一种利用链接信息检索关键资源的算法*)

顾 健 黄萱菁 吴立德
(复旦大学计算机科学与工程系 上海200433)

摘 要 随着互联网的发展,基于Web的信息处理技术越来越受到人们的重视,也是当前研究的前沿课题。本文探讨的是如何在现有检索技术的基础上,利用Web网页的链接信息,自动地得到更高质量的检索结果——关键资源。本文提出一种同时利用Web网页的结构和内容信息以及链接信息的新方法:先结合网页的结构信息和内容评分得到网页的文档评分,然后基于网页出链的文档评分计算网页的链接评分。实验表明,本文的方法减少了无用链接的干扰,比单纯利用链接信息的效果好得多。

关键词 链接信息,关键资源,Web检索

An Algorithm of Applying Linkage Information to Retrieving Key Resources

GU Jian HUANG Xuan-Jing WU Li-De
(Department of Computer Science and Engineering, Fudan University, Shanghai 200433)

Abstract With the development of the Internet, Web-based information processing is becoming more and more important, and it is one of the frontiers in current research. In this paper, we describe a new approach which is based on the traditional retrieval technology and can automatically obtain the high-quality retrieval results from the basic retrieval ones by utilizing the Web information (including structure information and link information): calculating the document score of a Web page by combining its structure information and content score, then computing the link score of the page based on the document scores of its out-linking pages. In the experiments, this approach reduces the influence of the "noisy" links and achieved fairly good results.

Keywords Link information, Key resource, Web retrieval

当今的世界正处于一个信息爆炸的时代,信息发布及更新的速度远大于整理、利用信息的速度。互联网就是一个最大的信息资源宝库。据文[1]的统计,在2000年,能够访问到的公开的Internet(Public accessible Internet)目前拥有超过21亿个不同的网页,网页平均大小为10kB左右,即总容量大约为21tB。

在这种背景下,搜索引擎应运而生,并且获得了巨大的成功,比较著名的如google,yahoo,infoseek等。其中,新一代的搜索引擎google合理地利用链接信息提高了Web检索的查准率和查全率。链接信息也成为当前Web信息检索的一个研究热点。

本文先介绍TREC会议中的Web检索任务。然后讨论现有的利用Web链接信息的方法,针对在Topic Distillation任务中的应用,提出采用基于文档评分的链接算法来提高链接算法的查准率,并给出其实验结果。最后,我们对实验结果进行分析,说明利用Web网页的结构和内容信息计算链接信息确实对Web检索的性能有很大的提高,并对Web链接信息的应用和接下来的工作做进一步的展望。

1 TREC及Topic Distillation子任务简介

文本检索会议(Text REtrieval Conference,简称TREC),是由美国国家标准技术局(NIST)和国防部高级研究计划局(DARPA)组织召开的一年一度的国际评测会议。TREC会议旨在促进大规模文本检索领域的研究,加速研究成果向商业应用的转化,促进学术研究机构、商业团体和政府部门之间的交流和协作。从1992年开始,迄今已举办了11次,是文本评测领域最权威的国际会议。

Web任务是TREC会议中重要的任务之一,始于1999年的TREC 8,其目的是针对Web环境建立一个框架,使得研究者能可靠地检验新的Web检索技术,并且可以重复各类实验[2]。TREC使用的语料集和实验数据已经成为了研究搜索技术的研究者们经常使用的实验语料。

Topic Distillation是TREC Web任务中的一个子任务,是TREC 11新设立的,目的是要找到和给定主题相关的关键资源(Key Resource)。关键资源有以下一些特征[3]:

1)关键资源比相关网页要少得多,它们是关于这个主题最有用的网页。

2)这个网页的链接比文字内容更重要,如一个站点的主页。

3)关键资源和分类目录是不同的,它更关注于给定的主题。

4)一个关键资源可能是以下几种网页:

a)一个和主题非常相关的网站的主页。

b)一个和主题相关的子网站的主页,如www.surgeon-

*)本文受国家自然科学基金项目(69935010,60103014)、863项目(2001AA114120,2002AA142090)资助。顾 健 硕士研究生,主要研究方向为Web信息检索。黄萱菁 博士,副教授,主要研究方向为自然语言处理与信息检索。吴立德 教授,博士生导师,主要研究方向为图像处理,模式识别,计算机视觉,计算机图形学和自然语言处理。

general.gov/topics/obesity/。

c)一个高度相关的 html，pdf，doc，ps 网页。

d)一个包含很多有用链接的网页，通常我们把这种网页称为 hub 页面。

e)一个提供相关服务的网页，如 http://www.nasa.gov/search/。

5)同一个网站不会包含很多关键资源，一般不超过两个。

6)在一个网站里往往存在正确的一级目录和主题相关，如 www.whitehouse.gov/history/ 对应于主题"white house history"。

关键资源是一个主观的但却是非常重要的概念。Topic Distillation 任务能帮助用户迅速找到相关主题的重要资源，在实际应用中有很大价值，也是一个全新的任务。

2 Web 链接信息的应用

实验证明，合理地利用链接信息对提高 Web 检索的查准率和查全率起着重要的作用[4~7]。Sergey Brin 和 Lawrence Page 提出的 PageRank 算法[4]，Kleinberg 等人提出的 HITS 算法[5]，Ron Weiss 等人提出的超链接相似度函数[6]以及 R.Lempel 等人提出的 SALSA 算法[7]是其中有代表性的几种算法。这些方法都是只考虑网页相互之间的链接关系，而不考虑这些网页本身的内容和结构信息。

但是现有的单纯利用链接信息的方法总是不能得到令人满意的效果，甚至使系统性能略有下降[8]。而根据 Topic Distillation 任务的要求，关键资源是一个和很多相关网页关联的网页，即通过它可以访问到很多相关网页。因此我们考虑把链接信息和网页本身的属性结合起来，利用网页的文档评分提高链接信息的精确率，从而提高整个系统的性能。

2.1 Web 网页信息的分类

Web 是一个巨大的半结构化的信息源，传统的信息检索技术不能满足如此巨大的信息量。相对于传统的文档，Web 网页除文本内容信息外，还蕴含更多的有用信息。这些信息主要分为两类：(1)结构信息，是页面本身所具有的结构化信息，如 TITLE，HEAD 等 HTML 标记。(2)链接信息，体现了网页之间的相互联系，构成了整个 Web 具有的拓扑结构。下面分别讨论这两类信息的具体内容。

2.1.1 *结构信息* 由于 Web 页面主要由 HTML 语言标记，因此含有丰富的 HTML 标记，这些标记也包含了许多文本内容以外的有用信息。

比较重要的结构信息有：

1)统一资源定位器：URL(Uniform Resource Locator)；2)标题；3)锚文本(Anchor Text)。

我们的实验中主要采用了标题和锚文本两种结构信息。因为这两种结构信息蕴含了丰富的信息，并且精确率很高。

2.1.2 *链接信息* Web 上的超文本网页相互之间的链接可以看成是引用，这种引用关系体现出的网页的重要性能够较好地符合人们的主观认识。对于每一个网页，它有两方面的链接信息：(1)入链：从别的网页指向它的那些链接。(2)出链：从它指向别的网页的那些链接。应用链接信息的假设就是：一篇网页会引用和它内容相关的网页。但是实际情况并非如此，会有很多无关链接(如广告等)，这也是应用链接信息的难点所在。

2.2 Web 网页结构信息的应用

在 Topic Distillation 任务中，我们把网页结构的信息也看作是网页的一部分，对标题和锚文本分别作了索引。在检索时我们除了对网页的内容计算评分以外，还同时对网页的标题和锚文本计算它们和主题的相似度。根据它们的重要程度分别给予权重，综合以后得到网页的文档评分。我们给标题出现全部查询词的文档打2分，随着查询词在标题中出现的减少，文档是相关的可能性也减小，并且减小的幅度非常大。网页标题及锚文本的形式和作用类似，因此相似度计算采用相同的方法，即它们所包含的查询项的个数的函数：

$$Score_T(d_i)=\begin{cases}2^{|T\cap Q|-|Q|} & |T\cap Q|>0\\ 0 & |T\cap Q|=0\end{cases} \quad (1)$$

$$Score_A(d_i)=\begin{cases}2^{|T\cap Q|-|Q|} & |A\cap Q|>0\\ 0 & |A\cap Q|=0\end{cases} \quad (2)$$

其中，T 和 A 分别为网页 d_i 的标题和锚文本包含的项的集合，Q 为查询项的集合。

我们把网页的标题和锚文本的相似度看作和网页内容的相似度同等重要，但在网页的标题和锚文本的相似度中只取其中较大的，因为我们认为标题和锚文本都是对网页内容的概括性描述。网页文档评分的计算公式如下：

$$Score_D(d_i)=Score_C(d_i)+\max\{Score_T(d_i),Score_A(d_i)\} \quad (3)$$

其中，$Score_D(d_i)$、$Score_C(d_i)$、$Score_T(d_i)$、$Score_A(d_i)$分别为网页 d_i 的文档评分、内容评分、标题评分和锚文本评分。$Score_C(d_i)$是通过常规的搜索引擎检索得到的文本内容评分，且 $Score_C(d_i)$、$Score_T(d_i)$和 $Score_A(d_i)$均归一化到同一量级，使计算得到的 $Score_D(d_i)$更合理。$Score_D(d_i)$体现了网页 d_i 本身和主题的相关程度。

2.3 基于文档评分的链接评分算法

为了提高链接评分算法的查准率，我们采用基于文档评分的链接评分算法。基于文档评分的链接评分算法把一个文档所有的链接看作具有不同重要程度，给予每个链接不同的权重，即为链接所指的网页的文档评分。根据关键资源的定义，我们认为，一个网页要么本身高度相关，要么通过它能很快地找到许多相关网页，一般不超过两级。因此，对一个网页，它的出链比入链重要，一级出链比多级出链重要。因此，我们只考虑网页的一级出链，计算公式如下：

$$Score_L(d_i)=\sum_{d_i\to d_j}Score_D(d_j) \quad (4)$$

$Score_L(d_i)$为网页的链接评分，等于 d_i 的所有出链的文档评分的总和。这是为了考虑到和主题相关的 hub 网页。$Score_L(d_i)$体现了网页 d_i 的链接信息和主题的相关程度。在这儿，d_i 的链接信息指 d_i 的所有出链的文档评分。

2.4 文档内容、结构评分和链接评分的综合

我们根据网页的文档评分和链接评分计算其综合评分。考虑到链接信息可能含有无关信息，而且实验也证明如此，我们在综合时更相信文档评分，因而给予文档评分较高的权重。综合公式如下：

$$Score(d_i)=a*Score_D(d_i)+(1-a)*Score_L(d_i) \quad (5)$$

为使 $Score(d_i)$更合理，我们把 $Score_D(d_i)$和 $Score_L(d_i)$化归为同一量级。$Score_L(d_i)$体现了综合了本身信息及其链接信息之后，网页 d_i 和主题的相关程度。既考虑到本身内容和主题高度相关的网页，也考虑到了和主题相关的 hub 网页。参数 a 确定 $Score_D(d_i)$和 $Score_L(d_i)$在 $Score(d_i)$中的贡献之比。因为这是一个全新的任务，我们没有训练语料，因此对参数 a 的估计不太准确。在我们的实验中，a 取0.8和0.6，发现前者比后者结果要好(见4.2节)。

3 实验

3.1 语料集[9]

TREC-2002 Web Track 用的语料集是.gov 语料集。这个语料集是从美国政府网站上获取的。它包括大约1M 个 HTML 文档和大约250K 个非 HTML 文档(*doc*,*pdf*,*ps* 等)。它的字节数达到了18.1 GB。

表1 .gov 语料集的详细信息

Number of pages	1,247,753
Number of pages by mime type:	
text/html	1,053,110
application/pdf	131,333
text/plain	43,753
application/msword	13,842
application/postscript	5,673
other (containing text)	42
Average page size	15.2 kB
Number of hostnames	7,794
Total number of links	11,164,829
Number of cross-host links	2,470,109
Average cross-host links per host	317

3.2 扩展基本的检索结果集

我们先用常规的搜索引擎根据网页的文本内容检索得到一个基本的检索结果集 B,然后根据一级入链进行扩展,得到一个扩展集 $E=B\cup I$。其中 I 为 B 的一级入链集,即 $I=inlink(B)=\{p \mid p\rightarrow q, q\in B\}$。因为我们的搜索引擎采用的是扩展布尔检索模型,得到的检索结果查准率比较高,查全率低一点[10],所以我们不丢弃任何结果,并且只扩展一级入链。扩展的过程如图1所示。

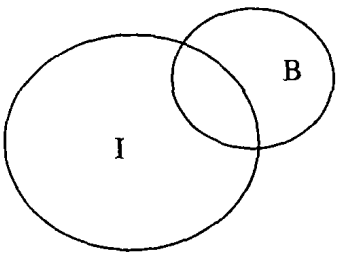

图1 扩展集 E = BUI

对得到的扩展集里的每一个网页的内容进行归一化:对于原来属于 B 的网页,我们把初始的内容评分,化归到[0, 1]区间上的一个评分;对于原来不属于 B 的网页,赋予它们归一化的内容评分为0。

3.3 应用链接信息

我们只考虑在扩展集 E 中的出链,因此定义网页的有效出链集为 $outlink(p)=\{q \mid p\rightarrow q, q\in E\}$。如果网页 p 本身具有 n 条出链,但其中只有 m 条指向 E 中的网页,那么此时有 $|outlink(p)|=m$。只有当网页的有效出链集个数 $|outlink(p)|\geq 2$时,我们才在网页的综合评分里加入链接属性。

在综合文档评分和链接评分时,我们给予文档评分较大的权重,在综合公式中取 $a=0.8$时效果较好。这样做的考虑是,网页本身的属性(包括标题,内容,锚文本)比较可靠,对提高检索结果的查准率作用很大。而链接包含许多无关信息。并且先计算文档评分,再用文档评分计算链接评分。其实质是,先尽量提高文档评分的查准率,再进一步选出其中的关键资源。

对得到的扩展集中的每一个页面,计算它们的文档评分,作为它的综合评分。经过网站过滤,即对同一个网站的网页进行筛选,一般根据它们的评分取一到两个网页,得到最终的检索结果。

3.4 实验结果

3.4.1 TREC 评测方法[11] TREC 目前采用的评价模式和评价标准被国际上广泛认可。它先在每一个单位提交的结果中抽取若干个结果,把所有这些结果合并以后作为一个评测集。NIST 的专业人员会审阅这些网页,对它们进行评价,判定其中哪些属于关键资源,哪些不是。然后利用这个评测集,对所有的结果自动进行评价。遇到未评价过的网页,则认为不是关键资源。这样得到的对每个网页的评价结果应该是很严格的,而且对每个系统也是比较公平的。

TREC 11中 Topic Distillation 任务的评价指标和以往有所不同,采用 $P@10$,即返回的前十个结果中属于关键资源的查准率。这是比较实用的,因为一般情况下,用户都希望能找到少数几个网页,却能包含涉及这个主题的大部分有用信息。

3.4.2 TREC 评测结果[9] 共有17家单位参加了 Topic Distillation 任务,提交了71个 run。NIST 评测人员根据上面介绍的方法对所有的综述进行了统一评测,一共评测了56,650个网页,其中1574个被评价为关键资源。

我们的布尔搜索引擎得到的基本结果集 *fduwt11b0*的 $P@10$性能指标为0.1510,而上面实验的两个结果 *fduwt11t2* 和 *fduwt11o2*分别为0.1939和0.1714。这两组结果都综合了网页的结构信息和链接信息。*fduwt11t2*中,$a=0.8$;*fduwt11o2*中,$a=0.6$,并用 WordNet 知识库进行了查询扩展,性能比 *fduwt11t2*略有下降。图2显示了这3个 *run* 的评测结果的比较,可以看出对基本检索结果集改进的效果还是很明显的。因为 *fduwt11t2*和 *fduwt11o2*都只提交了每个 Topic 的前30个检索结果,因此我们的性能指标只取到 $P@30$。横坐标为返回的前 N 个网页,纵坐标为全部50个 Topic 上的平均 $P@N$。

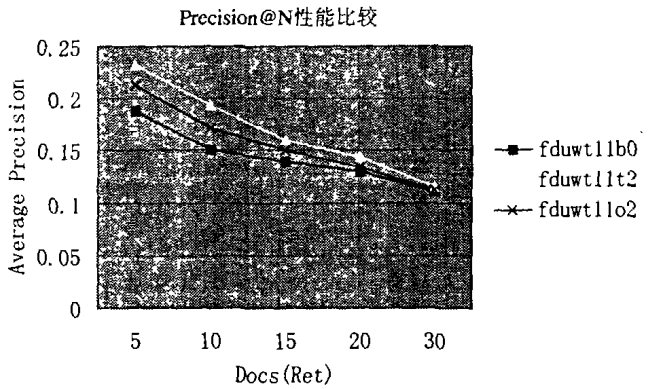

图2 Topic Distillation 任务的评测结果

3.4.3 进一步的实验 我们统计了 *fduwt11t2*和 *fduwt11o2*的 $P@10$随参数 a 变化的分布,发现 $a=0.9$时 $P@10$最高,$a=0.8$时次之,$a=0.6$时结果差一些。$a=1$时,即只考虑 $Score_D(d_i)$,而不考虑 $Score_L(d_i)$的贡献,$P@10$也能达到0.1796。并且参数 a 在0.8到1之间时,$P@10$比较稳定。网页的文档评分比我们原来想象的更重要[9],$P@10$随参数 a 的变化基本成线性上升关系($a=1$时除外),但链接信息对扩展基本的检索结果集,提高系统的查全率,起到了重要的作用。

因为我们的基本搜索引擎是基于扩展布尔模型的,因此

查全率较低，容易遗漏相关的网页，尤其是一些包含查询项较多的长查询。于是我们进一步在Okapi搜索引擎得到的基本结果集上进行了试验，并和原来的结果作比较。我们对动态确定参数 a 进行了探索，把50个查询中的前一半拿来训练，另一半用作测试。对训练用的25个查询分成5份进行交叉检验，然后在另25个查询上做测试。并且，我们评价了不做训练(取 $a=1$)时的结果，发现性能也很好。

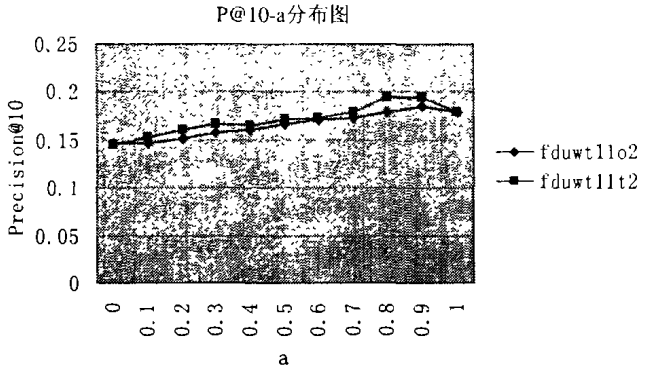

图3 参数 a 对 P@10的影响

Baseline 是搜索引擎的基本结果集，未作任何后续处理；自适应是应用交叉检验训练得到的结果。从表2中可以看出，运用这种算法可以比较明显地提高 $P@10$ 的查准率。以无参数($a=1$)为例，*fduwtllt2* 的 $P@10$ 提高了18.9%，*fduwtOkapi* 的 $P@10$ 提高了20.4%。搜索引擎的基本结果集对这种性能提高的影响不大。

表2 两组结果的P@10比较

Run	Baseline	不用参数	用参数	
		(a=1)	a=0.8	自适应
fduwtllt2	0.1510	0.1796	0.1939	0.1875
fduwtOkapi	0.2	0.2408	0.2204	0.2375

3.4.4 实验结果分析 由于 Web 本身具有的特点，通过它发布的信息极大丰富，但也包含有大量的垃圾信息。而对于用户来说，他们只希望返回的前几个网页都是非常相关的。在 Topic Distillation 任务中，采用的 $P@10$ 评价指标，也只注重返回结果前十个网页的查准率，对查全率没有什么要求。因此我们认为，Web 检索的查准率比查全率更加重要。

实验结果证明，这种方法对于 Topic Distillation 任务，效果是很明显的。相对于基本的检索结果，实验的两组结果的性能有明显的提高，尤其是提升了返回的前几个网页中的查准率。网页的标题、锚文本等结构信息对提高检索的查准率很有效，但它也依赖于基本搜索引擎的结果集，因此基本的检索结果集的查全率也是一个不能忽略的因素，不能太低。我们认为，如果语料足够大(相对于真正的 Web，18.1 GB 的语料还是很小)，我们的布尔搜索引擎得到的检索结果会比现在更好。

链接信息是非常有用的，但要和其他方法结合起来应用，对链接信息的查准率要求也比较高。如果能够准确地过滤掉网页中的无关链接，链接信息确实能有效地提高 Web 检索的性能。链接信息的结果也同样依赖于基本的检索结果集的查全率。

结论 检索的查准率对于 Web 检索的性能至关重要，因此要在不损失大量查全率的情况下，着重提高查准率。网页的结构信息，对提高 Web 检索的查准率作用很大。因此，我们还需要对网页的结构信息进一步细化，希望能系统化地处理网页中包含的结构信息。

过滤掉无关的链接，有效地应用相关的链接，有助于充分利用 Web 本身具有的特征，得到更高质量的检索结果。把网页的本身属性和链接信息结合起来，被证明确实是一条有效提高 Web 检索性能的途径。我们将对怎样把网页的结构信息和链接信息有机地结合在一起做进一步的探索。

Web 链接信息还可以应用于除 Web 信息检索以外的其他领域，如基于 Web 的文档分类、文档聚类、信息抽取、信息挖掘等。

参考文献

1 Murray B H, Moore A. Sizing the Internet. White paper, Cyveillance Inc., July 2000
2 Voorhess E M, Harman D. Overview of the Eighth Text Retrieval Conference (TREC-8). TREC 8, Gaithersburg, Maryland, Nov. 1999
3 TREC 2002 Web Track Guidelines. http://trec.nist.gov, 2002
4 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: Proc. of the Seventh Intl. World-Wide Web Conf. 1998
5 Kleinberg J M. Authoritative Sources in a Hyperlinked Environment. In: Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998
6 Weiss R, et al. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In: Seventh ACM Conf. on Hypertext, March 1996
7 Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In: 9th Intl. World Wide Web Conf. Amsterdam, Netherlands, May 2000
8 Hawking D. Overview of the TREC-9 Web Track. In: Proc. of the 9th Text Retrieval Conf. (TREC-9), National Institute for Standards and Technology, 2001
9 Craswell N, Hawking D. DRAFT Overview of TREC-2002 Web Track. http://trec.nist.gov/, 2002
10 Wu Lide, Huang Xuanjing, Niu Junyu, Xia Yingju, Feng Zhe, Zhou Yaqian. FDU at TREC2002: Filtering, Q&A, Web and Video Tasks. In: Proc. of the 11th Text Retrieval Conf. (TREC-2002), 2002
11 Voorhees E M. DRAFT Overview of TREC 2002. http://trec.nist.gov/, 2002

(上接第167页)

Springer-Verlar, 1999. 197～214
32 Singhai A, Sane A, Campbell H. Quarterware for Middleware. In: Proc. of the 18th Intl. Conf. on Distributed Computing Systems (ICDCS'98), Amsterdam, May 1998. 192～201
33 Cazzola W, Ancona M. mChaRM: A Reflective Middleware for Communication-based Reflection: [Technical Report DISI-TR-00-09]. DISI, Universita degli Studi di Milano, Milan, May 2000
34 Schmidt D C, Cleeland C. Applying Patterns to Develop Extensible ORB Middleware. IEEE Communications Magazine Special Issue on Design Patterns, 1999, 37(4): 54～63
35 Nahrstedt K, Chu H H, Narayan S. QoS-aware Resource Management for Distributed Multimedia Applications. Journal of High-Speed Networking, Special Issue on Multimedia Networking, 1998, 7: 227～255
36 Kiczales G, Lamping J, Mendhekar A, Maeda C, Lopes C V, Loingtier J M, Irwin J. Aspect-oriented programming. In: Proc. of 11th European Conf. on Object-Oriented Programming, Finland, LNCS 1241, Springer-Verlag, 1997. 220～242
37 Barga R, Pu C. A Practical and Modular Method to Implement Extended Transaction Models. In: Proc. Intl. Conf. on Very Large Data Bases, Zurich, Switzerland, 1995. 206～217
38 Barga R, Pu C. Reflection on a Legacy Transaction Processing Monitor. In: proc. of the Reflection'96 Conference, San Francisco, 1996
39 Wang Y M, Lee W-J. COMERA: COM Extensible Remoting Architecture. In: Proc. of COOTS, April 1998