词、短语、片段作为类别标签,然后对抽取的类别标签进行合理的评价与筛选,以抽取的标签为基础做进一步的文档聚类工作,主要工作包括文献[5-8]等.

1　层次化检索结果聚类

本文的层次化聚类方法主要包括以下几个步骤:

1) Snippet 的获取与整理

通过对搜索引擎的返回结果网页分析处理提取出 Snippet. 如果将聚类算法嵌入到搜索引擎系统后,可以直接从搜索引擎获取到 Snippet.

2) 层次化检索结果聚类

① 类别标签的抽取与选择

通过对 Snippet 的分析,抽取并选择出有价值的类别标签;

② 由类别标签进行层次化检索结果聚类

由抽取出的类别标签构造出基础类别,再对基础类别进行聚类,形成层次化聚类结果.

1.1　改进的标签选择算法

1) 候选标签评价的 DF 特征

DF 是指包含标签 l 的文档个数,把 DF 作为评价标签的一个统计指标是基于这样的假设:如果一个候选标签在很多文档中都出现,那么该候选标签具有比较广泛的代表性,因此可以作为一个类别的标签. 例如:输入查询词"猎豹",那么在检索返回的文档中,"汽车"、"动物"等词语会在大量的文档中出现,因此"汽车"、"动物",就会成为比较好的类别候选标签. 但利用 DF 评价有一个明显的问题,语言中的功能词如"一个"、"这个"等也会在大量的文档中出现,产生较高的 DF 值,从而影响 DF 评价的效果.

2) 候选标签评价的查询日志特征

查询日志记录了以往搜索引擎用户输入的查询词,从以往的查询中找到和本次查询相关的所有查询,并对这些相关的查询进行分析,会对类别标签的选择有较大的帮助. 例如:利用百度的"相关关键词搜索",输入"猎豹",可以得到以往的相关查询的结果如图 1 所示:

	Query	Total Submitted
1	长丰猎豹	
2	猎豹飞腾	
3	猎豹汽车	
4	猎豹奇兵	
5	猎豹雕像	
6	长丰猎豹汽车	
7	湖南长丰猎豹	
8	猎豹飞腾汽车	

Fig. 1　Query log of "liebao".

图 1　"猎豹"相关查询日志

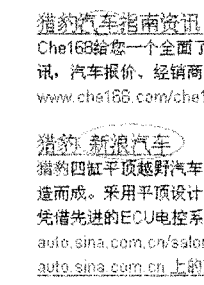
从图 1 可以看出,"长风"、"飞腾"、"汽车"这些非常有意义的标签都具有很高的"被搜索次数". 由于查询日志来自于以往搜索引擎的用户,因此集成了人的背景知识,实际上是一种非常有价值的背景知识库,因此利用查询日志对于检索结果聚类将有很强的指导意义.

3) 候选标签评价的查询词上下文特征

检索结果聚类与一般的文档聚类的一个重要的不同点是:对检索结果聚类时,有一个可以利用的重要信息就是查询词,这是一般的文档聚类问题所不具备的. 正是因为这一点,使得检索结果的聚类有了更加明确的指向性,也就是应该从查询词角度出发,对检索结果进行聚类. 查询词的上下文和查询的关系最为密切,因为上下文往往对查询词进行了限定或者解释,因此更加有利于确定查询词的语义. 例如:查询词是"猎豹",图 2 是利用百度查询"猎豹"得到的结果,可见"猎豹"的上下文中"汽车"具有较高的出现频率.

4) 多特征融合的候选标签评价算法 MILEM

DF 统计特征、查询日志特征及查询词的上下文特征是评价类别标签的有效特征. 这里我们提出了一种利用 DF 进行量化,把 3 种特征按照一定的权重进行融合的多特征融合模型 MILEM (multi-feature integrated label extraction model). 多特征融合模型


猎豹汽车指南资讯_che168
Che168给您一个全面了解猎豹汽车的平台。全新小排量汽车、讯，汽车报价、经销商电话、行情介绍，就在che168。
www.che168.com/che168/cardb/series/series ... 1K 2006-07 -

猎豹_新浪汽车
猎豹四缸平顶越野汽车,是在CFA6470G与CFA6470E的车型基础而成。采用平顶设计,更具越野车的刚毅气质。采用高性能的凭借先进的ECU电控系统,率先达到欧III标准,更加省油。..
auto.sina.com.cn/salon/LIEBAOMOTOR/LIEBAO ... 32K 2008
auto.sina.com.cn 上的更多结果

Fig. 2 Demonstration of query context.

图2 查询词上下文示意

MILEM 将 3 种特征按照不同权重加权进行排序,表示为

$$w_l = aw_{df} + bw_{log} + cw_{contex},$$

其中 a,b,c 表示不同特征权重的加权系数,w_l表示标签 l 的权重,w_{df}表示 DF 特征的权重,w_{log}表示查询日志特征的权重,$w_{context}$表示关键词上文的特征的权重. 为了计算方便,每一种特征的权重计算均采用标签的 DF 作为权重. 因此上面的权重可以进一步表示为:

$$w_l = (a + bf(l) + cg(l))DF(l),$$

$$f(l) = \begin{cases} 0, & \text{如果标签 } l \text{ 未出现在日志中,} \\ 1, & \text{如果标签 } l \text{ 出现的日志中,} \end{cases}$$

$$g(l) = \begin{cases} 0, & \text{标签 } l \text{ 没有出现在查询上下中文,} \\ 1, & \text{标签 } l \text{ 出现在查询上下文中,} \end{cases}$$

其中 $DF(l)$表示标签 l 的 DF 值.

在改进的标签选择算法中,通过将 DF 特征、查询日志特征以及查询词上下文特征融合到统一的模型 MILEM 中对类别标签进行评价.

1.2 层次化聚类的算法 GBCA

通过上面的标签抽取与排序,可以得到一些有价值的标签,由这些有价值的标签可以进一步形成一些基本的类别;下一步,再由这些基础类别相互合并形成更高层次的类别,从而建立起层次化的聚类结果.

定义 1. 基础类别. 基础类别是指层次化聚类结构中最底层的类别,基础类别不再包含其他类别而是直接包含文档.

在我们的层次化检索结果聚类中,基础类别的形成通过抽取的类别标签来构建,即包含某个标签的文档作为一个基础类别. 具体表示为

设 $D = \{d_1, d_2, \cdots, d_n\}$ 为搜索引擎返回的结果,$L = \{l_1, l_2, \cdots, l_m\}$ 为类别标签,函数 $f(d,l)$ 是二值函数,表示文档 d 是否包含标签 l,定义为

$$f(d,l) = \begin{cases} 1, & d \text{ 包含标签 } l, \\ 0, & d \text{ 不包含标签 } l, \end{cases}$$

那么类别标签 l 所形成的类别 $C(l)$表示为 $C(l) = \{d \mid d \in D, f(d,l) = 1\}$.

在层次化聚类中,我们提出了一种新的聚类算法,为了表示方便我们将这种算法命名为基于图的聚类算法 GBCA(graph based clustering method). 算法的主要思想是:通过对基础类别关系图的分析来形成层次化的聚类结果.

定义 2. 基础类别关系图. 以基础类别为顶点,基础类别间的相关性为边,基础类别间的相关度为边的权重的带权图 $G = (V, E, W)$称为基础类别关系图.

具体的基础类别关系图的构建过程如下:

设 $V = \{V_1, V_2, \cdots, V_m\}$ 是由基础类别构成的图 G 的顶点集,其中每个顶点 $V_i = \{d_1, d_2, \cdots, d_t\}$ 是由若干文档组成的基础类别,图 G 的边集 E 定义为:顶点 V_i 和顶点 V_j 间存在一条边 $e_{i,j}$,当且仅当 $V_i \cap V_j \neq \varnothing$,也就是说如果两个基础类别构成的顶点间存在一条边,当且仅当这两个基础类别的交集非空,边 $e_{i,j}$的权重表示为 $w_{i,j}$,权重集 W 表示为

$$W = \{w_{i,j} \mid w_{i,j} = \mid V_i \cap V_j \mid\}.$$

以查询'猎豹'为例,我们取 20 个基础类别生成的基础类别关系图如图 3 所示,图 3 的左边是 20 个基础类别标签,图中没有表示出边的权重,从图 3 可以看出基础类别间构成了一个复杂的关系网络.

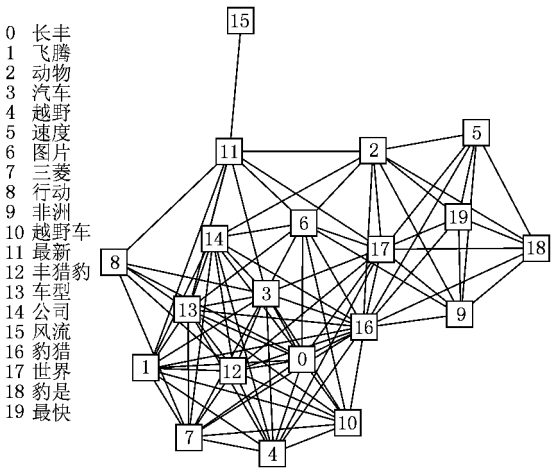


Fig. 3 Basic cluster relationship graph of "liebao".

图3 查询'猎豹'的基础类别关系图

基于图的聚类算法:

在得到了基础类别关系图 G 后,进一步就要通过对基础类别关系图 G 的分析,将顶点划分成若干

类.以往的研究中,采用图来聚类或对图进行划分也有一些方法[9-10],但这些方法通常需要比较大的计算开销,由于检索结果聚类是检索过程中实时进行的,因此计算开销大的方法往往不能适用,我们这里采用了一种非常简单的基于图的聚类算法,具体算法描述如下.

输入:基础类别关系图 G;

输出:对基础类别聚类的结果 $C=\{C_1,C_2,\cdots,C_u\}$.

1) 将图 G 的所有边按照边的权重进行排序;

2) 取出一条没有使用过的权重最大的边 $e_{i,j}$,并标注该边已经被使用过;

3) 如果 $e_{i,j}$的两个顶点 V_i 和 V_j 都没有标注类别,则给顶点 V_i 和 V_j 标注相同的类别;

如果顶点 V_i 标注了类别,V_j 没有标注类别,则把 V_j 标注成和 V_i 相同的类别;

如果顶点 V_j 标注了类别,而 V_i 没有标注类别,则把 V_i 标注成和 V_j 相同的类别;

如果顶点 V_i 和 V_j 都已经标注了类别则跳到4)执行;

4) 如果所有顶点都已经标注完毕或者所有的边都已经使用过,则停止,否则跳到2)执行;

5) 将所有标注了相同类别的顶点归为一类,如果仍有顶点没有标注类别,则独立成为一类,这样形成了基础类别的父类别:$C=\{C_1,C_2,\cdots,C_u\}$;

6) 对于父类别 C_i,取类别 C_i 中前 k 个基础类别标签权重最大的标签作为父类别的标签,$k=c\%\times|C_i|$ and $k<10$,这里令 $c=40$.

我们仍然以"猎豹"为例,图4是对图3应用GBCA算法后得到的结果图:

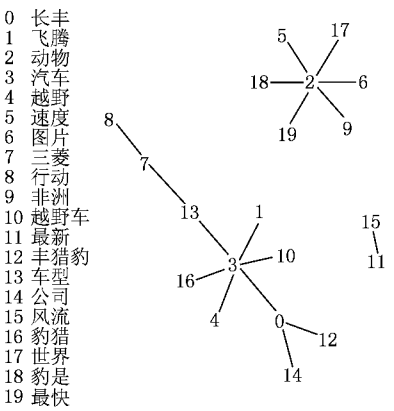


Fig. 4　Demonstration of GBCA.

图4　GBCA算法聚类结果展示

图4中左边是20个基础类别标签,右边是经过GBCA算法后形成的聚类结果,相互连接的顶点形成一类.从图4可以看出,查询"猎豹"的基础类别被划分成3个更高层次的父类别,作为汽车品牌的语义(结点3,1,4,0,…)和作为"动物"的语义(结点2,5,9,…)进行了有效的划分.

2　检索结果聚类的评价

2.1　评价方法

与一般聚类评价不同,检索结果的聚类需要对类别标签和文档在类别中的分布分别评价.以往的工作中,对于检索结果聚类的评价通常有4种方式:

1) 用户调查法

通过用户使用,给出一个系统的评价,缺点是主观评价、人工操作,评价代价高.

2) P@N评价类别标签抽取

采用类似信息检索中的P@N评价指标来对类别标签进行评价的方法.

3) F-Measure、信息熵

对聚类结果中文档分布进行评价,也是一般的聚类评价标准.

4) 利用检索结果评价

通过对最终检索结果质量的评价来评价检索结果聚类的效果,是一种间接评价.

本文采用了F-Measure[6]和P@N[7-8]指标对类别标签和文档分布分别进行评价.

2.2　评价数据集合构建

从目前了解的情况看,在检索结果聚类评价中,还没有一个标准的测试集.因此我们构建了自己的评价集合.具体构建方法是:我们找了30个不同类型的查询,如表1所示,查询的类型包括:歧义的查询、命名实体、一般关键词.

Table 1　List of Different Query Types

表1　不同类型的关键词列表

Query Type	Query Examples
Ambiguous Query	Beetles, Apple, …
Name, Entities	Mike, ICT…
General Query	Joke, Music…

对于每个查询都从"百度"搜索引擎下载了100个检索结果,共3000篇文档,人工对每个查询返回的结果进行了分类,并给每个类别标注适当的标签,利用人工标注的集合作为标准评测集.

2.3 实验结果

首先对聚类标签的抽取和排序进行评价,希望通过此实验验证多特征融合的类别标签抽取方法的有效性,按照多特征融合的算法,将 DF 特征、查询日志特征及查询词上下文特征,以不同权重进行组合,如表 2 所示:

Table 2 Feature Combination by Different Weight

表 2 加权系数变化形成不同特征组合

Weight Factor			Feature
a	*b*	*c*	
1	0	0	df
0	1	0	log
0	0	1	cnt
1	2.5	0	df + log
1	0	2.5	df + cnt
1	2.5	2.5	df + log + cnt

df:DF;log:guery log;cnt:query context.

对于聚类标签抽取和排序的评价采用 P@N 作为标准,评价的结果如表 3 所示:

Table 3 Evaluation of Feature Combination

表 3 不同特征排序算法评价

Feature	P@3	P@5	P@7	P@10
df	0.47	0.47	0.44	0.39
log	0.48	0.42	0.35	0.28
cnt	**0.54**	0.47	0.40	0.32
df + log	0.49	0.47	0.48	0.43
df + cnt	0.50	0.51	0.48	0.44
df + log + cnt	0.51	**0.54**	**0.49**	**0.46**

从表 3 可以看出,多个特征融合的方法要好于单独采用某一方面特征所取得的结果,同时采用 3 类特征时效果达到最好,可见 3 方面的特征能够很好的相互补充,本文所采用的 3 类特征刻画了类别标签的特点.

其次,将对层次化聚类结果进行评价. 实验中将我们的聚类结果与 STC 算法[1]和 Snaket 聚类算法[8]进行了比较,在比较中,STC 采用了原论文提出的标签抽取算法,Snaket 和 GBCA 算法采用了相同的标签抽取算法. 由于 STC 聚类的结果是扁平的,GBCA 算法和 Snaket 算法都是层次化聚类算法,因此在比较时,将层次化算法形成的最高层与 STC 的扁平结构进行比较. 实验中分别比较了类别标签和文档的分布,在类别标签比较中采用了 P@N 的标准,而文档的分布采用了 F-Measure 标准. 聚类标签的比较结果如表 4 所示:

Table 4 Label Extraction Evaluation of Different Clustering Algorithm

表 4 不同聚类方法标签抽取比较

Algorithm	P@3	P@5	P@7	P@10
STC	0.29	0.33	0.30	0.31
Snaket	0.43	0.47	—	—
GBCA	**0.46**	**0.47**	**0.48**	—

从聚类结果的标签评价可以看出,GBCA 算法效果最好,由于有的算法生成的类别数较少,因此有些评价值空缺. 与 STC 算法的结果相比较,Snaket 和 GBCA 算法的结果要明显提高,这说明本文的多特征融合的标签抽取算法的性能比 STC 的标签提取算法效果更好. 虽然 GBCA 和 Snaket 采用了相同的标签抽取算法,但聚类过程中会重新生成顶层类别的标签,而 GBCA 的顶层类别标签结果更好.

文档分布评价的 F-Measure 值如图 5 所示:

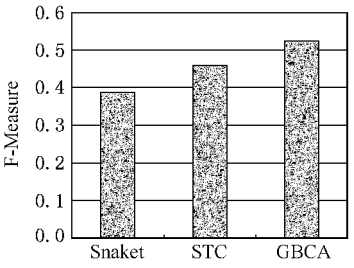


Fig. 5 F-Measure of different clustering algorithm.

图 5 不同聚类算法 F-Measure 比较

从 F-Measure 评价方面可以看出,GBCA 算法体现了比较好的聚类性能. 而 Snaket 算法主要利用了标签的特点而没有重视文档在类别中的分布因素,因此在性能上逊色于 STC 和 GBCA 算法.

3 总 结

本文针对检索结果聚类问题,给出了一种层次化的聚类算法. 该算法采用了基于标签的聚类思想,提出了采用 DF 特征、日志特征、查询词上下文特征相融合的类别标签抽取算法,对类别标签进行选择和排序. 进一步以抽取的类别标签为基础形成基础类别,通过对基础类别构成的基础类别关系图的分析,提出了 GBCA 算法来形成层次化的聚类结果. 在实验中,不仅验证了多特征融合模型的有效

性,而且通过与 STC,Sanket 检索结果聚类算法的比较,验证了我们提出的 GBCA 算法在类别标签的评价和 F-Measure 指标上也有明显的提高.

参 考 文 献

[1] Hiroyuki Toda, Ryoji Kataoka. A search result clustering method using informatively named entities [C]. In: Proc of the ACM Workshop on Web Information and Data Management. New York: ACM Press, 2005. 81-86

[2] M A Hearst, J O Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results [C]. In: Proc of the ACM Special Interest Group on Information Retrieval Conf. New York: ACM Press, 1996. 76-84

[3] F Giannotti, M Nanni, D Pedreschi. Webcat: Automatic categorization of Web search results [C]. In: Proc of the 11th Italian Symp on Advanced Database Systems. Italian: Rubbettino Editore, 2003. 507-518

[4] Wang Zhimei, Zhang Junlin, Li Qiushan. Research and realization about fast clustering method of Web search engine result [J]. Computer Engineering and Design, 2004, 25(12): 2231-2233 (in Chinese)
(王志梅,张俊林,李秋山. Web 检索结果快速聚类方法的研究与实现[J]. 计算机工程与设计, 2004, 25(12): 2231-2233)

[5] Oren Zamir, Oren Etzioni. Web document clustering: A feasibility demonstration [C]. In: Proc of the ACM Special Interest Group on Information Retrieval Conf. New York: ACM Press, 1998. 46-54

[6] Florian Beil, Martin Ester, Xiaowei Xu. Frequent term-based text clustering [C]. In: Proc of the 8th ACM Int'l Conf on Knowledge Discovery and Data Mining. New York: ACM Press, 2002. 436-442

[7] H Zeng, Q He, Z Chen, *et al*. Learning to cluster Web search results [C]. In: Proc of the ACM Special Interest Group on Information Retrieval Conf. New York: ACM Press, 2004. 210-217

[8] Paolo Ferragina, Antonio Gulli. A personalized search engine based on Web-Snippet hierarchical clustering [C]. In: Proc of the 14th Int'l Conf on World Wide Web. New York: ACM Press, 2005. 801-810

[9] X He, H Zha, C Ding, *et al*. Web document clustering using hyperlink structures [R]. Department of Computer Science and Engineering, Pennsylvania State University, Tech Rep: CSE-01-006, 2001

[10] Jianbo Shi, Jitendra Malik. Normalized cuts and image segmentation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

Zhang Gang, born in 1977. Ph. D. Research assistanter. His main research interests include information retrieval.
张　刚,1977 年生,博士,助理研究员,主要研究方向为信息检索.

Liu Yue, born in 1971. Ph. D., associate professor. Her main research interests include information retrieval.
刘　悦,1971 年生,博士,副研究员,主要研究方向为信息检索.

Guo Jiafeng, born in 1980. Ph. D. candidate. His main research interests include information retrieval.
郭嘉丰,1980 年生,博士研究生,主要研究方向为信息检索.

Cheng Xueqi, born in 1971. Ph. D. and professor. Member of China Computer Federation. His main research interests include information retrieval and network security.
程学旗,1971 年生,博士,研究员,计算机学会会员,主要研究方向为信息检索、网络安全.

Research Background

Conventional search engine returns long lists of ranked documents. Users have to go through the list and examine the title and snippets sequentially to find the relevant documents. This is a time consuming task especially when users submit an ambiguous or poor query. To address this problem, online search result clustering methods have been proposed. These methods organize search results into different groups and give each group a meaningful label, which enables users to identify their required documents at a glance. In this paper, we introduce a new hierarchical clustering method which uses multi-features for labels ranking and GBCA for clustering. Experiments indicate that the new method is quite effective. Our work is supported by the National Basic Research Program of China (2004CB318109, 2007CB311100).